SEMANTICALLY ENHANCED DOCUMENT CLUSTERING

IVAN DIMITROV STANKOV

A thesis submitted to the School of Engineering, Cardiff University

In partial fulfilment of the requirements for the degree of Doctor of Philosophy

2012

Abstract

This thesis advocates the view that traditional document clustering could be significantly improved by representing documents at different levels of abstraction at which the similarity between documents is considered. The improvement is with regard to the alignment of the clustering solutions to human judgement.

The proposed methodology employs semantics with which the conceptual similarity between documents is measured. The goal is to design algorithms which implement the methodology, in order to solve the following research problems: (i) how to obtain multiple deterministic clustering solutions; (ii) how to produce coherent large-scale clustering solutions across domains, regardless of the number of clusters; (iii) how to obtain clustering solutions which align well with human judgement; and (iv) how to produce specific clustering solutions from the perspective of the user's understanding for the domain of interest.

The developed clustering methodology enhances separation between and improved coherence within clusters generated across several domains by using levels of abstraction. The methodology employs a semantically enhanced text stemmer, which is developed for the purpose of producing coherent clustering, and a concept index that provides generic document representation and reduced dimensionality of document representation. These characteristics of the methodology enable addressing the limitations of traditional text document clustering by employing computationally expensive similarity measures such as Earth Mover's Distance (EMD), which theoretically aligns the clustering solutions closer to human judgement. A threshold for similarity between documents that employs many-to-many similarity matching is proposed and experimentally proven to benefit the traditional clustering algorithms in producing clustering solutions aligned closer to human judgement.

The experimental validation demonstrates the scalability of the semantically enhanced document clustering methodology and supports the contributions: (i) multiple deterministic clustering solutions and different viewpoints to a document collection are obtained; (ii) the use of concept indexing as a document representation technique in the domain of document clustering is beneficial for producing coherent clusters across domains; (ii) SETS algorithm provides an improved text normalisation by using external knowledge; (iv) a method for measuring similarity between documents on a large scale by using many-to-many matching; (v) a semantically enhanced methodology that employs levels of abstraction that correspond to a user's background, understanding and motivation.

The achieved results will benefit the research community working in the area of document management, information retrieval, data mining and knowledge management.

Acknowledgements

I would like to thank the supervisors of my studies, Professor Rossi Setchi and Dr Yulia Hicks, for their invaluable guidance and support throughout my work.

All members of the KES research group from School of Engineering in Cardiff University are thanked for their friendship and help.

My deepest gratitude is to my family who has given continuous support and encouragement to me.

TABLE OF CONTENTS

List of figures

List of tables

List of publications

1. Setchi R, Tang Q, Stankov I, ***Semantic-based information retrieval in support of concept design***, Advanced Engineering Informatics, invited paper for a Special Issue on Information Mining and Retrieval in Design , 25 (2) (2011) 131-146 ISBN/ISSN: 1474-0346 10.1016/j.aei.2010.07.006
2. Stankov I, Todorov D, Setchi R, ***Semantically Enhanced Text Stemmer (SETS) for Document Clustering***, In proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, San Sebastian, 10-12 Sep 2012, Spain
3. Stankov I, Todorov D, Setchi R, ***Towards enhanced crossed-domain document clustering with Semantically Enhanced Text Stemmer (SETS)***, Invited paper for a Special Issue on Innovation in Knowledge-Based and Intelligent Engineering Systems, submitted to Knowledge Engineering Systems Journal , September 2012

Chapter 1 : Introduction

1.1. Motivation

The success of the World Wide Web offers people the opportunity to share knowledge via textual documents contained in web sites, digital libraries and document re-positories. A large number of these documents have been made freely available and acces-sible, and there is a growing need for specific (vertical) and general document search and retrieval, which bene-fits the document browsing and the knowledge discovery across do-mains through more ef-fective document clustering (Huang, 2008, Grefenstette, 2009).

A problem in document clustering is the fragmentation of knowledge across do-mains, which results in specific and topic oriented approaches to grouping documents (Grefenstette, 2009). Therefore, clustering is traditionally used to produce specific groupings within pre-defined domains and cannot be used effectively by the search/retrieval algorithms in cross disciplinary tasks. This conflicts with the large variety of digital content consumption across domains (Andrews and Fox, 2007). As a result the desktop-based core search/retrieval market has begun to experience its first declines. The total number of core searches declined by 3% in 2012, driven primarily by a decline in searches per user (down 7%) despite growth in the number of searchers (up 4%) (Lipsman et al., 2013). The two reasons for the decline in the core search/retrieval intensity are (1) the shift towards vertical search/retrieval and (2) the shift to searching (retrieving) on mobile plat-forms, where the amount of mobile data traffic, speed and accuracy of retrieval are vital (Lipsman et al., 2013). With respect to vertical search/retrieval, users are increasingly likely to search for a product on dedicated market plat-forms such as Amazon or eBay, and search for people on Facebook, LinkedIn or Whitepages.

Meanwhile, vertical searches/retrievals are up by 8%, whereas core searches/retrievals are decreasing (Lipsman et al., 2013).

Since document clustering is focused in dealing with specific content within a domain of interest prevents its effective use in a general document search and/or retrieval across domains (Zhang et al., 2011). A vertical search/retrieval engine is distinct from a general document search/retrieval engine by its focus on a specific segment of content relevant to a predefined topic or set of topics. Therefore, document clustering needs improvement of the coherence of groupings produced by algorithms from documents that belong to various domains. Thus, the documents within a cluster will share higher degree of similarity unlike documents that belong to other clusters.

The degree of similarity is a measure, which formally characterises or recognises, either by processing text or through the use of ontologies, the contextual properties shared by two documents. The properties employed to measure the similarity between documents depend on the pragmatic context of the task they are used for (Grefenstette, 2009). Documents within a cluster are much more interconnected to each other and in this thesis they are considered *similar*. However, this definition does not exclude the possibility for a document from one cluster to share a certain degree of similarity with documents belonging to other clusters. Such documents are called in this work *related.*

Clustering algorithms can be divided into two types: model-based (Cadez et al., 2000), e.g. hierarchical and partitional, and similarity-based (Karypis et al., 1999). Most of them use the words in the documents as properties to measure pair-wise document similarity, ignoring their sequence or semantic relation, i.e. a relation explicitly stated in an external knowledge source (Li et al., 2008). However, it is proven that clustering algorithms, which incorporate background knowledge, achieve better performance than word-based algorithms (Hotho et

al., 2003b, Yong and Hodges, 2006). In the context of document clustering, background knowledge represents existing connections between terms in a document which indicate various entities, even if they do not exist literally (Hotho et al., 2003b).

Users of a document management system have certain background knowledge based on their previous experience. However, only domain experts have objective understanding of their domain of expertise and can actively contribute to the knowledge formalisation of that domain by explicitly discovering abstractions and existing relationships in it (Denaux et al., 2011). This process may result in a domain ontology, which is defined as an external representation of experts' subject-related knowledge (Engelbrecht and Dror, 2009). Therefore, different domain experts would create different domain ontologies based on the different perspective of the domain knowledge they have or different project specifications they follow (Wang et al., 2005, Denaux et al., 2011).

Ontologies are intended to be used by both machines and humans, yet they represent knowledge differently. The representations employed by humans are flexible and dynamic whereas ontologies are relatively static and contain fixed constructs. These constructs are established mental representations, which can be accessed by using knowledge elicitation techniques. However, evidence suggests that human internal representations of concepts are not stable entities but are the product of a dynamic, context dependant process (Barsalou and Neisser, 1987). Therefore, there is a mismatch between how people understand natural language and the assumptions inherent in formal logic. This may lead to using computational ontologies that contain contradictory statements; i.e. statements that do not comply with formal logic (Engelbrecht and Dror, 2009).

Cognitive psychology addresses this problem by focusing on the representation of the human knowledge in the process of creating ontologies. It analyses how human cognitive sys-

tem structures and processes conceptual information and suggests that these aspects can be used in knowledge elicitation as a model for structuring formal ontologies (Engelbrecht and Dror, 2009).

Cognitive psychology assumes that concepts have static mental representations in the human mind and can be retrieved from the long-term memory when needed (Barsalou and Neisser, 1987). On the other hand, concepts within text documents can encompass context-sensitive and context-independent information. The latter is considered to be highly accessible and relatively stable whereas the former is less accessible and is subject to interpretation (Barsalou and Neisser, 1987). Consequently, the cognitive processes that emerge during knowledge elicitation can be very subtle (Boroditsky, 2007). Therefore, efficient document clustering needs different perspectives for comparing objects by employing subjective criteria that allow for a diversity of views from which to look at the clustering task (Hotho et al., 2001).

Cognitive studies show that the comparison of similar objects makes them appear more similar, while comparing dissimilar objects makes them appear less similar (Boroditsky, 2007). As highlighted by Engelbrecht and Dror, 2009, this implies that certain knowledge elicitation methods may lead to the omission of identifying properties that are not shared, and increasing the actual similarity of the acquired properties.

Moreover, a controlled experiment, conducted in 2005 with a corpus of 50 short documents (Lee et al., 2005), reveals that existing clustering methods fail to emulate human expectations of similarity when comparing text documents. The results show that none of the clustering methods employed in the study could produce clustering solutions close to what the participant in the study expects. Further studies indicate that the problem might be due to the fact that all these methods are word-based and relate documents using identical terminology

(Hotho et al., 2003b). The evaluation of the proposed methodology employs the Reuters21578 corpus, which is tagged with words by linguists. The tags are used to emulate human judgement and provide objective evaluation.

This thesis advocates the view that document clustering could be improved by employing semantics to measure the conceptual similarity between documents. In addition to being semantic-based, the approach to be developed should provide different viewpoints of the document collection and consider the high computational complexity in large scale experiments where large memory footprints and CPU usage are setting challenges for high-dimensional vector space analysis (Zhang et al., 2010).

The main hypothesis of this research is that the effectiveness, which refers to the quality of document clustering in relation to human judgement, and the algorithmic efficiency, which refers to the speed of execution, can be improved by employing semantics. The improvement will result in providing better separation between and improved coherence within clustering solutions by organising large sets of documents into meaningful clusters. The clustering effectiveness, which refers to the quality of document clustering in relation to human judgement, and the algorithmic efficiency, which refers to the speed of execution, will improve as well. Producing clusters with improved coherency will enable the current state-of-the-art information retrieval algorithms to perform better across domains.

1.2. Aims and objectives

The goal of the research reported in this thesis is to develop an approach to producing coherent clusters from large-scale collections that provides multiple viewpoints to facilitate navigational, browsing, knowledge discovery and knowledge management tasks. The overall aim is to develop a semantically enhanced method, which generates multiple clusters that are

more topically homogenous and better aligned to human judgement[1]. The specific objectives of this research are as follows:

1. To develop a ***conceptual model*** that provides multiple deterministic clustering solutions and different viewpoints to a collection of documents, and enables large scale experiments;
2. To develop a semantically enhanced ***text stemming algorithm*** that provides reduced dimensionality and better separation between clusters;
3. To develop a ***method for measuring document similarity*** on a large scale by using many-to-many matching;
4. To design a ***methodology for semantically enhanced clustering*** that produces topically homogenous clustering solutions that are better aligned to human judgement;

The methodology proposed in chapter 3 (objective 1) outlines a general overview of the clustering and establishes a connection between the proposed functional blocks. An approach for multiple views to a document collection is suggested. Then (objective 2, chapter 4) an ontology driven dimensionality reduction is explored and tested on a large scale. The results reveal that concept indexing (Setchi et al., 2009) can be employed in the domain of document clustering and applied for all and not only pre-selected words contained in text documents when semantically enhanced text stemmer is used to normalise text prior to clustering. In addition, the approach to dimensionality reduction by replacing a group of words with a generic entity (Hotho et al., 2003b) is proven to work even when all words are replaced because the

[1] The tags assigned to the documents of the Reuters21578 corpus by linguists are assumed to be human judgement used in this thesis. The tags are used to cluster documents using the same algorithms as the proposed or used methodology in the relevant chapters and the clustering solutions produced are then compared.

concept index preserves the statistical information of word co-occurrence. The clustering solutions produced by the method proposed in chapter 5 (objective 3) is compared to human judgement on a large scale. The experimental investigation demonstrates that the methodology proposed in the literature fails on a large scale. Therefore, chapter 6 (objective 4) proposes a methodology for semantically enhanced clustering by using levels of abstraction which alleviates this problem.

1.3. Outline of the thesis

The rest of the thesis is organised as follows. Chapter 2 reviews clustering methodologies and techniques and contrasts methods for model-based and similarity-based discriminative clustering. A particular attention is devoted to acquiring document representation index, i.e. the process of feature selection and extraction, and how it is used by different strategies to measure pair-wise document similarity.

Chapter 3 addresses the first objective of this thesis, which is to develop a conceptual model that overcomes the problems and limitations of current state-of-the-art clustering algorithms with regard to their scalability. The model aims to provide multiple deterministic clustering solutions to the users. The chapter discusses semantic-based approaches to clustering as a prerequisite to producing clustering solutions that are consistent with and better aligned to human judgment. This chapter proposes a semantically enhanced document clustering model that provides multiple deterministic clustering solutions and different viewpoints to a document collection.

Chapter 4 addresses the second objective of the reported research by proposing a semantically enhanced text stemming algorithm. It discusses text normalisation techniques and approaches, and focuses on improving the clustering solutions produced by partitional cluster-

ing methods in terms of coherence within and separation between the clusters. The chapter firstly discusses approaches to document representation and techniques for document indexing. It then proposes a technique that improves clustering solutions by using a document index with reduced dimensionality. The proposed technique is compared to the word-based TF-IDF weighting system (calculated after the Porter stemmer normalises the document collection) and human judgement in a generic non-domain specific environment, through an analysis of the clusters' coherence. This chapter provides evidence if concept indexing as a document representation technique can be used to represent documents and successfully preserve the statistical information for words' co-occurrence on a large scale when semantically enhanced text stemmer is used for text normalisation prior to clustering. In addition, the clustering solutions produced will be aligned to human judgement for comparison.

Chapter 5 addresses the third objective of the thesis by proposing a method for measuring document similarity on a large scale by using many-to-many matching. It introduces the use of the Earth Mover's Distance (EMD) algorithm, used in image processing as a pair-wise document similarity measure as it offers a multidimensional approach to measuring similarity based on content distribution. In addition, the chapter outlines inadequacies and deficiencies of traditional document similarity measures. A comparison between the robust cosine and enhanced EMD measures in relation to human judgement is also conducted. This chapter proposes a method for measuring similarity between documents by using many-to-many matching on a large scale.

Chapter 6 addresses the fourth objective of this research, which is to develop a methodology for semantically enhanced clustering that improves the consistency and alignment of clustering solutions to human judgement. The methodology introduces levels of abstraction at which the similarity between documents is considered. The chapter firstly discusses tradition-

al clustering approaches and identifies areas for improvement. It then introduces the developed technique and evaluates it against a traditional clustering algorithm in comparison with human judgement. This chapter proposes a semantically enhanced methodology that employs levels of abstraction at which similarity between documents is measured.

Chapter 7 highlights the contributions of the thesis and discusses future research.

Chapter 2 : Literature review

This chapter reviews methods and techniques for text document clustering. First, clustering methodologies and techniques are reviewed. Then model-based approaches are compared to similarity-based discriminative methods. Particular attention is devoted to feature selection and extraction from text, used for indexing documents, and the different strategies for measuring similarity between documents.

2. 1. Clustering methodologies and techniques

Clustering, also known as numerical taxonomy (Xu and Wunsch, 2005), is unsupervised classification or exploratory data analysis carried out on unlabelled data (Jain and Dubes, 1988, Everitt et al., 2001). Categorisation, on the other hand, known as predictive modelling or supervised learning, constructs models to predict the value of a dependant variable using values of other known attributes (Ženko, 2007), i.e. it uses prior data assigned to the objects. Since clustering is not using such data it is regarded to be different from the predictive learning problems such as vector quantisation, probability function estimation, and entropy maximisation (Xu and Wunsch, 2005), even though predictive vector quantisation algorithms are used in non-predictive clustering analysis (Cherkassky and Mulier, 2007).

The goal of clustering is rather grouping unlabeled documents into finite sets, using an index, than providing inaccurate characterisation based on unobserved samples derived from the same probability distribution (Baraldi and Alpaydin, 2002, Cherkassky and Mulier, 2007). Document clustering is employed by many disciplines thus the approaches and assumptions used vary. Information retrieval defines users need for information as a query submitted to a search engine (Jain et al., 1999). In this scenario, which deals with text, the choice of words used in the query is important as it pre-determines the returned result. There-

fore, when users are not familiar with the terminology or the appropriate vocabulary in the topic of interest, they may commit to an inappropriate choice of words, which may lead to a poor search result. However, navigation within returned documents facilitates finding the information needed (Cutting et al., 1992). Furthermore, document clustering has a key role in refining the results returned from the search engines (Carpineto et al., 2009). Browsing a collection of documents and organising them into clusters to find specific information (Cutting et al., 1992, Carpineto et al., 2009) are of particular interest to the research reported in this thesis.

2.1.1. Clustering methods

Clustering methods can be divided into generative, also known as model-based (Cadez et al., 2000), and discriminative approaches also called similarity-based because of the use of a similarity measure (Karypis et al., 1999). The former approaches learn generative models from data, with each model corresponding to one particular cluster. They are driven by a predefined parameter, which sets the number of clusters. The discriminative approaches rely on a distance measure or a similarity function to determine the similarity (or dissimilarity) between documents and the most similar documents are then grouped together.

Selecting an appropriate methodology for grouping documents in a collection depends on the adopted document representation, which includes the assumptions made on the data to achieve certain abstraction (Jain et al., 1999). Data abstraction is a process of building a simple and compact representation of documents. The process aims simplicity from the perspective of automation analysis or/and ease of comprehension of the results from the human perspective. Data abstraction is purpose-oriented and subjective in nature. As a result even unsupervised classification such as clustering produces subjective results and disqualifies the ab-

solute judgment to the relative efficacy of all clustering techniques (Jain et al., 1999, Baraldi and Alpaydin, 2002). This finding is supported by the view that objects are grouped together into smaller homogeneous subgroups on a subjective basis, using a subjective measure of similarity, which provides the ability to create interesting clusters (Backer and Jain, 1981, Xu and Wunsch, 2005). However, clusters can still be described in terms of their internal homogeneity and external separation (Gordon, 1999), i.e. feature patterns within the same cluster should be similar to each other, whilst in different clusters they should be not, and yet it should be also possible to identify relation between patterns (Xu and Wunsch, 2005, Yang et al., 2008). This indicates the need to develop methods and techniques that provide multiple subjective views to a document collection where documents can belong to more than one cluster.

Fuzzy clustering (Zadeh, 1965) uses a degree of membership to assign a membership coefficient to a document, which belongs to more than one cluster. This coefficient satisfies certain constraints and makes every document a member of one or more clusters. However, a study conducted by cognitive scientists (Boroditsky, 2007) show that comparing two similar objects makes them appear more similar, while comparing dissimilar objects makes them appear less similar. The same study indicates that human judgement as a cognitive process during knowledge elicitation of comparing two categories leads to an increase in the perceived similarity between them even when the differences are listed. This finding suggests that certain knowledge elicitation methods which involve comparison of concepts in order to group them, may lead to omitting the attributes that are not shared by the categories been compared (Engelbrecht and Dror, 2009). Hence, it is difficult to define so called “gold standards” in clustering, except for document collections that belong to a narrow sub-domain (Jain et al., 1999).

2.1.2. Clustering techniques

Clustering techniques group documents together using similarity measures and thresholds. Model-based approaches measure similarity using distance functions such as Euclidean distance, cosine similarity, overlap measure, relative entropy, dice measure, Jaccard measure or itemset-based measure. In contrast, similarity-based clustering methods consider existing relationships between words or the internal structure of documents to calculate similarity by using multi-dimensional scaling and in particular OM-based (Optimal Matching) and EMD-based (Earth Mover's Distance) techniques (Wan and Peng, 2005a).

Xu and Wunsch (2005) state that clustering differs from multi-dimensional scaling, which goal is to depict all evaluated objects to minimise the topological distortion using as few dimensions as possible. However, Wan and Peng (2005b) have proven that statement wrong by employing EMD (Rubner et al., 2000) to measure the similarity between two documents.

Model-based clustering algorithms employ hierarchical or partitional clustering techniques (Jain and Dubes, 1988). The former techniques organise clusters into tree structures (dendograms), which allow identifying relationships between documents. Each intermediate level is either a combination of two clusters from the next lower level (agglomerative approach) or a breakdown of a cluster from the next higher level (divisive approach). These techniques produce nested sequences of partitions that contain an all-inclusive cluster at the top and singleton clusters at the bottom. The nodes inside the tree structure display the merging process and the intermediate clusters, thus providing a taxonomy (hierarchical index). The latter techniques create one-level (un-nested) partitioning of documents. The predefined number of clusters into which the documents are grouped drives document partitioning. Hierarchical clustering is considered to provide better-quality clustering. However, its implementations are

limited because of its algorithmic complexity, which is dependent on the number of documents. On the contrary, partitional clustering has complexity, which grows linearly with the number of considered features, but it produces inferior clusters. It has been proven that algorithms, which combine both techniques (e.g. 'bisecting' k-means algorithm) perform better than traditional partitional approaches and as well or better than the hierarchical approaches (Steinbach et al., 2000).

Similarity-based techniques first measure the similarity between all pairs of data samples, and then group similar ones together into clusters (Karypis et al., 1999). The main steps of similarity-based techniques are: 1) calculating the distance matrix between all pairs of documents; 2) using the distance matrix to merge the two closest clusters; and 3) modifying and rebuilding the distance matrix by treating the merged clusters as one object. The process stops if the number of desired clusters is reached, otherwise step 2 is repeated. These techniques are difficult for people to follow through due to the use of external knowledge and the high complexity of the word dimensional space employed (Punitha et al., 2011).

2.1.3. Clustering procedure

Selecting an appropriate clustering technique or any of its variants depends on the task it is employed for. The effectiveness and the efficiency of the selected technique depends on the chosen feature selection, which provides indices for the documents. Clustering algorithms use an index extracted from the documents to group them. Document indices used in document clustering include words, phrases, concepts, and topics.

Clustering analysis typically consists of four steps with a feedback pathway (Xu and Wunsch, 2005). The procedure is shown in figure 2.1.

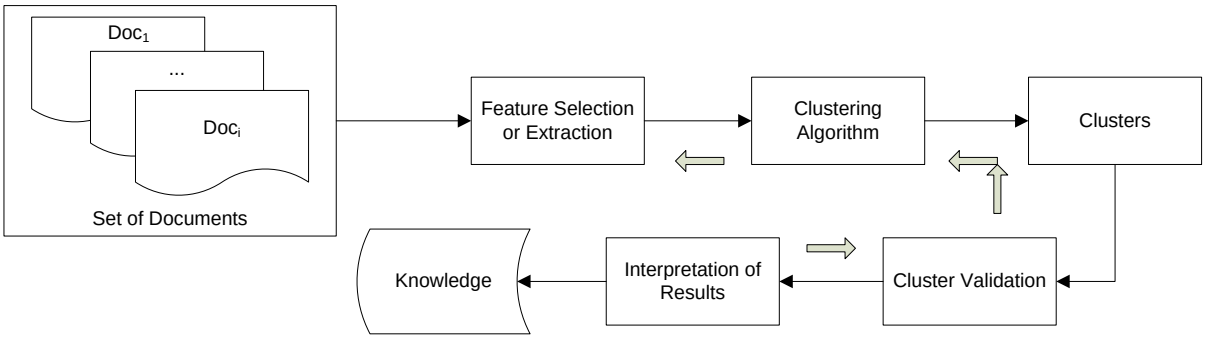


Figure 2.1 Clustering procedures

2.1.4. Feature selection

The first step of the clustering procedure is to distinguish subset of features from a set of candidates (Jain et al., 1999, Jain et al., 2000). The process is called feature selection and it differs from feature extraction as the latter utilises transformations needed to generate features from the original ones. In document clustering, document pre-processing can vary depending on the assumptions and abstractions made. Model-based algorithms normalise the words by employing stemming. The most commonly used stemming algorithm is the Porter stemmer (Porter, 1997). After the text is normalised, the statistical co-occurrence of words and phrases is calculated by weighting the indices produced (Lewis, 1992). Then, similarity-based algorithms are employed to generate an index, which takes into consideration existing relationships in an external knowledge resource (Setchi and Tang, 2007, Xiao, 2010).

Feature selection and extraction are crucial to effective and efficient clustering. Good feature selection or extraction can result in decreased workload and simplified subsequent clustering algorithm or/and improved clustering (Xu and Wunsch, 2005).

The rest of this chapter reviews methodologies for feature selection and extraction that use words/phrases, ontologies and semantics.

2.1.5. Clustering algorithm design and selection

The design of clustering starts with selecting a similarity measure and constructing a criterion function for measuring similarity between documents (Xu and Wunsch, 2005). Feature patterns are grouped together if they resemble each other, i.e. proximity measure over two feature patterns is applied and if the result corresponds to a pre-defined criterion function, they are placed in the same cluster. Therefore, the proximity measure, which can be defined in explicit or implicit manner, affects the formation and the quality of the clusters.

Clustering addresses problems associated with high dimensionality, scalability of the algorithms, measuring the accuracy and the quality of the produced clusters. The first problem is with regard to the spectral requirements of the documents, i.e. the large number of features used for document representation (Fung et al., 2005). The large feature set and the fact that every feature constitutes a dimension in the feature term-based space can be addressed by placing documents in a sub-space. However, this is a very challenging task and dimensional spaces which include all features are not used on a large scale.

The next problem is scalability. Algorithms which produce good results on a small data set (Fung et al., 2005) or in a specific domain (Zhang et al., 2011) fail to perform on larger scale or across domains. The first scenario considers algorithms with very high computational complexity which is impractical on a larger scale. The second scenario deals with the polysemy of words and the fact that many domains share common terms, which may contribute to low quality in the document groupings (Steinbach et al., 2000).

2.1.6. Evaluation of clustering solutions

The third problem refers to measuring the accuracy of the produced clusters. Following the similarity criteria for high homogeneity inside the clusters and diversity between clusters is used to project certain quality of the clustering structure produced by clustering algorithms from all documents, but have no practical value in terms of human judgement. Therefore, a methodology that aligns the produced clustering solutions to human judgement is discussed further in this section.

The accuracy of clustering depends on the quality of the document index (Facolta et al., 2008), which is selected in the view of a particular task (Lewis, 1992). Each index acquired from the same documents incorporates different assumptions and leads to a different clustering result. An important benefit of clustering is that it provides unseen groupings of features, but these groupings need to be viewed and evaluated from the perspective of human judgement. Thus, users will have a certain degree of confidence for the derived clusters and therefore, the validation of clustering becomes a crucial part of it. The evaluation should provide objective assessment of the derived clusters and have no preference to any algorithm. In addition, the evaluation standards and criteria should provide evidence whether the obtained clusters are meaningful to users or just a manifestation of the employed algorithm (Xu and Wunsch, 2005).

Generally, model-based clustering uses testing criteria based on external, internal and relative indices (Jain and Dubes, 1988). Criteria using external indices compare the new clusters to a pre-defined structure of the clustering data. This method for validation is used by partitional methods. Criteria using internal indices on the other hand, test the data without any prior knowledge by examining the clustering structure directly from the original data (used

by hierarchical methods). Relative criteria compare different clustering structures and provide a reference to decide which one is best (Xu and Wunsch, 2005).

Conversely, as reported by Wan (2007), there is no standard dataset for evaluation of document similarity, which can be used to validate the clustering structure produced. However, researchers adopt text classification experimentation corpora to validate their approaches (Hotho et al., 2003b, Hotho et al., 2003c, Wan, 2007, Wan et al., 2007) and the following two sections discuss similar approaches.

2.1.6.1. Evaluation methodology in information retrieval and cognitive psychology

In the domain of cognitive psychology the evaluation methodology usually involves people who participate in a study in which different algorithms or approaches are used to acquire or compare results with human judgement (Goldsmith et al., 1991, Lee et al., 2005). In 1991 a study assesses the cognitive representation of the structural knowledge of students by comparing it with that of their instructor for the purpose of constructing a predictive model (Goldsmith et al., 1991). This approach employs extensive analysis, which uses all features pre-defined by the researchers. Therefore, the study is comprehensive and accurate but requires a long time to manually analyse the results. This makes it impractical for large scale experiments.

Another study in the domain of information science evaluates existing document similarity methods in terms of their ability to emulate human judgement (Lee et al., 2005). This study reveals that existing similarity methods fail to emulate human expectations of similarity when comparing text documents. However, it presents very detailed analysis of the quality of the similarity measures on a small scale, where a small number of documents are manually analysed by the researchers. The analysis involves the researchers' subjective decision whether

the documents within the clusters are correctly grouped. It is noted that the similarity measure is an essential part of any clustering model. Therefore, these similarity measures if employed by clustering should perform similarly and the clustering solutions produced should have similar alignment to human judgement. However, the produced clustering solutions need to be compared for all documents in all clusters independently and then aligned to human judgement. A methodology which provides such comparison is discussed in the next section.

The quality of the similarity measures are further investigated with respect to the information retrieval domain. The measures employed by the evaluation methodology in this domain are divided into two groups. The first group includes measures of single-value metrics such as precision, recall, P@N (which considers precision and recall of the topmost results) and f-measure (Blair, 1979), which also involves precision and recall. These measures are based on the complete list of documents returned by the algorithms. However, in this case all relevant and retrieved documents must be known prior to execution for every query upon which documents are retrieved. The corpora used for evaluation must have a list of predefined and analysed queries against which the comparison of the results is evaluated.

The second group includes measures which return a ranked sequence of documents. These measures consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, the precision-recall curve can be plotted, where the precision is a function of the recall. In information retrieval the average precision measure is widely used (Voorhees, 1998). It computes the average precision of the retrieval as a value in the interval from 0 to 1. The positive predictive value or precision of the results is used to indicate the retrieval accuracy. Mean average precision (MAP) measures the retrieval performance of algorithms for a set of queries, where the mean is the average precision scores for each query. The second group of measures evaluates

the retrieval value for a given number of the top ranked documents. Document (web) search engines retrieve a certain number of documents, which are sorted in descending order of their rank. Hence, the algorithms can be evaluated by the quality of their retrieval based on a pre-selected number N from the top ranked documents. Then, for N number of documents is measured as the number of correctly retrieved documents using the following equation:

$$AveP = \frac{\sum_{k=1}^{n}(P(k)x\ rel(k))}{number\ of\ relevant\ documents}$$

where rel(k) is an indicator function, which equals to 1, if the item at rank k is a relevant document, or zero otherwise. The average is calculated for all relevant documents and the relevant documents not retrieved get a precision score of zero.

The mean average precision (MAP) is calculated using the equation below:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

where Q is the number of queries.

The aforementioned methods for measuring correctness of retrieval are useful for testing similarity functions or retrieval power of information retrieval algorithms. However, it is not enough to address the overall performance of clustering algorithms because (i) it is impractical for large scale experiments (ii) to consider the quality of all groupings for the entire collection. Therefore, a comprehensive evaluation approach to a clustering is obtained if the entire structure of the clustering solutions produced is explored. This approach is similar to the method for evaluation of clustering solutions with silhouettes.

2.1.6.2. Evaluation methodology employed

It is important to note, that there are no gold standards (Jain et al., 1999) for evaluation of clustering solutions and researchers use different corpora and methods. The evaluation meth-

odology used in this thesis targets objective evaluation of the structure of the clustering solutions produced with all the documents from a collection. The evaluation needs to represent the alignment of the produced results to human judgements. A condition to the methodology is to carry out the evaluation on a large scale.

The selected corpus for the evaluation is Reuters21578. This collection consists of 21578 news articles on different economic subjects published in 1987. An important property of the corpus is the presence of tags (the total number of which is 445) manually assigned to every article (up to 29 tags per article) by linguists. The tags are separated in different categories. The tags that indicate the topics of the documents are 140 and a few additional sets of total 305 unique tags are used to indicate properties that articles convey with regard to entities such as people, places, dates, orgs, exchanges and companies. Since the tags are morphological words or named entities provided by the linguists (not necessarily contained in the text) they are all used in the evaluation of the document groupings to provide a perspective of human judgement. All tags are used to cluster the corpus and compare the results against other methods. The evaluation demonstrates how the clusters produced by different clustering techniques align to human judgement (i.e. clusters produced by using the Reuters collection's tags). The tags are considered by the reported research as a judgement provided by humans. This is as a consequence of the fact that the linguists have the same background and motivations. Therefore, the assigned to the documents tags are consistent in representing the content of articles. There are no constraints applied for the choice of tags. All tags are assigned manually. In addition, the motivation of the task the linguists are assigned with is to represent an article with words. They were not given the task to group similar documents. Therefore, the clustering results obtained by using the tags are not manipulated by the task.

In figure 2.2 is shown a comparison between clustering solutions. Clustering approach T employs the tags of the Reuqters21578 collection to cluster all or a pre-defined set of articles. This clustering solution represents how the linguists would have clustered the articles by employing the selected clustering algorithm. Then, the same collection (the same set of articles) is clustered two more times by clustering approaches A and B. Documents X, Y and Z are used as centroids for the clusters in the relevant clustering solutions. Then, the evaluation measures the number of similarly grouped documents for the individual clusters in per-cents [%]. The clusters build up around document X by clustering approach A and B are compared with the cluster X produced by the clustering approach T. The number of documents, which are clustered the same by clustering approaches A and B in comparison to T are calculated. Finally, when the same procedure is applied for every cluster (X, Y and Z) the calculated results are summed up. The equation below explains the procedure:

$$\text{comparison (A | T)} = \frac{1}{l+n+m} 100 \sum_{k=1}^{l} a_k + \sum_{j=1}^{n} a_j + \sum_{i=1}^{m} a_i \text{, [\%]}$$

where, comparison (A | T) is the result of similarly clustered documents by clustering solutions A and T [%], *l*, *n* and *m* are the total number of documents respectively in clusters X, Y and Z and $a_{k|j|i}$ is an article from the relevant cluster. The total number of documents in the collection is l+n+m. Thus, in fig. 2.2 are presented clustering solutions of 3 clusters produced by clustering approaches T, A and B. The higher number of documents is clustered similarly to the clustering solution produced by approach T, the better alignment of the clustering results to human judgement.

An objective evaluation of different document representational techniques in relation to human judgement, for example, is obtained by firstly, producing a clustering solution using

the tags, and then compare that solutions with the clustering solution produced by the same clustering algorithm but using other document representation techniques (such as vector space model and concept indexing). Then, the comparison will demonstrate the difference of the document representation techniques in clustering in relation to human judgement.

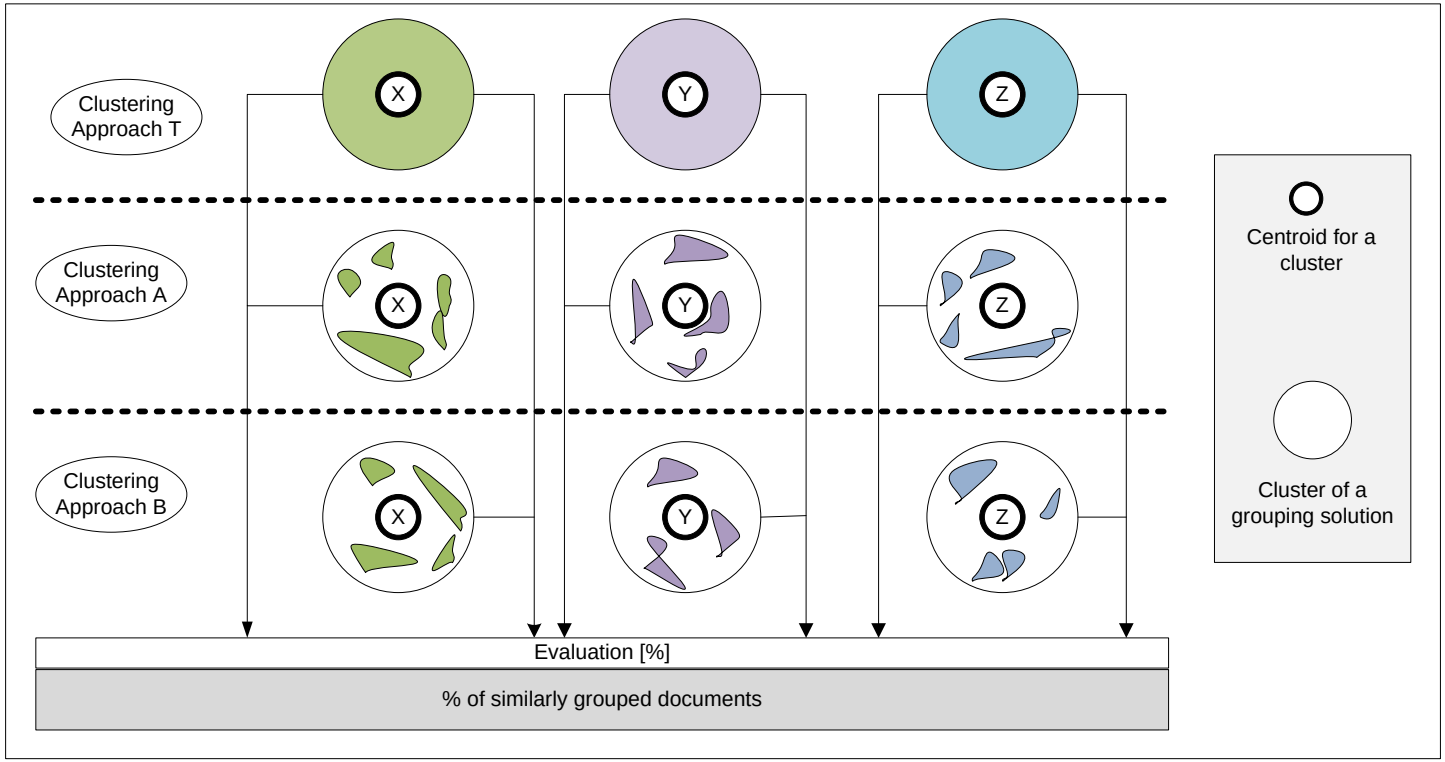


Figure 2.2. Comparison of clustering solutions

Section 2.2 reviews approaches which employ words and phrases as features used to index documents and acquire model-based patterns of the features to calculate the similarity between documents. In section 2.3 similarity-based approaches, which rely on external knowledge, are discussed and reviewed.

2. 2. Model-based document clustering

Model-based algorithms use various document representations, such as the vector space model (VSM), that treat documents as a bag of words (BOW) (Salton and Mcgill, 1986), set of repetitive words (Wang et al., 1999, Beil et al., 2002) or word-sequences (Li et al., 2008). Syntactic and semantic information (Yun et al., 2010) including word senses (Peng and Choi, 2005) is outside the scope of model-based clustering. This section reviews partitional and hierarchical clustering, which are the two main types of model-based approaches used. Document grammar is outside the scope of this research and is not reviewed.

2.2.1. Partitional approach to clustering

Partitional clustering, also known as hard partitioning, creates flat, non-hierarchical clusters, whose number is controlled by a value given to the algorithms prior to execution. As highlighted by Fung et al. (2005), the number of clusters k drives the process of partitioning documents in k clusters by employing the standard k-means algorithm or any of its variants. However, selecting the number of clusters without any domain knowledge in the area of interest may worsen the results. In addition, if documents cover a broader thematic area, the clusters produced would be inferior. Kernel-based partitional methods such as kernel k-means algorithms, which consider mapping of the input prior to clustering (Karatzoglou and

Feinerer, 2006) using string kernels (Huma et al., 2002) or word-sequence kernels (Cancedda et al., 2003), perform better than the standard k-mean partitional algorithms.

Partitional algorithms use VSM document representation. They have two main disadvantages. Firstly, they do not consider semantic relations between words. Not only words with similar meaning (Yun et al., 2010) but also relationships between words, which share similar semantic context, are treated as irrelevant features (Hotho et al., 2003b, Hotho et al., 2003a). Secondly, same words with different meaning in different context are not considered either (Yun et al., 2010). Furthermore, the number of dimensions has to be of the same length for all vectors, i.e. short and long text documents should have the same number of words representing them in the vocabulary space of the document collection (Steinbach et al., 2000).

VSM represents each document **d** as a vector **D** in the vocabulary space. It represents a document using terms co-occurrence (TF–term frequency) within a document so that $D_{tf} = (tf_1, tf_2, tf_3, \dots, tf_n)$, where tf_i is the frequency of the i^{th} term contained in the document. However, not all terms have the same discriminative power, and determining what the discriminative power of the words is can be considered as a two-stage process. Firstly, stop words are removed and then words that are used often in the documents within the collection are given less discriminative power. The second stage employs the common practice of weighting the words' significance within the document collection. This is achieved by calculating IDF (inverse document frequency) for every word, or classifying words with eliteness (Robertson, 2004). Words that are very frequent in the collection gain less discriminative power (less IDF weight) than the more unique words. The presumption is that words with higher statistical value are more relevant to the topic of a document. Before IDF is calculated, words that occur in different grammatical forms are normalised to their canonical form using a stemming algorithm (Porter, 1997). Then the weights of the document indices are calculat-

ed by multiplying TF and IDF. However, computing the weight of all words within all documents leads to high computational complexity (Beil et al., 2002), which motivates considerable interest in low-dimensional document representation that overcomes this particular issue (Matveeva, 2006).

The k-means algorithm uses the robust cosine measure to compute the similarity between documents. It is defined as $cosine(d_1, d_2) = (d_1 \bullet d_1)/||d_1||\ ||d_2||$, where • indicates the vector dot product and $||d||$ is the length of the vector. The k-means algorithm computes randomly a $\boldsymbol{k}$ number of vectors in the feature space to identify the closest documents to the centroids and then uses these vectors to form clusters. The algorithm iteratively refines the randomly chosen initial k centroids, minimising the average distance (homogenising the clusters by increasing the similarity within clusters).

Improvement of the standard k-means algorithm is the "bisecting" k-means algorithm proposed by Steinbach et al. (2000). It randomly selects k documents and creates k initial clusters, which are incrementally updated with every consecutive document rather than at the end of the assignment pass. The basic "bisecting" step is when a cluster is selected, using basic k-means algorithm, to find two sub-clusters and to repeat that step until the k number of clusters is reached. The "bisecting" approach produces better overall similarity and lower entropy and has better accuracy and efficiency (Zhao and Karypis, 2002). However, both algorithms are found to be not only relatively efficient and scalable but also sensitive to noise (Fung et al., 2005). The authors state that not only the noise, which can be easily introduced in the preparation step and will influence the construction of centroids, can cause poor performance but also if the number of k clusters is incorrectly estimated. Although, the noise problem is addressed by the k-medoids algorithm (Krishnapuram et al., 1999), its computational cost

makes it impractical. In addition, these algorithms are considered not suitable for discovering clusters of varying sizes, which is the case with document clustering.

Alternative strategy for achieving better performance and scalability of the k-means family of algorithms is addressed by the dimensionality reduction techniques (Xiao, 2010). Latent Semantic Analysis (LSA) is one of the best known dimensionally reduction algorithms in information retrieval (Deerwester et al., 1990, Zhang et al., 2011). It is an algebraic indexing method, which provides a mechanism for low dimensional document representation. The algorithm uses statistical data of word co-occurrence (TF-IDF), but it further employs a higher-order document- term (semantic) structure. This approach aims to find the best sub-space approximation to the original space in terms of minimising the global reconstruction error. Using singular vector decomposition, LSA projects the representing vectors into an approximate sub-space. The sub-space represents the original feature space with fewer dimensionality, which enables the cosine similarity to compute semantic similarity between documents accurately. The semantic structure is acquired from external knowledge and is used to improve the detection of relevant documents (Zhang et al., 2011). This is necessary because word-based document representation is challenged by word polysemy and the fact that in information retrieval, relevant documents might be indexed with words that users with perspective different to the encoded knowledge would not use in their retrieval queries. Semantic-based approaches to clustering that address this shortcoming are reviewed in section 2.3.

The kernel-based technique is another method for dimensionality reduction. It is successfully used for document ranking and filtering in information retrieval and text classification when dealing with large collections. Exploring kernel methods such as kernel k-means and spectral clustering (Ng et al., 2001) in the area of document clustering is needed due to the inadequacy of the standard k-means algorithms to separate clusters that are not linearly sepa-

rable in the input space. Kernel algorithms first map the input data into a high dimensional non-linear space and then a kernel function places the result of the mapping implicitly into a pre-selected feature space. Then Euclidian distance is used to measure the distance between the properties. The spectral clustering uses affinity matrix and leads to easier to solve clustering problems since points tend to form tight clusters in the eigenvector subspace (Karatzoglou and Feinerer, 2006). However, this is an application oriented (specific) approach and therefore, full string kernel is considered as a more general technique for clustering. The comparison between spectral clustering with string matrix, kernel k-means with full string matrix, and simple k-means with term matrix demonstrates that spectral clustering with string kernel performs better than all other methods (Ng et al., 2001). Computing the kernel matrix is a very time consuming task and the performance of the kernel-based algorithms strongly depends on the length of the string.

2.2.2. Hierarchical approach to clustering

There are two types of hierarchical approaches: agglomerative and divisive. The first family of algorithms builds the cluster hierarchy bottom-up by computing iteratively the similarity between every two pairs of clusters and merging the most similar pair (Kaufman and Rousseeuw, 2005). The difference in the variants of this family of algorithms is in the selected function for calculating similarity between documents (Zhao and Karypis, 2001). The second family of algorithms builds the hierarchy top-down starting from the top with all documents in one cluster, as the similarity measure considers the global distribution of the document representation (Manning et al., 2008). The cluster is split by using a flat clustering algorithm with a certain similarity measure. This procedure is applied recursively until each document is in its own singleton cluster.

Although top-down clustering is conceptually more complex, it can be more efficient if the complete hierarchy of the tree structure is not generated. In addition to that, the evaluation of the results carried out by Steinbach et al. (2000) using the f-measure shows that the divisive approach produces more accurate hierarchies when combined with partitional clustering. The main disadvantage of both approaches is their high computational complexity in similarity calculation. In addition, early decisions cannot be undone, i.e. previous splitting or merging of clusters cannot be adjusted, which lowers their clustering accuracy (Fung et al., 2005).

The algorithms that implement the agglomerative techniques maintain very high homogeneity within the clusters. Alternatively, the intra-cluster similarity technique (IST) uses the agglomerative approach to merge a pair of clusters that results in slight decrease of the homogeneity within the merged cluster. This technique measures the homogeneity by comparing all members of a cluster; it has a quadratic complexity to the number of documents. Another technique, the centroid similarity technique (CST), reduces the complexity by measuring the similarity between clusters based only on their centroids. It uses cosine similarity measure and is faster than the IST. UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is another technique, which is similar to IST, but it uses altered cosine measure for calculating the similarity. Steinbach et al. (2000) prove that UPGMA and IST perform equally, although UPGMA's performance is better when the number of clusters is high. UPGMA is later proven not scalable and unsuitable for large data sets because of its complexity (Fung et al., 2003).

The same study (Fung et al., 2003) compares the standard k-means algorithm, "bisecting" k-means approach (partitional approach, which produces hierarchies) and UPGMA, which is the best agglomerating hierarchical technique. The results indicate that "bisecting" k-means is better that the standard k-means algorithm and as good as or even better than UPGMA. The

comparison is according to the entropy and the overall similarity measures of the cluster quality. The authors indicate that the time needed to execute the partitional approaches is significantly shorter, which suggests that these algorithms are scalable and can be used in large data sets.

Besides pair-wise similarity, which is document-centred distance measure (Fung et al., 2005), documents can be grouped using clustering transactions based on frequent itemsets (Wang et al., 1999). This similarity measure places documents in the same cluster if they share many frequently repeating items and sustain homogeneity. This approach treats a word as an item and a document as a transaction. The authors argue that for transactions made of sparsely distributed items, pair-wise similarity is neither necessary nor sufficient for a cluster of transactions to be similar. This approach does not meet the spectral clustering requirements but offers a mechanism for dynamic clustering with substantial influence on efficient and quality clustering, which achieves good consistency with human judgement.

Hierarchical frequent term-based clustering (HFTC) (Beil et al., 2002) and frequent itemset-based hierarchical clustering (FIHC) (Fung et al., 2003) address the clustering spectral requirements by using the notion of frequent itemsets. HFTC considers the low-dimensional frequent term sets only, whilst FIHC uses the global frequent itemsets that appear in more than minimum fractions of the document, which drastically reduces dimensionality. The former algorithm forms clusters by minimising the overlap between them in terms of shared documents. The latter creates a cluster for each itemset and if a document belongs to more than one cluster it is placed in the cluster, which is the best match. HFTC produces accuracy comparable to the "bisecting" k-means, but is experimentally proven to be not scalable, whilst FIHC besides proven to be scalable, fast and very accurate, generates a tree, a.k.a.

pruning tree, which is easy to browse and navigate among the documents (Fung et al., 2003). The pruning tree is based on inter-cluster similarity.

Subspace text clustering is another methodology for reducing dimensionality by discovering clusters embedded in the subspaces of a high dimensional data such as text documents (Jing, 2008), through bottom-up or iterative top-down search. The main difference between the searches is how the locality measure used is determined in the evaluation of the subspaces (Parsons et al., 2004). The simultaneous keyword identification and clustering of text documents (SKWIC) algorithm (Frigui and O. Nasraoui, 2004) is associated with the bottom-up methods whilst the adaptive subspace iteration (ASI) (Li et al., 2004) employs top-down search.

SKWIC is a unsupervised clustering algorithm based on cluster-dependent keywords weighting. The algorithm automatically identifies clusters that are the most dissimilar in their best keyword sets and assigns different weights to the keywords used in each cluster (Fountain et al., 1991). The algorithm locates clusters by using a special keyword set rather than the entire keyword space. Furthermore, it benefits from richer feature relevance representation by not tolerating the terms equally, which means that a term can have different weights in different clusters. The experiments conducted demonstrate that the feature relevance of SKWIC is very high and reflects the general theme of the category (Frigui and O. Nasraoui, 2004).

ASI allows explicit modelling of the subspace structure associated with each cluster. It achieves data reduction by assigning data points to a cluster and conducts simultaneous subspace identification by identifying the subspace structure associated with each cluster. However, the data points and their attributes are utilised in an iterative optimisation manner for achieving canonical duality contained in the point-by-data representation. ASI performs bet-

ter than the k-means algorithm by achieving high clustering accuracy and meaningful description of each cluster.

2. 3. Similarity-based document clustering

In this section clustering methods and techniques that employ words context and/or rely on external knowledge for feature extraction and aggregation are reviewed. The overall aim is to exclude from document representations those words that are irrelevant to their theme by employing word sense disambiguation (WSD) or enriching their representation with concepts providing more abstract indexing (Peng and Choi, 2005). Once the features are selected, the algorithms use hierarchical or partitional approaches, or simple heuristics (Hotho and Staab, 2003).

2.3.1. Word Sense Disambiguation (WSD)

Identification of the words relevant to the document theme can be achieved by employing WSD, which addresses the issue of words polysemy (Ide and Veronis, 1998). Disambiguation algorithms use a variety of resources such as external knowledge resources and supervised or unsupervised techniques (Dwivedi and Parul, 2009). Supervised WSD methods utilise a labelled training set by training the sense detection model on a sense-tagged corpora. By linking contextual features to the word's sense, WSD is reduced to a classification problem. These methods need prior tagged corpus or interaction with an operator and are therefore excluded from the scope of this thesis. On the other hand, unsupervised methods identify patterns in large data sets, without the benefit of using manually tagged data. They group patterns together, so that patterns within one group have more in common than patterns in other groups. Unsupervised approaches are very powerful and scalable as they require little compu-

ting time. The Latent Semantic Analysis (LSA) employed by the Latent Semantic Indexing (LSI) in the area of document retrieval is one of the unsupervised methods of particular interest to this research.

The knowledge-based approach to WSD uses machine readable dictionaries (Lesk, 1986, Cowie et al., 1992), thesauri (Yarowsky, 1992), ontologies (Hotho and Staab, 2003), lexicons (Leacock and Chodorow, 1998) or heuristics. These methods capture words' meaning by matching their context to external sources. Some techniques, such as those based on machine readable dictionaries, are impractical (Cowie et al., 1992) due to their high complexity. Since words are typically replaced with their definitions in the document representation this approach is suitable for disambiguating single words only.

The thesauri-based approaches have practical scalability and high accuracy. They overcome the knowledge acquisition bottleneck by using the semantic structure of thesauri (Yarowsky, 1992) and exploiting the explicit synonymy relations between words' meanings, which allows dealing with polysemy. Other research using thesauri is based on the observation that polysemous words that appear more than once in text share the same meaning (Gale et al., 1992), although that is not always true (Wan, 2007). The similarity measure between words is based on their distance in the semantic structure of the thesaurus used and is therefore called semantic distance (Jarmasz and Szpakowicz, 2003b).

Lexicons organise the mental vocabulary in the speaker's mind. Lexicon-based word sense disambiguation algorithms use semantic similarity resemblance between concepts that words in text belong to. Then, they acquire the semantic relatedness between them. This approach is scalable and achieves high precision (Dwivedi and Parul, 2009). The semantic similarity measure used is based on path length (Wu and Palmer, 1994), the shortest path between two

concepts (Leacock and Chodorow, 1998), or on the information content relatedness (Resnik, 1995).

An heuristic approach to WSD that uses "all concepts" is proposed by Hotho et al. (2003a). Instead of discriminating words' senses, this approach generates alternative document representations based on words' meaning (Hotho and Staab, 2003). As a result of using a background knowledge encoded into a domain specific ontology, this method provides multiple output views and a very specific perspective to documents. A similar approach, which uses both a general and a domain-specific ontologies, is used for large scale concept indexing of web pages (Setchi and Tang, 2007). It considers all possible meanings of the words whereas a word with multiple meanings shares its weight (significance) equally among the concepts it belongs to. In the end, the weight of every possible concept is calculated and the highest-ranked ones are used in a concept index. Similarly, a method that employs ontology to "enrich the term vector with concepts" is proposed by Hotho et al. (2003a). The approach significantly reduces dimensionality and computational complexity; it provides better scalability and improved clustering by using concepts in VSM representation (Hotho et al., 2003b, Hotho et al., 2003a). It also provides a generic perspective in establishing the similarity between topics.

2.3.2. Document representations

Different assumptions made lead to different approaches to feature selection and extraction, which result in different document representations and different preparation processes. Well-selected features contribute to reduced dimensionality and noise in the final document representation. They also allow easy browsing and navigation of document collections. Re-

duced dimensionality diminishes the computational requirements, whilst reduced noise improves clustering efficiency (Fung et al., 2005, Li et al., 2008, Mugunthadevi et al., 2011).

Document representation can employ either external information source or rely on heuristic rules for feature extraction from the documents' context to represent them. The multi-word approach to document representation captures words context using statistics or linguistics approaches (Zhang et al., 2009), but the algorithms do not employ external knowledge. The idea is that a word is characterised by "the company it keeps" (Firth, 1957). Any method that uses a sequence of two or more words with meaningful content is a multi-word approach (Chen et al., 2006). Zhang et al. (2011) use syntactic rules to extract the context (as 2 to 6 word sequences) that resembles a predefined regular expression, and count its co-occurrence in the document. Then a similarity function uses multi-word representations of the documents and seeks through the collection counting the occurrences of these sequences in the documents. The study concludes that multi-word performance is strongly dependent on the type of genre, and the approach is effective in narrow sub-domains, where fixed expressions and terminology are used. This indicates the robustness of the approach when proper heuristics or rules are used. However, this method is dependent on the wording of documents.

LSI is another method for capturing words context and reducing the document representation dimensionality by using word co-occurrences (Deerwester et al., 1990). It uses the LSA technique, which is developed to retrieve documents on the basis of their conceptual content instead of their meanings and is successfully used in WSD (Katz and Pinkham, 2006). In contrast to the multi-word approach, LSI uses external knowledge, which provides an implicit higher-order structure and relations between terms ("a semantic structure") within the documents aiming to find its best subspace approximation. The use of semantic structures allows users in different contexts or with different needs, knowledge or linguistic habits to retrieve

relevant information using different terms (Berry et al., 1995, Dwivedi and Parul, 2009). LSI resolves synonymy and polysemy but at high computational cost.

A comparative study conducted by Zhang et al. (2011) on text representation achieved by TF-IDF, LSI and multi-word approaches claims that the multi-word approach and LSI have better semantic quality, while TF-IDF has better statistical quality. The evaluation of text representation in term of semantic and statistical quality is conducted by intuition rather than systematically, using measures. The reason is the lack of standard data set or suitable measures for evaluation. LSI produces index, which achieves better discriminative power and performance over the index produced by TF-IDF. The authors evaluate the performance and the robustness of the techniques and conclude that LSI and the multi-word approach outperform TF-IDF (Zhang et al., 2011).

2.3.3. Similarity measures

Several document similarity measures have been proposed in the literature. They include the cosine measure (Salton and Buckley, 1998), the dice measure, Jaccard measure, the overlap measure (Blair, 1979, Baeza-Yates and Ribeiro-Neto, 1999) and the information-theoretic measure (Aslam and Frost, 2003). However, cosine similarity is very robust (Dhillon and Modha, 2001) and used most (Wan and Peng, 2005b). All these similarity measures define similarity between two documents as they are positively related to their commonality and negatively related to their differences in a common feature space (Lin, 1998b). The similarity-based approach seeks commonality in the shared context and themes of the documents, taking advantage of their structure and subtopic distribution (Wan and Peng, 2005b, Wan, 2007).

A method called TextTiling is proposed for capturing the document structure by subdividing texts into multi-paragraph units that represent subtopics (Hearst, 1993). Text tiles are used to capture the lexical pattern distribution of subtopics contained in the text. The approach uses three algorithms: (i) lexical analyses based on TF-IDF, (ii) information retrieval measurement to determine the extent of the tiles, and (iii) a statistical disambiguation algorithm which relies on thesaural information. This method provides segmentation that is aligned well to human judgments (Hearst, 1997).

A document similarity search algorithm that employs the TextTiling technique to capture the document subtopic structure in plain text, find documents similar to a given query document and return a ranked list of similar documents, is proposed by Wan and Peng (2005b). The similarity model takes into consideration the structure of the document subtopics and computes the similarities between different pairs of text segments. Then the overall similarity between the documents is measured by combining the similarities of different pairs with the optimal matching (OM) method. Experimental results show that TextTiling is effective and performs better than the cosine measure, which also reveals that the OM-based matching is appropriately applied.

Wan (2007) argues that a subtopic can be matched to more than one topic but with different weight and the one-to-one matching is limiting compared to many-to-many matching. Such matching is proposed by Wan and Peng (2005b) in response to the need for measuring the semantic similarity between any two words based on a lexical database. The semantic distances between words measured by a context vector using WordNet establishes relatedness between the words by calculating the angle between the vectors (Patwardhan, 2003). The author further states that the semantic distance measured by combining statistical information of the words derived from a large corpus and external knowledge produce clusters close to hu-

man judgement, because, unlike other methods, this method considers the context of the words. This approach can be applied to any domain and number of documents and has no constraints on the kind of words processed (nouns, verbs etc), which is a problem for some other approaches (Resnik, 1995). The context vector approach considers words with the most similar senses. The more similar two senses are, the smaller the semantic distance between the words is. Thus, the semantic similarity between two documents relies on measuring the distribution of the semantic distance between the words containing them.

Wan and Peng (2005a) propose using the Earth Mover's Distance (EMD) (Rubner et al., 2000) for measuring similarity between documents using "many-to-many" matching. The matching computes dissimilarity between two multi-dimensional distributions in a feature space (words). EMD uses a distance measure between every two single features called ground distance and defined by a function or matrix, to measure the distance between two multi-dimensional distributions. Wan and Peng (2005a) use a function to measure the semantic relatedness between words based on the WordNet structure. However, a thesaurus-based matrix distance (Jarmasz and Szpakowicz, 2003b) or ontology-based concept tree distance function (Lakkaraju et al., 2008) can be utilised instead. Furthermore, a custom similarity function can be used to measure the distance between any two features using an external information resource and the semantic distance quality will depend on the quality of the external source.

A comparison study evaluates similarity measures based on words and phrases, such as the cosine, dice, and Jaccard similarities, overlap and information-theoretic measures, and compares them to context-based similarity measures such as those based on OM and EMD (Wan, 2007). The OM-based and the EMD-based approaches in the study use the TextTiling algorithm to decompose documents into subtopics (a.k.a. tiles). The author adopts non-

interpolated Mean Average Precision (MAP) and the precision (P) at the top N results (P@N) to evaluate the different measures. According to the study, the context-based similarity measures provide better accuracy than those using statistical measures, with EMD outperforming OM.

The same (Wan et al., 2007) study further analyses the context-based approaches since they rely on the subtopic structure of the documents. The author investigates how documents structure influences their performance. A hierarchical agglomerative clustering algorithm, which groups sentences with similar subtopics together to form a document, is employed to produce structurally different sets of documents. The empirical results show that the execution time of the cosine measure is 3.49 times shorter than the OM-based and 3.68 times shorter than the EMD-based approaches. This is explained by the complexity of the structure-dependant algorithms where a graph structure has to be built and mathematical computations to be completed. Deeper analysis reveals that the EMD approach performs more accurately than the OM approach in all cluster sets. This experimental evidence supports the independence of the EMD approach from the employed text decomposition technique (Wan, 2007). However, the fact that the OM approach produces close results to human judgement by taking advantage of the words' context (Wan and Peng, 2005b) leads to the conclusion that the results produced by the EMD-based approach are similar. Moreover, the EMD-based algorithms rely on measuring the similarity between two multi-dimensional distributions of subtopics to calculate the similarity between two documents (Wan, 2007). Therefore, employing EMD as a similarity measure in clustering may be the key to improving the clustering solutions in terms of their coherence with human judgement. A similar hypothesis was discussed by Patwardhan (2003).

Another technique that considers document context to measure similarity relies on ontology. The similarity measure uses a variety of methods such as: (1) similarity between the properties of concepts; (2) semantic distance between concepts; (3) hierarchy depth of the concepts; and (4) domain dependant adjustment of weights (Yang et al., 2008). The first method computes the number of properties every two concepts share. The more properties they share the closer they are. The second method computes the distance based on the shortest distance between two concepts using thesauri or lexicons (Jarmasz and Szpakowicz, 2003b). The hierarchy depth factor considers the depth of the ontology tree. Hence, the shorter the distance between two concepts, the greater the semantic similarity between them. This method for measuring similarity is considered to represent how abstract the measured similarity between the concepts is. Finally, the last method considers domain-dependant adjustment of words or concepts weights. This method enables increase of the semantic similarity of concepts that occur in an auxiliary ontology. This method is used by Setchi et al. (2009) to increase the weight of concepts that occur in a domain-specific ontology used along side a general ontology. The relations between concepts in the specific ontology augment the similarity measured with the help of the general ontology.

2.3.4. Clustering techniques

Clustering relies on a document representation, a similarity measure and a clustering technique to group documents in clusters. Clustering methodology faces the practical problems of dealing with a high computational cost causing implications on a large scale, high dimensionality, and complex similarity measures (Yang et al., 2008).

Distributional clustering techniques address the aforementioned problems by considering the distributions of the words in documents and in a collection. Distributional clustering is

rather addressing the high dimensionality problem by reducing the dimensions than seeking different feature extraction approach. It provides a mechanism for feature reduction of the original feature space, transforming it into a space of new features represented by word clusters (Baker and Mccallum, 1998). The typical similarity measure employed by the distributional techniques is based on the information theoretical divergence criteria (Lin, 1991). Distributional clustering techniques, such as the information bottleneck, aim to provide a more compact representation of the data by maintaining maximum mutual information between the joint probability distribution of two variables by "compressing" one of the variables (Slonim and Tishby, 2000, Slonim et al., 2002). The results of using the information bottleneck algorithm demonstrate that the produced clusters are inferior unless word clustering is employed (Slonim and Tishby, 2000).

Contextual document clustering (CDC) is a clustering technique, which uses the context words to address the problem of complex similarity measures. Word context is a term, which describes the probability distribution of a set of words that co-occur with a given word in a document. The approach is based on distributional clustering and identifies documents, which belong to highly specific contexts (Mcdonald et al., 2004). It relies on the distribution of the subject related words, with a narrow context, and uses them as meta-tags for that subject. These words, called contextual words, form the basis for creating thematic clusters of documents (Baker and Mccallum, 1998) by providing a mechanism for grouping semantically related documents together (Mcdonald et al., 2004).

The purpose of the contextual document clustering is not to provide a compact representation of the documents but a mechanism that automatically, in unsupervised manner, discovers contextual words of narrow scope. Then documents are partitioned by using the context words without any use of pre-defined categories or labels into a large number of relatively

small thematic homogeneous clusters. Clustering is completed regardless of the clustering criteria used (Mcdonald et al., 2004). The large number of clusters is a reflection of the complex thematic structure of the text documents, which cannot be adequately expressed by classifying documents into a small number of categories or topics. Contextual document clustering has lower complexity and experimentally is proven to be applicable on a large scale. It demonstrates high quality of clustering and coherency over time (Rooney et al., 2006). The algorithm organises documents into a minimum spanning tree enabling their topical similarity to be assessed. In addition, the contextual document clustering technique is proven to be suitable for identifying important and stable themes.

In the literature, specific ontologies are created automatically from text to capture relations between words and their context (Lin, 1998a, Lin, 1998b, Khan and Luo, 2002, Khan et al., 2003, Lee et al., 2007). The established relations in an automatically created ontology are specific for the collection of documents that is used to create the ontology. Since an ontology defines basic classes of words in the domain of knowledge, it also outlines their semantic structure (Gruber, 1993). Therefore, a connection exists between the concept meaning of a word and its semantic structure. The relations between words in an ontology are used by Hotho et al. (2001) to enrich the original term vector with concepts before employing a model-based clustering algorithm. The ontology-based approach performs better than a baseline approach when WSD and feature weighting are used. For that reason, there are different strategies for compiling ontology into text representation focusing on concepts, disambiguation and hypernyms (Jing, 2008). The concept-based strategies for using ontology in document representation include adding related concepts into the term vector, replacing terms with more general concepts or replacing terms with concepts only. The problem with the last strategy is that all statistical data for terms co-occurrence is discarded. However, Setchi et al.

(2009) suggest a document indexing approach that uses a concept vector and relevant concepts weight, which is computed using statistical term co-occurrence data. This approach has never been tested for document clustering, but demonstrates good retrieval results.

Other algorithms employ words polysemy by providing alternative document representations. These algorithms use different words meanings in document representations and exploit various relationships that exist between word senses (Hotho and Staab, 2003, Yang et al., 2008). This alternative representation is supported by the view that a document refers to multiple topics and it is important to avoid confining it to a single cluster (Hearst, 1999). An algorithm that uses ontology-based heuristics rules in feature selection and aggregation is proposed by Staab and Hotho (2003). In addition to suffix stripping, the authors group the words into sets of synonyms and calculate their similarity. The pre-processing allows users to select between the results, whereas the selection constructs alternative text representations, using different background knowledge. This method is also known as COSA (Concept Selection and Aggregation). Once a group of synonyms is selected, a corresponding text representation is aggregated. The k-means algorithm is employed then to conduct clustering. Thus, different clustering results are explained by using different selections of words meanings, i.e. different concept relations from the used ontology. As a result, the proposed method performs better than the k-means algorithm without word groupings in the index aggregation step.

Another approach in support of Hearst's (1999) idea of different perspectives, i.e. multiple views, to a document collection is proposed by Yang et al. (2008). It is based on Slonim and Tishby's (2000) technique of pre-clustering of the words in the collection and then the clusters formed to guide document clustering. The authors use words relations explicitly defined in an ontology to compute word similarities and by using the measured similarities to group words in clusters. Unlike the algorithm proposed by Slonim and Tishby (2000), where words

are grouped by their properties, Yang et al. (2008) group words using their relations in the ontology. The experiments demonstrate that the ontology-based method has better precision and F-score, whilst the term-based method has higher recall. The authors explain the results with the type of corpus used by them, where the words are semantically related to each other. They conclude that if a general corpus with a large variety of topics is used, the recall of the term-based approach will significantly deteriorate.

2.3.5. External semantic source

In this section, the hierarchical structure of the ontology used in this thesis namely OntoRo is presented. The semantically enhanced feature extraction algorithm, which originated from concept indexing (Setchi et al., 2009) and is discussed in detail in chapter 4, employs the same lexical knowledge source. In addition, the algorithms for text normalisation and document representation proposed in chapter 4 employ the same ontology for the reasons outlined in this section.

The most commonly used general lexical sources, which in the context of the research presented in this paper are called lexical ontologies, are WordNet and various thesauri. WordNet is a large lexical database of English words (Miller et al., 1990), which groups words into sets of cognitive synonyms called synsets. Every synset expresses a distinct concept. Concepts can be interlinked by means of a conceptually established semantic link, or defined lexical relation. The resulting structure is a network of meaningfully related words and concepts that superficially resemble a thesaurus.

WordNet and the thesauri group words together, based on pre-defined criteria, and have important distinctions. Firstly, WordNet interlinks not just the word forms, but also their specific senses. As a result, the words that are found in close proximity to each other in the net-

work structure are semantically disambiguated. Secondly, the semantic relations between words in WordNet are labelled, whereas the groupings of the words in the thesauri follow the explicit pattern of being grouped by the similarity of the ideas they express. The words in the thesauri are grouped into concepts (fig. 2.2). In this context, the concepts represent entities that refer to broad ideas, which are used to group the words together. Therefore, the concepts and the pattern followed to group words in WordNet and in the thesauri are different, as the thesauri's concepts have a more generic structure.

Therefore, by taking the thesaurus structure into account, we see (i) a tree such that every class is a root of a separate tree (ii) that provides a very rigid and robust structural organisation and (iii) the generic nature of the conceptual organisation of the words. Therefore, the OntoRo is selected as an external knowledge source. It must be noted that the thesaurus structure provides tree organisation of the concepts it contains, whilst words might be linked to concepts that belong to different trees. Concept indexing takes into account the tree-based structural organisation of the concepts. In addition, the reduced number of concepts in the OntoRo will address the spectral requirement of the clustering. The hierarchal structure of the thesaurus is shown in figure 2.2.

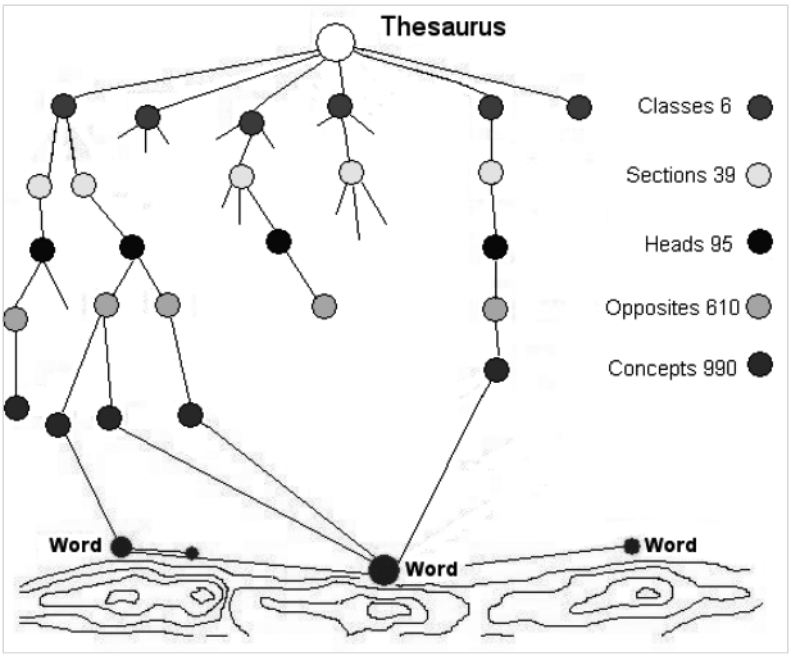


Figure 2.3 . The hierarchical structure of a thesaurus is resembled by the OntoRo

In the context of this research, and as a result of analysing the differences between WordNet and the thesauri, it is concluded that the thesauri are the lexical source of more generic knowledge in terms of grouping words by ideas, and not by synonymy-based semantic relations. The specific conceptual interconnectivity of WordNet pre-defines specific relations between the words represented in the semantic-based network structure. Any other knowledge source used manifests different word organisation and hence, different result.

2. 4. Summary

This chapter has reviewed model-based and similarity-based document clustering methods and techniques as well as similarity measures. It has also outlined different feature extraction and aggregation methods used in document representation. The review has highlighted the following points:

1) Clustering solutions, produced by clustering algorithms that employ cosine measure in a collection of documents with various topics, are found to be inconsistent and poorly aligned to human judgment. This limitation can be overcome by using words' context and words' meaning in measuring document similarity.

2) Document index aggregated by using ontology enables semantic relations between words to be considered. The aggregated index reduces dimensionality of document representation and clustering solutions produced with it are closer to human judgement. A change of the ontology used to aggregate document index, changes clustering solutions towards the specificity of the newly used ontology.

3) The Document clustering domain currently lacks a methodology that employs document structure for measuring document similarity and many-to-many similarity measure on a large scale.

4) In order to improve the quality of the clustering solutions and make them consistent with human judgment, traditional clustering algorithms need to provide multiple views to document collection, i.e. multiple clustering solutions.

5) A clustering methodology that incorporates reduced dimensionality of document representation, provides multiple views to document collection, produces clustering solutions close to human judgment, and is scalable for a large scale clustering is needed.

Chapter 3 : Conceptual model of semantically enhanced document clustering

This chapter presents problems and limitations of the current state-of-the-art clustering methods and techniques. Then, semantic-based approaches to clustering are discussed as pre-requisites for obtaining clustering solutions that align well to human judgment. Particular attention is devoted to methods and approaches that overcome limitations of the current algorithms. Finally, an enhanced semantic-based conceptual model for document clustering is proposed.

3. 1. Limitations of traditional document clustering

Traditional document clustering has the limitations to produce clustering solutions that are inconsistent and poorly aligned to human judgement, as a result of not considering user's information needs, and to use computationally expensive and restrictive similarity measures in order to improve that alignment of the results across domains.

3.1.1. Clustering solutions generated are inconsistent and poorly aligned to human judgement

The main task of document clustering is to discover groups of documents, which represent topics contained in a document collection. The main limitation of the current state-of-the-art methods is that the most meaningful grouping is not always produced (Andrews and Fox, 2007). Clustering solutions produced by the traditional approaches do not meet the expectations of users who retrieve documents, search document collections, and explore different domains of interest. This thesis considers the main reason for these limitations to be that users do not have control over behind-the-scene grouping process, which forms the clusters, e.g.

the centroids in clustering solutions produced by the k-means algorithm are randomly selected (Andrews and Fox, 2007).

Traditional clustering algorithms provide clustering solutions closer to human judgment when they are used to group documents belonging to the same domain of knowledge. However, knowledge cannot be limited to a domain of strictly pre-defined number of documents (Burkey and Kuechler, 2003). Often, the information users seek is scarce and is found in various information sources and domains (Sánchez et al., 2011). Therefore, clustering algorithms need to produce well aligned to human judgment clustering solutions from documents with various topics belonging to different domains. The fast growing number of documents freely available to users emphasise the importance of finding a solution to that problem.

Current research investigating different similarity measures, which are used to enhance clustering performance in cross-domain environment (Wan and Peng, 2005b, Wan, 2007), indicate that the clustering improves when a similarity measure able to identify relations between documents that are otherwise omitted by the traditional measures, is used (Andrews and Fox, 2007).

3.1.2. Document similarity across domains

The effectiveness of the clustering algorithms in the context of the reported research is measured through the consistency of the automatic clustering solutions generated in relation to human judgement. Clustering effectiveness is impeded by the limitation that results obtained on one corpus are not necessarily reproduced on different corpora. Therefore, clustering algorithms experience difficulties across domains or in a collection of documents with various topics (Steinbach et al., 2000, Andrews and Fox, 2007, Sánchez et al., 2011). As stated in the above studies, this problem can be explained with the specific requirements every

domain has on the number of clusters, properties and relations between documents. For example, the properties of a document are used by similarity functions to measure similarity between documents prior to clustering. On the other hand, properties utilised by a document representation technique on one corpus may be irrelevant when applied to different corpora. The requirements towards the solutions change across domains and for that reason the clustering produced for different collections differ from each other in their alignment to human judgement. This limitation outlines a need for a clustering methodology that can perform equally well on different document repositories.

Similarity measures that considers words' meaning (Hotho et al., 2003b) and documents' context (Hearst, 1997) compute closer to human judgement similarity between documents. These similarity measures overcome the limitations of the traditional ones since words' meaning acquired is relevant for the context of document the words occur. Therefore, the change of domain is captured and relevant similarity between documents is measure. However, these algorithms are computationally expensive and restrictive because they require use of external knowledge source. As a result of these constraints these similarity measures are not used on a large scale.

The quality of the labels assigned to clusters by the current state-of-the–art algorithms needs improvement. Current labels generated set limitations to neither facilitate document browsing nor provide acceptable description of the clusters (Andrews and Fox, 2007). As a consequence of the poor labelling, users cannot navigate efficiently between clusters. This indicates a need for better organisation of the clustering solutions that help users to understand document groupings better and provides ease of browsing. The poor labelling is in the focus of several research studies investigating different ways of computing the similarity between documents, which is otherwise undetected by traditional approaches (Hotho et al.,

2003b, Wan and Peng, 2005a, Andrews and Fox, 2007). Traditional similarity measures typically consider words' order and frequency to measure similarity (Hammouda and Kamel, 2004, Eissen et al., 2005). An improvement in accuracy is achieved by utilising word synonymy (Hotho et al., 2003a). Nevertheless, clustering solutions obtained using traditional similarity measures achieve alignment to human judgement, which does not exceed 40% (Lee et al., 2005). Therefore, improving this alignment is considered in this research a key to a more efficient clustering.

3.1.3. Meaningful clustering solutions

A meaningful grouping of documents is a clustering solution that matches user's information needs and is easy to comprehend. Instead these needs to be taken into account by the clustering, documents are grouped together by algorithms, which follow a certain model of knowledge. The model followed is relevant for a domain and therefore, a limitation of the clustering solutions produced across domains by traditional algorithms is that they are not well aligned to human judgement. The limitation of poorly aligned clusters to human judgment is explained in the literature with the fact that users with greater understanding, i.e. richly structured knowledge, have the characteristics of more experienced users, e.g. domain experts, whilst inexperienced users have the characteristics of a novice (Chi et al., 1981, Chi et al., 1988, Glaser, 1991). The domain experts can foresee relations between concepts, e.g. facts, event, and objects. On the other hand, the less experienced and/or knowledgeable users struggle to establish such relations (Novak, 1990, Wandersee, 1990). This yields a need for multiple viewpoints to clustering, which to consider different relations between facts, events and objects. This thesis advocates the view that more comprehensive clustering solutions will

be provided by assigning more control to users in clustering. The involvement of user in the clustering is a step forward to satisfying user's personal information need (Wan et al., 2010).

Clusters produced by the current state-of-the-art algorithms, which do not consider users need, are inferior compared to human judgement (Lee et al., 2005). This inconsistency of grouping documents can be explained by the observation that "clustering is ultimately in the eye of the beholder" (Estivill-Castro, 2002), i.e. the individual understanding of users (Norvig, 1987, Mccarthy, 2009). Therefore, document clustering needs to consider the understanding of users in order to improve clustering accuracy and effectiveness (Sánchez et al., 2011). This thesis considers efficient clustering to be the one that provides clustering solutions consistent with those produced manually by people using their judgement.

Traditional partitional clustering approaches produce a pre-defined number of clusters, which are unlikely to be meaningful to the users. The effectiveness of the produced clustering solutions depends on the aggregated document index and the similarity measure used (Fung et al., 2005, Li et al., 2008, Mugunthadevi et al., 2011). Different clustering solutions are produced when the index representing the document collection is modified or different similarity measure is employed. An index is aggregated following a certain model of text representation (Salton and Mcgill, 1986) and/or assumptions made for the text (Hotho et al., 2003b). Therefore, acquiring a different/modified index is a time consuming and goal-oriented task.

3. 2. Requirements towards the methodology

This section discusses requirements towards techniques and approaches in document clustering that inform the development of a methodology that will provide meaningful document

groupings and intuitive browsing through multiple viewpoints and deterministic clustering solutions positioned in reduced number of dimensions.

3.2.1. Reduced Dimensionality

The approaches used by search engines and partitional clustering to organise documents are considered by this thesis similar in terms of grouping documents around a string of words. The information retrieval algorithms organise documents around a query of words, i.e. short and ambiguous text (Belkin, 2000, Jansen et al., 2000, Kelly and Fu, 2007). The clustering algorithms, on the other hand, organise documents around other documents. In contrast to the information retrieval approaches, the document clustering algorithms do not rank documents in clustering solutions by their similarity to the centre of the cluster (Craswell et al., 2006). However, a central document of a cluster can be considered as a word query since the query and the central document are used by relevant algorithms to organise documents around them by similarity. Therefore, if a central document is used instead of a query the returned documents will represent a cluster from the perspective of the search algorithm employed. The difference between a query and a document is that the latter is topically less ambiguous since it represents a complete idea.

The information retrieval algorithms benefit from low dimensionality of queries, i.e. a query contains a few keywords. Thus, the retrieval algorithms benefit from computationally more expensive algorithms. Reduced dimensionality is a step forward for clustering to employ more sophisticated algorithms for measuring similarity and partitioning documents into clusters. Dimensionality reduction techniques elaborate document representation and discard certain amount of statistical information by minimising the restoration error. The reduction of dimension is empirically modified and depends on the quality of the documents, i.e. words

used in the documents, domain of knowledge, and particular task. Therefore, a document index that represents documents in full feature space but still enables computationally expensive algorithms to be employed is required.

3.2.2. Multiple viewpoints to clustering solutions

The research presented in this thesis acknowledges the fact that there are many correct ways to group documents (Vladimir, 2002). This section considers that a clustering methodology, which produce multiple clustering solutions and offers to a user a choice to select a solution that meets her/his expectations of document groupings, will enable clustering solutions in close relation to human judgement (i.e. document groupings close to how a person would cluster the documents).

It is possible a document to belong to more than one cluster. In a scenario when different but close by meaning queries are submitted to a search engine, some documents are returned for more than once. This resembles to a certain extend the fuzzy clustering approach (Zadeh, 1965) where one document is likely to be a member of more than one cluster. Documents clustered by fuzzy clustering algorithms have a different degree of membership assigned to them for any particular cluster they are a member of. The degree of membership is considered to be the similarity of the document to the cluster. This indicates how close a document is to the centre of a cluster. As a result, if a document has higher ranking to one cluster than to another, then this ranked value represents its closeness to the particular cluster. Therefore, a document that causes least change in the cluster's coherence (silhouette value) (Kaufman and Rousseeuw, 2005), i.e. introduces least noise and distortion to the cluster, becomes a member of that cluster. A threshold for noise is an optional parameter used to increase the efficiency of clustering. However, in contrast to the information retrieval, clustering algorithms form

clusters using all documents from document repositories. In that cases, even when a document has very low similarity to any of the clusters it is added to the one, which has highest similarity to it. Thus, the newly added document introduces noise and reduces cluster coherency.

Introducing the notion of documents that introduce noise can extend the similarity between the in-formation retrieval and the clustering algorithms. Documents that introduce noise are not returned by the information retrieval algorithm since they do not share concepts with the sub-mitted query. On the other hand, partitional clustering algorithms have no technique or mechanism to recognise documents that introduce noise and they have to place documents in the clusters in which they introduce least noise. The hierarchical clustering would produce separate clusters for these documents and those algorithms have a mechanism in place to handle such documents, although they do not have specific name for them. Nevertheless, the coherency of that cluster is yet disrupted. The notion of documents that introduce noise in clustering solutions is introduced in this thesis to explain the poor consistency of clusters with human judgement and impeded document browsing across domains. The explanation is that these documents are not removed from the clusters and they introduce noise in the clustering solutions.

The notion of documents that introduce noise to a clustering solution could further provide an explanation why clustering solutions produced by traditional partitional algorithms perform better on documents from narrow and specialised domains than from domains with a wider variety of topics. The explanation is that such domains contain fewer documents, which introduce noise. Therefore, it is believed that if these documents are removed from the representation of a document collection, the clustering solutions will align better to human judgement, i.e. documents that share uncommon concepts with the rest of the documents in

the collection will be not considered and will be excluded from the clustering. The documents that introduce noise to a clustering solution are believed to be positioned far away from the centre of the cluster (together with other documents with low ranking in terms of their similarity to the centroids of the clusters). The documents that are not close enough to the seed documents introduce noise and a decrease in the silhouette value of the clusters. The conceptual model presented in section 3.4 employs semantics to detect the documents that introduce noise to a collection and remove them from the collection representation.

The practical problem of identifying documents that introduce noise is approached by this thesis from the perspective of different users, i.e. different groupings of documents are meaningful to different users. Users ("beholder's eye") recognise many meaningful groupings of documents according to their understanding, motivation and background (Chi et al., 1981, Chi et al., 1988, Glaser, 1991). Therefore, a threshold, which defines how close a document needs to be to the centre of a cluster to become its member, can be used. When a distance between a document and a cluster is greater than this threshold, then the document will be considered to introduce noise in to the groupings.

This threshold is further assumed to improve the alignment of clustering solutions to human judgement. The greater value the threshold has, the greater number of documents would be considered to introduce noise and excluded from the clustering solution. Thus, the documents left for clustering share more conceptual similarities, i.e. they are positioned closer to the centre of the clusters. As a result, the clusters produced would be better aligned to human judgement. On the other hand, a threshold with a low value, would lead to more files being clustered. In this case, the clusters generated should provide broader, less consistent, clustering solutions, which differ substantially from human judgment. The average distance of documents to the centre of a cluster measures its coherence. Therefore, small threshold values

will refer to clustering solutions that contain more noise. This will make these solutions inferior to human judgement.

3.2.3. Consistent to human judgement clustering solutions

This section reports techniques and approaches in both document clustering and information retrieval. The algorithms in both domains automatically acquire information encoded in documents and retrieve or group them together. A successfully completed task is considered the one which groups documents in close relation to human judgement.

The information retrieval algorithms retrieve documents by comparing their similarity to a query of words. In contrast, the clustering algorithms group documents in clusters by measuring their similarity to central documents. These central documents are called centroids or medoids as the name depends whether the algorithms employed are deterministic (i.e. PAM and its variant) or non-deterministic (k-means and it variants). The process of measuring the similarity between documents yields a distance matrix (i.e. $O(n^2)$ complexity) for all documents in a collection. Clustering algorithms follow the procedure of adding a document to the cluster in which the least distortion of its coherence is introduced. Nevertheless, the use of traditional one-to-one similarity measure proves ineffective in terms of their alignment with human judgment (Lee et al., 2005) and alternative methods for measuring document similarity on a large scale are required.

This thesis considers a problem the fact that partitional clustering algorithms split all documents from a collection into a pre-defined number of clusters. Therefore, the alignment of the clustering solutions to human judgement depends on k and on the user's understanding, motivation, background and experience. As a result, the algorithms performance deteriorates if the number k is not properly selected (Steinbach et al., 2000). Although there are methods

which estimate the most appropriate number of clusters via indirect measures such as silhouettes (Rousseeuw, 1987), they do not remove the documents that make the clustering solutions worse from the representation of the collection that has to be clustered. In case these documents are removed from the collection's representation, which is used to generate clustering solutions, the quality of the results will improve. Information retrieval algorithms do not return documents that do not resemble enough similarity with a query submitted. This benefits the scalability of the algorithms, since they work with fewer documents. In addition, the information retrieval algorithms benefit from reduced dimensionality of the queries, i.e. a query usually contains a few keywords.

This thesis presents work towards identifying documents that introduce noise to clustering solutions and recognises the fact that many correct ways exist to organise documents into clusters. Therefore, any grouping of documents could be better or less well aligned to user's judgement. This fact is addressed by introducing the notion a level of abstraction, which is a conceptual organisation of cognitive structures that enables a knowledge representation in a wide range of data granularity. It is aimed to provide insights of how clustering results change when the level of abstraction, used to identify heterogeneous documents, is modified. The level of abstraction defines the perspective from which users perceive/understand the clustering solutions. A low level of abstraction refers to narrow and very specific clusters, i.e. a high-threshold value is needed to identify documents that would worsen clusters prior to clustering. A high level of abstraction is defined by a small value for the threshold, which allows more documents with a higher variety of topics to be used in the clustering. The use of a greater number of documents will include, to certain extend, more documents that would spoil the produced solutions. Then the resulting documents in the clusters would be topically more diversified and the clusters produced would be more inferior.

3.2.4. Meaningful clustering solutions and intuitive browsing

Although traditional clustering algorithms produce clusters, which demonstrate good results according to the typical measures of quality used, clustering solutions generated are neither intuitive nor clear (Andrews and Fox, 2007). Clustering systems need to enable users to browse and to navigate between documents and clusters efficiently (Wan et al., 2010). For that reason, a requirement to clustering systems is to facilitate users in selecting the clustering solution, which is useful to them. The selected solution is considered a meaningful grouping for that user, who will be able to browse the documents and the clusters more intuitively.

The use of levels of abstraction that aligns clustering solutions well to human judgement will also help users control the process of creating meaningful clusters. The threshold value, i.e. the level of abstraction, will be adjusted until document groupings produced by traditional clustering algorithms align well to user's judgement. The level of abstraction supports the first requirement (3.2.1) to a methodology by enabling traditional algorithms to produce multiple clustering solutions for different levels of abstraction. In this process user's personal background, motivation and understanding will be important. In addition, this technique requires no preliminary information such as profile preferences that pre-defines a certain level of abstraction. Users will modify clustering solutions via increase or decrease of the threshold value until document groupings start being understandable to them. Thus, users will be provided with a mechanism to produce different groupings using the same index aggregated in the document representation stage. Each grouping will differ from the others by the number of documents removed from the collection, because they are considered to introduce noise to the clustering solutions. This technique will enhance document browsing between and within clusters making the navigation in the document collections more intuitive and efficient. The

clustering effectiveness will improve as a consequence from better understanding of document groupings and the reduced number of documents in the clusters. Once a suitable level of abstraction for a particular user is achieved, the centroids of the clusters can be used as their labels. These documents will then provide more detailed labelling on the clusters since they are not as ambiguous as the keywords contained in labels.

The notion 'levels of abstraction' is explained in this thesis from the perspective of the cognitive science with different relations between documents. They can be established by using various associations that exist between objects, events and facts. A combination of these associations is considered to represent users' personal understanding for these objects, events and facts that are conveyed in documents. Therefore, it is important to acquire various knowledge structures from documents, which to be used to measuring the pair-wise similarity between them (Hearst, 1997, Xia and Lewis, 2007). Different levels of abstraction allow establishing relations between documents that can change over time. This is important characteristic since the understanding of the user also changes in time. A change in the relations between objects, events and facts leads to a change in the measured pair-wise document similarity (Barsalou and Neisser, 1987). Therefore, the change of the user's understanding triggers a need to change the document groupings. In the same way in the domain of information retrieval the query submitted to a document retrieval system changes over time until documents returned to the user contain the needed information (Lin et al., 2006). Replacing, adding or removing words from it until the documents advance to a meaningful grouping that suits user's needs changes. Thus, the users have control over the results and they are involved in the process of producing meaningful (Hotho et al., 2003c, Shih et al., 2011) document groupings.

3.2.5. Deterministic clustering solutions on a large scale

A disadvantage of the fast partitional clustering algorithms, e.g. the k-means algorithm and its modifications (Steinbach et al., 2000), is that they are non deterministic and provide multiple clustering solution by creating clusters around randomly selected documents, i.e. the centroids. On the other hand, the deterministic partitional algorithms such as PAM (clustering around medoids) (Kaufman and Rousseeuw, 1990), have high computational complexity and are impeded to perform on a large scale due to high dimensionality of text. Both partitional approaches need in advance the number of clusters, which reveals the underlying knowledge structure of the domain of interest. Nevertheless, one of the main tasks that clustering has to accomplish is to find unseen before relations between documents by grouping them together. Deterministic algorithms provide clustering solutions, which can be reproduced retrospectively for known k. On the contrary, non deterministic partitional clustering algorithms produce non deterministic clustering solutions and reveal existing relation between documents, but they perform poorly when k is selected regardless of the knowledge in the domain of interest. Both approaches produce clusters that do not demonstrate good alignment with human judgement (Lee et al., 2005). Therefore, a requirement to the model is to produce deterministic clustering solutions aligned well to human judgement on a large scale for a wide range of numbers of clusters.

For that purpose the model is required to reduce dimensionality of document representation and employ deterministic clustering algorithms. The advantage of deterministic algorithms such as PAM is that they are more robust than the standard k-means algorithm since they minimise a sum of dissimilarities in clustering instead of a sum of squared Euclidean distances. The reduced dimensionality will address the main disadvantage of the deterministic clus-

tering algorithms which is slow performance in high number of dimensions (Kaufman and Rousseeuw, 1990).

3. 3. Towards advanced document clustering

Clustering analysis employs various techniques and methods to group documents in clusters. This section aims to explore the first and the second steps of the four-step clustering model (see fig. 2.1) (Xu and Wunsch, 2005). The former addresses document representation techniques and the latter approaches to measuring pair-wise document similarity. The clustering process is presented from the view point from which the limitations of the traditional clustering algorithms discussed in the previous section are overcome.

3.3.1. Advanced document representation

Documents are represented from features acquired in the feature extraction and selection processes. The used features are taken from a pool of candidates. The candidates are words selected or extracted from a document collection. A document is presented with a subset of all candidates only acquired from within the document. A weighting system in place indicates which words are the most representative ones (Lewis, 1992). There are no restrictions on the number of words that to be considered per document or what the minimal weight of a word needs to be in case not to be ignored by the feature selection algorithm. Therefore, all words in a document are usually used in its representation. Since, document are likely to contain large number of words dimensionality reduction techniques are employed to reduce the features (dimensions) to adjacent feature space with minimal lost of accuracy. The selected features (or components when projected in reduced feature space) represent documents (Jain et al., 1999, Jain et al., 2000).

The features (words or components) with the highest weight are very commonly selected to represent the documents (Xu and Wunsch, 2005). However, there are more advanced document representation techniques in the literature that improve clustering (Hotho et al., 2001, Yang et al., 2008, Zheng et al., 2009). These techniques use external knowledge such as ontologies (Hotho et al., 2003b) to acquire relationships between words that do not exist explicitly. The established relationships through the external knowledge source represent the background knowledge of a user (Hotho et al., 2003c). Thus, the semantic relations between words are used to cluster a document collection from the perspective of a potential user. Therefore, in document clustering an ontology reflects a particular viewpoint, which originates from the knowledge structure in the user's mind (Chi et al., 1981, Chi et al., 1988, Glaser, 1991) and is represented by the relations established in the ontology. For that reason, established relations in different ontology provides different perspectives to the clustering results in terms of user's understanding.

Researchers agree that document representation that employs external knowledge and exploits established relationships between words and phrases uses words meaning(s). Since an ontology represents a specific conceptualisation of a domain knowledge, the understanding of the ontology creator for the words' meaning is used in the clustering. Therefore, clustering algorithms are employed to cluster documents from the perspective of the user who has created the ontology.

The ontology used by Hotho et al. (2003c) is WordNet. It is a knowledge source built over a few years by a group of experts. A creation of a personal ontology is limited in terms of amount of information used to build and organise it. The process of building ontologies further set more constrains for the accuracy of the external source. The first reason is that ontology creation is challenging and time consuming process. The second reason is that such

knowledge sources, even as general as WordNet, are used to achieve particular result, i.e. the algorithms that use them are designed for a specific task and exploit specific relations between words. The third reason is that users' understanding is very complex and they experience difficulties to reproduce it in a formal conceptual manner (Goldsmith et al., 1991, Hsien-Hsun et al., 2005). The clustering algorithms that rely on external knowledge have a task to group similar documents by using common relations between words. The produced clustering solutions by ontology enhanced algorithms obtain better aligned results with human judgement (Hotho et al., 2003b). A different grouping of documents is produced once a different set of relationships, corresponding to different user's understanding, is provided.

User's understanding, which is defined by relations between concepts (Goldsmith et al., 1991), is an associative model of memory (Shavelson, 1974). The model outlines a network of concepts connected by relations that exist between them. Some of the concepts are connected via a direct relation whilst others are linked together via associative relation through one or more concepts. This model explains why experts, who possess richly structured cognitive structure, foresee relations between entities such as objects, events, and fact, whilst the inexperienced users do not. The memory structure is characterised by a distance between any two concepts. Therefore, user's formal conceptual understanding, i.e. represented by an ontology, can be used to produce a personal (ontology-based) distance function or a distance matrix that measures the distance between concepts (from within an ontology) (Jarmasz and Szpakowicz, 2003a).

In the document clustering domain concepts, i.e. generic entities that unite word terms under common criteria, are used to enhance clustering results by enriching document-term representative vectors with concepts (Hotho et al., 2003c) or to create a diversity of views from which to look at the clustering task (Hotho et al., 2001). The first approach targets improved

effectiveness by using fewer and more generic features in representing documents. The second approach targets better efficiency by reduced dimensionality and improved effectiveness by introducing subjective criteria which enable a diversity of views onto the produced clustering solutions.

The first approach for enhancing clustering employs adding concepts, replacing terms and concepts only techniques to achieve improved effectiveness. The first technique extends each term vector by adding new entries (concepts), which are relevant topics and appear in the document collection. Then documents are represented by concepts and terms simultaneously. The technique which replaces terms with concepts expels all terms from the representation vector for which at least one corresponding concept exists. This technique mixes concepts and terms in the representation vector, since terms that do not appear in the used external knowledge source are not discarded. The third technique uses only concepts and terms that do not appear in the external knowledge source. These techniques slightly improve clustering efficiency and effectiveness. The evaluation is carried out on the pre-categorised Reuters21578 corpus using standard measures such as purity and F-measure.

The second approach improves clustering efficiency by reducing dimensionality similarly to the first approach by replacing terms with concepts. The difference is that the latter reduces the dimensions in two steps and employs a simple, core ontology for restricting the set of relevant document features. The core ontology defines the background knowledge used for preprocessing and selection of relevant views (i.e. aggregations) onto the set of texts (Hotho et al., 2001). The first step in dimensionality reduction replaces all terms with concepts. The second step uses an “agenda”, which describes pre-selected features used in the concept vectors that represent a part/section of the core ontology. Thus automatic aggregations of features are generated by modifying the “agenda”. The experimental results of this approach

show that a structure in the clustering solutions can be found in a low dimensional space. In addition, the "agenda" provides the user with an explanation to a certain extent for the process of forming the groupings.

Another document representation approach which replaces some of the words with concepts in VSM representation is proposed by Peng and Choi (2005). The purpose of this representation is to remove words irrelevant to the meaning of the documents. This technique uses words and concepts simultaneously. It enriches document representation with generic concepts and thus, provides a more abstract indexing method. However, all statistical information of concept co-occurrence is lost and even if only concepts are used to measure distance, the document representation will be binary (e.g. 0 – concept is not present for the document and 1 – concept is present for the document).

The problem with the aforementioned representational technique is that statistical information is lost once a word is replaced by a concept. This problem is overcome by document representational technique called concept indexing (Setchi et al., 2011). Concept indexing is an analytical process of identifying document entities and relations between them that represent the knowledge conveyed in documents. It is a machine understandable index of entities and concepts, where a concept is defined as "abstract or physical information about entities or relations between them". Concept indexing is designed to be used in the document retrieval domain. It relies on generic concepts to represent documents. The representation of each concept within a document is a statistically computed real number. Therefore, a distance function or a distance matrix acquired from an external knowledge source can be used along with concept indexing.

This method has restrictions on the terms that are replaced with concepts. The terms are pre-selected and all other terms are discarded. The concepts from the representative vectors

however, are given a relevant representative value, which is calculated by using the TF-IDF value. This method demonstrates good retrieval results, but has never been tested in the domain of document clustering. Nevertheless, concept indexing follows approaches similar to Hotho et al. (2001, 2003c) and Peng and Choi (2005). Therefore, the clustering solutions produced should be generic and provide structural information for the clustering solutions in reduced dimensionality. The restriction on the used terms needs to be removed. Then, concept indexing has to be calculated for all terms, which appears in the external lexical source.

The advantage of the concept indexing over the other approaches is that it employs an ontology which resembles the structure of a thesaurus, where all words in the English language (~230.000) are grouped by the ideas they express in 990 concepts (see section 2.3.5). Thus, the dimensionality of the produced index is limited and there is no need of further restrictions or assumptions for reduction. A challenge to the use of concept indexing in the domain of clustering though, would be to produce multiple groups of documents from which to look at the clustering task as a personal view.

A benefit of using concept indexing in document driven tasks is that it supports automatic knowledge extraction, efficient reasoning, rich lexical and semantic representation, good scalability is necessary for large document collections, flexibility in conducting queries, and automatic content categorisation. This support cannot be provided by the conventional text representation approaches (Setchi and Tang, 2007). In addition, the concept indexing meets all of the requirements for knowledge representation of explicit knowledge such as the one encoded in text documents. Concept indexing provides good scalability needed to perform on a large scale and automated routines for text representation and categorisation. In addition, this technique has the flexibility necessary to conduct querying and to provide fast clustering solutions since it works in reduced dimensionality. The improved speed of the clustering al-

gorithms makes it possible to use measures, which are too computationally expensive otherwise.

3.3.2. Advanced document similarity

The use of concept indexing to represent documents enables a pair-wise similarity between documents to be measured by analysing commonly shared concepts between documents. The relations that are likely to exist between concepts can vary in terms of abstractions and this corresponds to certain level of abstraction which two documents share. The concept indexing requires an external lexical source using which to identify the concepts for every document in the collection.

Document representation with concept indexing keeps statistical data for word co-occurrence, which is a prerequisite for employing a many-to-many similarity matching between two multi-dimensional distributions (Patwardhan, 2003). The semantic distances between two multi-dimensional distributions of concepts can be measured by context vector and a many-to-many matching that uses an external knowledge source. The distance establishes relatedness between documents by measuring the angle between two concept vectors (Patwardhan, 2003). The many-to-many semantically measured distance is proven more accurate when statistical information for words co-occurrence is derived from a large corpus and a generic external knowledge source is used. In addition, it aligns clustering solutions close to human judgement and can be applied across domains, to any number of documents and has no restrictions on the kind of words calculations based on their relations as long as they exist in the external knowledge source.

The many-to-many similarity measure is semantic since it relies on external knowledge source for measuring the similarity between documents. It considers document structure at

concept level by measuring similarity between distributions of concepts. An algorithm of a particular interests that implements many-to-many matching is the Earth Mover's Distance (EMD) (Rubner et al. 2000). It implements many-to-many comparison to evaluate the dissimilarity between two multi-dimensional distributions in a feature space and needs a distance measure between every two single features. The similarity of two documents is equal to 1 minus the dissimilarity (distance). The EMD algorithm is computationally expensive and has never been tested in the domain of document clustering neither on a large scale.

3. 4. Conceptual model

This section presents a conceptual model of a system that enables semantically enhanced document clustering. Firstly, a general overview of the model is presented. Next, a text representation that enables reduced computational complexity of the entire model is introduced. It then describes the use of an external lexical knowledge source and its utilisation in the text normalisation. Finally, a mechanism for using multiple viewpoints, i.e. different clustering solutions, by employing levels of abstraction is presented.

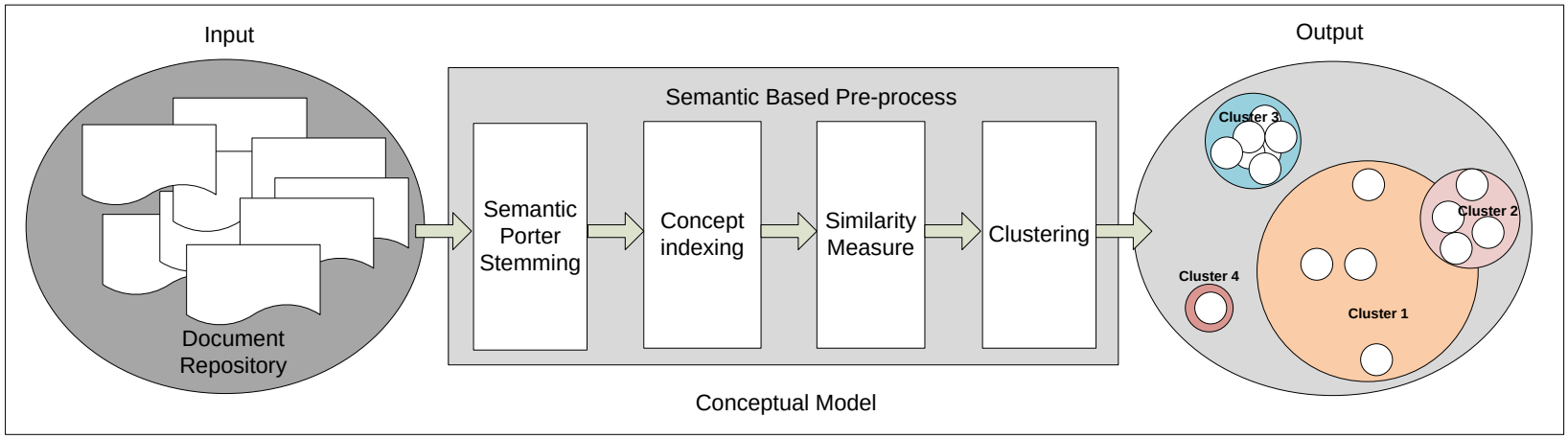


Figure 3.1 Conceptual model of a semantically enhanced document clustering system

The proposed conceptual model, shown in fig 3.1, improves the generic model for document clustering with a feedback pathway presented by Xu and Wunsch (2005) (see fig. 2.1). The improvement is introduced by linking the feedback pathway back to the similarity measure and not the clustering, which allows multiple clustering solutions to be produced. Multiple viewpoints are produced by measuring similarity between documents at a different (modified) level of abstraction. Two clustering solutions differ from each other by the level of abstraction used to measure a pair-wise document similarity. The proposed model consists of four main functional blocks: (i) semantic-based stemming, (ii) semantic indexing, (iii) semantic-based similarity, and (iv) traditional clustering. The input to the system that implements the model is a collection of unstructured documents, i.e. documents, which have no pre-defined structural organisation.

The feedback mechanism in the traditional model allows the users to alter the clustering solutions by changing the number of produced clusters (Fung et al., 2005). The conceptual model for semantically enhanced document clustering (shown in fig. 3.1) also implements similar feedback mechanism for altering the document groupings. However, in contrast to the generic model, the semantic clustering model implements a feedback pathway between the similarity measure and the clustering results. Then, the alteration of the document groupings involves measuring similarity between documents from different prospective (abstraction). The purpose of this computationally expensive feedback pathway is to enable multiple deterministic clustering solutions closely aligned to human judgement. Thus, users will be provided with a mechanism to alter the document groupings not only by changing the number of clusters, but also by providing limitations to the similarity measure, i.e. below a threshold documents

are considered not similar. The threshold will be with regard to certain extend to their judgement. In the next section the mechanism of setting the limitations is explained.

3.4.1. Pair-wise similarity measure

The traditional similarity techniques explored in the literature for measuring similarity between documents employ VSM to represent documents. These techniques use one-to-one approach to measure similarity between documents (Blair, 1979, Salton and Buckley, 1998, Aslam and Frost, 2003). The technique that has the most robust performance of all one-to-one techniques is the cosine similarity (Dhillon and Modha, 2001). Therefore, when a many-to-many similarity technique is proposed (Wan and Peng, 2005b) based on the text-tiling algorithm (Hearst, 1993) it is compared to the cosine measure (Wan, 2007). The results, theoretically predicted by Patwardhan (2003), are supported by the experiments conducted on a small scale by Wan (2007). Scientific evidence support the conclusion that many-to-many approach to measuring pair-wise similarity between documents aligns clustering solutions better to human judgment than one-to-one measures. However, that is never proved on a large scale due to the high computational complexity of the algorithm.

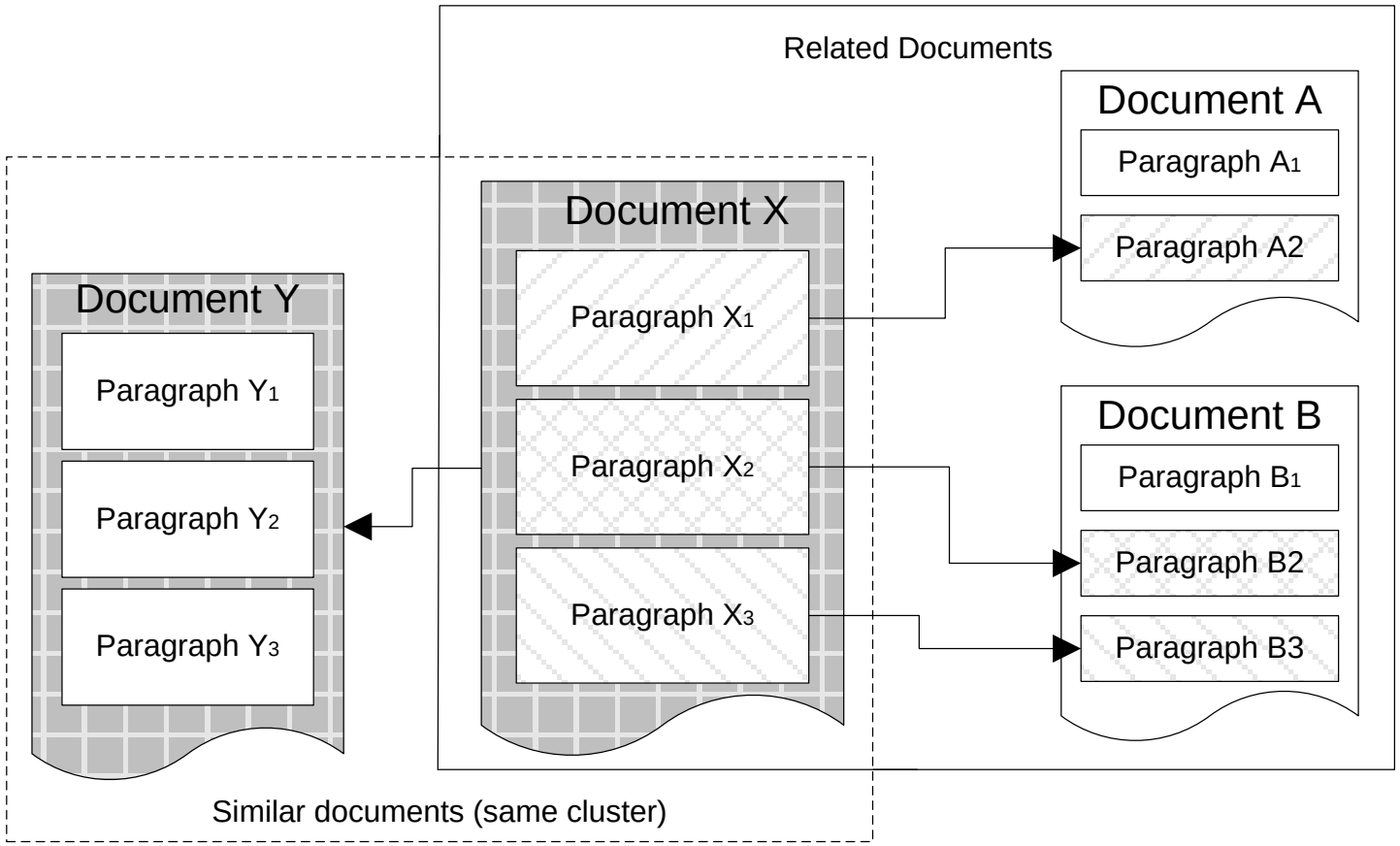


Figure 3.2 Measuring similarity between documents

A main objective of this thesis is to develop an algorithm, which produces clustering solutions consistent and well aligned to human judgement. Therefore, the first step towards achieving that is to use many-to-many similarity measure. The selected measure is based on the Earth Mover's Distance (EMD) (Rubner et al., 2000) and evaluates dissimilarity between two multi-dimensional distributions in the feature space of words or concepts. This algorithm considers the structure of the documents in measuring the distance between any two documents. An illustrative example is shown in figure 3.2. Similarly to the TextTiling algorithm document similarity is measure by matching structural blocks from within the documents. The granularity of these blocks varies from words and phrases to paragraphs. A change in the level of abstraction modifies this granularity of the blocks. Thus, multiple clustering solutions are generated.

3.4.2. Concept indexing in clustering

The complexity of the EMD algorithm has a quadratic (O^2) complexity. Therefore, the traditional document representation techniques will be irrelevant on a large scale. An alternative document representation is concept indexing (Setchi et al., 2009). It represents documents in smaller number of dimensions. The total number of concepts is defined by externally used knowledge source employed by the concept indexing. As a result of the reduced dimensionality, the text representational model will have good scalability needed to perform on a large scale. In addition, concept indexing provides flexibility on running queries on a large scale. Flexibility and reduced di-

mensionality are features needed by the presented model, since similarity measured between documents needs to be re-calculated for different levels of abstraction.

The EMD similarity technique requires a ground distance between any two features to measure the distance between a pair of their multi-dimensional distributions. Therefore, a distance measure between any two concepts is needed. A distance matrix can be acquired from the external knowledge source. Fig. 2.2 shows the structure of a knowledge source used by Setchi and Tang (2009) to acquire index for representing web pages. It is called OntoRo and is based on the "Roget's Thesaurus of the English Words". The OntoRo structure will be used to create a distance matrix for features contained in the knowledge source. Similar matrix is created by exploiting the tree structure of a thesaurus for the purpose of producing semantic chain of words (Jarmasz and Szpakowicz, 2003). The same idea will be used to acquire a distance matrix from OntoRo.

3.4.3. Text normalisation

The accuracy of the distance measured between a pair of documents is defined by the quality of the index, i.e. the precision of computing the concept indices. Setchi and Tang (2009) use the suffix stripping algorithm of Porter (Porter, 1997) to normalise words. However, the quality of the index they need is based on 132 semantic adjectives and not all the words occurred in documents. Therefore, the acquired concept indices for representing documents will be ambiguous if their approach is used (Table 3.1). The quality of the indices will increase when less ambiguous concept information is acquired from the OntoRo. Therefore, the first step of the conceptual model

is to provide a text normalisation algorithm, which alleviates the problem with ambiguity of the concept indexing.

Table 3.1 Occurrence of a word in OntoRo after stemming

Word	Stem	OntoRo Occurrence
Struggle	struggl	7 concepts; (6v); (5n); 11 occurrences
Struggled	struggl	none
Struggles	struggl	none
Struggling	struggl	2 concepts; (2 adverbs); 2 occurrences

The conceptual model will addresses that problem by introducing semantically optimised word normalisation algorithm. The algorithm is presented as a semantically enhanced text stemmer, which will capture as much disambiguated semantic information from text as possible. In view of the fact that different word forms have different positions in the OntoRo structure they respectively have different meanings. Therefore, certain position of a word in the OntoRo structure defines a particular semantic meaning for it. The semantic distance between a pair of words, within the OntoRo tree structure, is measured with the path between two meanings of these words. The approach of measuring a pair–wise distance between words will be extrapolated to measure document pair-wise similarity by using the words' meaning.

The words "studying" and "student" give an example of measuring word pair-wise similarity. After the porter stemmer processes the words their stemmed forms are "studi" and "student" as the latter one remains unchanged. The word "studi" can either originate from the word "studying" or the word "study". In OntoRo the word "study" occurs in 20 concepts whilst the word "studying" refers to 3 concepts. In total, the root word "studi" is more ambiguous than "student" since it refers to 23 con-

cepts (similarly to the example shown in Table 3.1). This example shows that Porter stemmer makes the words simpler and more consistent, but it introduces a semantic distortion, which makes words stems more ambiguous. Therefore, the Porter stemmer needs to be semantically enhanced to capture as much clear semantic information contained in text as possible.

3. 5. Summary

This chapter proposes a conceptual model that allows multiple viewpoints, i.e. clustering solutions, of a document collection to be produced. The model employs semantics to acquire generic and abstract document index. The acquired index will be used to measure similarity between documents by using computationally expensive many-to-many matching. The design of the conceptual model proposes innovative approach to clustering by employing semantics in every step, which enable a feedback from clustering to be re-used as an input to the system to enhance the grouping results. The feedback represents an increase or decrease in the level of abstraction for which a similarity between documents is measured. Thus, levels of abstraction will be used to produce clustering solutions that meet the expectations of the user for existing relations between documents.

Chapter 4 : Semantically enhanced text normalisation

This chapter presents a semantically enhanced text stemming algorithm (SETS) that provides reduced dimensionality and better separation between clusters. Discussion on text normalisation techniques considered promising for obtaining clustering solutions with improved separation between and coherence within clusters is presented.

4. 1. Improvement of clusters coherency

Cluster analysis solves the general problem of forming groups of similar objects, called clusters. The properties employed to measure pair-wise object similarity are defined by the domain of application and the pragmatic context of the task they are used for (Grefenstette 2009). The objects within a cluster are more similar to each other than the objects belonging to other clusters (Karypis et al. 1999; Cadez et al. 2000). The quality of the produced clustering solutions is measured by the separation between clusters and the similarity of documents within the clusters (Rousseeuw, 1987, Kaufman and Rousseeuw, 2005). Rousseeuw (1987) suggests this quality of clusters to be measured by a number between 0 and 1 and calls it a silhouette of a clustering solution. The higher the silhouette value is, the higher the cluster's coherency (quality) is. In the scope of document clustering particular attention is devoted to the coherency of the clusters produced across domains. Text normalization (stemming), is used to reduce inconsistency in text introduced by the different inflections of words and to improve the silhouettes of the produced clusters.

4.1.1. Document normalisation – text stemmers

Document clustering is employed in various domains such as information retrieval and data mining, to group a set of similar documents that resemble a query of words to certain extent (Jain et al., 1999). Research carried out in recent years in these domains indicates a growing need for more effective document search and retrieval as well as document browsing and knowledge discovery through more efficient clustering (Huang 2008; Grefenstette 2009). Users who interact with information retrieval systems, e.g. document search engines, submit queries of words to retrieve the information required. The search engine returns a set of documents that resemble the query to certain extent (Jain et al. 1999). In this scenario the choice of words used in the query is crucial as it determines the quality of the returned documents. Users who are unfamiliar with the domain terminology are likely to formulate inadequate queries. Their choice of words leads to poor search results. One method to alleviate this difficulty is to enable users to find information through effective navigation and browsing within clusters of documents (Cutting et al. 1992; Carpineto et al. 2009). This can be achieved with more abstract indexing of text document, which can be obtained by employing a semantic text normalisation technique and semantic hierarchies in text representation (Peng and Choi, 2005, Setchi et al., 2011). This section is focused on methods and techniques that increase the homogeneity of documents within clusters, i.e. providing document groupings with better silhouettes.

Document clustering relies on features acquired from texts to measure pair-wise document similarity. The features, called word stems, are obtained from documents after text normalisation. The traditional approaches to text normalisation achieve text

consistency by employing affix stemming, statistical approaches or mixed techniques (Jongejan and Dalianis 2009). Affix stemming achieves normalisation via rule-based transformations, which aim to remove known prefixes or suffixes from words by relying on language morphology. Statistical stemming is independent from language knowledge. This stemming technique analyses distribution of root morphological elements in corpus. A set of various techniques used to obtain the root elements of words can be combined into so-called 'mixed approaches'. The efficiency of the stemming algorithms depends on their computational complexity, as well as on the quality of the corpus. The computationally expensive stemming algorithms are brute force, lemmatisation and production technique. These algorithms output real words. However, the algorithms that are of particular interest to the research presented in this thesis are those with shorter execution time, such as affix stemming, stochastic algorithms, n-gram analysis, hybrid approaches and matching stemming algorithms. These algorithms do not necessarily output real words. They belong to the family of rule-based stemming algorithms and their output aims to provide basic text consistency. The rule-based normalisation recognises as similar those words that share a common grammatical root. These algorithms also produce errors as a result of over- and under-stemming. The former refers to words from which the morphological ending is too far removed, whilst the latter refers to words that are not reduced to their root elements (Xu and Croft 1998). As a base line algorithm for all stemmers is used the suffix-stripping algorithm of Porter (Porter 1980). The Porter stemmer provides a good trade-off between speed, reliability and accuracy, and is usually used as a base-line algorithm for comparison purposes. State-of-the-art algorithms, which perform slightly better and provide a small advantage over the Porter stemmer, are also slower and

more difficult to implement (Smirnov 2008). The stemming efficiency, by means of separation between clusters, depends on the computational complexity and the quality of the corpus. The SETS algorithm that is investigated in this chaper is an alternative to the Porter stemmer, which is used as a base line stemmer in the evaluation.

The Porter stemmer achieves text normalisation by employing suffix stemming (Porter, 1980). It involves rule-based transformations, which remove known suffixes of words by relying on morphological rules of the language. The words are stemmed to their morphological root form. Rule-based normalisation recognises as similar those words that share a common grammatical root form. However, rule-based stemming may produce errors as a result of over- or under-stemming. Stemming is used to reduce inconsistency in the text introduced by different inflections of a word with the same stem.

4.1.2. Document representation in reduced dimensionality

Model-based clustering employs various document representations such as the vector space model (VSM), a set of repetitive words (Wang et al. 1999; Beil et al. 2002) or word-sequences (Li et al. 2008). The VSM representation extracts a bag of words (BOW) (Salton and McGill 1986) from documents and treats them as representative features for these documents. Each document in the collection is then represented as a vector of certain weighted word frequencies. The weight of the words stands for their representativeness for the document in the context of the collection (Robertson 2004). In addition, VSM representation, when used in conjunction with higher level indexing, i.e. semantic hierarchy, provides better coherency for the clusters (Peng and Choi 2005). The VSM outperforms the other document representative methods by speed

(Patwardhan 2003) and therefore, the VSM approach is selected to represent documents in this chapter twice: firstly, documents are represented in VSM by using the Porter stemmer, and secondly, by using the proposed algorithm. The evaluation section 4.5 provides a comparison between the proposed semantic algorithm and the traditional approach. Both algorithms use the spherical k-means partitional algorithm (Dhillon et al. 2002) to cluster the documents. The quality of the clusters produced by both algorithms is evaluated with silhouettes, which is a graph-based technique for interpretation and validation of clusters (Rousseeuw 1987).

This chapter focuses on increasing the homogeneity of documents within of clusters and providing better document groupings by developing a semantic-based text normalisation algorithm. The normalisation is achieved by using semantic hierarchies, contained in ontologies. Semantically normalised text allows more abstract features to be used in document indexing. The produced clustering solutions are evaluated against human judgement.

4. 2. Document representation – challenges, limitations and advantages

This section firstly discusses the feature extraction of word stems from text documents using the Porter stemmer. Then the traditional vector space model representation is analysed. Finally, the standard k-means algorithm with cosine similarity measure, i.e. spherical k-means algorithm, used in the evaluation, is described.

4.2.1. Feature selection and feature extraction

The first step of a clustering procedure is to distinguish a subset of features from a set of candidates (Jain et al. 1999; Jain et al. 2000). This process is called feature se-

lection and it differs from feature extraction, as the latter utilises transformations needed to generate features from the original ones. In clustering, documents are pre-processed by employing stemming algorithms like the Porter stemmer. After text normalisation, a statistical co-occurrence of all words is calculated (Lewis 1992). The index that represents the documents in a collection is then aggregated. A document index comprises of sets of pairs of word stems and weights, <s, w>. By contrast, semantic-based algorithms take into consideration pre-established relations in an external knowledge source. The relationships used to aggregate the index (Setchi and Tang 2007; Xiao 2010) are established between the word stems and a higher order structure of entities in the structural organisation of the lexical source. The relationships established in the external knowledge source pre-determine the index and the document groupings, respectively. Feature selection and extraction are crucial to efficient clustering. The aggregation of good quality features leads to decreased workloads and simplified subsequent clustering or/and improved document groupings (Xu and Wunsch 2005).

Porter stemmer

The Porter stemming algorithm is the word normalisation technique used most widely in the information retrieval community (Smirnov 2008). It provides normalisation at document level by removing common morphological and inflectional endings of words. In other words, the technique resembles the suffix stemming used to produce root elements (Porter 1980). The stemmer improves the precision and the recall of the information retrieval systems for two reasons. Firstly, the stemmer produces a reduced number of root elements (reduced dimensionality) by conflating a group of

words into a single root element, through the removal of various suffixes like "–en", "-ing", "-ion", "-ions". An example is shown in table 4.1. Secondly, the root elements are believed to convey the same topic. Although the morphological forms of the words produced are not necessarily real words, the documents retrieved indicate good quality (Wessel et al. 1996).

Table 4.1 Suffix stripping algorithm by the Porter Stemmer

Word	Porter (root) form
connect	connect
connected	connect
connection	connect
connecting	connect
connections	connect

The Porter stemmer employs various suffix-stripping rules and does not rely on external lexical sources. As a result, the accuracy of the stemmed words will never be absolutely accurate, irrespective of the evaluation. In addition, there are cases of word stemming that demonstrate the inadequacy of the algorithm to cope with the words that follow no morphological rules of inflection, such as the irregular verbs and the words that shift from their root form when suffix is added – for example, the words 'sand' and 'wand'. The words 'sand' and 'sand-er' are correctly stemmed to 'sand', as they share a common syntax stem. However, in the case of the words 'wand' and 'wander', as well as the words 'experience' and 'experiment', the algorithm wrongly conflates the words to 'wand' and 'experi'. The problem with the ending "-er" of the word 'wander' is that it is considered a suffix and is stripped off. As a consequence, the meaning of the word is changed. Instead, the algorithm should leave the ending and consider the whole word as a part of the stem. In the case of 'experience' and 'experiment', the change in the meaning of the words occurs as a result of ambiguity,

rather than because of a wrong meaning. According to the Porter stemmer rules, both words conflate to the syntax stem ‘experi’, without considering the fact that both words have different meanings. Nevertheless, when the words are gathered together by meaning, the word ‘experiment’ is placed in the group of words ‘experiential’, ‘experimental’, ‘experimentation’ and ‘experimenter’, whilst ‘experience’ and ‘experienced’ share a common meaning. Any attempt to improve the performance of the suffix stripping in one area of the vocabulary causes deterioration of the performance in another area. This problem also reveals the challenge in the clustering across domains. In particular, it is difficult to foresee rules that cope with a rare change of the root form of a word when a suffix is added, as in the case of ‘prescribe/prescription’, ‘deceive/deception’ and ‘resume/resumption’. Therefore, the approach that resolves inconsistent stemming behaviour and performs well on one corpus, fails when applied to a different corpus.

The Porter stemmer uses a set of transformation rules applied in a sequence of steps, namely 60 rules in 6 steps. The implementation of the classical Porter stemmer needs no recursion and the individual steps of the algorithm are as follows:

Step 1: Remove plural of known suffixes, e.g. “–s”, and “-ed” or “-ing”;

Example: possesses ⇨ possess; ponies ⇨ poni; interesting ⇨ interest

Step 2: Replace terminal “-y” with “-i” when there is another vowel in the stem;

Example: coolly ⇨ coolli; furry ⇨ furri; fly ⇨ fly;

Step 3: Map double suffixes to single ones: “-isation”, “-ational”, etc.

Example: optional⇨option; possibly⇨ possibli ⇨ possible; playfulness ⇨ playful

Step 4: Remove suffixes of the words, e.g. such as “-ful” and “-ness” etc.

Example: largeness ⇨ large; playful ⇨ play; practical ⇨ practice; felicity ⇨ feliciti ⇨fe-lic

Step 5: Take off suffixes such as "-ant", "-ence", etc.

Example: precedent ⇨ preced; operational ⇨ operate; controllable ⇨ controll

Step 6: Remove the final "–e" and "-l"

Example: controllable ⇨controll ⇨ control; deflate ⇨ deflat

An example of three words reduced to their root forms by the Porter stemmer is shown in table 4.2: 'semantically', 'destructiveness' and 'recognizing'. The example follows the stemming process through the six steps of the stemmer. Finally, at step 6, the words are in their root forms.

Table 4.2 Illustrative example through the porter steps

Step #	Word 1	Word 2	Word 3
Step 1	Semantically	Destructiveness	Recognizing
Step 2	Semantically	Destructiveness	Recogniz
Step 3	Semanticalli	Destructiveness	Recogniz
Step 4	Semantical	Destructive	Recogniz
Step 5	Semantic	Destructive	Recogniz
Step 6	Semant	Destruct	Recogn

4.2.2. Improvement of the suffix stripping stemming

These shortcomings of the Porter stemmer illustrate how rudimentary it is and state the need for improvement. Stripping words' suffixes without using word sense disambiguation and/or part of speech disambiguation, introduces the problem of shifting the meaning of the words after stemming. In addition, the stemmer needs to handle or avoid the rare inflected irregular forms of the words produced using conjunctions. This paper aims to improve text normalisation by considering the words' grammatical structure, which affects their semantic representation, or the meaning of the words, in

corpus-based stemming. Therefore, approaches where stemming is corrected by searching in a dictionary (Krovetz, 1993) or using statistical properties of the corpus (Xu and Croft, 1998) are proposed. These approaches resolve the inconsistent behaviour of stemming on a small scale, where algorithms perform well on one corpus but fail when applied on a corpus from a different domain. The statistical properties of a new corpus need to be acquired. Therefore, this approach is not suitable for large scale document collections.

This context-sensitive stemming (Lee, 1999) is used for document searches. The corpus is analysed prior to clustering, to establish the distributional similarity of words. The next step is to apply the Porter stemmer to the candidates acquired from the documents, that is to say word similarity by distribution, to remove all possible grammatical inflections of pluralisation. The stems obtained from the words are used in query expansion on non-transformed indices to retrieve documents. For example, the words ‘experiment’, ‘experiments’, ‘experimental’, ‘experimentation’ and ‘experimenter’ are produced, but only ‘experiments’ is retained for pluralisation purposes, which allows query expansion of ‘experiment’ to ‘experiments’.

The derivational and the inflectional stemmers improve Porter's algorithm by adding a dictionary check after each iteration (Krovetz, 1993). The aim is to stop further stemming if forms of the words are found, as well as enabling the processing of irregular forms. The resulting stemmer, however, performs worse than the original Porter stemmer at an additional computational cost. The semantically enhanced text stemmer (SETS) aims to acquire disambiguated semantic information by searching every stem in a dictionary, without the need to produce real words.

4.2.3. Document representation

This section is focused on the document representation that uses the TF-IDF weights in the document vectors.

Partitional clustering employs the vector space model (VSM), which treats documents as a bag of words (BOW) (Salton and Mcgill, 1986). The documents in a collection are transformed into VSM by using TF-IDF weighting, to build the document-term matrix. The weight of stems that do not occur in a document (the row in the matrix) is 0. This trans-formation into VSM yields a matrix, where each row is the vector representation of a document from the corpus in the TF-IDF vector space. The dimensionality of the matrix is usually very high and makes the scalability of clustering algorithms difficult (Fung et al., 2005). The scalability problem is typical for algorithms that produce good results on a small dataset (Fung et al., 2005) or in a specific domain (Zhang et al., 2011), but which fail to perform on a larger scale or across domains. The first case considers algorithms with very high computational complexity, which is impractical on a larger scale. The second refers to word ambiguity and the fact that many domains share common terms, which may contribute to low quality in grouping documents (Gliozzo et al., 2004) and poor separation between the produced clusters, resulting in a low value for the clusters' silhouettes.

A strategy for achieving better performance and scalability of the clustering is achieved by applying traditional feature selection, enhanced by a semantic-based approach for dimensionality reduction. A semantic approach, which relies on a general ontology, is employed for large-scale indexing of web pages with concepts. It uses a higher order semantic hierarchy in the document representative vectors and is regard-

ed as concept indexing (Setchi et al., 2011). Concept indexing considers all possible meanings of words. A word with multiple meanings shares its weight (significance) equally among the concepts to which it belongs. Eventually, an accumulated scoring result for every possible concept is calculated (see equation 1) and a document-concept matrix is produced.

$$w_c(d_j) = \sum_{i=1}^{n} \left(w_{tf-idf}(t_i, d_i) \frac{1}{C_{(t_i)}} \right) \quad (1)$$

where n is the number of terms in the document that contains a concept C, while $w_{tf\text{-}idf}$ denotes the significance of a stem. The coefficient $1/C_{(t_i)}$represents the idea, based on empirical observations, that monosemic words are more domain-oriented than polysemic ones, and provide a greater amount of domain information (Setchi et al., 2011). The index aggregated for the document representation comprises of sets of pairs of concepts and weights, <c,w> for every document. The size of the vectors does not exceed 990, which is the number of concepts in the ontology used (see section 3.4.2 in chapter 3).

A discussion on the representation of the external knowledge source OntoRo, mentioned in chapter 3 section 3.3.1, is extended towards concepts coverage and their discriminative power in terms of the concepts that do not appear in the ontology. In addition, the ad-vantages and disadvantages of the OntoRo are covered and how the results of the document representation will be affected by a change in the ontology used in document representation.

Similarly, a method which employs an ontology to enrich the term vector with concepts by partially or entirely replacing terms with concepts is proposed by Hotho et al. (Hotho et al., 2001, Hotho et al., 2003b). The use of a higher order topic structure to

replace the words in the representative vectors with concepts significantly reduces the dimensionality of document representation and the computational complexity (Hotho et al., 2001) on a small scale. In addition, the higher order hierarchy used to reduce dimensionality provides better scalability and improved clustering as well as a generic perspective in measuring similarity between documents when used in clustering. The produced clusters are reported to have better homogeneity of documents represented by higher silhouette values.

A significant difference between the approaches proposed by Setchi et al. (2009) and Hotho et al. (2001) is that respectively the former employs general (large) ontology, while the latter relies on a specific core ontology. In addition, the approach of Setchi et al. (2009) is shown to be scalable and is successfully used on a large scale in the domain of information retrieval. In contrast, in 2003 Hotho et al. (2003c) propose a modified version of the same algorithm, which is applied on a larger in comparison to the corpus used in 2001. The modification consists of a use of an additional specific ontology, namely 'Agenda', which is used to produce specific and profiled results. Therefore, the approach which employs a general ontology (Setchi et al. 2009) needs to be tested in the domain of document clustering. The evaluation parameter that will be of specific interest is the separation between the produced clusters (i.e. silhouette values). Hotho et al.'s (2001, 2003c) approaches obtain better separation in comparison to base line algorithms, which use VSM and BOW document representation. Therefore, the evaluation has to compare the clustering solutions produced with different document representation, i.e. concept indexing that relies on a general ontology and BOW.

4.2.4. Traditional approach to document clustering

Partitional clustering, also known as hard partitioning, creates a flat, non-hierarchical structure of clusters, the number of which is controlled by a value given to the clustering algorithm prior to execution. The number of clusters k drives the process of partitioning all documents from a collection into k clusters (Fung et al., 2005). However, selecting the number of clusters without initialising the domain knowledge in the area of interest may worsen the results. Also, if documents cover a broader thematic area, the clusters produced would be inferior.

Kernel-based partitional methods such as the kernel k-means algorithms, which consider mapping of the input prior to clustering (Karatzoglou and Feinerer, 2006) using string kernels (Huma et al., 2002) or word-sequence kernels (Cancedda et al., 2003), perform better than the standard k-means. Nevertheless, the standard k-means algorithm is selected in the presented evaluation (section 4.5) because of its good trade-off between the speed of execution and the quality of clusters produced.

In order to apply a partitional clustering algorithm, like sk-means, the document collection used needs to be represented in a document-term matrix, by using VSM. Each document is then represented as a vector d in the vocabulary space. The position of the vectors in the multidimensional space is defined by the co-occurrence of every term from the collection within the documents, the TF–term frequency, multiplied by the inverse document frequency (IDF) of the terms. Thus, TF-IDF defines the representative weight of each term encountered in the collection within each document:

$$D_{tf} = (tf_1, tf_2, tf_3, \dots, tf_n)$$

where $tf - idf_i$ is the weight of the token with index i. However, not all terms have the same discriminative power, and determining the discriminative power of the words can be considered as a two-stage process. Firstly, stop words, such as words that are common and would make no difference if they are considered or not in the clustering, are removed. The second stage employs weighting of the word's significance within the document collection, or calculating word's IDF weight. Thus, those words that are frequent in the collection are given less discriminative power, or less IDF weight, which classifies them with less eliteness (Robertson, 2004) in the collection than the rarer words. A word that is a representative token of the collection but not representative in a particular document is represented with weight equal to 0 for that document. Before the TF-IDF value is calculated, words that occur in different grammatical forms are normalised to their canonical form (Porter, 1997), thus reducing text inconsistency. The weights of the document indices are calculated by multiplying TF and IDF. However, computing the weight of all words across all documents leads to high computational complexity (Beil et al., 2002), which motivates considerable interest in low-dimensional document representation that overcomes this particular issues (Matveeva, 2006). The SETS algorithm addresses this problem by reducing the dimensions in the document-term matrix, but it still relies on TF-IDF values to measure a word's eliteness.

The position of the vectors in the multidimensional space is defined by the co-occurrence of every term from the collection within the documents, i.e. TF–term frequency, multiplied by the inverse document frequency (IDF) of the terms. Thus, TF-IDF defines the representative weight of each term encountered in the collection within each document in the collection:

$$d_{tf-idf} = (tf - idf_1, tf - idf_2, tf - idf_3, \dots, tf - idf_n) \quad (2)$$

where $tf - idf_i$ is the weight of the token with index i. Words, which are frequent in the collection, are granted with less discriminative power (less IDF weight) than the more rare words. A word that is a representative token of the collection but is not in a specific document is represented with a weight of 0. However, computing the weight of all words across all documents leads to high computational complexity.

Selecting the number of clusters without a priori domain knowledge in the area of interest may worsen the results. In addition, if documents cover a broader thematic area, the cluster can be inferior (Steinbach et al., 2000). The spherical k-means algorithm uses the robust cosine measure to measure the similarity between documents. It is defined as

$$cosine(d_1, d_2) = \frac{d_1 \cdot d_1}{||d_1||} \left||d_2|\right| \quad (3)$$

where · denotes the vector dot product and |(|d|)| is the length of the vector. The sk-means algorithm randomly selects k centroid vectors to identify the closest documents to the centroids and forms clusters around them. The algorithm iteratively refines the randomly chosen initial k centroids, minimising the average distance between them. In other words, it performs homogenisation of the clusters by increasing the similarity within clusters.

4. 3. Semantically enhanced stemming

This section firstly discusses the hierarchical structure of the used ontology in this research (OntoRo). Then, the semantically enhanced feature extraction algorithm is introduced. Finally, a document representation with reduced dimensionality is dis-

cussed. Syntactic and semantic information (Yun et al. 2010) including word senses (Peng and Choi 2005), which are outside the scope of the Porter stemmer, are considered by the proposed in this section SETS stemmer.

4.3.1. External knowledge source (OntoRo)

An external knowledge source is selected according to the task it is employed to complete. In many scenarios external knowledge source is domain specific, i.e. it depends on the input data and is used for achieving certain output. The output of the proposed algorithm aims to group documents with various topics in coherent clusters by placing documents that share common or similar topics in the same cluster. As a consequence of topic diversity, the proposed algorithm needs a general ontology. It needs to provide a generic perspective to the relations between concepts (i.e. predefined topics).

The organisation of the words in concepts according to OntoRo as per the discussion presented in section 2.5 is generic and can be used to provide a general perspective to document similarity. The organisation of words and phrases in OntoRo can be used to measure the distance between concepts and respectively between text documents, when employed by a similarity measure. It is assumed that the use of generic relations established between the words (i.e. the ideas around which words are grouped together) in OntoRo will enable the similarity measure used to establish relations between documents which are more similar to human judgement. It is important to point out that in OntoRo a concept is a virtual entity which groups together words expressing similar ideas. Unlike WordNet or other ontologies, where a concept is understood as a word or a meaning of an entity, the concepts in OntoRo are established

around designated ideas which can change over time. Respectively, the number of concepts which outline a particular word organisation may vary in different versions of the thesaurus. For instance, Roget's thesaurus from 2003 contains 990 concepts (organising 230.000 words) and the thesaurus from 1911 contains 1000 concepts, which organise ~40.000 words. On the other hand, the thesaurus from 1985 has 1185 concepts, which organise ~56.000 words. However, the total number of concepts in OntoRo, which is 990, organise ~170.000 words and ~60.000 phrases. Therefore, concepts in OntoRo are assumed to have organised all common words in the English language. For a reference the latest version of Oxford dictionary contains 171,476 words and according to the Global Language Monitor (GLM), which analyses and tracks trends in language over the world, with a particular emphasis upon global English and is supported by Google, recognises ~ 1,022,000 words. A distinct difference (and the first reason for selecting OntoRo as an external knowledge source) of the thesaurus from 2003 year is its high number of words and phrases organised in its structure in comparison to the other available thesauri.

A document similarity measure which relies on a thesaurus to compute a similarity between documents will measure different values between the same documents when a different version of a thesaurus is used. The difference is in the different organisations of the words and the phrases in the thesaurus. The second reason for selecting Roget's thesaurus of 2003 for an external knowledge source in this chapter is due to its associative nature of organisation of the words and the phrases. The structure of the other thesauri also group together words and phrase that express similar idea, but the thesaurus from 2003 is built around associations and extends that characteristic even further. Therefore, a similarity measure which employs OntoRo to compute sim-

ilarity between documents is assumed to align the similarity between text documents better to human judgement.

Despite the rich vocabulary of OntoRo, it contains no named entities. This disadvantage of the lexical source can be overcome by merging OntoRo with other available ontologies which contain named entities such as ontologies used in the automobile industry or ontologies from the bio-science domain. However, this is outside the scope of the presented research and is outlined in chapter 7 as a future work which needs to be thoroughly investigated. In addition, the pure lexical nature of OntoRo allows an objective comparison between the BOW document representation, which uses named entities alongside other words, and entirely semantic document representation. Thus, the evaluation will reveal the positive change in the quality of semantic representation and will outline a future need for combining both approaches for a new kind of document representation.

The proposed algorithm employs concept indexing as a text representation technique and OntoRo as an external lexical source. For the purpose of the stemming OntoRo is stemmed using the Porter stemmer. Thus, all words in OntoRo are stemmed to produce a second lexical source regarded as stemmed OntoRo. An example for the word "connect" is shown in Table 4.3. The column 'Word' contains inflectional forms of connect as they exist in OntoRo by design (Setchi et al., 2009). The column 'Root Form' contains the relevant root forms of the words from column 'Word' produced by the Porter stemmer. The column 'Word' represents OntoRo whilst the column 'Root Form' represents its stemmed alternative.

The column "Semantic Meanings" demonstrates the semantic polysemy of words represented with OntoRo. Since the words in column "Word" conflate to the morpho-

logical root "connect", stemmed OntoRo source will only contain the word "connect". The semantic ambiguity refers to 12 unique meanings for all forms of connect. For that reason, the semantic stemming proposed in the next section aims to conflate words in the stemming to less ambiguous morphological forms.

Table 4.3 Semantic representations of word stems in OntoRo

Word	Root Form	Semantic Representation[2]	Senses
connect	connect	$C_9, C_{45}, C_{62}, C_{71}, C_{202}, C_{305}$	6
connected	connect	C_9, C_{45}, C_{50}	3
connection	connect	$C_9, C_{11}, C_{45(2)}, C_{47}, C_{48}, C_{706}$	7
connecting	connect	C_{45}	1
connective	connect	C_{45}, C_{47}	2

4.3.2. Semantically Enhanced Text Stemming (SETS) algorithm

This section proposes a stemming algorithm, which relies on semantics to achieve text consistency. It is called a *Semantically Enhanced Text Stemmer* (SETS). A semantic approach is used to address the problem of words, which are conflated to the same morphological stem, but which have different semantic representations. The stems produced by the Porter stemmer are more semantically polysemy and when used in clustering the clusters produced are inferior. The proposed algorithm aims to improve cluster coherency by keeping the words less semantically polysemy.

[2] The semantic representation $C_{<number1>(<number2>)}$ stands for concept (C), concept number in OntoRo ($< number1 >$) and the number of senses in the concept ($< number2 >$) (in case of one senses the number is omitted), i.e. $C_9 - relation, C_{11} - consanguinity, C_{45} - union, C_{47} - bond, C_{48} - coherence$, $C_{50} - combination,$ $C_{62} - arrangement, C_{71} - continuity, C_{202} - contiguity, C_{305} - passage, C_{706} - cooperation$;

The errors from stemming described above can be minimised by using a semantic approach with external knowledge, i.e. such as OntoRo. The approach is similar to using a dictionary for searching words' stems after every step of the Porter stemmer (Krovetz, 1993). The difference with the algorithm proposed by Krovetz (1993) is that not the form of the word is important, but the semantic information acquired from the external knowledge source. Before a rule-based stemming to be applied on the words, the semantic meaning of words is considered. Thus, the words in the examples given for under-stemming ("experiment" and "experience") and over-stemming ("adhere" and "adhesion") will be stemmed by their meaning, which will alleviate these problems.

The proposed algorithm (fig. 4.1.) relies on semantics to achieve text consistency with the words' meaning. The semantics is used to address the problems regarding words, which conflate to a different semantic stem. Additionally, semantics is used to obtain the grammatical structure of the words and acquire the necessary information, which will conflate the words by meaning. Therefore, the algorithm recognises as similar those words that share common meaning. The algorithm implements the six steps of the Porter stemmer and after every step the produced stem is searched in the OntoRo. In the end if the stem is not found, the stem is searched in the stemmed OntoRo. If the stem does not occur in the stemmed OntoRo it is considered a named entity.

```
in:   w ... word
out:  s ... set of semantic meanings
s = ontoro_search_for_occurrence(w)
for(step = 1; s is {} and step <= 6; step = step + 1)
   w = porter(w, step)
   s = ontoro_search_for_occurrence (w)
end
if s is {}
   s = ontoro_stemmed_search_for_occurrence (w)
end
# if s is {}, w is a named entity
# otherwise s contains the semantic meanings of W
return s
```

Figure 4.1 Semantically enhanced text stemming algorithm

The proposed algorithm implements the six steps of the Porter stemmer with semantic enhancement by searching first for every word in every document in the OntoRo for occurrence of the morphological forms of the words. If a word is found in OntoRo it is considered for semantically stemmed and the algorithm proceeds to the next word. If the word does not occur in the lexical source, the algorithm proceeds with the first step of the Porter stemmer. After a stem is produced by the first step of the Porter stemmer it is sought for occurrence in OntoRo. If the stem is found in OntoRo, the word is semantically stemmed. This process is repeated for each of the six steps of the Porter stemming algorithm. Note that at this point, if the word is not found in OntoRo, it is in a Porter stemmed form. This form of the word is checked for occurrence in stemmed OntoRo. Finally, if the algorithm does not find any of the forms of the word in the lexical sources, it is considered for a named entity. Other-

wise, the algorithm returns the concepts with which the word is associated (the senses of the word).

The next stage is to use the semantic stems produced by the SETS algorithm to represent the documents in the VSM. For this purpose, the weights (TF-IDF) for all stems in the collection are calculated for every document. Then, using equation (1) the documents are represented in the higher order hierarchy of concepts yielding a matrix of concept indices $< concept, weight >$ (Setchi et al., 2011). The concept indexing reduces dimensionality, since the concept number is limited to the number of concepts in OntoRo.

This approach is similar to using a dictionary for searching the word stem after every step of the Porter stemmer (Krovetz, 1993). However, the aim of the proposed algorithm is to acquire less ambiguous semantic information and not real words. This is achieved by considering the senses of the words before applying rule-based stemming. The errors from under- and over-stemming are thus alleviated. The proposed algorithm still relies on TF-IDF to measure the discriminative power of words.

4. 4. Illustrative example

This section presents an illustrative example, which is part of the preliminary experiments conducted on the Wikipedia collection in the process of testing the SETS algorithm. The number of Wikipedia articles analysed to acquire statistical data of co-occurrence is 2,694,787. For illustrative purposes, the stemming example does not

contain a full article, but only the first sentence of two articles: Transport[3] and Cargo[4]. The first sentence of article "Transport" is shown in table 4.4. The results of stemming the sentence (document) with the Porter stemmer are shown in the next column of the same table. The sentence is stemmed one more time by the SETS algorithm (see table 4.4 third column). A representative weight value (respectively TF-IDF and concept indexing values) for each word/concept in the sentence is calculated by considering the sentence as a whole document (see table 4.5 and table 4.7). Then, the words are placed in a document-term matrix (see table 4.6). The document-term matrix produced by the Porter stemming algorithm is sparse.

Table 4.4 Article "Transport" from Wikipedia

Article "*Transport*"	Stemmed with the Porter stemmer	Stemmed with the SETS algorithm
Transport or transportation is the movement of people, animals and goods[5] from one location to another.	Transport transport movement peopl anim good locat	Transport transportation movement people animal goods location another

Table 4.5 Words co-occurrence in article "Transport" stemmed with the Porter stemmer (see Table 4.4)

Word	Occurrence	TF-IDF	TF	IDF
transport	2	1.15121	0.28571	4.02925
movement	1	0.49085	0.14286	3.43595
peopl	1	1.50040	0.14286	10.5028
anim	1	0.63437	0.14286	4.44055
good	1	0.43355	0.14286	3.03486
locat	1	0.61941	0.14286	4.33584

[3] http://en.wikipedia.org/wiki/Transport
[4] http://en.wikipedia.org/wiki/Cargo
[5] The word goods is hyperlinked to the article Cargo – http://en.wikipedia.org/wiki/Cargo

Table 4.6 Document-term matrix for article "Transport" stemmed with the Porter stemmer

	...[6]	transport	movement	peopl	anim	good	locat	...[7]
Doc	...[8]	1.15121	0.49085	1.5004	0.6344	0.4336	0.6194	...[9]

Table 4.7 Words co-occurrence in article "Transport" stemmed with the SETS algorithm

Word	Occurrence	TF-IDF	TF	IDF
transport	1	0.50366	0.125	4.02925
transportation	1	0.55868	0.125	4.46942
movement	1	0.42949	0.125	3.43595
people	1	0.26619	0.125	10.5028
animal	1	0.57115	0.125	4.44055
goods	1	0.60051	0.125	3.03486
location	1	0.43393	0.125	4.33584

Table 4.8 Morphological forms of the word *good*

word	porter
good	good
goodness	good
goods	good

The statistical data of co-occurrence, i.e. TF-IDF, calculated after the SETS algorithm is processed text, differs from the same data calculated after the Porter stemmer normalises the text. The first difference is that 'transportation' remains the same word and it is not conflated to 'transport'. The word 'transportation' is less ambiguous since it refers to 4 concepts only. On the other hand, the word 'transport' refers to 35 concepts. Every concept represents a different meaning or idea a word represents or is associated with. Hence, the SETS algorithm preserves less ambiguous word mean-

[6] Other words, which occur in documents from the same collection
[7] Other words, which occur in documents from the same collection
[8] Relevant weight of the words that occur in other documents
[9] Relevant weight of the words that occur in other documents

ings. The rest of the words people, animal, and location remain words and are not replaced with corresponding concepts. The word ‘goods’ is not stemmed to ‘good’, which otherwise would be a wrong conflation and would result in a shift of meaning. Similarly, table 4.7 shows all morphological forms of the word transport to demonstrate the ambiguity of that word. The Porter stemmer conflates any of the variants of these words to the same root stem transport. However, table 4.8 shows the semantic ambiguity of the grammatically conflated forms from table 4.7. The shift in meaning for the words good, goods, and goodness are displayed in table 4.9. The concept index for the same document yields another matrix, where the row represents the same article but presented by concepts – document-concept matrix.

Table 4.9 Semantic ambiguity of grammatically inflected forms of the word good

N	concept	word	porter
1	34	goodness	good
2	164	goods	good
3	272	goods	good
4	390	good	good
5	575	good	good
6	615	good	good
7	615	good	good
8	638	goodness	good
9	640	goodness	good
10	640	good	good
11	640	good	good
12	640	good	good
13	644	goodness	good
14	644	good	good
15	650	goodness	good
16	660	good	good
17	680	good	good
18	694	good	good
19	730	good	good
20	739	goodness	good
21	739	good	good
22	777	goods	good
23	795	goods	good
24	826	good	good
25	884	good	good
26	897	good	good
27	901	good	good
28	913	good	good
29	929	goodness	good
30	929	good	good
31	933	goodness	good
32	933	good	good
33	933	good	good
34	950	good	good
35	965	goodness	good
36	979	goodness	good
37	979	goodness	good
38	979	good	good

The semantic ambiguity of the word *good* is shown in table 4.9. According to OntoRo *good, goods and goodness* have 38 meanings together, i.e. the meanings are defined by the number of concepts. The Porter stemmer conflates these words from Table 4.9 to *good* (see Table 4.8) and makes the stem more ambiguous since it refers to

all 38 semantic meanings. Nevertheless, the SETS reduces the ambiguity for *goods* (to 4 – concepts 164, 272, 777, 795) and yet, it does not explicitly state which meaning should be used in the context of the document. Instead, the concept index is calculated to represent documents by following equation (1), i.e. the TF-IDF value for the word *goods* (0, 0.6005) from table 4.6 is divided by the number of semantic meanings for the same word (*goods* has 4 meanings see table 4.9). Then, the result (0, 6005 ÷ 4 = 0, 1501) represents the statistical representation of the concepts (equally) acquired for this particular word. The concept index for the document is calculated in the end by accumulating the statistical representation for every concept acquired from all words. Then, the statistical co-occurrence for all concepts is shown in table 4.10. The concepts are sorted in descending order of their representation to the document.

The normalisation even of a sentence demonstrates the differences in the document representation. The practical difference in both approaches is that the Porter stemmer normalises documents and a TF-IDF weighting represents them in VSM (see table 4.5) by a sparse matrix (document-term matrix). On the other hand, the use of the SETS and concept indexing to represent documents outputs a dense compressed by design matrix (document-concept matrix). The matrix is compressed since the number of concepts is pre-defined by the OntoRo. In addition, short documents contain a few words and their representation is limited to the use of these words only. Therefore, clustering algorithms are limited to group them in clustering solutions with higher density, i.e. clusters with prevailing number of documents. On the contrary, even the short document represented by concept index produce a dense row in the document-concept matrix allowing better separation of the documents.

Table 4.10 Concept indexing of the first sentence of article "Transport" (only the first 11 out of all 51 concepts shown)

N	Concept Number	Concept Name	Occur	SETS
1	272	transference	4	0.34576
2	265	motion	4	0.24335
3	365	animality	3	0.15578
4	371	humankind	2	0.05324
5	191	dweller	2	0.05325
6	38	addition	2	0.15374
7	186	situation	2	0.10848
8	963	punishment	2	0.16765
9	944	sensualism	2	0.10385
10	187	location	2	0.08086
11	305	passage	2	0.16765
...	...	...	...	...
51	722	combatant	1	0.02798

Article "Cargo" from Wikipedia:

Cargo (or freight) is goods or produce transported, generally for commercial gain, by ship or aircraft, although the term is now extended to intermodal train, van or truck

A demonstration of stemming two documents (sentences) and measuring similarity is presented in table 4.11. The article "*Cargo"* from Wikipedia is not selected randomly. The first sentence of article "*Transport"* is hyperlinked to article "*Cargo"* via the word "goods". This states a logical link between the two articles and yields certain similarity. The two articles, according to the traditional approach, are similar only by the word "good", i.e. the Porter stemmer produces that word after stemming the articles. However, for article "*Cargo"* this stemming yields shift in the meaning of the word.

Table 4.11 Semantic ambiguity and similarity of the words *goods, transport, cargo,* and *ship*

N	concept	goods^{Transport}	transport^{Transport}	cargo^{Cargo}	ship^{Cargo}
1	32	☐	☐	☒	☐
2	164	☒	☐	☐	☐
3	193	☐	☐	☒	☒
4	272	☒	Occurs 2 times ☒	☒	Occurs 2 times ☒
5	777	☒	☐	☒	☐
6	795	☒	☐	☒	☐

☒ - has an occurrence in OntoRo for the particular concept (sense)
☐ - lacks an occurrence in OntoRo for the particular concept (sense)
a word is likely to have more than one occurrence in OntoRo, i.e. occurrences represent different part of speech (POS) words (verb, noun, adjective etc.) and/or different meanings within the same POS

In table 4.11 is shown the mechanism of detecting similarities by using the concept indexing. Both articles "*Transport"* and "*Cargo*" share the word "goods" (which has semantic meanings referring to concepts 164, 272, 777, and 795). The word "transport" (a word from the first article) contributes to the scoring results of concept 272 for the same article. On the other hand, from the second article words "cargo" and "ship" contribute to the scoring result for representing the article by concept 272, 777 and 795. Thus, both articles are similar by three concepts 272, 777, and 795 represented by different weight and thus, the logical connection through the word "goods" is revealed, i.e. a combination of three concepts with relevant weight. The weight of the concepts can be considered as a context for the article. The traditional stemming approach and representation by chance manage to relate both documents through the word stem "good". However, the word good occurs in documents with different meaning such as "good" and "goodness", which as a result makes the clustering inferior.

4. 5. Evaluation of clustering solutions with silhouettes

The evaluation of the proposed SETS algorithm is performed on the Reuters-21578 text categorization test collection. For the purpose of the evaluation, the corpus was transformed into VSM three times. Firstly by using the proposed SETS normalisation, secondly by using the classical approach of TF-IDF weights, where tokens are transformed into normalised forms (stems) by using the Porter stemming algorithm, and thirdly by using the tags of the articles[10]. The implementation used in the evaluation is the Common Lisp version of the algorithm made available by Porter. This transformation into VSM yields three matrices where each row is the vector representation of a Reuters news article in TF-IDF or respectively OntoRo vector space or tags space. The TF-IDF matrix is a sparse matrix with 18457 rows and 44293 columns – one for each unique word in the corpus. The SETS matrix has 18457 rows and 990 columns, one for each concept in OntoRo and this matrix is dense. The matrix used for the human judgement contains 18457 rows (observations) and 445 columns (attributes). The former TF-IDF matrix was produced for 8.25 minutes and the later (SETS) for 43.03 minutes. The matrix that represents the Reuters tags is used only for quality measure of the clustering and no time measures are collected. The same matrix is produced by weighting the tags using $\frac{1}{[number\ of\ concepts]} = [weight\ per\ concept]$ formula. Thus, the matrix represents the 18457 observations into 445 dimensions, and every dimension does not have a binary presence but a continuous one.

[10] 1862 articles in the Reuters21578 corpus have no tags. These articles are removed from the collection. In addition, 1259 articles have no concept index. Therefore, they are also removed from the collection, which number of articles for the experiments is reduced to 18457.

The three matrices are clustered using the spherical k-means algorithm (Dhillon et al., 2002), which is available in the CRAN repository. This version of the algorithm is fast and requires as input the number of clusters. Experiments are performed to split the data in 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 clusters. The quality of the clusters is assessed with the silhouette measure proposed by Rousseeuw (1987). This measure is defined as

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (4)$$

where $a(i)$ is the average dissimilarity of the object (row) i to all other objects of a cluster A and $b(i)$ is the average dissimilarity of the object i element A to all objects in the cluster nearest to i. To assess the overall quality of a clustering model (M), the measure $p(M)$ was averaged over all objects:

$$p(M) = \frac{1}{n}\sum_1^n s(i) \quad (5)$$

where n is the total number of news articles. The measure is a number between 0 and 1 with higher number suggesting stronger cluster coherence.

The first evaluation of the clustering solutions produced on the three matrices use natural dimensionality of the document representation, i.e. no dimensionality reduction techniques are used. The SETS represents documents in reduced dimensionality by design (990 is the number of concepts in OntoRo). Although, the dense matrix produced after text normalisation with the SETS algorithm and represented by concept index is a thin matrix with reduced dimensionality s-kmeans performs faster (time-wise) clustering on the sparse document-term matrix, produced after the Porter stemmer normalises text and TF-IDF weighting is used, where 99.99% of it is popu-

lated with zero values. In addition, to produce the dense document-concept matrix the SETS algorithm requires 5.11 times more time than the standard Porter stemmer to normalise text. This is as a consequence of searching the stems in OntoRo after every step of the Porter's algorithm. The memory footprint of the dense matrix is 7.02MB, whilst the other TF-IDF-weighted matrix in dense representation is 1.7GB and 25MB in sparse representation.

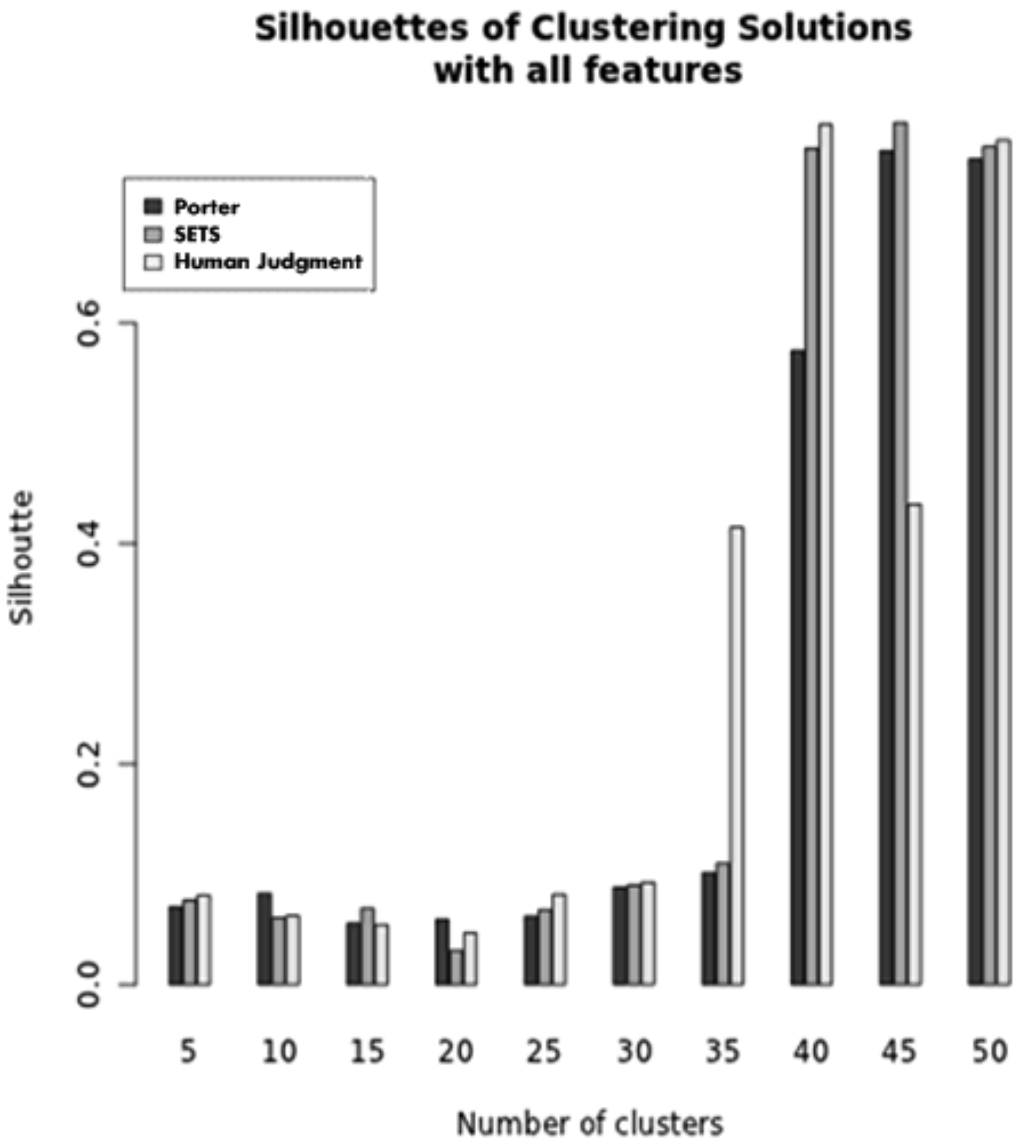


Figure 4.2 Performance evaluation of the Porter stemmer, SETS, and Human Judgement (full-feature space)

The first of a few series of experimental results are summarised in Fig. 4.2. The evaluation of the performance demonstrated by SETS shows that the algorithm is outperformed by the Porter when the number of clusters is 10 – the reduced representation provided by SETS cannot separate clusters in this particular solution. However, the separation provided by SETS for 10 clusters is closer to human judgement. For all

series of clustering SETS normalises the documents in closer relation to human judgement than the Porter does. In figure 4.2 can be noted that a clustering solution for all 18.578 document clustered in 45 clusters do not separate clusters well, although the documents normalised by the Porter stemmer and by the SETS algorithm provide good results in terms of clustering silhouettes. This clustering is the only one for which document normalised by the Porter stemmer align better to human judgement than the SETS's. The overall performance of the SETS document normalisation is better than the Porter's and thus, the separation of the clustering solutions is improved. Therefore, the conclusion that semantic stemming of the words, i.e. semantic-based word disambiguation, provides clustering solutions with better coherence, which aligns better to human judgement, can be made.

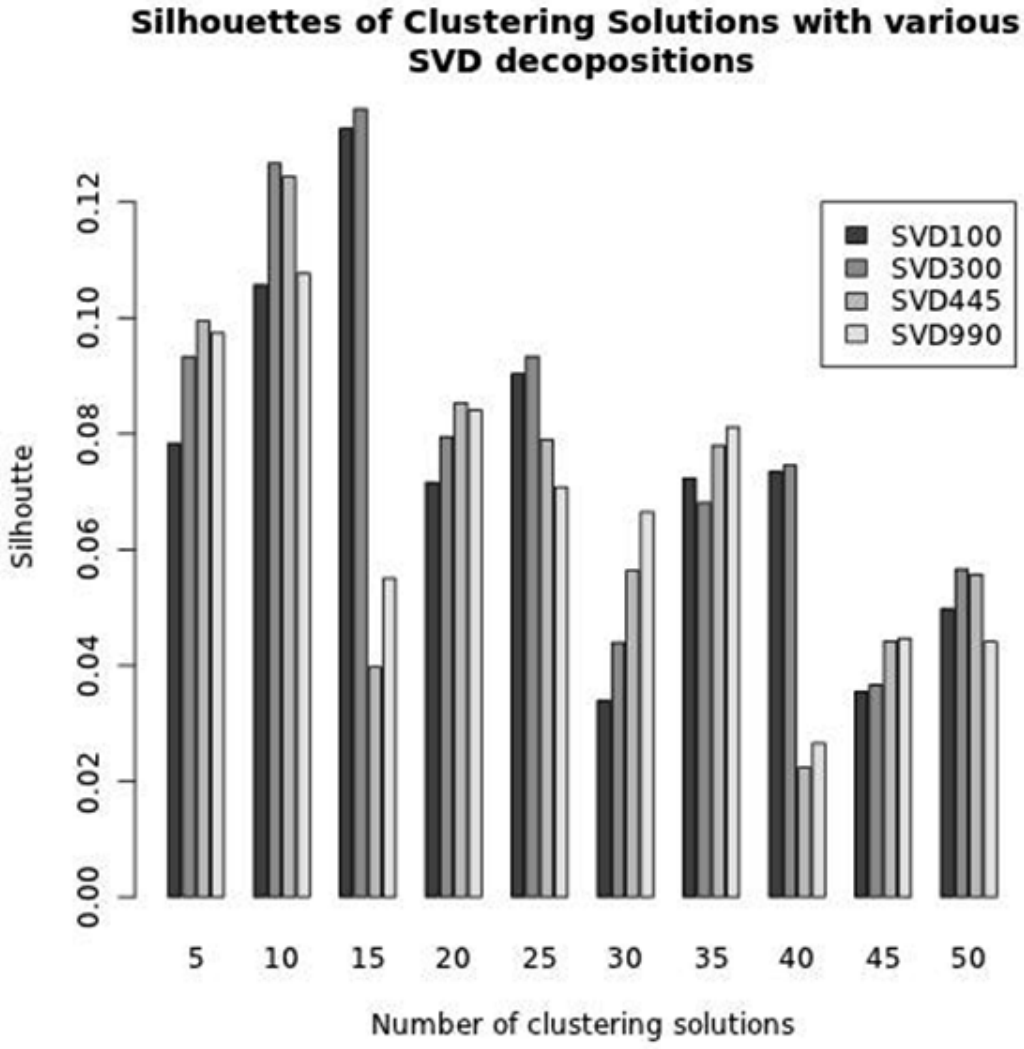


Figure 4.3 An example of clustering silhouettes

Clustering solutions in reduced dimensionality theoretically should produce similar results with minimal reconstruction error. Therefore, series of experiments of the same data but with reduced dimensionality are conducted. The two most commonly used techniques for dimensionality reduction are Principal Component Analysis (PCA) and Singular Vector Decomposition (SVD). In the presented evaluation is used the SVD approach. The reason is because the three matrices are different as the matrices of SETS (i.e. 18,578x990) and TAGS (18,578x445) are "thin" whilst the Porter produces a "fat" matrix (i.e. 18,578x44,134). PCA employs different component

analysis techniques to reduce dimensionality of “thin” and “fat“ matrices whilst the SVD approach uses the same techniques. Therefore, the latter is the more consistent choice for dimensionality reduction. In fig. 4.3 is shown that s-kmeans performance is relatively consistent over matrices produced by the Porter stemmer for various dimensions, i.e. reducing dimension is relatively stable and the loss of information is insignificant. Exception is only clustering solutions that split the data into 15 and 40 clusters. The inconsistent performance is only for reduced dimensionality above 300 components. The usual dimensionality reduction considers between 100 and 300 components (Berry et al., 1995).

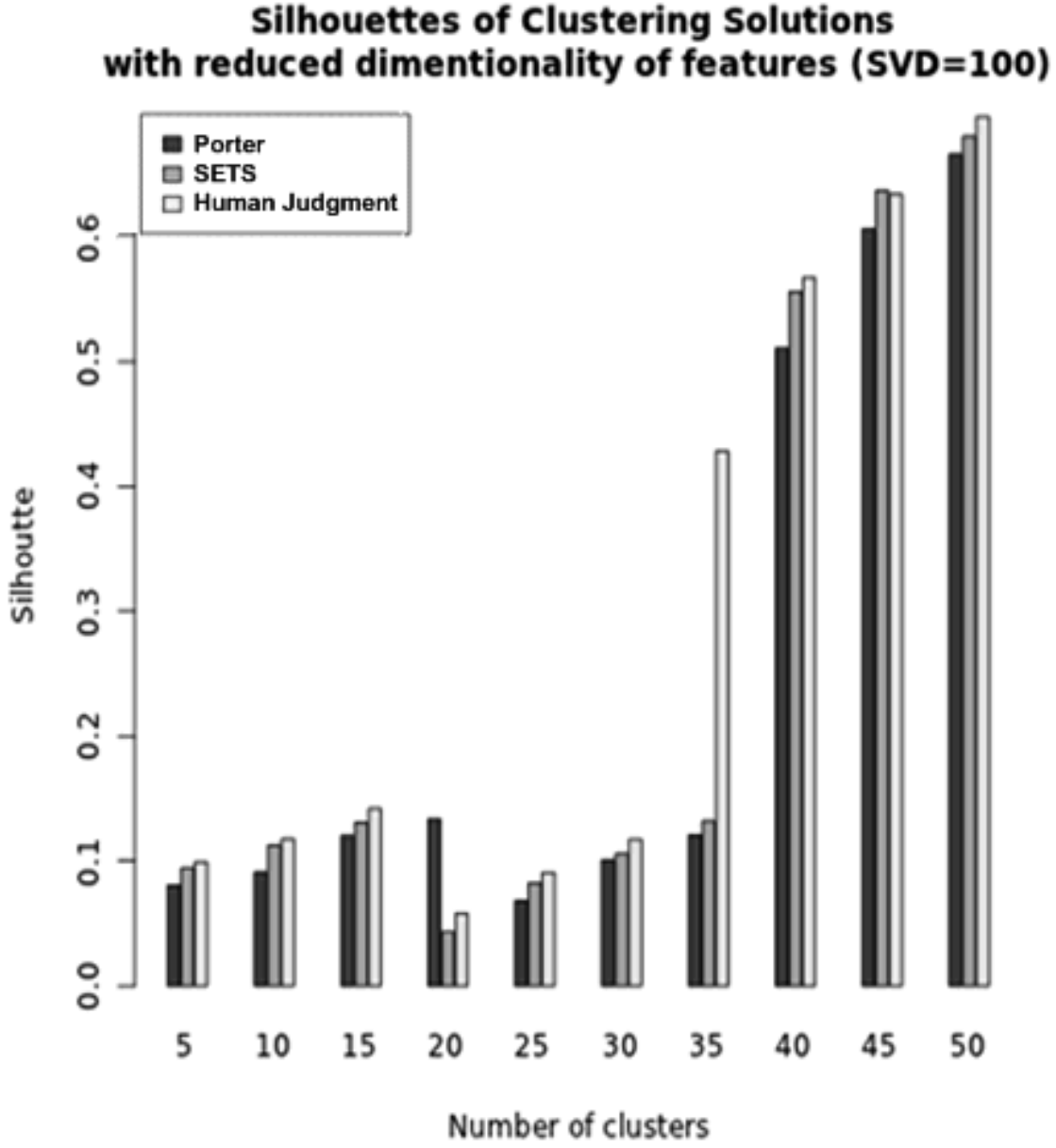


Figure 4.4 Clustering silhouettes (one hundred dimensions)

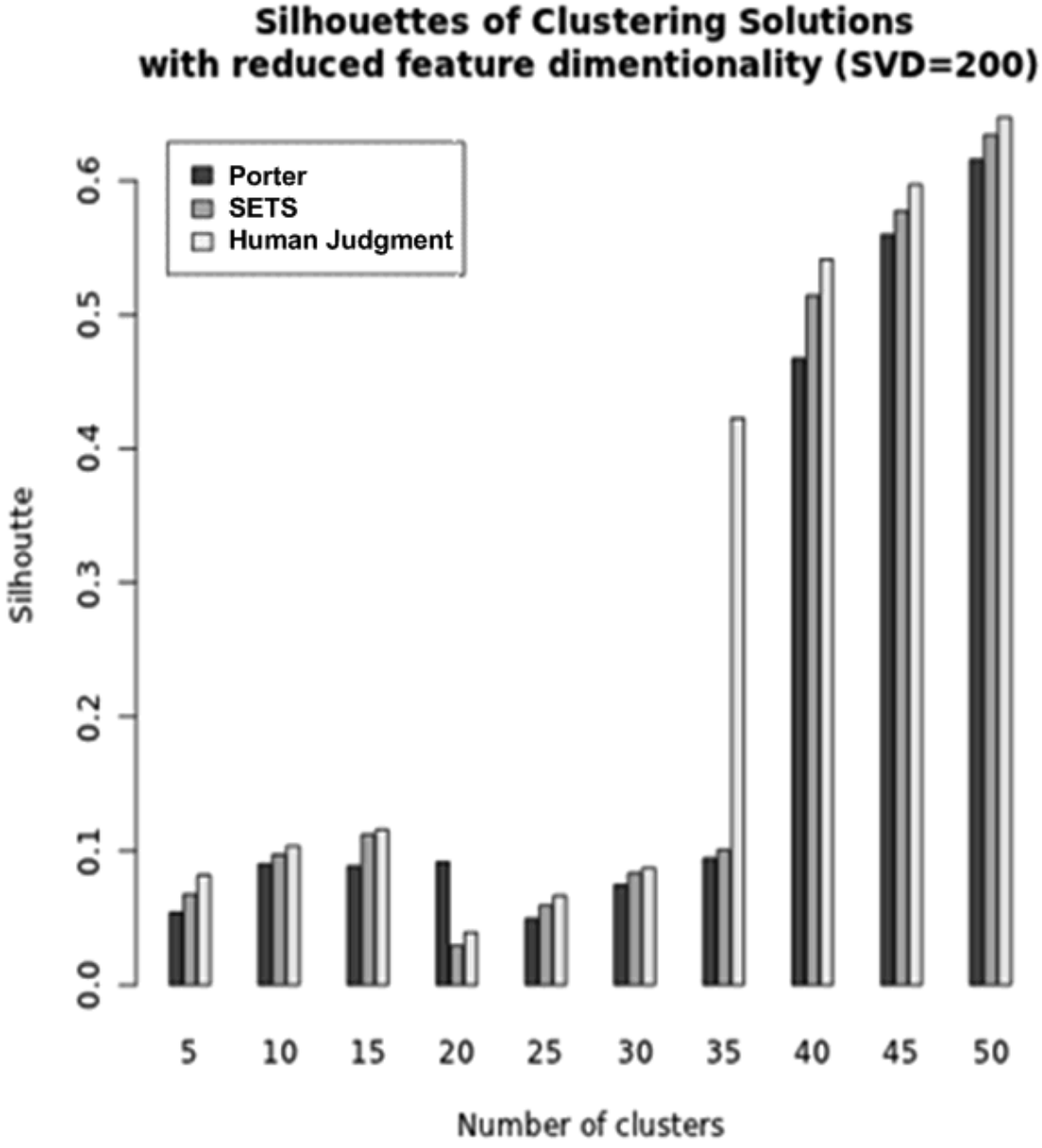


Figure 4.5 Clustering silhouettes (two hundred dimensions)

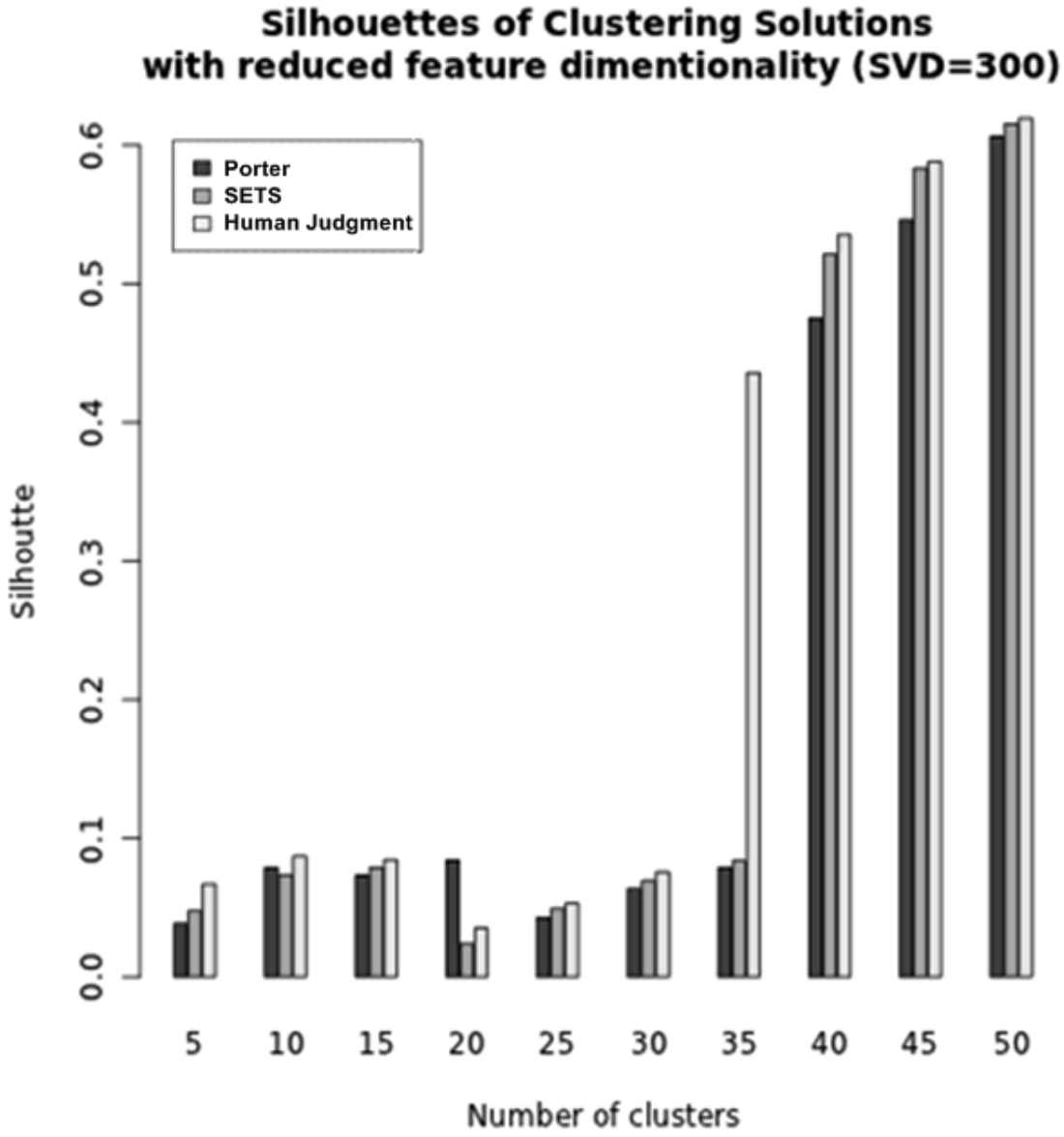


Figure 4.6 Clustering silhouettes (three hundred dimensions)

Figures 4.4 to 4.6 demonstrate that the silhouette values for all clustering solutions produced by using the SETS document normalisation always align better to human judgement than the Porter stemmer's normalisation. However, the only exception for that alignment is when more components (300) are considered. Therefore, if the dimensionality is reduced to less than 300 components, document representation with the SETS algorithm will always provide better clusters.

This evaluation clearly shows that the proposed document representation based on concept indexing, which employs general ontology, provides better separation between clusters than the traditional BOW representation. The results support the view stated by Hotho et al. (2001, 2003a) that an ontology improves document clustering. In contrast to the literature, though, the evaluation in this chapter proves that better separation between clusters can also be achieved by using a general ontology. A significant difference between the proposed clustering approach and the clustering suggested by Hotho et al. (2001, 2003a) is that the former produces objective and generic clustering solutions and is not biased towards the specificity of the used ontology (core). Concept indexing relies on no further dimensionality reduction and is successfully deployed on a large scale.

4. 6. Summary

This chapter presents a comparison of the Porter and the proposed SETS stemmers when employed in document clustering. The performance of stemmers is compared against human judgement. In addition, the chapter provides evaluation of the concept indexing as a document representation approach in the domain of clustering.

The proposed stemmer has the advantage to work well across domains. In collections, where documents are topically grouped based on named entities, the algorithm is expected to perform worse. The performance can be improved by using a domain specific ontology. Clustering solutions obtained by the partitional clustering algorithm k-means demonstrate, according to the results, better separation between and improved coherency within the clusters generated. The clustering solutions generated

with document normalised by SETS are better aligned to human judgment than those normalised by traditional stemmers (e.g. the Porter stemmer).

The number of clusters throughout the clustering solutions is made with the purpose to explore a broad range of the parameter space without considering the optimal number of clusters for the Reuters21578 data. Thus, the experiments demonstrate that the SETS algorithm enhances better separation between and improved coherence within clusters, even when the best number of clusters is unknown.

The presented evaluation demonstrates that SETS in conjunction with concept indexing provides reduced dimensionality in document representation, which allows more advanced clustering algorithms that are impeded by the high dimensionality of documents, to be used to produce clustering solutions on a large scale. In addition, the approach can employ a domain specific ontology in conjunction with OntoRo to produce clustering solutions from the perspective of the domain knowledge. The scalability and the flexibility of the approach in terms of number of documents and different perspectives to a collection of documents will enable the document clustering domain to seek new approaches to grouping documents and discovering relationships between them.

Chapter 5 : Evaluation of document similarity measures to human judgement

This chapter presents a method for measuring document similarity on a large scale by using many-to-many matching. Firstly, a discussion is presented on inadequacies and deficiencies of the traditional similarity measures used for measuring pair-wise similarities between documents. Secondly, approaches which alleviate the limitations of the traditional measures are presented. Finally, a multidimensional approach for measuring document similarity based on content distribution is proposed to be used for clustering.

5. 1. Similarity measure and consistency with human judgement

Similarity measures such as the cosine measure (Salton and Buckley, 1998), the Dice measure (Zobel and Moffat, 1998), the Jaccard measure (Jones and Furnas, 1987), the Overlap measure (Blair, 1979, Baeza-Yates and Ribeiro-Neto, 1999) and the information-theoretic measure (Aslam and Frost, 2003) proposed in the literature do not consider the contextual meaning of documents when measuring their pair-wise similarity. The traditional techniques use collectively identified sets of keywords to represent the content of the documents in a given dimensional term space.

5.1.1. Document representation towards human judgment

The vector space model (VSM), which is the most commonly used document representation technique, positions every document in a multidimensional term space. The purpose of using such a complex and sophisticated dimensional space is that any spa-

tial proximity also stands for a relevant semantic proximity. Therefore, documents positioned close to each other are assumed to share certain commonality.

Documents are positioned in a multidimensional term space by algorithms that use VSM to represent documents. Therefore, the accuracy of the document position in the term-space depends on the quality of the coordinates. The coordinates (words and/or concepts) are acquired from the documents by using statistical term weighting techniques. The most common weighting techniques used for identifying the representative values of the term sets are TF-IDF, LSI and sets of multi-words (Zhang et al., 2011). The LSI technique performs generally better than both the TF-IDF and the multi-words approaches and provides good statistical and semantic discrimination power for the representative sets (Zhang et al., 2011). In the previous chapter it is shown that semantically enhanced text stemmer, which enables concept indexing to be used for document representation performs better than the Porter stemmer and TF-IDF weighting document representation in the domain of document clustering. The performance of concept indexing over other document representation techniques is measured in terms of the separation of clusters and the similarity between documents within clusters. The clustering solutions obtained by employing concept indexing as a document representation technique is more consistent with human judgement. In other words, semantically enhanced text normalisation and concept indexing provide better silhouette values than the text normalisation provided by the Porter stemmer and represented by the TF-IDF weighting.

5.1.2. Similarity measure towards human judgment

The similarity measures establishes the similarity between two documents. The similarity increases with greater commonality and decreases with greater differences in a common feature space (Lin, 1998b). However, the aforementioned traditional similarity measures do not seek commonality in either a shared document context or in the structural similarity between documents (Wan and Peng, 2005b). Nevertheless, the cosine similarity measure is robust with good quality even across domains to certain extend (Dhillon and Modha, 2001) and is used in the domains of data mining and document retrieval as a base-line similarity measure (Wan and Peng, 2005b).

The TextTiling technique (Hearst, 1993) measures similarity by capturing document structures. This technique subdivides a text document into multi-paragraph units that represent passages or subtopics called text tiles. This approach considers that one document contains more than one topic from a generic topic set, which is to assign topics to documents. Thus, a set of subtopics identified by the algorithm constitutes a context that can be associated with the document. This approach is recognised by the research reported in this thesis as more scalable. The trade-off made in favour of scalability is to represent a document with a subset of topics. Approaches that consider one document to constitute one topic (Witten, 2010) are disregarded by this thesis.

The identified text tiles are used to capture patterns of subtopics contained in text and their distribution across documents of a collection is used for measuring similarity. The approach uses three algorithms: (i) lexical analysis based on TF-IDF, (ii) information retrieval measurement to determine the extent of the tiles, and (iii) a statistical disambiguation algorithm which relies on information from a thesaurus. The au-

thor of this technique reports good text segmentation, which is consistent with human judgments (Hearst, 1997).

The TextTiling technique is employed by a document similarity search algorithm to capture document subtopic structure in plain text and find documents similar to a query document. Then it returns a ranked list of similar documents (Wan and Peng, 2005b). The similarity model considers the outlined subtopics from the document structure and calculates the similarities for different pairs of text segments. Finally, the overall similarity between the documents is measured by combining the similarities of different pairs with the optimal matching method (OM-based method).

5.1.3. Matching of features for measuring similarity

The optimal matching (OM) and maximal matching (MM) are graph theoretic problems. These matching approaches are suitable for measuring dissimilarities between documents because two documents can always be represented as a bipartite graph (or bigraph) with its vertices divided into two disjoint sets. As shown in Fig. 5.1 the disjoint sets are document $X = \{x_1, x_2, x_3, x_4\}$ and document $y = \{y_1, y_2, y_3\}$. Both documents consist of segments, i.e. text tiles. Every edge shown in Fig. 5.1 and Fig. 5.2 connects a vertex in X to one in Y and X and Y are independent. Bigraphs are often denoted as $G = \{X, Y, E\}$, where $E = \{e_{ij}\}$ is a set of edges connecting X and Y. A matching M in G is defined as an independent pair-wise and non-adjacent set of edges, which shares no common vertices.

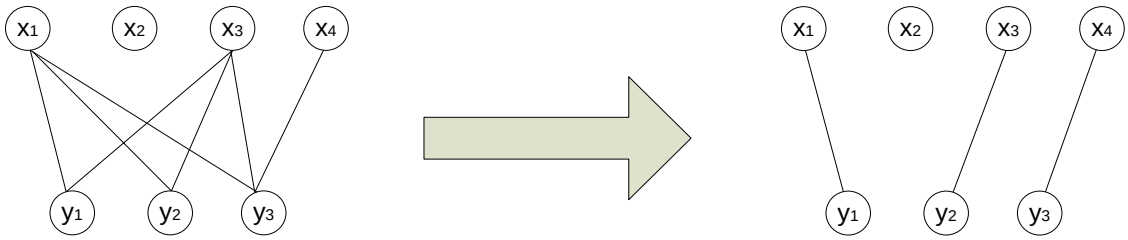


Figure 5.1 Maximal matching

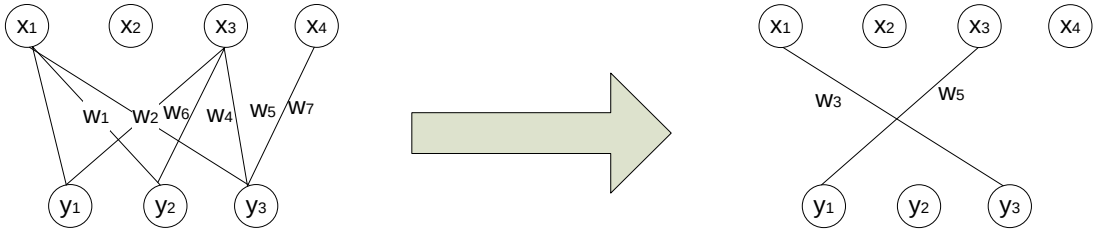


Figure 5.2 Optimal matching

Optimal matching (Fig. 5.2) is an extension of the maximum matching (Fig. 5.1). The latter is usually used on an un-weighted bigraph (Fig 5.1) with the goal to find the matching M with the maximum number of edges. Optimal matching on the other hand is used to assess dissimilarities between two independent sets X and Y, which represent a weighted bigraph (Fig. 5.2 weights $W = \{W_1, W_2, W_3 \dots\}$ are assigned to the edges). The task of OM is to find a matching, which calculates the maximal weight. Both measures are used successfully in measuring the similarity between documents in context-based retrieval.

Experimental results in the domains of document search and document retrieval demonstrate that the TextTiling approach in conjunction with the OM technique for measuring pair-wise document similarity is a very effective approach. This approach outperforms other state-of-the-art retrieval models such as Okapi's BM25 model, Smart's vector space model with length normalisation as well as the cosine measure (Wan and Peng, 2005b). As a result the OM-based pair-wise similarity approach with

TextTiling decomposition of the text documents is a prominent method for improving the consistency of the current state-of-the-art document retrieval and document clustering algorithms with human judgement.

5. 2. Multi-dimensional approach to pair-wise document similarity

This section discusses a multi-dimensional approach – the Earth Mover's Distance (EMD) algorithm – to measuring similarity between documents. The EMD-based similarity measure demonstrates better document retrieval results when compared with human judgement than the traditional similarity measures such as the cosine measure, the Jaccard measure, the Dice measure, the overlap measure, and the IT-slim measure (Wan and Peng, 2005a).

5.2.1. Similarity between documents using distributions

The advantage of the EMD-based measure over the traditional similarity measures is that it considers the documents as multi-dimensional distributions. Therefore, a pair-wise similarity between documents is computed as a similarity between two distributions. The process is regarded as many-to-many matching and includes matching between a pair of words (features) contained in separate documents (distributions). A pair of documents is regarded as two sets U and V (see Fig 5.1 and 5.2), where the members of the sets are the words from the relevant document. Wan and Peng (2005a) suggest WordNet as a knowledge source for a ground distance needed by the EMD to measure the similarity between words. Then the algorithm scales up the individual distances between the words to full distributions by means of documents. The knowledge source provides semantic information that is used by the algorithm to dis-

ambiguate the words and measure the distance between them. This distance is regarded as a semantic distance. The EMD calculates similarity ($sim = 1 - EMD$) between any two words regardless of their context. In related literature (Wan and Peng, 2005b, Yang et al., 2008) it has been demonstrated that pair-wise similarity between documents improves when the context of the words is used.

Concept indexing (Setchi et al., 2009) considers the context of the documents in their representation. In addition, it provides a statistical relevance of the document context with respect to the other documents in the collection. A methodology for measuring document similarity based on EMD and concept indexing is suggested in section 5.5 and is evaluated in section 5.7.

5.2.2. Advantages of distributional similarity

Many-to-many matching overcomes the limitations of the traditional one-to-one matching provided by the VSM representation. The semantic distance between a pair of words is measured with a context vector by using a semantic knowledge source. The distance measured with a context vector establishes relatedness between words by measuring the angle between vectors. This approach of measuring semantic distances combines statistical information about the words derived from a large corpus and external knowledge. The results reported by Patwardhan (2003) demonstrate very close word similarities to human judgement. Since the approach is not domain specific it can be used for measuring similarity on any number of documents with no restrictions on the words, i.e. nouns, verbs etc, which are a problem in other approaches (Resnik, 1995).

The properties of the EMD-based metric are investigated in the literature for the purpose of content-based document retrieval (Patwardhan, 2003, Wan, 2007). The EMD algorithm is based on a solution to the transportation problem from linear optimisation. It calculates the minimal amount of work that must be done to transform one distribution into another in a precise sense. Therefore, it enables natural partial matching. The algorithm successfully operates on representations that vary in length. It provides a true metric when distributions with the same overall mass (significance in the representation) are compared. The algorithm constructs a weighted graph $G = \{A, B, D\}$ of two documents $A = \{(t_{a1}, w_{a1}), (t_{a2}, w_{a2}), \dots, (t_{am}, w_{am})\}$, where t_{ai} is a unique word for document A and w_{ai} is the relevant statistical weight. Analogically, document B is presented as $B = \{(t_{b1}, w_{b1}), (t_{b2}, w_{b2}), \dots, (t_{bn}, w_{bn})\}$, where t_{bj} is a unique word for document B and w_{bj} is the relevant statistical weight. Then a distance matrix $D = \{d_{ij}\}$ or a function that returns the distance between words i and j is employed. The graph G has vertices $V = A \cup B$ and edges $D = \{d_{ij}\}$. The transportation problem is defined as the minimal flow problem (minimal overall cost or work that needs to be done to equalise two distributions) $F = \{f_{ij}\}$, where f_{ij} is the flow between features i and j. The constraints on the flow are:

$f_{ij} \geq 0$ for $\forall \quad 1 \leq i \leq m; \quad 1 \leq j \leq n;$

$\sum_{j=1}^{n} f_{ij} \leq w_{ai} \quad 1 \leq i \leq m;$

$\sum_{i=1}^{m} f_{ij} \leq w_{bj} \quad 1 \leq j \leq n;$

And the goal function which is to be minimised is:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} d_{ij} = \min\left(\sum_{i=1}^{m} w_{ai}\ , \sum_{j=1}^{n} w_{bj}\right)$$

The distance EMD is defined as $EMD(A,B) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}d_{ij}$. The final result may need normalisation in order to avoid favouring shorter documents in case of partial matching. Finally, the similarity between documents is defined as $SIM(A,B) = 1 - EMD(A,B)$. The more similar two documents are, the shorter the distance is.

5. 3. Analysis of EMD, OM, and cosine similarity measures

The similarity measures based on words and phrases such as the cosine, the Dice and the Jaccard similarities as well as the overlap and the information-theoretic measures are compared to the context-based similarity measures such as the OM-based and the EMD-based (Wan, 2007). The OM-based and the EMD-based approaches use the TextTiling algorithm to decompose the documents into subtopics. The evaluation adopts the non-interpolated Mean Average Precision (MAP) and the precision (P) at the top N results (P@N) to compare different similarities.

5.3.1. Performance of similarity measures

The results demonstrate that the EMD-based measure outperforms the others including the OM-based measure. An observation, pointed out by Wan (2007), is that from all VSM-based measures, the cosine measure achieves the highest result on the MAP value and comparable values to the Jaccard and the Dice measures at the P@5 and the P@10. This supports the Wan and Peng's (2005a) conclusion that the cosine measure performs better than the other VSM-based measures in relation to human judgement. In addition, Wan (2007) analyses the context-based approaches, which outperform the other similarity measures. Since these approaches rely on the subtopic

structure of documents, Wan (2007) investigates the influence of the different document structures to the performance of the similarity measures. A hierarchical agglomerative clustering algorithm is used to group sentences with similar subtopics together and to produce separate documents with different structure. Another set of documents, for the same purpose, is produced by a sentence clustering algorithm. The difference in processing documents from the different sets is that the TextTiling approach processes consecutive sentences, which makes it suitable for the first set of documents. In the second set of documents the sentences may not be consecutive. The sentence clustering algorithm, at the beginning, considers each sentence for an individual cluster. Then, a cluster with the largest similarity value to another cluster is both merged to form a new cluster until. This iteration continues until the similarity values measured between clusters are all under a pre-defined threshold. The different sets of documents are produced using a different threshold similarity value. A sentence clustering algorithm then derives a subtopic structure for the document. However, Wan (2007) omits that this approach can be considered as a subtopic summarisation and the result of it is a document with higher density of similarly closed subtopics. Also skipping sentences from text, i.e. sentences that are insignificant in the context of the other sentences can be interpreted as noise reduction in the documents. Sentences, which contain insignificant or less information to the subtopics of the document, are ignored. Furthermore, the threshold (P@5 and P@10) modifies the measure of importance to noise ratio. A higher value for the threshold contributes to less noise in the clustering solutions. The analysis of the described approach support the manifested in chapter 3 idea that different thresholds can correspond to different abstraction levels in the view of documents. Therefore, a given level of abstraction corresponds to certain data granularity.

5.3.2. Complexity of similarity measures

The time complexity of the Cosine, OM-based, and EMD-based measures is empirically evaluated (Wan, 2007). The execution time of the cosine measure is 3.49 times less than the OM-based measure and 3.68 times less than the EMD-based measure. This result is explained with the complexity of the algorithms, which need to build a graph structure for the OM and EMD based measures and solve optimisation problems, i.e. to measure similarity between two distributions EMD builds matrix and solves the transportation problem. There are two strategies for reducing the time complexity and computation times for EMD, which improve the retrieval effectiveness and provide users with a real-time response to their queries.

The analysed results demonstrate that the EMD approach outperforms the OM approach in all cluster sets produced by the sentence clustering algorithm and also show the high dependency of the latter on the TextTiling decomposition. This concludes that the EMD approach is very robust because the results are independent from the text decomposition technique employed. However, the fact that the OM approach employs TextTiling (Wan and Peng, 2005b), which uses a context vector for measuring word similarity and produces good results into human judgement (Patwardhan, 2003), and the fact the OM-based approach with optimal matching outperforms the other similarity measures in document retrieval (Wan and Peng, 2005b) leads to the conclusion that the OM-based approach is the best approach in measuring document similarity. In addition, it could perform well into human judgement by taking advantage of the context of words (Patwardhan, 2003). On the other hand, the fact that the EMD approach outperforms the OM-based approach leads to the conclusion that EMD

could also produce results closer to human judgement. Moreover, the EMD-based algorithms does not rely on the TextTiling technique, but on measuring the similarity between two multi-dimensional distributions of subtopics to calculate the similarity between two documents (Wan, 2007). This technique considers the context of the words in emerging subtopics of documents and therefore, the produced clustering results are closer to human judgement.

5. 4. A Similarity measure based on external knowledge structure

This section explains an approach for creating a distance matrix from an ontology with a tree structure. A distance matrix is created in the domain of semantics for the purpose of producing semantic chains (Jarmasz and Szpakowicz, 2003a). A semantic chain relies on a distance matrix to provide a semantic reference. The semantic reference is used to measure similarity between words topologically located next to each other in a chain.

An ontology provides an explicit conceptualisation and taxonomy for knowledge model of a domain. Therefore, by considering the introduced ontology OntoRo, the established relationships between classes (see Fig 2.2) in the ontology can be used to measure the similarity between any two concepts (see Table 5.1). The process of computing this similarity depends on factors such as (1) similarity of the properties between concepts; (2) semantic distance between concepts; (3) hierarchy depth of the concepts; and (4) regulatory parameters (Yang et al., 2008).

The first factor computes how many similar properties every two concepts share. The more properties they share the closer topologically in the ontology structure they are. The second factor computes the distance based on the shortest edge path between

every two concepts. This kind of semantic similarity can also be computed on thesauri or ontologies with a hierarchical tree structure(Jarmasz and Szpakowicz, 2003a)(Jarmasz and Szpakowicz, 2003b)(Jarmasz and Szpakowicz, 2003b)(Jarmasz and Szpakowicz, 2003b). The factor with regard to the hierarchy depth considers the depth of the ontology tree structure so that the semantic similarity of a concept increases by the total of its hierarchy. This factor represents how abstract the concept in the ontology is with regard to other concepts. The last factor considered is a specific parameter. This parameter allows a given semantic similarity to be increased over others. A regulatory factor is used by Setchi and Tang (2011) when a specific and general ontology are employed in parallel to provide more discriminative power to the domain specific concepts contained in the specific ontology. The relations contained in the domain

specific ontology magnify the measured similarity towards the domain specific knowledge.

Table 5.1 Concept distance matrix produced using the OntoRo's structure

C(990)	1	2	3	4	5	...	990
1	0	0.12	0.25	0.5	1	...	1
2	0.12	0	0.12	0.25	0.5	...	1
3	0.25	0.12	0	0.12	0.25	...	1
4	0.5	0.25	0.12	0	0.12	...	1
...	...	...	...	...	...	...	...
990	1	1	1	1	1	...	0

The distance matrix in Fig. 5.1 is built from the OntoRo's structure (see Fig. 2.2) by following the semantic distance between concepts (factor 2) and hierarchy depth of the concepts (factor 3). Factor 1 is considered in the selected document representation SETS, where the words in the documents are replaced with related concepts. On the other hand, factor 4 is reserved for future work to augment the approach proposed in the next section for measuring similarity between multi-dimensional distributions in a specific domain. Hence, a general approach to measuring pair-wise document similarity is proposed. It relies on a distance matrix acquired from a generic ontology (e.g. OntoRo) to measure similarity between documents. The method employs the same ontology used in the stages of document normalisation and document representation to produce concept index. Factors 2 and 3 are used to produce table 5.1 by using a distance between a pair of concepts from a tree structure in the OntoRo. The longer the path connecting two concepts on a tree branch, the grater is the distance between

the concepts. Two concepts have distance 1, i.e. they are unreachable, when they belong to different classes (see Fig. 2.2).

5. 5. A distributional approach for measuring document similarity

This section proposes a multi-dimensional approach to measuring similarity between documents based on the EMD algorithm, i.e. optimal matching – OM, that employs SETS document normalisation and concept indexing for representing documents. The document similarity measured by the proposed multi-dimensional similarity approach is used to produce a distance matrix for the documents in a collection (Table 5.2). A distance matrix represents a document collection before clustering. Then, the matrix is clustered by a deterministic version of the standard k-means algorithm called PAM, a.k.a. Partitioning Around Medoids (Kaufman and Rousseeuw, 1990). The clustering solutions produced by PAM using the document distance matrix are evaluated against documents normalised by SETS and represented by concept indexing and clustered by the same PAM algorithm but by using cosine similarity measure. Thus, the difference in the produced two clustering solutions is a result of the difference in the similarity measures only used by the clustering algorithm. An evaluation on both similarity measures, i.e. the EMD-based with semantic distance matrix (many-to-many) and the cosine (one-to-one) measures, is presented in section 5.7.

A cognitive study (Lee et al., 2005) outlines the performance of retrieval of similar documents by the cosine measure to be poorly aligned to human judgement when compared to similar documents retrieved by the EMD measure (Wan and Peng, 2005a, Wan, 2007). Since the proposed method uses the EMD similarity measure in

the clustering, then the clustering solutions produced with the same similarity measure should theoretically align the same to human judgement. Therefore, the results produced by both approaches are compared with human judgement. The results of this comparison are presented in the evaluation section 5.7 of this chapter.

Table 5.2 Distance matrix

	D1	D2	D3	D4	D5	D6
D2	1.1554					
D3	1.0309	1.0297				
D4	1.2360	1.0554	1.2449			
D5	1.0811	1.0283	1.0104	1.0240		
D6	1.3066	1.0891	1.1104	1.0736	1.4494	
D7	1.090	1.0923	1.1243	1.0474	1.7990	1.4641

The use of concept indexing, which is never tested in the domain of document clustering, for representing text documents, benefits the measuring of pair-wise document similarity in two ways: First, the documents are represented in fewer dimensions. The dimensionality is defined by the number of concepts in the external knowledge source employed for the task. Setchi and Tang (2009) use a general ontology (OntoRo), which contains 990 concepts. Thus, the dimensionality of the concept indexing representation is limited to 990. In comparison, the Reuters21578 corpus contains 44261 words (dimensions of representation). Second, the ground distance needed by the EMD algorithm, which measures any two features members of multi-dimensional distributions, can be measured by a distance matrix produced from the OntoRo's structure (Table. 5.1).

Similarly to Wan and Peng's (2005b) semantic approach, where WordNet is used to measure the distance between a pair of words, the proposed approach relies on the general ontology OntoRo to measure the distance between two concepts. Therefore, the proposed algorithm is also semantic-based. In contrast to the EMD-based approach, the proposed semantic approach uses semantic information in two different stages. First, semantics is used to represent documents with concept indexing in combination with statistical data of word co-occurrence (SETS). Since, it was shown in the previous chapter that SETS provides better separation between clusters, in this chapter only SETS is used as a document representation technique. Second, semantics is used in measuring the distance between concepts by relying on the structure of the ontology (Fig. 2.2) to acquire a concept distance matrix (5.2). Thus, similarity between documents is semantically measured by using the commonly shared concept that represents the context of documents. The EMD algorithm use optimal matching of many-to-many similarity. Therefore, the similarity measure returns a similarity value even if concepts are not the same, i.e. which is a limitation of one-to-one similarity matching. Therefore, the approach aims to improve the similarity measure between documents based on their internal structure and subtopic distribution, which both are proven consistent to human judgement (Wan and Peng, 2005b, Wan, 2007).

In theory, using of OM-based similarity principle in EMD must outperform the cosine similarity measure. In addition, since the matching is many-to-many and is optimal not all 990 dimensions need to be considered. Instead, the concept indexing, which represents an array of pairs $< concept, weight >$ (see section 4.2.3), is sorted in descending order by the weight. Thus, only a predefined number of concepts can be considered instead. Experimentally it was established that clustering solutions are im-

proved with up to 75 concepts. The experiment presented in the evaluation uses 40 concepts. This number is chosen with the purpose of producing results within a feasible computation time.

5. 6. Illustrative example

This section presents an illustrative example, which is a part of preliminary experiments conducted on the Wikipedia collection in the process of developing and implementing the EMD many-to-many matching algorithm. The example demonstrates the ambiguous (with respect to the author's judgement) pair-wise document similarity produced by the many-to-many matching. Since Wikipedia does not have the characteristics of the Reuters21578 corpus (i.e. manually assigned tags to every article, which are used in section 5.7 to evaluate the EMD measure for consistency with human judgement), the evaluation presented in table 5.4 might be biased towards the author's understanding and motivation. The evaluation of document similarity produced by the EMD measure into author's judgement is conducted on forty documents as described in the following.

The Wikipedia collection contains 2,694,787[11] articles. The collection is split into 27 folders so that each folder contains up to 99,997 articles (see table 5.3). Statistical data of co-occurrence (TF-IDF values for every word) is calculated from the entire corpus. However, for illustrative purposes the experimental results presented in this section (see table 5.4) show the pair-wise document similarity of one document (AY-

[11] A Wikipedia archive dump, used in the preliminary experiments, was downloaded on 06 / March / 2009

wiki00011.html) to the top 40 most similar of all 99.997 documents from folder AY (see table 5.3). All articles from this folder are accessible online[12]. The full list of the pair-wise document similarity of article AYwiki00011.html to all other documents in the same folder is also available online[13]. The Wikipedia collection is normalised by the SETS algorithm and two one dimensional arrays, as described in Chapter 4, are printed to produce a concept index for all files from that folder[14]. Statistical data similar to the data presented in the illustrative example from chapter 4 is not included.

Section 5.7 presented at the end of this chapter compares clustering solutions produced on the same corpus by using the same index and the same deterministic clustering algorithm (PAM). The difference between the clustering solutions is established by different similarity measures used to group similar documents. Therefore, the illustrative example presented in the current section aims to demonstrate how the EMD similarity measure used in section 5.7 works with real documents. Assumptions needed to be made on the data are also explained.

[12] http://kescrunch.engin.cf.ac.uk/keswiki/AY/

[13] http://kescrunch.engin.cf.ac.uk/ch5/illustrative_exmpl_AYwiki00011.csv

[14] The processed files are accessible at http://kescrunch.engin.cf.ac.uk/keswikiprocessed/AY/, i.e. individual files are addressed as wiki<five digits [00000, 99998]>.html

Table 5.3 Concept indexing of the Wikipedia collection

No	Folder Name	Num of files not indexed	Num of concepts Min	Num of concepts Max	Files in Total	Concepts in Total used	Avg num of concepts per file
1	AA	1	2	979	99,997	49,386,589	493
2	AB	2	1	984	99,997	47,118,084	471
3	AC	1	1	971	99,997	41,766,453	417
4	AD	4	1	974	99,997	39,119,877	391
5	AE	4	1	977	99,997	37,700,868	377
6	AF	1	1	986	99,997	36,274,838	362
7	AG	1	1	981	99,997	33,973,032	339
8	AH	4	1	978	99,997	32,970,629	329
9	AI	2	1	985	99,997	31,183,810	311
10	AJ	3	1	983	99,997	30,906,153	309
11	AK	1	1	974	99,997	29,996,397	299
12	AL	6	1	976	99,997	27,528,901	275
13	AM	4	1	979	99,997	26,928,452	269
14	AN	5	1	979	99,997	27,906,753	279
15	AO	2	1	979	99,997	27,587,020	275
16	AP	3	1	975	99,997	26,206,691	262
17	AQ	9	1	980	99,997	25,432,158	254
18	AR	6	1	980	99,997	20,450,714	204
19	AS	13	1	972	99,997	20,528,811	205
20	AT	6	1	973	99,997	23,036,802	230
21	AU	2	1	976	99,997	18,849,623	188
22	AV	24	1	971	99,997	17,671,467	176
23	AW	5	1	979	99,997	22,000,543	220
24	AX	20	1	982	99,997	18,178,203	181
25	AY	2	1	980	99,997	19,063,664	190
26	AZ	11	1	977	99,997	20,898,513	208
27	BA	10	1	978	94,812	17,587,104	185
In Total		152			2,694,734	770,252,149	285

The first column of table 5.4 contains observation number; the second column contains information with regard to the files for which an EMD distance is measured, i.e. number of concepts assigned to documents and number of words contained in documents, which will provide information on how the EMD measure copes with measuring similarity of documents with different length; the third column displays the actual distance measured between files in the interval [0 (similar),1 (not similar at all)]; and the last column contains the author's judgement for the similarity of the documents.

Table 5.4 Document similarity measured for AYwiki00011 (top 40[15])

N	Files $c_{1/2}$ – num of concepts for document ½ $w_{1/2}$ – num of words for document ½			EMD distance [asc ↓]	Evaluation by the author
1	AYwiki00011(c1=90-w1=156)	→	AYwiki00011(c2=90-w2=156):	0	calibration
2	AYwiki00011(c1=90-w1=156)	→	AYwiki74450(c2=58-w2=12):	0.055046	bad
3	AYwiki00011(c1=90-w1=156)	→	AYwiki97072(c2=65-w2=10):	0.055785	good
4	AYwiki00011(c1=90-w1=156)	→	AYwiki01800(c2=90-w2=221):	0.061834	excellent
5	AYwiki00011(c1=90-w1=156)	→	AYwiki56672(c2=90-w2=89):	0.064348	excellent
6	AYwiki00011(c1=90-w1=156)	→	AYwiki45194(c2=90-w2=79):	0.067633	excellent
7	AYwiki00011(c1=90-w1=156)	→	AYwiki17349(c2=90-w2=136):	0.070313	excellent
8	AYwiki00011(c1=90-w1=156)	→	AYwiki37929(c2=90-w2=100):	0.072193	excellent
9	AYwiki00011(c1=90-w1=156)	→	AYwiki35697(c2=62-w2=14):	0.075228	bad
10	AYwiki00011(c1=90-w1=156)	→	AYwiki42562(c2=41-w2=14):	0.076466	bad
11	AYwiki00011(c1=90-w1=156)	→	AYwiki33493(c2=90-w2=103):	0.076743	excellent
12	AYwiki00011(c1=90-w1=156)	→	AYwiki49753(c2=90-w2=38):	0.076842	excellent
13	AYwiki00011(c1=90-w1=156)	→	AYwiki53968(c2=75-w2=33):	0.077769	excellent
14	AYwiki00011(c1=90-w1=156)	→	AYwiki17299(c2=90-w2=103):	0.082038	excellent
15	AYwiki00011(c1=90-w1=156)	→	AYwiki61621(c2=58-w2=16):	0.085056	not too bad
16	AYwiki00011(c1=90-w1=156)	→	AYwiki41362(c2=90-w2=109):	0.086535	excellent
17	AYwiki00011(c1=90-w1=156)	→	AYwiki49572(c2=90-w2=24):	0.087078	excellent
18	AYwiki00011(c1=90-w1=156)	→	AYwiki76183(c2=63-w2=24):	0.088543	bad
19	AYwiki00011(c1=90-w1=156)	→	AYwiki88595(c2=90-w2=142):	0.088902	excellent
20	AYwiki00011(c1=90-w1=156)	→	AYwiki31628(c2=90-w2=42):	0.089882	excellent
21	AYwiki00011(c1=90-w1=156)	→	AYwiki55901(c2=77-w2=25):	0.090664	good
22	AYwiki00011(c1=90-w1=156)	→	AYwiki22311(c2=90-w2=146):	0.091427	very good
23	AYwiki00011(c1=90-w1=156)	→	AYwiki88591(c2=90-w2=107):	0.09222	excellent
24	AYwiki00011(c1=90-w1=156)	→	AYwiki47675(c2=90-w2=118):	0.092289	excellent
25	AYwiki00011(c1=90-w1=156)	→	AYwiki40937(c2=90-w2=29):	0.092745	good
26	AYwiki00011(c1=90-w1=156)	→	AYwiki61065(c2=90-w2=37):	0.094005	very good
27	AYwiki00011(c1=90-w1=156)	→	AYwiki29206(c2=71-w2=13):	0.094137	not too bad
28	AYwiki00011(c1=90-w1=156)	→	AYwiki06090(c2=41-w2=10):	0.096664	excellent!!!
29	AYwiki00011(c1=90-w1=156)	→	AYwiki04783(c2=51-w2=24):	0.096989	good
30	AYwiki00011(c1=90-w1=156)	→	AYwiki16648(c2=80-w2=12):	0.097372	very bad
31	AYwiki00011(c1=90-w1=156)	→	AYwiki90818(c2=90-w2=72):	0.099062	good
32	AYwiki00011(c1=90-w1=156)	→	AYwiki53698(c2=74-w2=31):	0.099277	very poor!!
33	AYwiki00011(c1=90-w1=156)	→	AYwiki36547(c2=90-w2=22):	0.100201	very good
34	AYwiki00011(c1=90-w1=156)	→	AYwiki38975(c2=90-w2=119):	0.100352	good
35	AYwiki00011(c1=90-w1=156)	→	AYwiki32856(c2=62-w2=13):	0.100568	very poor
36	AYwiki00011(c1=90-w1=156)	→	AYwiki56708(c2=90-w2=139):	0.101164	excellent
37	AYwiki00011(c1=90-w1=156)	→	AYwiki61613(c2=70-w2=10):	0.105201	very poor
38	AYwiki00011(c1=90-w1=156)	→	AYwiki40850(c2=90-w2=52):	0.105474	very good
39	AYwiki00011(c1=90-w1=156)	→	AYwiki24422(c2=56-w2=10):	0.105914	very poor
40	AYwiki00011(c1=90-w1=156)	→	AYwiki49472(c2=90-w2=137):	0.106308	excellent

[15] The content of all files are shown in Appendix A

The evaluation shown in table 5.4 represents similarity that documents have to document AYwiki00011 according to the EMD measure. The author's understanding of the document is that it refers to a person who is a politician and is affiliated with the United Nations. Therefore, documents that convey political topics or refer to policy making, politics, or activity of the United Nations are regarded as similar. It is observed that documents with few words (up to 25 or 30) are often wrongly classified as similar. An exception is document AYwiki06090 (c2=41-w2=10). It is correctly regarded as similar and yet contains only 10 words. This indicates the importance of the quality of the words used to convey an idea.

The experimental results shown in table 5.4 comply with two constraints: The first constraint regards the difference (gap) between two sequential weight values for the assigned concepts. Table 5.5 shows the top 10 concepts that the relevant files in the table are indexed with. It is established experimentally that the gap can be unlimited (i.e. no constrains at all). However, the clustering solutions are slightly improved when a gap valueis used. The results presented in table 5.4 are produced using an unlimited gap. The file AYwiki52756 is ranked in 15916 place by similarity to AYwiki00011 whilst AYwiki29756 is in 15320 place for similarity to the same document.

Table 5.5 Concept index and relevance of concepts

AYwiki52756(c2=90-w2=74)		AYwiki29756(c2=90-w2=172)	
465	1.6114888	465	1.6977179
524	0.45973605	106	0.45834875
551	0.4443297	579	0.44807023
590	0.3343826	590	0.44807023
548	0.2717258	635	1390/3807
32	0.23153867	54	0.35110027
505	100/469	981	0.33604532
586	100/469	693	0.33172125

The second constraint considered on the experimental data is the minimum number of concepts that a document needs to be tagged with in order to be considered for measuring pair-wise similarity to any other document from the collection. Experimentally it is established that documents with at least 30 assigned concepts produce good results when used in measuring pair-wise document similarity. On the other hand, pair-wise document similarity measured with 75 or more concepts does not change considerably, and yet the complexity of the EMD algorithm considerably slows down the performance in terms of processing speed. The experimental results presented in Table 5.4 are conducted with a fixed number of concepts of up to 90 with the minimum number of concepts set to 40. Thus the original number of documents in folder AY is reduced from 99,997 to 71,793 documents.

The presented illustrative example demonstrates that similarity measured with the EMD similarity measure is poorly aligned to the author's judgement. As a result of this the clustering solutions produced by using the EMD as a similarity measure are expected to be inferior.

5. 7. Evaluation

This section presents an evaluation of clustering solutions produced by the same algorithm but once groupings are produced with clustering algorithm which employs the standard cosine similarity measure and the other time with the distributional similarity measure enhanced by semantic optimal matching. The evaluation is performed on the Reuters21578 corpus. The main objective of this evaluation is to compare the alignment of the produced document groupings to human judgement. Another objective is to analyse the separation between the clustering solutions produced by both measures. The document representation used to produce the cluster solutions is concept indexing. The clustering solutions are produced by a modified version of the standard k-means partitional clustering algorithm, which builds the clusters around medoids (Kaufman and Rousseeuw, 1990).

This section presents experimental results on the separation between clusters obtained by the PAM clustering algorithm. This algorithm is used to produce three of clustering solutions as follows: series 1 – documents are represented by concept indexing and cosine similarity measure is employed to compute similarity between documents; series 2 – documents are represented by concept indexing and EMD similarity measure is employed to compute similarity between documents; series 3 – documents are represented by the documents tags from the Reuters21578 corpus and cosine similarity measure is used to compute similarity between documents. Once the clustering solutions are produced separation between clusters is measured.

Another set of experiments demonstrates patterns of similarly grouped documents obtained by the aforementioned series of clustering solutions. These experimental

results are obtained by following the evaluation methodology discussed in section 2.1.6.2. In the end, a percentage is presented of similarly grouped documents.

The evaluation carried out in the chapter 4 employs all 445 tags of the Reuters21578 corpus to establish and represent the objectivity of human judgement. The prerequisites for the human judgement are the same as in the previous chapter, i.e. linguists with similar background, motivation and understanding of the task are employed to assign tags to every document in the test collection. Their experience is disregarded as a factor for objectivity. The used corpus is a sub-set of the Reuters21578 corpus and consists of 18,457 articles. The reduction of the corpus is explained in the evaluation section of chapter 4.

The evaluation is conducted in two stages. In the first stage the separation that the EMD-based and the cosine similarity measures provide to clustering solutions is evaluated. Since the clustering algorithm and the document representation are the same for both series, the separation between the clusters, which is measured with silhouettes (Rousseeuw, 1987), is a function of the quality of the similarity measures. The results of this evaluation are presented in table 5.6. The last column in that table, which is named “HJ & Cosine”, provides a reference to the quality of the clustering solutions in terms of silhouettes to human judgement. Negative values for the silhouettes of the clustering solutions address documents that are placed in a wrong cluster. The negative values are explained with documents scattered far away from the centre of the clusters. Therefore, the topological representations of the clusters overlap each other (Rousseeuw, 1987).

The presented in table 5.6 clustering results are obtained by the PAM algorithm, which was selected over the standard k-means approach because (i) it accepts a dis-

similarity matrix which is clustered with the k-means clustering algorithm performed around medoids, i.e. the medoids make the PAM algorithm a deterministic version of the k-means approach; and (ii) it also provides robustness since it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances and (iii) the deterministic solutions of PAM's enable the evaluation to be conducted in simulated cross-domain environment. The PAM algorithm searches for a set of good initial medoids so that there is no other single medoid, which will provide better objectiveness. The medoids represent the internal structure of the collection. The deterministic nature of PAM and the robust performance enable the evaluation to focus on clustering across domains. This cross-domain clustering is simulated by randomly selecting 20 sub-collections, where each sub-collection consists of 1.000 documents. That is the reason for using mean and standard deviation in the experimental series presented in table 5.6.

Table 5.6 Silhouette values obtained with different document representations and similarity measures

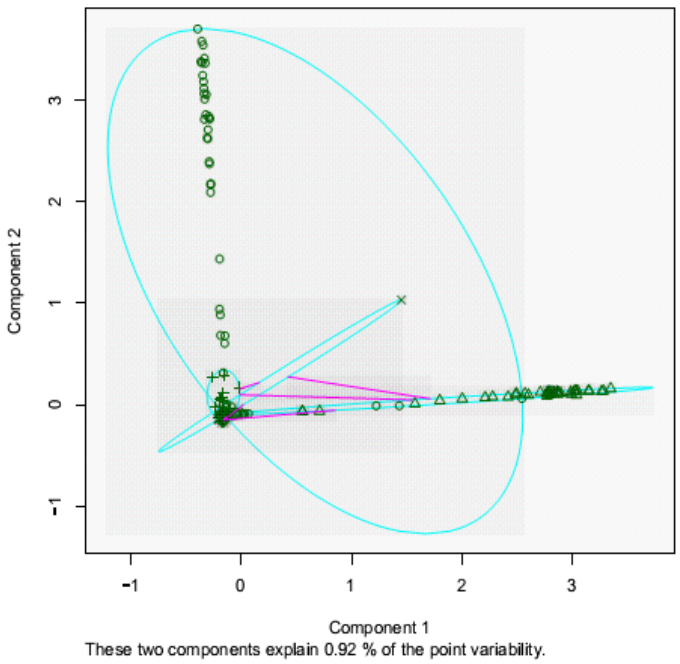
#	SETS & EMD		SETS & Cosine		HJ & Cosine	
	Mean	SD	Mean	SD	Mean	SD
5	-0.14982	0.04613	0.88605	0.03495	0.03036	0.01771
10	-0.17731	0.02072	0.84877	0.03496	0.07993	0.01691
15	-0.20210	0.02331	0.80332	0.05678	0.11610	0.01778
20	-0.22030	0.02047	0.76699	0.06213	0.14710	0.02153
25	-0.23671	0.02508	0.71357	0.06770	0.18231	0.02522
30	-0.24260	0.02370	0.66602	0.08644	0.21330	0.02693
35	-0.24758	0.02529	0.62423	0.09947	0.24408	0.02937
40	-0.24640	0.02307	0.60567	0.08601	0.27533	0.03221
45	-0.24899	0.02207	0.56894	0.08920	0.30633	0.03482
50	-0.24922	0.01862	0.56262	0.09181	0.33795	0.03585
Total	-0.22210	0.02485	0.70462	0.07095	0.19328	0.02583
EMD - Earth Mover's Distance (in conjunction with SETS)						
SETS - Semantically Enhanced Text Stemmer						
HJ - Human Judgment (Reuters Corpora onTopics tags)						

The first step of the evaluation is to initialise the clustering medoids by using the human judgement from the Reuters21578 corpus. The human judgement is used to identify the underlying structure of the sub-collections. Then, the series of clustering solutions, i.e. the rows of Table 5.3, are created by using the objectiveness of the human judgement in minimised sum of dissimilarities of the observations, i.e. the outlined medoids are passed as parameters to the other clustering solutions. Once a clustering solution of human judgement is completed, the sub-collection is clustered two more times by employing the SETS & EMD, i.e. concept indexing and EMD as a similarity measure, and the SETS & Cosine, i.e. concept indexing and cosine similarity measure. The difference in the latter two clustering solutions is that they use the "identified" by the human judgement structure of the clustered sub-collection, i.e. they use the medoids initialised by human judgement to cluster a sub-collection.

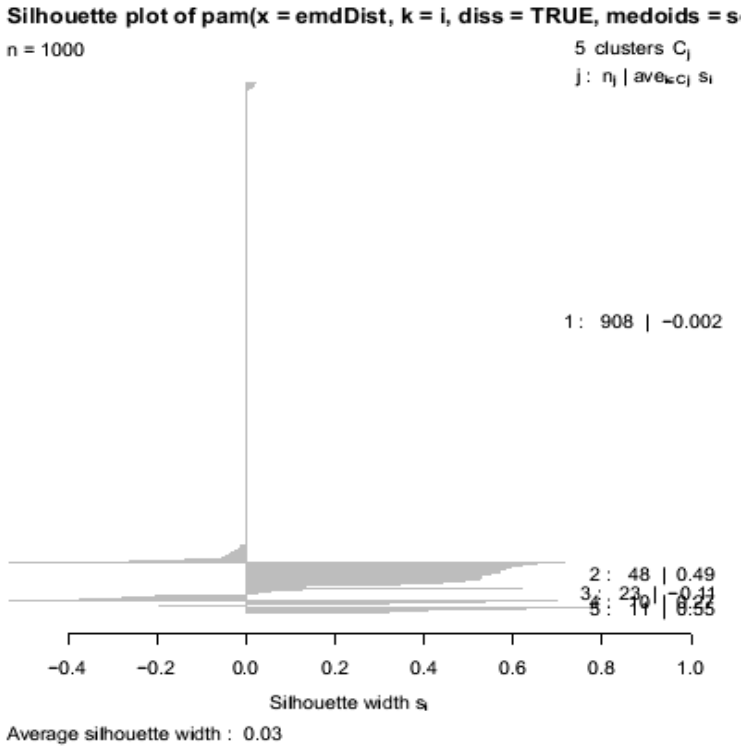
However, the specified order of the medoids is irrelevant in general, since PAM is designed not to depend on the order of the observations.

The results shown in figure 5.3.b demonstrate that the SETS & EMD series of experiments place documents in wrong clusters (indicated by the negative values – Table 5.6). The wrong clustering provided by the EMD similarity measure is a result of similarity returned for any two distributions. On the other hand, the silhouette values produced by SETS & Cosine demonstrate very high separation between the clusters and at the same time very high coherence inside the clusters. However, the silhouette values consistently degrade when the number of clusters increases. In contrast to human judgement the silhouette values consistently provide larger values when the number of clusters increases. It is noted that the overall performances of the SETS & EMD and HJ & Cosine medoids are comparably similar whilst the base-line approach, i.e. SETS & Cosine, is 3.5 times more inconsistent (see Table 5.6 – Total value for SD). This fact prompts a further investigation of the clustering results.

A detailed analysis of the clustering solutions produced by the PAM algorithm is presented in figures 5.3 - 5.5. The extended analysis includes a visualisation of the clusters and an investigation of the number of files in every clustering solution. Figures 5.3 through 5.5 show the clustering solution from the first sub-collection.



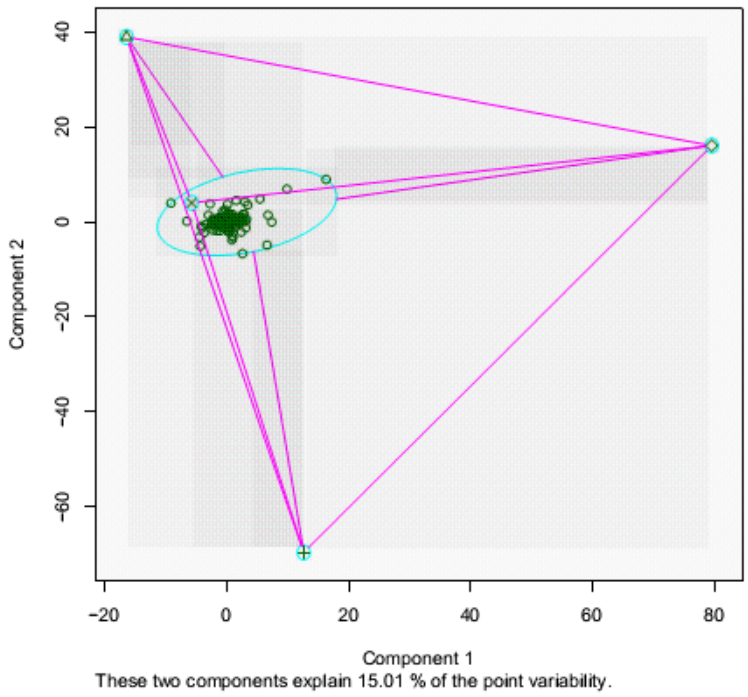
a)



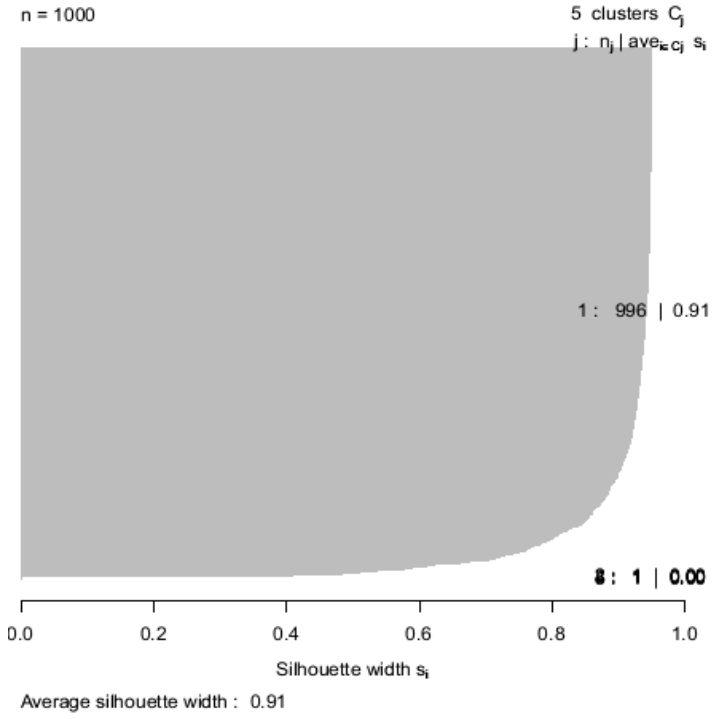
b)

a) topological groupings: document representation – SETS; similarity measure - EMD
b) silhouettes of the clusters and distribution of files per cluster

Figure 5.3 Plot of a clustering solution of 5 clusters with 1000 files



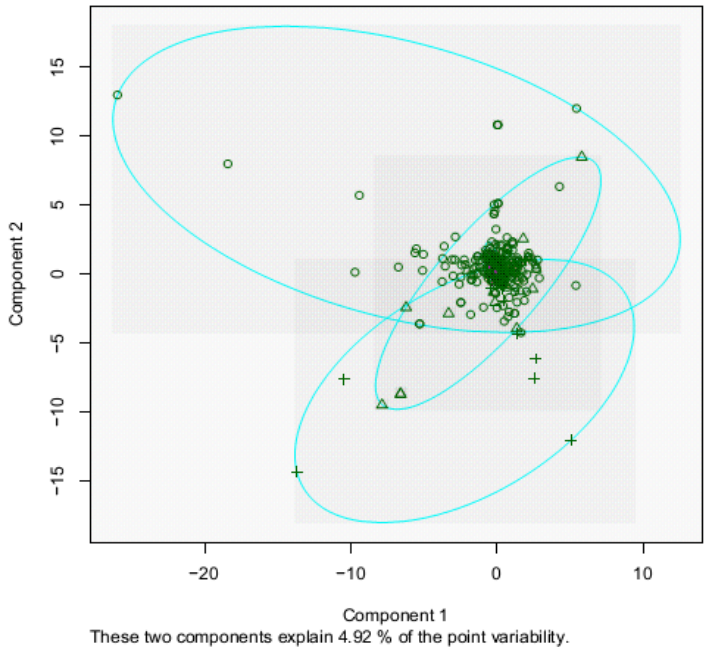
a)



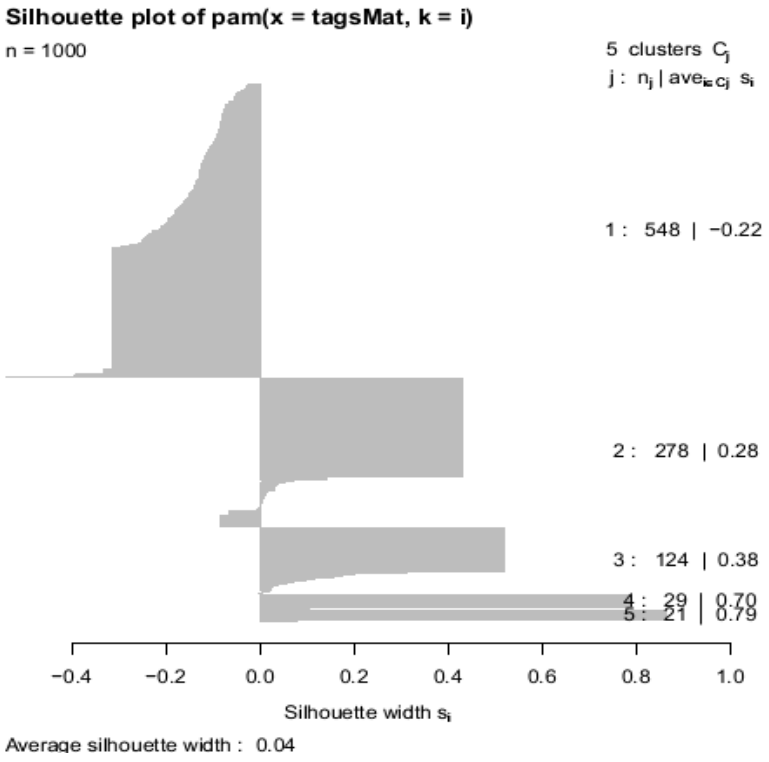
b)

a) topological groupings: document representation – SETS; similarity measure - cosine
b) silhouettes of the clusters and distribution of files per cluster

Figure 5.4 Plot of a clustering solution of 5 clusters with 1000 files



a)



b)

a) topological groupings: document representation – Human Judgement (tags from the Reuters21578); similarity measure - cosine

b) silhouettes of the clusters and distribution of files per cluster

Figure 5.5 Plot of a clustering solution of 5 clusters with 1000 files

Figures 5.3.a - 5.5.a visualise topologically produced clustering solutions. For simplicity of the visualisation, the figures present solutions for 5 clusters. The first observation is the topological visualisation of the base-line similarity measure, i.e. the cosine similarity, which creates 4 clusters that consist of only one document. The rest of the documents are placed in one huge cluster. Therefore, the silhouette values of these solutions indicate very good separation of the clusters (silhouette values close to 1) (Fig. 5.4.b). In comparison, the distribution of the documents per clusters according to the human judgement is more balanced (Fig. 5.5.b).

A commonality shared by both clustering solutions, produced with the Cosine and the EMD measures, is that both produce one significantly larger than the others cluster. And yet, the EMD is remedied since it reduces the size of this large structure by 10% and no clusters that consist of one document exist. At the same time, the conclusion that silhouette values cannot be a distinct characteristic for the quality of the results, since the silhouette values for the human judgement are surprisingly low is made (Fig. 5.5.b).

The other objective of the evaluation is to the question whether the human judgement indeed produces clusters with poor separation. For that purpose, the document clustering solutions produced in the first stage of the presented evaluation is further evaluated in the second stage. In the second stage the produced document groupings are compared against the groupings produced by the human judgement. This is achieved by establishing the percentages of documents clustered in one clusters, i.e. percentage of overlapping clustering solutions. The second stage of the evaluation is to run this evaluation over the same 20 sub-collections of 1.000 files each. The complete results are presented by using average mean and standard deviation.

Table 5.7 Values in percentages of overlapping clustering solutions

#	HJ & Cosine OVERLAP SETS & EMD		HJ & Cosine OVERLAP SETS & Cosine	
	Mean	SD	Mean	SD
5	17.27%	1.20%	38.81%	2.16%
10	10.96%	1.01%	32.68%	1.95%
15	8.49%	0.89%	28.37%	1.49%
20	7.20%	0.95%	25.66%	1.36%
25	6.29%	0.88%	23.38%	1.30%
30	5.67%	0.82%	21.59%	1.07%
35	5.23%	0.78%	19.85%	0.95%
40	4.84%	0.70%	18.32%	0.89%
45	4.53%	0.62%	17.18%	0.91%
50	4.28%	0.60%	16.32%	0.91%
Total	7.48%		24.22%	

Although, the EMD similarity measure produces extremely incoherent clusters, which is explained by the negative values of the first column in Table 5.6, the document groupings for 5 clusters recover 16% to 18.5% of the groupings obtained by using human judgement. On the other hand, the robust cosine measure reaches ~41% recovery of these groupings. Therefore, the cosine similarity measure performs better in relation to human judgement. An interesting fact is that Lee et. al (2005) present the same consistency of the cosine measure with human judgment, i.e. consistency of 40%. The results in Table 5.7 support the experimental observations obtained by the cognitive researchers although a different document representation technique is used and a different evaluation technique is designed to evaluate the clustering solutions.

The results in Table 5.7 yield an increase of the consistency of the cosine similarity measure with human judgement of 1% ÷ 1.5%, which can be assumed as a result of the different evaluation method used to conduct the experiments. However, the 40% consistency of the cosine similarity measure is not realistic result since 99% of the

documents are in one large cluster and the other 4 clusters contain only a document. Therefore, considering the large cluster of 548 files (Fig. 5.5.b), produced by the clustering algorithm with the Reuters21578's tags, it can be concluded that the consistency of the cosine similarity measure is defined by the size of the largest cluster in the clustering solution produced with human judgment. This observation can provide an explanation for the continuous deterioration of the clustering results produced with the cosine measure in relation to human judgement. The explanation is that with the increase of the number of clusters causes fewer documents to be placed in a large cluster. On the other hand, the EMD similarity measure demonstrates 1.5 to 2 times better consistency (SD value) in cross-domain clustering.

5. 8. Summary

This chapter proposes a methodology to measuring pair-wise document similarity, where documents are presented as multi-dimensional distributions. The evaluation of this approach is conducted against the robust cosine similarity measure, which is used as a base-line algorithm. The clustering solutions obtained with both measures are evaluated for consistency with human judgement with the purpose to explore a broad range of clusters without considering the optimal number of cluster for the Reuters21578 collection.

The clustering silhouette values obtained demonstrate better separation of the clusters enhanced by the robust cosine measure than the EMD-based measure. This separation is a valid observation across domains. However, this result is not objective since most of the documents are placed in a large cluster, which is the reason for the high silhouette value. Besides, tags assigned to every document by linguists, which

are supposed to be better quality than the others, produce low silhouette values when used to cluster documents. On the other hand, all silhouette values obtained on clusters produced by the PAM algorithm with EMD similarity measure produce only negative values. The negative values indicate that documents are placed in wrong clusters, i.e. documents within the clusters are positioned far away from each other, which makes clusters topologically very wide and they overlap each other.

The second evaluation of the clustering, conducted on groupings produced by both similarity measures, demonstrated that the EMD similarity measure performs 2.25 times worse in consistency with human judgement than the cosine measure. On the other hand, the EMD measure performs 1.8 times more consistent across all experimental clustering series. The consistency of the clustering results achieved by the cosine measure and the k-means algorithm with human judgement is ~41%, which is a repetition of the experimental results obtained by Lee et. al (2005). Then, in a cognitive study is concluded that none of the traditional clustering methods achieves more than 40% consistency of the produced clustering solutions with human judgement. Although, concept indexing provides better separation between clusters it provides a slight improvement of the results to human judgement, i.e. the improvement is within the range of 1÷1.5%.

A significant achievement, supported by the repeated results from the cognitive study, demonstrates that the assumptions with regard to the role of the linguists' background, motivation, and understanding of the task that is made in the evaluation process are correct. In addition, the linguists' experience is correctly disregarded as a prerequisite for objectiveness of the evaluation.

The evaluation section of this chapter disproves the theory (Wan and Peng, 2005a) that suggests better alignment of document similarity, which is measured based on the feature distributions, to human judgement. This conclusion is motivated by the experimental evidence obtained on a large scale from sets of real articles. In comparison, Wan and Peng (2005) use a small corpus of 132 sentences, where each sentence is considered for a separate document.

In addition, the experimental results yield no existing correlation of the alignment of clustering solutions and their silhouette values to human judgement. After a thorough analysis of the clustering solutions, it is noted that the good separation between clusters, i.e. high silhouette values of clustering solutions, is due to a bad clustering decision of the algorithm, which forms clusters containing only one document. In 19 out of 20 cases, the PAM algorithm creates four clusters (in a five-cluster clustering solution) that contain only a single document each and the rest of the documents (99%) are placed in one large inferior cluster. Therefore, chapter 6 discusses a methodology of identifying heterogeneous for the separate clusters documents which alleviates this problem. The main objective of this methodology is to improve the alignment of clustering solutions produced with cosine similarity measure to human judgement by excluding clusters consisting of a single document.

Chapter 6 : Methodology for semantically enhanced clustering

This chapter proposes a semantically enhanced methodology to clustering that improves the alignment of clustering solutions to human judgment. Firstly, traditional clustering approaches are discussed to identify areas for improvement. Then a technique is proposed that improves the alignment of clustering results to human judgement by reducing the introduced noise caused by clusters that contain only one document. For that purpose, the entire corpus is scanned and a pair-wise document similarity for all documents is measured. A document, that has a similarity for all documents below a predefined threshold value namely a level of abstraction, is not considered in the produced clustering solution. Finally, the alignment of the clustering solutions produced by an algorithm that implements that methodology and a base-line algorithm are compared to human judgement.

6. 1. Evaluation of traditional clustering approach

A disadvantage of the partitional clustering algorithms is that they produce finite number of clusters to reveal existing relations between documents. The relations are limited to the extent of commonly shared between documents features used by document similarity measures. The quality of the established relationships defines the effectiveness of clustering. Nevertheless, any grouping of documents can be meaningful to users (Estivill-Castro, 2002) but will not exceed 40% consistency with human judgement (Lee et al., 2005).

The partitional clustering model produces pre-defined number of clusters and typically comprises of four compulsory functional blocks (see Fig. 3.1). First, a (i) docu-

ment representation technique is selected with the purpose (II) to index all documents in a collection. However, not necessarily the more information document index contains the better quality clusters algorithms produce (see section 4.5). Fig. 4.3 shows that partitional clustering algorithms produce clusters that are inconsistent with human judgment regardless of the number of dimensions. The document index employed to represent documents is dependent on the size of the collection, and the dimensionality reduction technique used. A typical clustering model provides insights referring to documents with regard to their similarity to each other. The pair-wise document similarity measured by (iii) the employed similarity function is used (iv) to cluster the documents in fewer than the number of documents clusters. The result of clustering represents a clustering solution that is usually evaluated by the internal homogeneity and external separation of the clusters. However, this measure is proven to be irrelevant when clustering solutions are evaluated for consistency with human judgment. The aim of this chapter is to develop and evaluate a methodology that improves the alignment of clustering solutions generated with human judgement.

6. 2. Improvement of clustering algorithms

Clustering algorithms employ document representation techniques, which distinguish a subset of features from a set of candidates. Chapter 4 provides evidence that concept indexing when used as a document representation technique by traditional clustering algorithms provides better separation between clusters and simultaneously improves the consistency of clustering solutions produced with human judgement. Since concept indexing provides reduced dimensionality by design, this simplifies the document representation and as a consequence more advanced and sophisticated clus-

tering algorithms can be employed for clustering large document collections. Computationally expensive similarity measures are traditionally explored with the purpose to improve clustering solutions in relation to human judgement (Wan and Peng, 2005a, Wan, 2007).

6.2.1. Strategies to improve model-based clustering

Model-based clustering approach traditionally employs scalable stemming algorithms such as the Porter stemmer, which is dependant only on morphological language rules. The obtained text normalisation acquires word stems to represent documents in large collections. Then, clustering algorithms calculate a statistical co-occurrence of words and phrases. This information is used to produce a weighted index and construct a document-term matrix (VSM representation), where rows are observations, i.e. every row represents a document, and every column is a stemmed word from the documents. VSM is highly restrictive representational model, since it relies on a high dimensional matrix. As a result of using a matrix, algorithms that implement VSM have quadratic complexity. The high complexity is not suitable for large number of documents and therefore, documents are never represented in a natural feature space of words, but in a reduced adjacent space. The reduced space is produced by dimensionality reduction techniques (such as PCA or SVD) with the intention to discard least meaningful information and achieve least restoration error. Then, clustering solutions are produced from a reduced document-term matrix by employing clustering algorithms. This approach to clustering is simple and scalable but is proven inefficient to produce consistent with human judgement clusters (see section 4.5).

Partitional clustering algorithms group similar documents by organising them in a flat, non-hierarchical, and restrictive structure of clusters. The main restriction of this clustering is due to a pre-defined number k, which controls the clustering. Thus, constraints on the document groupings are imposed in terms of establishing similarity driven relations between documents. As a result, clustering algorithms produce inaccurate and/or inferior clustering solutions. The number of clusters k pre-defines the quality of relations established within clustering solutions. This limits the diversity of relations that can be established between documents. Chapters 4 and 5 demonstrate that clustering solutions can be improved for a wide range of values for parameter k, i.e. even when k is not the most favourable for a document collection.

6.2.1.1. Kernel-based clustering

Kernel-based clustering approaches alleviate to certain extend the limitations in establishing relations between documents. A kernel represents a sequence of words that are considered by the representational technique as a single feature. Thus, words along with their context of occurrence are taken into consideration by the document representation techniques (Lee, 1999). Kernel-based techniques acquire certain amount of documents context. Capturing context of text brings information to the representation of documents with regard to their meaning and not only statistical information of co-occurrence. The kernel-based clustering algorithms map the document index to a sub-space of the original features prior to clustering (Karatzoglou and Feinerer, 2006) using string kernels (Huma et al., 2002) or word-sequence kernels (Cancedda et al., 2003). Since these algorithms rely on the context of the words they provide better clustering than the traditional partitional algorithms according to the

typical measures. Nevertheless, all partitional algorithms do not consider semantic relations between words and words with different meaning are not considered as such. This causes words with similar meaning and/or similar semantic context to be treated as irrelevant features by the clustering process. In addition, chapter 5 provides evidence (see section 5.7) that not all clustering approaches, which consider the context of words or the structure of documents, produce clusters that align well to human judgement, although, theoretically they should (Wan and Peng, 2005a, Wan, 2007).

6.2.1.2. Computationally expensive similarity measure

Clustering solutions produced by the partitional clustering are enhanced besides by improving document index and similarity measure, but also by improving the process of partitioning documents. An improvement of the standard k-means algorithm is demonstrated by the "bisecting" k-means algorithm (Steinbach et al., 2000). The improvement of the partitioning process consists of random selection of k documents and incremental update of the initial k clusters with every consecutive document. The "bisecting" approach produces better overall similarity and lower entropy and has better accuracy and improved efficiency (Zhao and Karypis, 2002). However, the "bisecting" partitioning is relatively effective and scalable and sensitive to noise (Fung et al., 2005).

6.2.1.3. Noise reduction

Document indexing is the stage of clustering in which noise is usually introduced and often influences clustering solutions. The partitional approach is improved by addressing the noise problem with the PAM (k-medoids) algorithm (Krishnapuram et

al., 1999). However, its computational cost makes the algorithm impractical for large collections, unless dimensionality of their representation is reduced. Chapter 5, though, presents experimental results of clustering a few thousands files with the algorithm in reasonable time. However, the evaluation provides evidence that partitional clustering is not suitable for discovering clusters of varying sizes that align well to human judgement (see Fig. 5.5).

6.2.1.4. Context-aware dimensionality reduction

A strategy for achieving better speed performance and scalability of the k-means family of algorithms is obtained by the dimensionality reduction techniques such as the Latent Semantic Indexing (LSI). This algebraic indexing method enables a mechanism for low dimensional document representation based on word co-occurrence. This approach relies on a higher-order structure of the words (Deerwester et al., 1990). This structure is referred to as semantic although, no external knowledge source is used. Semantic structure is derived from a document-term matrix on the entire collection by considering the top 100 to 300 components. In the evaluation section of chapter 4 is shown that SETS normalisation with concept indexing document representation provides better clustering solutions than the Porter stemmer and TF-IDF weighting in VSM representation. In addition, concept indexing not necessary needs dimensions reduction due to its design. And yet, concept indexing outperforms the base-line method with k-means clustering algorithm for selected series of 100, 200, and 300 component sub-spaces. The achieved approximation with SVD on document normalisation obtained with concept indexing to the original document space is

better than the base-line since algorithm since the clustering solutions deviate less across dimension reductions.

6.2.1.5. Kernel-based document representation

The dimensionality reduction is further improved by the kernel-based document representation techniques. This representation is successfully used for document ranking and filtering. It is particularly useful for partitioning large collections. The kernel methods such as kernel k-means and spectral clustering (Ng et al., 2001) are used to deal with the inadequacy of the standard k-means algorithms to separate clusters that are not linearly separable in the input space. Kernel algorithms first map the input data into a high dimensional non-linear space and then a kernel function places the result of the mapping implicitly into a pre-selected feature space. Then, the Euclidian distance measures the distance in the projected results. The spectral clustering forms tight clusters in an eigenvector subspace (Karatzoglou and Feinerer, 2006). The kernel-based approaches are application-oriented. Nevertheless, the full string kernel technique is usually used to construct the matrix prior to partitioning, since this technique is more generic than the other kernel-based techniques. The evaluation of the spectral clustering with string kernel demonstrates very strong time performance and produces better clustering solutions than the standard k-means algorithm (Ng et al., 2001). However, to compute the kernel matrix requires long execution time, which makes the performance of the kernel-based algorithms strongly dependable on the length of the string. The shorter string kernel, the better speed and the poorer performance. Therefore, it is difficult to find a good trade-off between these characteristics for a specific collection for a particular task.

6.2.1.6. High number of clusters

Hierarchical clustering approach alleviates the restrictive similarity driven relations between documents produced by the partitional clustering. Hierarchical clustering algorithms produce as many clusters as necessary to separate documents with unique context. The hierarchical clustering employs agglomerative and divisive algorithms to cluster documents. The former builds a hierarchy of clusters bottom-up. It measures iteratively the similarity between every two pairs of clusters and merging the most similar (Kaufman and Rousseeuw, 2005). The latter builds a hierarchy of clusters top-down. It starts from the top with all documents in one large cluster. The variants of this family of algorithms differ from each other by the similarity measure (Zhao and Karypis, 2001). The similarity measure considers the global distribution of the document representation and splits recursively the cluster by using a flat clustering algorithm until each document is in its own singleton cluster (Manning et al., 2008). The latter approach can be considered as a variant of the partitional clustering, which does not consider pre-defined number of clusters. Clustering completes when documents are clustered together by similarity without any restrictions imposed on the relations between documents, i.e. number of clusters.

The top-down approach is more efficient than the bottom-up when the complete hierarchy of the tree structure is not generated. The approach of not considering parts of the clustering algorithm is later utilised in an improved clustering methodology presented in section 6.3. The divisive approach produces more accurate hierarchies according to the typical f-measure when is used in combination with partitional clustering (Steinbach et al., 2000). The agglomerative algorithms maintain high homogenei-

ty within the clusters by employing different techniques for measuring similarity between documents that rely on typical hierarchical clustering algorithms. Therefore, a good trade-off between the complexity of the similarity measure and the homogeneity of the produced clustering solutions is important for the clustering solutions.

6.2.1.7. Transactional clustering

A clustering approach, which is not document centric, but provides substantial influence on effectiveness is transactional clustering. This clustering is based on frequent itemsets (Wang et al., 1999). Clustering solution produced by itemset-based algorithms produce close to human judgement clustering solutions. The transaction-based similarity approach groups documents together if they share many frequently repeating items, which provide sustainable homogeneity, i.e. cluster-centred similarity. This approach does not perform well if itemsets are sparsely scattered and do not satisfy spectral clustering requirements. However, it provides mechanism towards dynamic transactional clustering, which is a foundation for multiple viewpoint perspective to document clustering.

6.2.1.8. Itemset clustering

The itemsets approach provides better clustering and meets the spectral requirements of the clustering, e.g. HFTC (Beil et al., 2002) and FIHC (Fung et al., 2003) by using the simple frequency of co-occurrence. The difference in co-occurrence is that the HFTC considers low-dimensional frequent term sets, whilst the FIHC uses the global frequent itemsets that appears in more than minimum fractions of the document. The multiple viewpoint perspective to document clustering uses the co-

occurrence of the itemsets locally or globally to create a particular single-view, i.e. one clustering solution. HFTC is not suitable for large document collections but produces accuracy comparable to the "bisecting" k-means. On the other hand, the FIHC is proven to be scalable, fast and very accurate, and eases browsing and navigation among documents (Fung et al., 2003) based on inter-cluster similarity. A change in global and local distribution trade-off would provide different clustering solutions.

The itemsets approach is used to discover clusters embedded in sub-spaces of a high dimensional data (Jing, 2008). This methodology includes bottom-up, e.g. simultaneous keyword identification and clustering of text documents (Frigui and O. Nasraoui, 2004), and iterative top-down, e.g. adaptive subspace iteration (Li et al., 2004), search approaches. The difference between the methods is in the local measure that determines the evaluation of the subspaces. The SKWIC is unsupervised algorithm that uses cluster-dependent keywords weighting to identify clusters that are the most dissimilar in particular keyword sets, i.e. the terms are not tolerated equally (Fountain et al., 1991). The algorithm allows clusters to be located by a special keyword set, therefore, if the keywords change the clusters will change and depending on the perspective defined by the keywords a different clustering solution will be produced. The flexibility of the algorithm is benefitted from richer feature relevance representation. SKWIC demonstrates that the feature relevance corresponds to a generic cluster theme (Frigui and O. Nasraoui, 2004). Therefore, every cluster is not identified only by unique terms but is defined also by the degree of representation of the terms. This characteristic is used by the ASI, which allows explicit modelling of the subspace structure associated with each cluster.

6.2.1.9. Strategies for improvement of clustering

A comparison of different clustering solutions according to entropy and overall quality of the similarity measures is conducted on the results presented in chapter 5. These clustering solutions (see Fig. 5.5) can be improved if there are no clusters that contain only one document, i.e. excluding these documents from clustering solutions should alleviate the problem. This approach of excluding documents from clustering solutions is similar to the hierarchical top-down approach. The similarity between both approaches is that hierarchical algorithms with top-down approach will place such documents in separate clusters, i.e. clustering will complete when certain coherence is obtained. Then, the produced clusters will have high internal homogeneity and external separation.

The overall performance of the hierarchical clustering is better than the partitional clustering but the high computing complexity in measuring similarity between documents makes it the second choice for large collections. The standard k-means algorithm is a good starting point for further development of the partitional algorithms, but all enhancement of this algorithm reported in the literature achieve no more than 5% improvement of the clustering solutions at very high computational complexity.

6.2.2. Similarity-based Clustering

The model-based document clustering is challenged by word polysemy. The fact that in information retrieval the relevant documents might be indexed with words that a user with a perspective different from the domain knowledge would not use in retrieval queries pushes clustering to explore more advanced clustering techniques. The

similarity-based clustering is more advanced approach to clustering than the model-based. The reason is that it take into consideration relationships, which exist in an external knowledge resource to aggregate index and/or use it in the similarity measure (Setchi and Tang, 2007, Xiao, 2010). And yet, the similarity-based clustering employs model-based algorithms to produce clustering solutions.

6.2.2.1. Word sense disambiguation improves clustering

A step towards improved clustering is using the words' meaning. Therefore, external knowledge sources such as ontologies, lexicons and dictionaries, are employed by algorithms for disambiguating words' sense (WSD). This will enable algorithms to deal with word polysemy (Ide and Veronis, 1998). The purpose of WSD is to acquire index using the context of the documents that is relevant to the document theme (Ide and Veronis, 1998). The unsupervised WSD algorithms are of a particular interest to this thesis. They identify text-based repetitive patterns in a large data set, without the benefit of using pre-tagged data, to acquire meaning for the words contained in text. The algorithms then grouped together similar patterns. Document index is aggregated using the acquired patterns. Documents that share common patterns are grouped together in clusters by maintaining as high coherence as possible. The unsupervised approach to WSD is powerful and scalable.

The concept indexing is considered by this thesis to disambiguate the words meaning on concept level. SETS employs concept indexing, which is a thesauri-based approach to document representation. SETS is scalable approach, which demonstrates improved clustering coherence and consistency to human judgment on a large scale. In addition, it overcomes the knowledge acquisition bottleneck by using the semantic

structure of a thesaurus (Yarowsky, 1992) and by exploiting the explicit synonymy relations between words' meanings allows dealing with polysemy. The similarity between words is measured by their distance in the semantic structure of an external knowledge source (Jarmasz and Szpakowicz, 2003b). Thus, a concept distance is measured on the distance between two concepts in the thesaurus structure (Wu and Palmer, 1994). Chapter 5 shows clustering results obtained with a distance matrix and a many-to-many similarity matching based on an optimal matching for two multi-dimensional distributions. This clustering approach produces worse clustering solutions according to human judgement than the base-line algorithm. The reason for this poor consistency to human judgement is that the used many-to-many matching measures the distance between any two distributions, i.e. to much finely granulated information is considered. However, it performs very consistently for a wide range of k. In addition, it is observed that document distributions across different clustering solutions are closer to the distribution of documents according to human judgement (see Fig. 5.3 and Fig 5.5).

The evaluation presented in chapter 5 demonstrates the inadequacy of the typical one-to-one similarity matching proposed in the literature (Blair, 1979, Salton and Buckley, 1998, Baeza-Yates and Ribeiro-Neto, 1999, Aslam and Frost, 2003) to measure the distance between any two documents with various topics (Wan, 2007), i.e. poor performance of one-to-one measure across domains. The EMD-based similarity measure overcomes the disadvantages of the one-to-one measure by seeking document commonality in shared by the documents common context. This approach provides more scalable distance measure that performs better in a collection of documents with various topics (Wan and Peng, 2005b, Wan, 2007). Although, this meas-

ure provides segmentation that is aligned well to human judgments (Hearst, 1997) the base-line algorithm produces clusters that align better to human judgement than clustering solutions produced by clustering algorithms that employs the EMD-based similarity measure. Even statistical information derived from a large corpus used in conjunction with external knowledge, which mixture theoretically should align clustering solutions close to human judgement (Patwardhan, 2003), demonstrates poor results (see section 5.7).

6.2.2.2. Reduced number of observations simplifies clustering

Similarity-based and model-based clustering methodologies face a practical problem to dealing with high dimensional data, i.e. large document-term matrices. Therefore, clustering solutions cannot rely on complex similarity measures that produce consistent with human judgement clusters (Yang et al., 2008) which are used and tested on a small scale. The distributional approach to document clustering, which is based on word/concept distributions over the documents, provides a more compact representation of the data by maintaining maximum mutual information between the probability distribution (Slonim and Tishby, 2000, Slonim et al., 2002). Important observation is that clustering solutions produced by distributional clustering are inferior if word clustering is not performed prior to document clustering (Slonim and Tishby, 2000).

The Contextual Document Clustering (CDC) describes the probability distribution of a set of words that co-occur with a given word in a document. It is based on distributional clustering and identifies documents, which belong to highly specific contexts (Mcdonald et al., 2004). Since CDC relies on a distribution of subjects related to

words they are used to form the basis for creating thematic clusters of documents (Baker and Mccallum, 1998). In contrast to the literature, this thesis uses the idea of this method not to group semantically related documents together (Mcdonald et al., 2004), but to identify documents that do not share enough similarity with the rest of the documents in a collection, i.e. document that form one-document cluster (see Fig. 6.1). The purpose of using this idea is not to provide a compact representation of documents but to provide a mechanism that automatically, in unsupervised manner, discovers one-document-clusters and removes these documents from the document collection prior to clustering. Thus, the dimensionality remains the same, but the number of observations, i.e. the rows in document-term matrix, decrease.

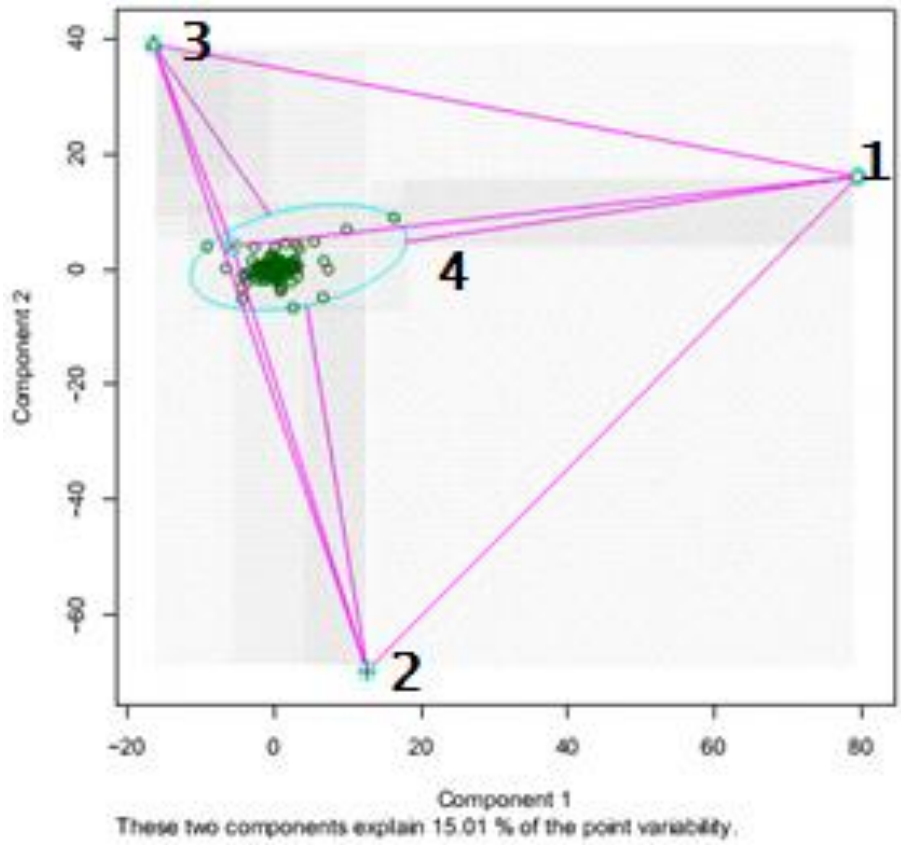


Figure 6.1 Clustering solution of 5 clusters

This task is completed without any use of pre-defined categories or labels, regardless of the clustering criteria. Thus, this approach to clustering will alleviates the problem of partitional clustering, which partitions documents in a pre-defined number of clusters. The thematic complexity of document will be reduced since documents left in the representation of the collection will topically vary less. This complexity cannot be adequately expressed by the traditional partitional algorithms. The aim of suggested approach is to produce low complexity mechanism, which to be suitable for large scale clustering.

6.2.2.3. Semantic relations between words improve clustering

An important requirement to the methodology proposed in the next section is to consider a document representation that uses various relationships between words to measure pair-wise document similarity (Hotho and Staab, 2003, Yang et al., 2008). Thus, documents that refer to various topics can be re-grouped in many meaningful clusters. Therefore, it is important to avoid placing large number of documents into a single cluster (see section 5.7). Nevertheless, the experimental results shown in chapter 5 of this thesis, demonstrate that the traditional clustering algorithms cluster large number of documents in one cluster for small values of k (see Fig. 5.4.b). Alternative document groupings can be produced by employing a similarity measure that commits to different relations between words/concepts established in the ontology employed by the algorithm. Since, the relations between words are explicitly defined in the ontology, a different clustering perspective can be produced by using a different set of relations. Theoretically, the proposed approach should demonstrate higher precision,

whilst the term-based method higher recall and yet, the results are dependent on the collection used to evaluate.

6. 3. Methodology for improved clustering solutions

This section proposes a methodology, which will enable clustering solutions generated by traditional algorithms to be consistent and well aligned to human judgement. The methodology will achieve that by excluding from the collection representation documents that do not share enough similarity with the rest of the documents for a given level of abstraction documents (see Fig. 6.1). The excluded documents are assumed to introduce noise to clustering solutions. Therefore, a clustering solution produced from the representation of a collection that has fewer of these documents in it is believed to be better aligned to human judgement. Different levels of abstraction derive different representations of the same collection. Every representation of a collection provides different perspective to the relations between the documents within the collection and thus, multiple viewpoints are generated.

6.3.1. Document concept similarity

Chapter 5 provides evidence in support of the poorly aligned clustering solutions produced by the optimal-matching similarity measures from the view point of human judgement. Therefore, an alternative mechanism for measuring the document similarity is proposed in this section.

The similarity measured between two documents is likely to be greater than zero in relation to human judgment, i.e. human judgment, according to the cognitive science, always finds certain resemblance between two objects when the comparison task is to

establish similarities between them. Therefore, theoretically any two documents could share features that will enable establishing a relation of similarity between them. In figure 6.2 is shown an algorithm for measuring document concept similarity between two documents represented by the concept indexing. Both documents are represented with concept indexing so that $C_{d_{1n}} > C_{d_{2m}}$, which represents documents that have different length by means of words. A concept similarity function is proposed to measure similarity between two documents using the algorithm shown in Fig. 6.2 by using equation (6).

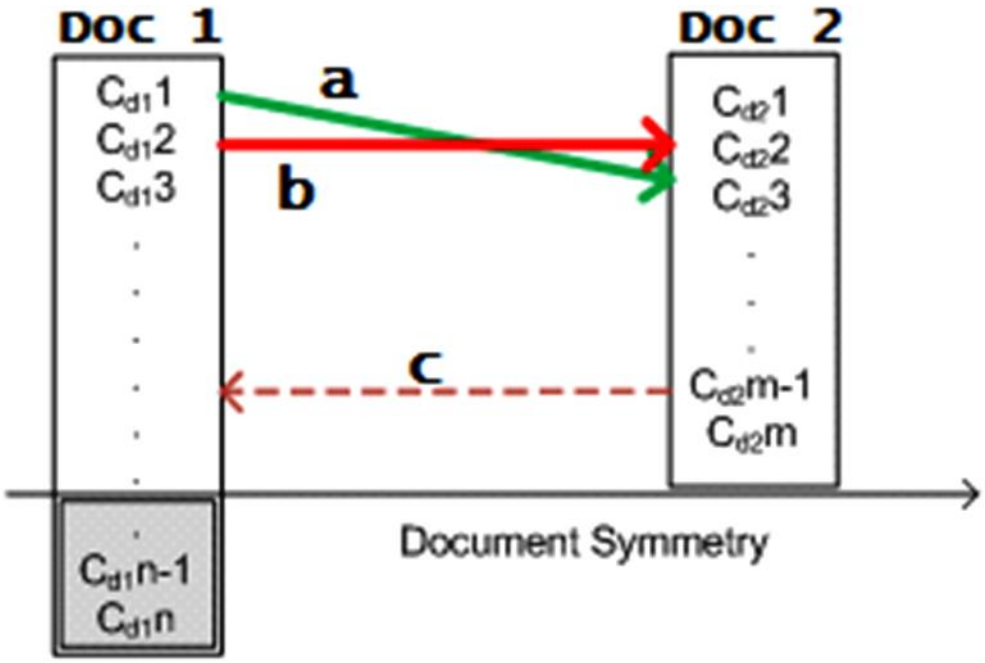


Figure 6.2 Presentation of documents with concept indexing: a - same concepts different place (SCDP); b - same concepts same place (SCSP); c - different concept most appropriate place (DCMAP)

The document concept similarity uses equation (6) to calculate similarity produced by the separate matching of the presented in Fig. 6.2 concept index. A detailed explanation of the matching is presented below.

$$DCS = (3 * SCSP + 2 * SCDP + DCMAP) / m \quad (6)$$

The proposed document concept similarity (DCS) measures similarity between two documents by considering matching relations a to c from Fig. 6.2. The equation is shown to take into consideration the concepts' number and position from the document representation structure. Relations between documents based on Fig 6.2.b, i.e. same concept same place – *SCSP*, contribute to the similarity between documents most. The number of concepts that are shared by the two documents through *SCSP* relation is multiplied by 3. On the other hand, the number of concepts that are shared by the two documents through relation $SCDP$ (see Fig. 6.2.a) is multiplied by 2. The last relation that documents are related through is *DCDP*, i.e. these concepts are not shared by the two documents, but their similarity to each other is calculated through the distance matrix from Table 5.1. The optimal matching is used to measure similarity between these concepts. One concept is used only once in the matching and the concepts with a higher rank, i.e. a grater representative weight, are considered by the optimal matching with priority over the others. The similarities measured for concepts that establish relations between the documents through $DCDP$ are added up. In the end, similarity measured through $SCSP$, $SCDP$, and $DCDP$ relations are summed and added to the total score of DCS. To avoid documents with greater number of concepts in their representative index to have advantage over the shorter documents, the final result is normalised by $min(m, n)$, i.e. the smaller number of concepts in the representative index for the two documents.

6.3.2. Levels of abstraction

The DCS measures the concept similarity between documents by using the many-to-many matching technique. Nevertheless, the evaluation section in chapter 5 shows that EMD similarity measure, which is based on the many-to-many matching, produces inconsistent to human judgement clustering solutions. According to the summary section 5.8 in chapter 5, the many-to-many matching needs a limitation for which a similarity between two documents to be measured. Since, the DCS matching compares similarities between concepts; the limitation for which a similarity is returned is called level of abstraction (concepts organise words in higher order of knowledge structure and provide abstract representation of documents). The level of abstraction is marked with *LoA* so that a high value of this parameter corresponds to low level of abstraction and vice versa. A restriction manifested by the LoA is considered to impose limitations on the pair-wise document similarities used by clustering algorithms to produce clustering solutions (Fig. 6.3). In case DCS value is greater than LoA, the EMD-based similarity between documents is returned, otherwise, the similarity returned is 0, i.e. the distance between the two documents is set to 1 (see Table. 6.1). The difference between the distance matrix used in chapter 5 and the matrix built by using the algorithm in Fig. 6.3 is shown in Table 6.1.

```
If DCS > LoA then
      Pair-wise Document Similarity = EMD
Else
      Pair-wise Document Similarity = 0
```

Figure 6.3 LoA and matching limitation

LoA defines abstraction through which two documents are considered similar. A discrete LoA value set prior to clustering, enables a level of abstraction to be used in

measuring similarity between documents. Clustering firstly measures a similarity between a central for a cluster document and a candidate document, and places the latter in the closest by similarity to the central document cluster. If a similarity measured between two documents is 0, i.e. the distance between them is 1, then the document that is not central for a cluster is considered not similar enough with the rest of the documents in that cluster for the pre-selected level of abstraction. In case a document is considered not similar enough with the rest of the documents from all other clusters, then this document is excluded from the entire clustering solution for the pre-selected level of abstraction.

Thus, a user is enabled to set a level of abstraction for which a similarity between a pair of documents is measured. Table 6.1 shows the distance matrix updated from chapter 5 (see Table 5.2.). A row and a column in red illustrate one-document-clusters, i.e. a document, which is found to introduce noise to a clustering solution. Therefore, if rows and columns coloured only in red are excluded from a clustering solution, then clusters produced by traditional algorithms should be consistent and well aligned to human judgement. The consistency of the produced clustering solutions with human judgement is for a LoA value selected prior to clustering. A modification of the level of abstraction produces different clustering solutions, respectively with lower or higher level of abstraction.

Table 6.1 Distance matrix with level of abstraction

	D1	D2	D3	D4	D5	D6
D2	**1.1554**					
D3	**1**	**1**				
D4	**1.2360**	**1**	**1.2449**			
D5	**1**	**1**	**1**	**1**		
D6	**1.3066**	**1**	**1.1104**	**1**	**1.4494**	
D7	**1.090**	**1**	**1.1243**	**1**	**1.7990**	**1.4641**

An objective for the evaluation of the proposed methodology is to observe the consistency of clustering solutions with human judgement for different levels of abstraction. The evaluation presented in section 6.5 repeats methodologically the evaluation from chapter 5. The evaluation of the proposed methodology investigates whether a clustering solution produced by clustering algorithms with a larger value for LoA, i.e. low level of abstraction, aligns better to human judgement.

6. 4. Illustrative example

This section presents an illustrative example, which is part of preliminary experiments conducted on the Wikipedia collection in the process of developing the presented in this chapter clustering methodology. The methodology aims to produce clustering solutions in closer relation to human judgement. The presented example demonstrates how the ambiguous pair-wise document similarity produced by the many-to-many matching (i.e. the EMD similarity measure) has improved the consistency of similarity measured between documents with author's judgement by using a document concept similarity (DCS), i.e. a level of abstraction for which a similarity between two documents is measured. The level of abstraction (LoA) used for the example in table 6.2 is 1.5 and above, whilst for the example shown in table 6.3 the LoA is less than 1.5. The former removes documents that introduce noise. The latter

demonstrates the inconsistency of the similarity measured between documents that are believed to introduce noise to a collection. Similarly to the illustrative example in chapter 5, the author of this thesis evaluates the similarity measured between documents. Therefore, the evaluation presented in tables 6.2 and 6.3 might be biased towards author's understanding and motivation.

The illustrative examples shown in tables 6.2 and 6.3 represent similarity that documents from folder AY[16] have to document AYwiki00011 (only the top 40 documents are displayed). The author's understanding of the document is that it refers to a person who is a politician, affiliated with the United Nations. Therefore, documents that convey political topics and refer to policy making, activity of the United Nations etc, are regarded as similar. Respectively, documents are considered to introduce noise if they do not convey these topics. It is observed that documents with fewer words are likely to be wrongly classified as similar.

The first column of table 6.2 contains observation number; the second column contains information with regard to the documents for which an EMD distance (forth column) and DCS (fifth column) are measured, i.e. number of concepts assigned to documents and number of words contained in documents, which will provide information of how the two measures cope with measuring similarity between documents with different length; the last column contains the author's judgement for the similarity between documents.

[16] All files from folder AY are available online at http://kescrunch.engin.cf.ac.uk/keswiki/AY/

The constraints applied on the documents are the same as those in the presented illustrative example in chapter 5. The minimum number of concepts per document is 40, the maximum is 90, and the gap between concepts is set to unlimited. All documents in table 6.2 are sorted in descending order by DCS value. It can be observed that all 40 documents in table 6.2 have 90 concepts. The number of concepts is important for measuring similarity when LoA is employed. The numbers of documents that comply with all of the described constraints reduce the original number of files in folder AY from 99,997 to 220, i.e. 220 documents have DCS $\geq$ 1.5. After the results are analysed by author's judgement, all 40 documents are found to be relevant to the document AYwiki00011, which is a significant improvement in comparison to the example presented in chapter 5.

The constraints on the document presented in table 6.3 are the same as those applied on the documents in table 6.2 with the only difference that the LoA is set to less than 1.5. The number of documents is reduced from 99,997 to 71,571. The EMD and DCS measures perform inconsistently with author's judgement. This manual evaluation provides evidence that LoA adequately removes documents that reduce the consistency of clustering solutions with authors' judgement. Section 6.5 provides objective evaluation of the same approach but conducted independently from the author's understanding and motivation.

Table 6.2 Document similarity measured (LoA ≥ 1.5, top 40)

N	Files $c_{1\|2}$ – num of concepts; $w_{1\|2}$ – num of words			EMD	DCS	Similar-ity
1	AYwiki00011(c1=90-w1=156)	→	AYwiki00011(c2=90-w2=156):	0	3	calibration
2	AYwiki00011(c1=90-w1=156)	→	AYwiki88537(c2=90-w2=117):	0.1241	1.811	excellent
3	AYwiki00011(c1=90-w1=156)	→	AYwiki17349(c2=90-w2=136):	0.0703	1.7667	excellent
4	AYwiki00011(c1=90-w1=156)	→	AYwiki56672(c2=90-w2=89):	0.0644	1.711	excellent
5	AYwiki00011(c1=90-w1=156)	→	AYwiki01800(c2=90-w2=221):	0.0618	1.7	excellent
6	AYwiki00011(c1=90-w1=156)	→	AYwiki37929(c2=90-w2=100):	0.0722	1.7	excellent
7	AYwiki00011(c1=90-w1=156)	→	AYwiki04013(c2=90-w2=133):	0.1119	1.7	excellent
8	AYwiki00011(c1=90-w1=156)	→	AYwiki02748(c2=90-w2=181):	0.2048	1.7	excellent
9	AYwiki00011(c1=90-w1=156)	→	AYwiki40347(c2=90-w2=120):	0.1473	1.689	excellent
10	AYwiki00011(c1=90-w1=156)	→	AYwiki88591(c2=90-w2=107):	0.0922	1.678	excellent
11	AYwiki00011(c1=90-w1=156)	→	AYwiki46929(c2=90-w2=63):	0.1642	1.667	excellent
12	AYwiki00011(c1=90-w1=156)	→	AYwiki88514(c2=90-w2=101):	0.1174	1.656	excellent
13	AYwiki00011(c1=90-w1=156)	→	AYwiki88603(c2=90-w2=98):	0.1207	1.656	excellent
14	AYwiki00011(c1=90-w1=156)	→	AYwiki37567(c2=90-w2=97):	0.1417	1.644	excellent
15	AYwiki00011(c1=90-w1=156)	→	AYwiki33109(c2=90-w2=72):	0.1601	1.644	excellent
16	AYwiki00011(c1=90-w1=156)	→	AYwiki69411(c2=90-w2=145):	0.2012	1.644	excellent
17	AYwiki00011(c1=90-w1=156)	→	AYwiki63598(c2=90-w2=116):	0.3855	1.644	excellent
18	AYwiki00011(c1=90-w1=156)	→	AYwiki88595(c2=90-w2=142):	0.0889	1.633	excellent
19	AYwiki00011(c1=90-w1=156)	→	AYwiki47225(c2=90-w2=99):	0.1195	1.633	excellent
20	AYwiki00011(c1=90-w1=156)	→	AYwiki88619(c2=90-w2=144):	0.1689	1.633	excellent
21	AYwiki00011(c1=90-w1=156)	→	AYwiki48318(c2=90-w2=159):	0.2171	1.633	excellent
22	AYwiki00011(c1=90-w1=156)	→	AYwiki47608(c2=90-w2=130):	0.2288	1.633	excellent
23	AYwiki00011(c1=90-w1=156)	→	AYwiki33154(c2=90-w2=94):	0.4717	1.633	excellent
24	AYwiki00011(c1=90-w1=156)	→	AYwiki33493(c2=90-w2=103):	0.0767	1.622	excellent
25	AYwiki00011(c1=90-w1=156)	→	AYwiki49472(c2=90-w2=137):	0.1063	1.622	excellent
26	AYwiki00011(c1=90-w1=156)	→	AYwiki86367(c2=90-w2=95):	0.1382	1.622	excellent
27	AYwiki00011(c1=90-w1=156)	→	AYwiki37788(c2=90-w2=166):	0.1387	1.622	excellent
28	AYwiki00011(c1=90-w1=156)	→	AYwiki47559(c2=90-w2=103):	0.1705	1.622	excellent
29	AYwiki00011(c1=90-w1=156)	→	AYwiki17486(c2=90-w2=131):	0.1750	1.622	excellent
30	AYwiki00011(c1=90-w1=156)	→	AYwiki38614(c2=90-w2=78):	0.1782	1.622	excellent
31	AYwiki00011(c1=90-w1=156)	→	AYwiki02605(c2=90-w2=121):	0.3725	1.622	excellent
32	AYwiki00011(c1=90-w1=156)	→	AYwiki56708(c2=90-w2=139):	0.1012	1.611	excellent
33	AYwiki00011(c1=90-w1=156)	→	AYwiki24175(c2=90-w2=179):	0.1460	1.611	excellent
34	AYwiki00011(c1=90-w1=156)	→	AYwiki52137(c2=90-w2=79):	0.1464	1.611	excellent
35	AYwiki00011(c1=90-w1=156)	→	AYwiki56683(c2=90-w2=160):	0.1532	1.611	excellent
36	AYwiki00011(c1=90-w1=156)	→	AYwiki45167(c2=90-w2=139):	0.1556	1.611	excellent
37	AYwiki00011(c1=90-w1=156)	→	AYwiki60119(c2=90-w2=61):	0.1620	1.611	excellent
38	AYwiki00011(c1=90-w1=156)	→	AYwiki88523(c2=90-w2=127):	0.2099	1.611	excellent
39	AYwiki00011(c1=90-w1=156)	→	AYwiki01669(c2=90-w2=114):	0.2668	1.611	excellent
40	AYwiki00011(c1=90-w1=156)		AYwiki58368(c2=90-w2=101):	0.2984	1.611	excellent

Table 6.3 Document similarity measured (LoA < 1.5, top 40)

N	Files ($c_{1\|2}$ – num of concepts; $w_{1\|2}$ – num of words			EMD	DCS	Similarity
1	AYwiki00011(c1=90-w1=156)	→	AYwiki09439(c2=77-w2=15):	0.1335	1.493506	not bad
2	AYwiki00011(c1=90-w1=156)	→	AYwiki78585(c2=75-w2=20):	0.2785	1.493333	bad
3	AYwiki00011(c1=90-w1=156)	→	AYwiki43743(c2=73-w2=16):	0.2842	1.493151	bad
4	AYwiki00011(c1=90-w1=156)	→	AYwiki56864(c2=69-w2=18):	0.3258	1.492754	bad
5	AYwiki00011(c1=90-w1=156)	→	AYwiki71198(c2=55-w2=14):	0.3058	1.490909	bad
6	AYwiki00011(c1=90-w1=156)	→	AYwiki71215(c2=55-w2=14):	0.3058	1.490909	bad
7	AYwiki00011(c1=90-w1=156)	→	AYwiki71216(c2=55-w2=14):	0.3058	1.490909	bad
8	AYwiki00011(c1=90-w1=156)	→	AYwiki71218(c2=55-w2=14):	0.3058	1.490909	bad
9	AYwiki00011(c1=90-w1=156)	→	AYwiki71223(c2=55-w2=14):	0.3058	1.490909	bad
10	AYwiki00011(c1=90-w1=156)	→	AYwiki71224(c2=55-w2=14):	0.3058	1.490909	bad
11	AYwiki00011(c1=90-w1=156)	→	AYwiki71228(c2=55-w2=14):	0.3058	1.490909	bad
12	AYwiki00011(c1=90-w1=156)	→	AYwiki71229(c2=55-w2=14):	0.3058	1.490909	bad
13	AYwiki00011(c1=90-w1=156)	→	AYwiki40861(c2=49-w2=20):	0.1606	1.489796	not too bad
14	AYwiki00011(c1=90-w1=156)	→	AYwiki21592(c2=49-w2=17):	0.2289	1.489796	bad
15	AYwiki00011(c1=90-w1=156)	→	AYwiki30283(c2=47-w2=9):	0.3122	1.489362	excellent
16	AYwiki00011(c1=90-w1=156)	→	AYwiki43325(c2=47-w2=9):	0.4675	1.489362	bad
17	AYwiki00011(c1=90-w1=156)	→	AYwiki49572(c2=90-w2=24):	0.0871	1.488889	good
18	AYwiki00011(c1=90-w1=156)	→	AYwiki61065(c2=90-w2=37):	0.0940	1.488889	very good
19	AYwiki00011(c1=90-w1=156)	→	AYwiki46990(c2=90-w2=58):	0.1354	1.488889	very good
20	AYwiki00011(c1=90-w1=156)	→	AYwiki26551(c2=90-w2=65):	0.1574	1.488889	good
21	AYwiki00011(c1=90-w1=156)	→	AYwiki65344(c2=90-w2=68):	0.1697	1.488889	very good
22	AYwiki00011(c1=90-w1=156)	→	AYwiki47945(c2=90-w2=64):	0.1736	1.488889	not too bad
23	AYwiki00011(c1=90-w1=156)	→	AYwiki06377(c2=90-w2=74):	0.1754	1.488889	not bad
24	AYwiki00011(c1=90-	→	AYwiki57510(c2=90-	0.1762	1.488889	very good

	w1=156)		w2=138):			
25	AYwiki00011(c1=90-w1=156)	→	AYwiki57308(c2=90-w2=133):	0.1772	1.488889	very good
26	AYwiki00011(c1=90-w1=156)	→	AYwiki89695(c2=90-w2=55):	0.1776	1.488889	not bad
27	AYwiki00011(c1=90-w1=156)	→	AYwiki21282(c2=90-w2=235):	0.1905	1.488889	not bad
28	AYwiki00011(c1=90-w1=156)	→	AYwiki07675(c2=90-w2=128):	0.2008	1.488889	very good
29	AYwiki00011(c1=90-w1=156)	→	AYwiki41990(c2=90-w2=122):	0.2019	1.488889	good
30	AYwiki00011(c1=90-w1=156)	→	AYwiki23714(c2=90-w2=177):	0.2030	1.488889	good
31	AYwiki00011(c1=90-w1=156)	→	AYwiki44234(c2=90-w2=53):	0.2040	1.488889	good
32	AYwiki00011(c1=90-w1=156)	→	AYwiki46912(c2=90-w2=79):	0.2101	1.488889	good
33	AYwiki00011(c1=90-w1=156)	→	AYwiki54659(c2=90-w2=99):	0.2149	1.488889	good
34	AYwiki00011(c1=90-w1=156)	→	AYwiki58447(c2=45-w2=10):	0.2181	1.488889	not too bad
35	AYwiki00011(c1=90-w1=156)	→	AYwiki50662(c2=90-w2=26):	0.2284	1.488889	very bad
36	AYwiki00011(c1=90-w1=156)	→	AYwiki60132(c2=90-w2=41):	0.2449	1.488889	excellent
37	AYwiki00011(c1=90-w1=156)	→	AYwiki36033(c2=90-w2=44):	0.2635	1.488889	not too bad
38	AYwiki00011(c1=90-w1=156)	→	AYwiki29912(c2=90-w2=32):	0.2645	1.488889	interest-ing
39	AYwiki00011(c1=90-w1=156)	→	AYwiki79569(c2=90-w2=92):	0.2774	1.488889	very good
40	AYwiki00011(c1=90-w1=156)	→	AYwiki97312(c2=90-w2=430):	0.2817	1.488889	very good

The presented illustrative example demonstrates that documents that introduce noise for a given LoA value can be detected and excluded from a clustering solution when pair-wise document similarity is measured. As a result, the similarity measured between documents improves. Therefore, the clustering solutions produced after employing the methodology proposed in this chapter are expected to have relevant to a given level of abstraction consistency with human judgement.

6. 5. Evaluation

This section presents an evaluation of the proposed in this chapter clustering methodology. Clustering solutions presented in this section are produced by the PAM algorithm and they include experiments on clustering solutions produced with different levels of abstraction. The purpose of the experiments is to prove that clusters produced with a greater value for the LoA (a lower level of abstraction); will align better to human judgement than vice versa. In addition, silhouette values of clustering solutions obtained for the two series of experiments (each series is produced with a different value for LoA) are analysed. A conclusion for the performance of the PAM algorithm is analysed with regard to human judgement and silhouette values.

The Reuters21578 corpus benefits from manually assigned tags to documents. The tags are used in the evaluation under the consideration that they represent an expert opinion. The expert opinion is believed to be objective since there is no restriction imposed on the tags. They can be any word(s) that represents the content of documents. An assumption is made that document tags in the Reuters21578 corpus are accurate and objective. Since these tags do not change over time, human judgement for the meaning of the documents does not change as well. In the conducted experiments only the level of abstraction changes.

Table 6.4 Comparison of silhouette values of clustering solutions (LoA ≥ 1.5)

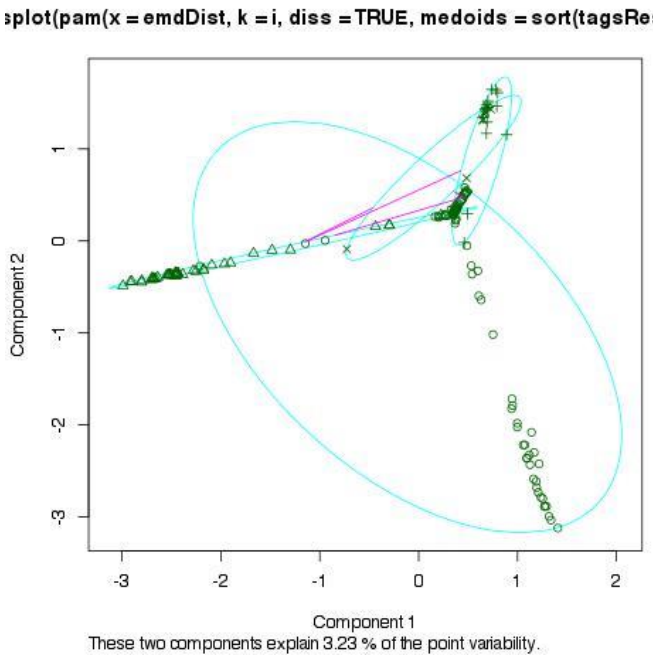
#	SETS & EMD		SETS & Cosine		HJ & Cosine	
	Mean	SD	Mean	SD	Mean	SD
5	0.14991	0.03054	0.82259	0.03916	0.62223	0.23924
10	0.15224	0.02844	0.75836	0.04212	0.55877	0.19984
15	0.16907	0.02920	0.68598	0.09119	0.61894	0.15447
20	0.17107	0.03512	0.62659	0.07123	0.69031	0.13809
25	0.18996	0.03150	0.61175	0.06955	0.74755	0.14460
30	0.21655	0.03711	0.54613	0.09686	0.72535	0.19856
35	0.23813	0.03808	0.51042	0.08195	0.75643	0.22058
40	0.26093	0.03461	0.42050	0.09910	0.69930	0.25483
45	0.27739	0.03553	0.34644	0.08825	0.51211	0.30740
50	0.29232	0.03396	0.33963	0.09695	0.38727	0.25976
Total	0.21176	0.03341	0.56684	0.07764	0.63183	0.21174
EMD - Earth Mover's Distance (in conjunction with SETS)						
SETS - Semantically Enhanced Text Stemmer						
HJ - Human Judgment (Reuters Corpora onTopics tags)						

The experimental results shown in Table 6.4 are conducted with a LoA value equal to 1.5 or above. The silhouette values for the SETS & EMD are still very low in comparison to the SETS & Cosine that is used as a base-line algorithm. The deviation in total of the base-line results is more than twice larger than the total value for SETS & EMD. The last column of Table 6.4 shows that the silhouette values obtained for HJ & Cosine significantly increase in comparison to the experimental values presented in Chapter 5 Table 5.6 (in Total from 0.19328 to 0.63183). This suggests that documents excluded from the clustering solutions for the relevant LoA value introduce noise to the clustering solutions analysed in chapter 5. After the documents, which are found to introduce noise, are excluded from the clustering solutions they are better separated, more consistent and better aligned to human judgement. The improvement of the clusters coherency and their consistency with human judgment is for a wide range of values for k. This concludes that by modifying the levels of abstraction clustering so-

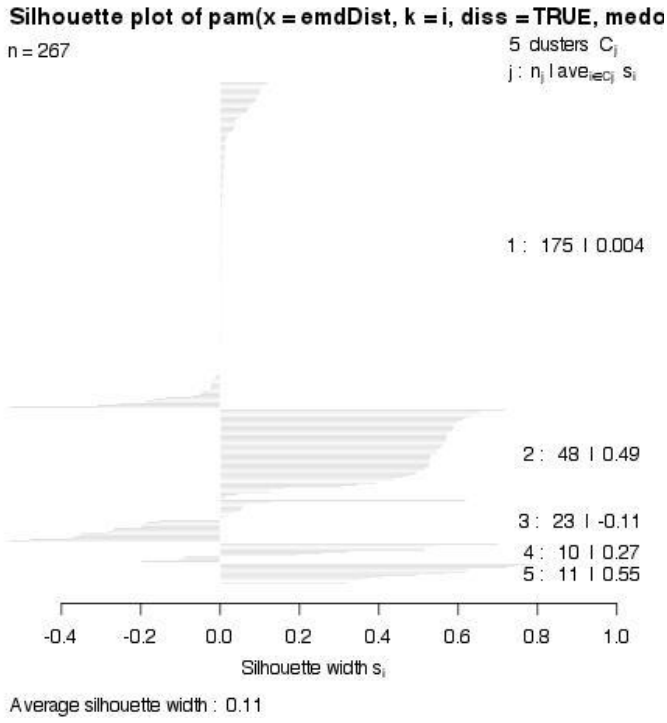
lutions well aligned to human judgment can be produced. The overall coherence of the clusters produced with human judgment is higher than the other two series of clustering. Table 6.4 provides evidence that after documents that are found to introduce noise are removed from clustering solutions clustering silhouette values are an adequate measure for the quality of the clustering results in relation to human judgement.

Table 6.5 Consistency of clustering solutions with human judgement (LoA ≥ 1.5)

#	HJ & Cosine OVERLAP SETS & EMD		HJ & Cosine OVERLAP SETS & Cosine	
	Mean	SD	Mean	SD
5	33.99%	8.03%	64.56%	18.64%
10	20.03%	3.54%	46.61%	12.93%
15	15.52%	1.90%	39.69%	9.00%
20	12.63%	1.41%	34.91%	5.00%
25	11.12%	1.23%	32.99%	4.65%
30	10.10%	1.22%	29.89%	4.20%
35	9.49%	1.38%	27.19%	3.79%
40	9.15%	1.58%	24.18%	3.27%
45	8.75%	1.65%	21.81%	3.11%
50	8.21%	1.68%	20.32%	3.20%
Total	13.90%		34.22%	



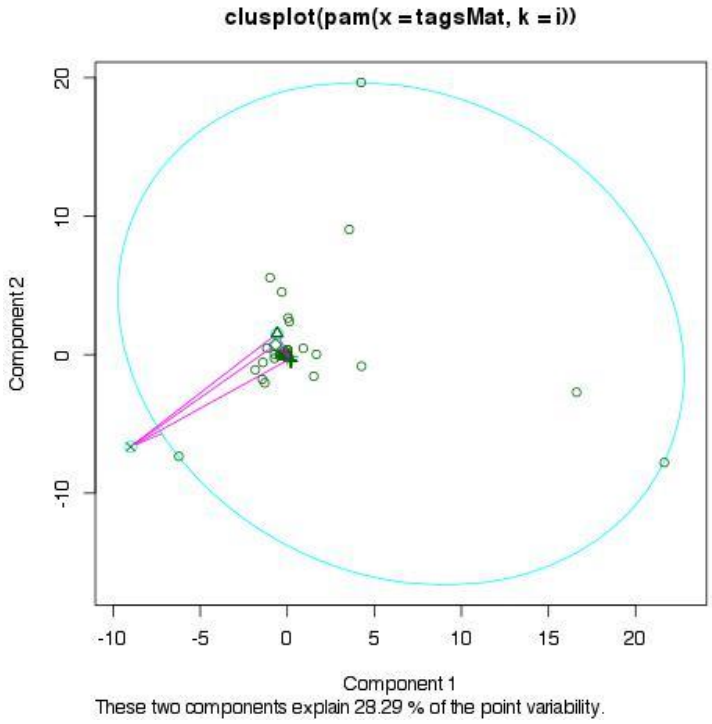
a)



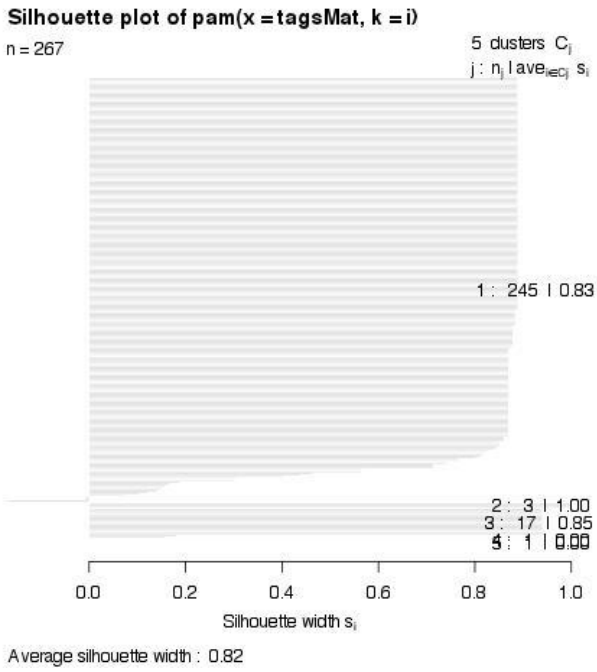
b)

a) topological groupings: document representation – SETS; similarity measure - EMD
b) silhouettes of the clusters and distribution of files per cluster

Figure 6.4 A clustering solution according to the proposed algorithm - LoA ≥ 1.5



a)

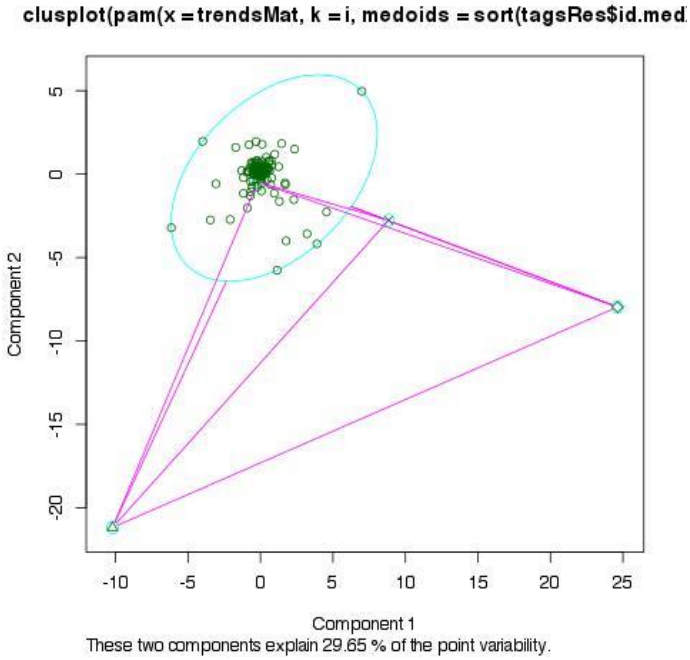


b)

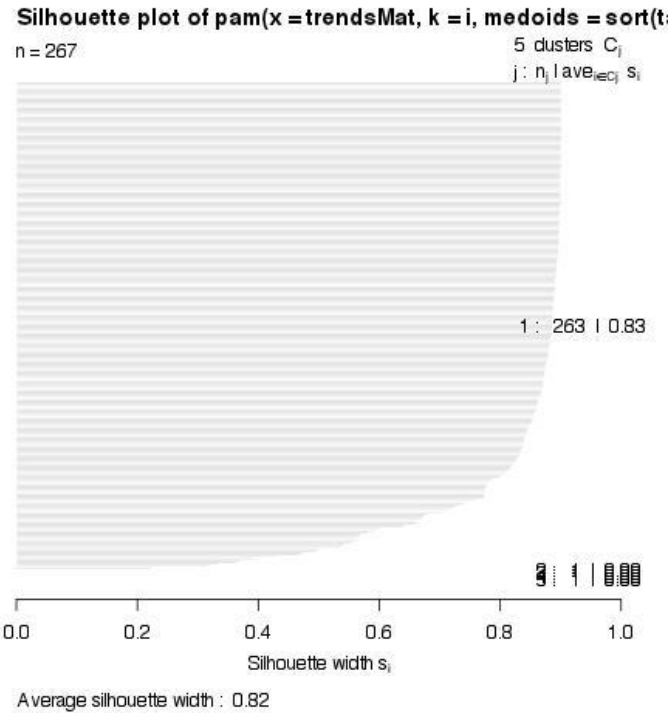
a) topological groupings: document representation – Human Judgement (tags from the Reuters21578); similarity measure - cosine

b) silhouettes of the clusters and distribution of files per cluster

Figure 6.5 A clustering solution according to Human Judgment - LoA ≥ 1.5:



a)



b)

a) topological groupings: document representation – SETS; similarity measure - cosine
b) silhouettes of the clusters and distribution of files per cluster

Figure 6.6 A clustering solution according to a base line algorithm - LoA ≥ 1.5

The second part of the evaluation is shown in Table 6.5. It presents evidence that documents clustered by human judgement and the base-line method and SETS & EMD improve in comparison to the evaluation results in chapter 5. The experimental results demonstrate that a large number of documents are grouped together by the base-line approach similarly to human judgement (Figures 6.5.b and 6.6.b). The percentage reported in the literature and repeated in Chapter 5 of 40% is improved more than twice for one of the sub-collections (see Table 6.5). The improvement is obtained for all 20 sub-collections of the Reuters21578 corpus. The best clustering solutions out of the twenty clusters exceeds a little 83%. On the other hand, the worst clustering solution is 45.92% of the documents grouped similarly to human judgement. In addition, clusters produced with the k-means algorithm that uses SETS & EMD have increased their consistency with human judgement almost twice compared to the results presented in chapter 5. Therefore, excluding documents that introduce noise from the representation of a document collection improves the consistency of clustering solutions with human judgement.

Evaluation with lower level of abstraction (LoA ≥ 1.9)

An experiment is conducted to evaluate how clustering solutions generated with a lower level of abstraction (larger value for LoA) influence the clustering solutions in relation with human judgement. A suggestion is made that by increasing the value of the LoA threshold, the results should be closer to human judgement. The increase of LoA (from ≥ 1.5 to ≥1.9) means that documents must be more similar to each other according to the DCS value. Therefore, a value for LoA of ≥1.9 will include only documents in clustering solutions that have similarity value of DCS higher than 1.9, otherwise documents will be ignore and marked that introduce noise to the clustering

solution. Larger value for the LoA threshold defines closer similarity relations between documents. Therefore, the abstraction of the clustering solutions will decrease and relations between documents will increase.

Table 6.6 Comparison of silhouette values of clustering solutions (LoA ≥ 1.9)

#	SETS & EMD		SETS & Cosine		HJ & Cosine	
	Mean	SD	Mean	SD	Mean	SD
5	0.23171	0.05766	0.79892	0.03754	0.83457	0.17183
10	0.23242	0.06130	0.65571	0.10831	0.81714	0.21416
15	0.25083	0.05086	0.57296	0.11473	0.61793	0.35399
20	0.26419	0.05861	0.48256	0.12807	0.29213	0.21830
25	0.29000	0.07063	0.42014	0.12493	0.20791	0.06640
30	0.28705	0.05698	0.39606	0.10202	0.19576	0.10289
35	0.28071	0.05658	0.33276	0.09147	0.19305	0.12723
40	0.26577	0.04926	0.29173	0.06475	0.18560	0.15927
45	0.23914	0.05740	0.25571	0.05767	0.17635	0.17255
50	0.21223	0.05981	0.24847	0.05422	0.15898	0.19658
Total	0.25540	0.05791	0.44550	0.08837	0.36794	0.17832
EMD - Earth Mover's Distance (in conjunction with SETS)						
SETS - Semantically Enhanced Text Stemmer						
HJ - Human Judgment (Reuters Corpora onTopics tags)						

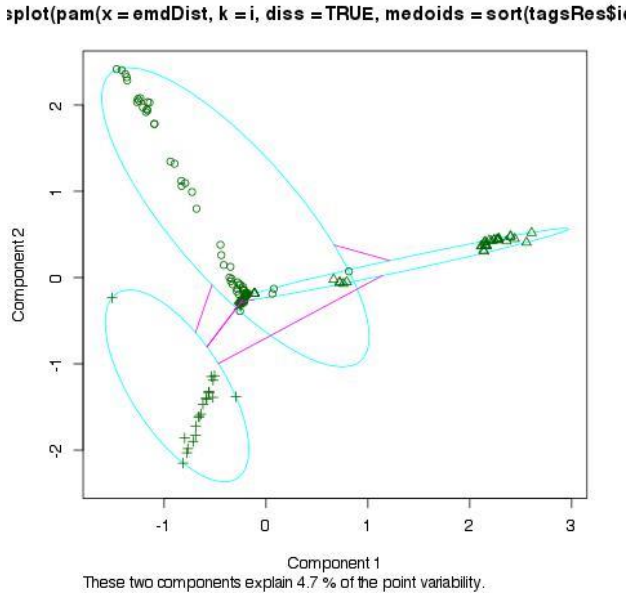
The results of the second series of experiments with regard to silhouette values are shown in Table 6.6. The quality of the clustering solutions produced by human judgement according to the silhouette values is improved. The base-line algorithm and the human judgement worsen their clustering solutions in total. The latter performs better than the base line for up to 20 clusters, whilst for 20 and more clusters the performance of the clustering in relation to human judgement significantly drops. On the other hand, SETS & EMD series are improved by 5% in total. However, SETS & EMD series show consistency in their silhouette values to the number of clusters – standard deviation of 5,7% compared to 8.8% for base line algorithm and 17,8% for human judgement. The consistency of clustering solutions produced by the clustering

algorithm with SETS & EMD and SETS & Cosine shown in Table 6.6 are more consistent, i.e. the SD for both is down by respectively 2.45% and 3,38%.

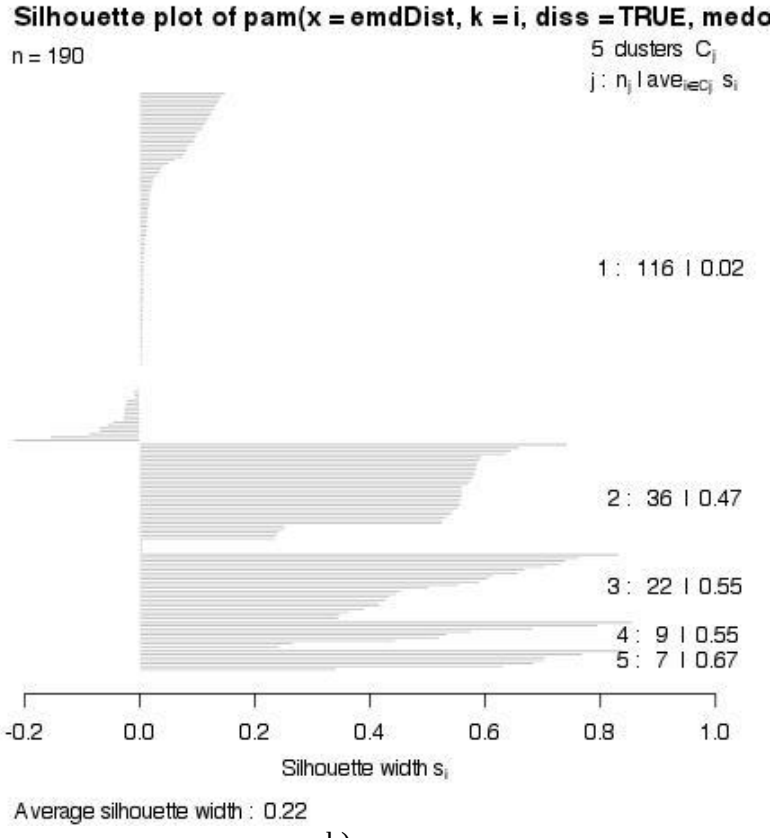
Experimental results, which demonstrate the consistency of the clusters produced by the SETS & EMD and the SETS & Cosine perform with human judgement, are shown in Table 6.7. The results support the predicted quality of clustering solutions in their closeness to human judgement. The average consistency of the produced clusters (for 5 clusters) is improved by 6.18%. However, the total consistency of the clusters is down by 1,08%.

Table 6.7 Consistency of clustering solutions with human judgement (LoA ≥ 1.9)

#	HJ & Cosine OVERLAP SETS & EMD		HJ & Cosine OVERLAP SETS & Cosine	
	Mean	SD	Mean	SD
5	26.23%	5.58%	70.84%	15.26%
10	16.78%	3.80%	51.51%	10.52%
15	12.79%	2.30%	41.24%	5.87%
20	10.94%	2.56%	35.62%	6.53%
25	9.72%	2.50%	31.44%	5.51%
30	8.61%	1.84%	27.96%	5.70%
35	7.37%	1.75%	22.98%	5.50%
40	6.27%	1.57%	19.60%	5.38%
45	5.56%	1.29%	15.98%	3.87%
50	5.00%	1.10%	14.25%	3.41%
Total	10.93%		33.14%	



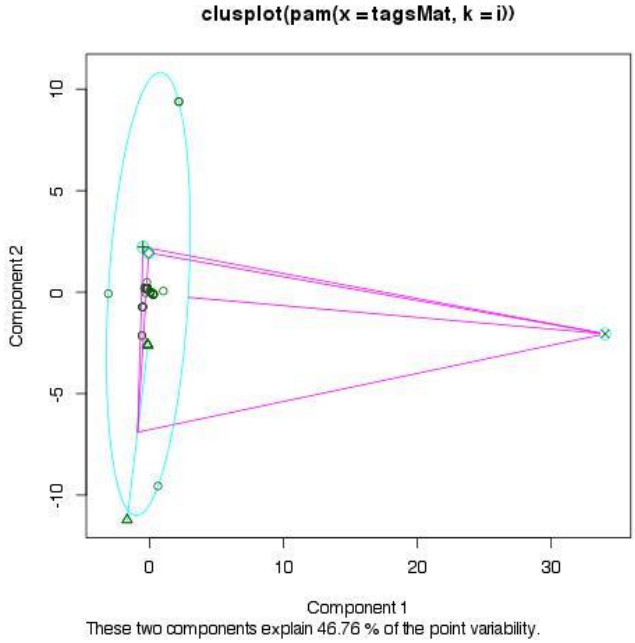
a)



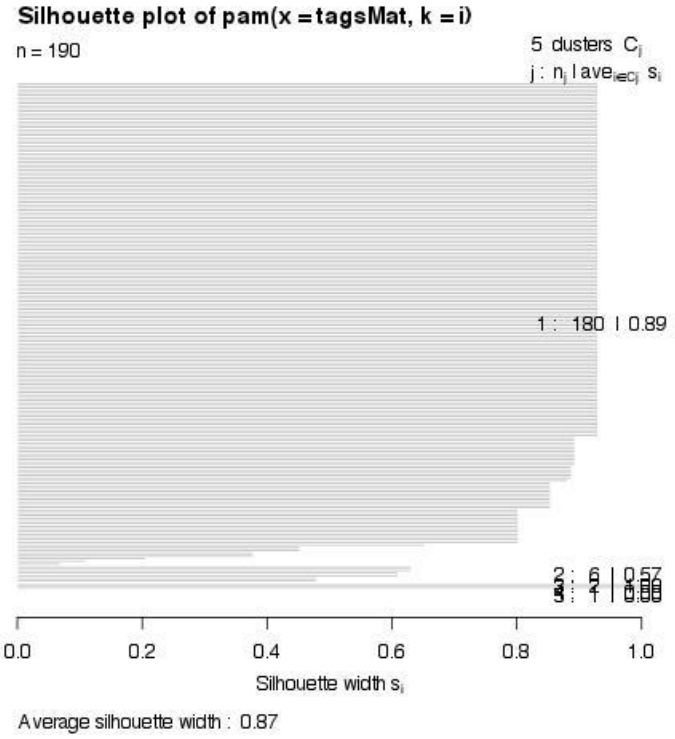
b)

a) topological groupings: document representation – SETS; similarity measure - cosine
b) silhouettes of the clusters and distribution of files per cluster

Figure 6.7 A clustering solution according to the proposed algorithm - LoA ≥ 1.9



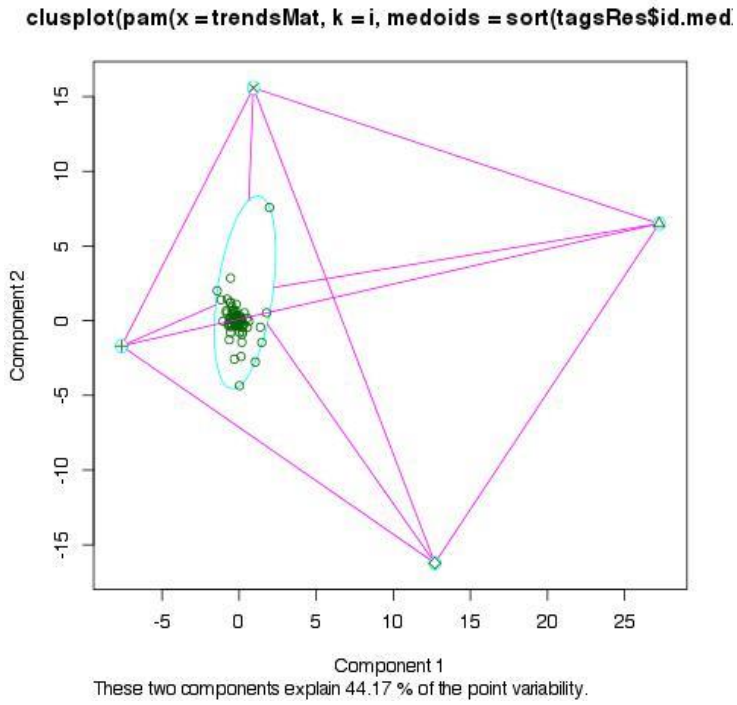
a)



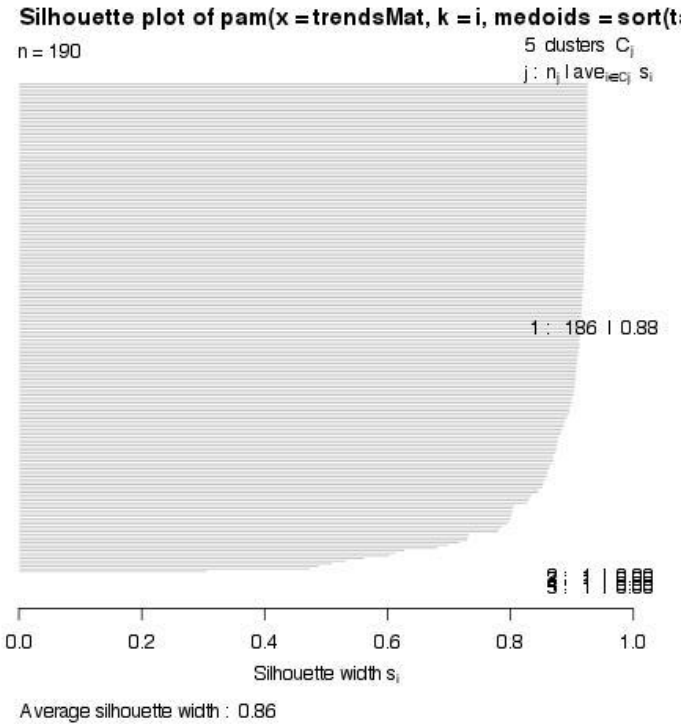
b)

a) topological groupings: document representation – HJ; similarity measure - cosine
b) silhouettes of the clusters and distribution of files per cluster

Figure 6.8 A clustering solution according to Human Judgment - LoA ≥ 1.9



a)



b)

a) topological groupings: document representation – the Porter; similarity measure - cosine
b) silhouettes of the clusters and distribution of files per cluster

Figure 6.9 A clustering solution according to a base line algorithm - LoA ≥ 1.9

6. 6. Summary

The proposed in this chapter clustering methodology aligns clustering solutions produced by the traditional clustering techniques better to human judgement. The improved performance of the traditional methods is achieved by introducing a level of abstraction (LoA) for which the produced clustering solutions are meaningful, i.e. aligned well to human judgment. Different values for the LoA enable traditional algorithms to produce different clustering solutions without changing the document representation, the similarity measure, or the clustering approach.

The proposed methodology utilises a technique that removes documents from the representation of a collection that introduce noise for a given level of abstraction. The LoA sets a threshold so that a similarity (sim) measured between a pair of documents is set to 0 if $LoA > sim$. Thus, if a selected value for LoA is set to a high value, the produced clustering solutions have low level of abstraction. Relations between documents, defined by the little abstraction of the clustering solutions, are very rigorous. On the other hand, a small value for the LoA defines more abstract relations between documents. Abstract relations produce abstract clustering solutions.

The evaluation of the methodology suggests that clustering solutions produced with a high level of abstraction (small value for the LoA) align worse to human judgement. In contrast, clustering solutions produced with a low level of abstraction (large value for LoA) align better to human judgement. In the second case, documents that are remained in the collection and are clustered are thematically closer to each other and refer to specific knowledge. Therefore, an observation that for smaller number of clusters the clustering solutions generated are more consistent to human judgement.

However, if the number of clusters that a collection has to be partitioned in is very large, i.e. 50, 60 ... 100, the multi-dimensional similarity measure will produce more consistent to human judgement clusters.

Chapter 7 : Contributions and conclusions

7. 1. Contributions

The main contributions of this thesis is the development of an approach to semantically enhanced clustering, which produces on a large scale coherent multiple clustering solutions that are more consistent and better aligned to human judgment than solutions produced by traditional clustering. The specific contributions are listed below.

1. A conceptual model of semantically enhanced document clustering provides multiple deterministic clustering solutions and different viewpoints to a collection of documents by employing semantics, many-to-many matching, and levels of abstraction.
   (a) The model does not rely on dimensionality reduction techniques, which benefits its simplicity.
   (b) The model enables large scale experiments in a full feature space due to the reduced dimensionality of the document representation it uses.
   (c) The model enables users to select the most meaningful grouping, which is consistent with a given level of abstraction by utilising the clustering solutions through a feedback pathway that is linked to the similarity measure and not the clustering step.
   (d) The model enables different representations, i.e. multiple viewpoints, of a document collection by using levels of abstraction.

2. Introducing concept indexing, which uses a general knowledge source, as a document representation technique to clustering? Concept indexing employs semantic relations between words, statistical data of co-occurrence, and a pre-defined number of dimensions.
   (a) Concept indexing provides reduced dimensionality of document representation
   (b) The reduced dimensionality of document representation provides simplicity to implementing sophisticated similarity measures.
   (c) Concept indexing provides more generic document representation.
   (d) Concept indexing enables the use of a computationally expensive many-to-many matching for measuring similarity between documents
3. A semantically enhanced text stemmer (SETS) represents semantically enhanced version of the Porter stemmer, which is a benchmark stemmer for document normalisation, and provides a semantically enhanced stemming, which normalises text by using external knowledge source.
   (a) SETS improve the separation between and the coherence within clustering solutions generated by traditional clustering algorithms.
   (b) SETS semantically disambiguate words by employing external knowledge source.
   (c) Clustering solutions produced from documents normalised by SETS and represented by concept indexing, align better to human judgement than those produced by traditional clustering for a wide range of a number of clusters.
   (d) Clustering solutions generated across domains from documents normalised by SETS and represented by concept indexing, align more consistently to human

judgement for a wide range of numbers of clusters than those produced by traditional clustering.

4. A method for measuring similarity between documents, which employs semantic relations between words established in an external knowledge source and a many-to-many matching.
   (a) A many-to-many matching measures similarity between document by considering distributions of features representing them. This approach to measuring document similarity is a step forward to better alignment of the clustering solutions generated to human judgement.
   (b) The method enables the employment of the computationally expensive many-to-many matching for large scale experiments.
   (c) The use of external knowledge source for measuring similarity between documents enables multiple viewpoints to a document collection to be generated.
5. A semantically enhanced methodology that employs levels of abstraction, i.e. thresholds, at which similarity between documents is measured.
   (a) The semantically enhanced clustering methodology provides clustering solutions that are more consistent and better aligned to human judgment for a given level of abstraction than those produced by traditional clustering. The higher level of abstraction, the more inconsistent clustering solutions generated to human judgment.
   (b) The methodology provides deterministic clustering solutions on a large scale without using dimensionality reduction.
   (c) The methodology enables higher speed of clustering for clustering solutions generated with lower level of abstraction.

7. 2. Conclusions

Clustering solutions, produced by traditional clustering algorithms that employ a co-sine measure in a collection of documents with various topics, are found to be inconsistent and poorly aligned to human judgment. The evaluations carried out in chapters 4, 5 and 6 demonstrate that these limitations can be overcome by using the proposed conceptual model (proposed in chapter 3, which addresses the first objective) and a document representation technique (such as concept indexing for which a methodology was proposed in chapter 4, which addresses the second objective) that considers the context and meaning of a word in measuring document similarity.

The SETS algorithm (proposed in chapter 4, which addresses the second objective) enables distributionally driven similarity measure between representative features for documents by using many-to-many matching (discussed in chapter 5, addresses the third objective). Experimental evidence in chapter 5 demonstrates that statistical information acquired from a large collection used to represent documents, semantic word disambiguation and semantically enhanced many-to-many matching for measuring similarity between documents are prerequisites for aligning clustering solutions closer to human judgment, but are just about enough to exceed a threshold of 40% stated in related literature. The experimental results from chapter 5 motivate the research presented in chapter 6, which addresses the fourth objective of this thesis.

The research presented in chapter 6 considers that concept indexing provides document representation, which enables the traditional partitional clustering to produce clustering solutions across domains with better separation between and improved coherence within clustering solutions. The methodology proposed in chapter 6 uses the

fact that semantically enhanced text stemming improves clustering solutions generated for a wide range of numbers of clusters (k) to align them better to human judgement even when k is not suitable for the collection. In addition, the proposed methodology in the same chapter takes advantage of the fact that concept indexing provides reduced dimensionality for document representation to employ computationally expensive similarity measures such as EMD on a large scale. Chapter 6, as well as chapter 5, discuss clustering results in the view of their silhouette values. However, the experimental analysis demonstrates that this measure is inadequate to indicate the consistency of the clustering solutions generated to human judgment, unless levels of abstraction for the document representation are employed.

Levels of abstraction are used to target the fourth objective of this thesis, which is to produce multiple viewpoints provided by deterministic clustering solutions to a document collection. Although, different views to document clustering are a prerequisite for efficient browsing and navigation within and between clusters, the consistency and the alignment of clustering solutions produced for a given level of abstraction depend on the background, understanding and motivation of the user.

Nevertheless, conceptually organised cognitive structures provide a knowledge representation that enables traditional clustering algorithms to produce clustering solutions across domains consistent and well aligned to the user's judgement. Experimental evidence is given in chapter 6 showing that clustering solutions produced with a higher level of abstraction are more inconsistent and poorly aligned to human judgement than clustering solutions produced with a lower level of abstraction. This observation provides more information for the fourth objective of this research and

outlines the use of the architectural model of a document clustering system, which is the fifth objective.

7. 3. Future work

With regard to document concept similarity, a possible research direction is improvement of its accuracy in specific domains and the integration of a specific ontology in the process of measuring similarity. This would enhance the performance of the semantically enhanced clustering methodology across domains.

With regard to semantically enhanced clustering methodology, the mechanism for representing conceptually organised cognitive structures by levels of abstraction can be further enhanced by developing mental models of human cognition and perception. This would enhance and facilitate the finding of meaningful clustering solutions.

With regard to the external knowledge source OntoRo, which groups together words that express similar ideas, any other knowledge source is likely to have a different structural organisation. The new organisation will produce different clustering solutions, which need to be evaluated following the methodology proposed in this thesis. A comparative analysis of the newly obtained results to the results obtained by following this methodology is required. The different results will be due to the different structure of the used knowledge source.

The evaluation conducted in chapters 4, 5 and 6 shows some of the weaknesses of the proposed methodology. It has quadratic complexity of the number of documents $O(n^2)$. For real life scenarios where algorithms process millions of documents the methodology needs to have linear complexity. The linearity enables algorithms that

implement the methodology to work in parallel. Therefore, more work is needed towards linearizing the proposed nonlinear methodology.

With regard to the confidence of the produced clustering solutions in relation to human judgment a study is needed, the purpose of which would be to utilise the relation be-tween silhouette values and the consistency of the clustering solutions with human judgment, i.e. a range of values for which the clusters produced will align well to human judgment. A limitation of the algorithm is that it is designed to work well on generic cross-domain document collections. In collections where documents are topically grouped according to named entities, the algorithm is expected to perform less satisfactorily. However, the performance can be improved by using a domain-specific ontology, which provides specific semantic in-formation referring to the terms of the domain of interest or named entities. Therefore, further research is needed towards combining word-based clustering algorithms with the proposed methodology. The size of the collection is relevant with regard to the quality of the clustering solutions. The size matters with regard to the efficiency of the algorithm that implements the proposed methodology, but not the effectiveness of the groupings.

As for a user interface enabling interaction with the end user, research is needed towards developing a mechanism for intuitive modification of the level of abstraction of clustering results.

Another important issue worth further investigation is the applicability of the semantically enhanced clustering methodology developed initially for document clustering to other areas such as document search and retrieval, and intelligent tutoring systems for producing personalised educational materials.

References

Andrews, Nicholas and Edward Fox (2007) Recent Developments in Document Clustering. Computer Science, Virginia Tech.

Aslam, Javed A. and Meredith Frost (2003) An Information-Theoretic Measure for Document Similarity. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* Toronto, Canada.

Backer, Eric and Anil K. Jain (1981) A Clustering Performance Measure Based on Fuzzy Set Decomposition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* PAMI-3**,** 66-75.

Baeza-Yates, Ricardo A. and Berthier Ribeiro-Neto (1999) *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc.

Baker, L. Douglas and Andrew Kachites McCallum (1998) Distributional Clustering of Words for Text Classification. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* Melbourne, Australia, ACM New York, NY, USA.

Baraldi, A. and E. Alpaydin (2002) Constructive Feedforward Art Clustering Networks. Ii. *Neural Networks, IEEE Transactions on,* 13**,** 662-677.

Barsalou, L. W. and U. Neisser (1987) The Instability of Graded Structure: Implications for the Nature of Concepts. *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge University Press.

Beil, Florian, Ester Martin and Xu Xiaowei (2002) Frequent Term-Based Text Clustering. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* Edmonton, Alberta, Canada, ACM.

Berry, Michael W., Susan Dumais and Gavin O'Brien (1995) Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review,* 37**,** 573-595.

Blair, David C. (1979) Information Retrieval, 2nd Ed. C.J. Van Rijsbergen. London: Butterworths. *Journal of the American Society for Information Science,* 30**,** 374-380.

Boroditsky, Lera (2007) Comparison and the Development of Knowledge. *Cognition,* 102**,** 118-128.

Burkey, J. and W. L. Kuechler (2003) Web-Based Surveys for Corporate Information Gathering: A Bias-Reducing Design Framework. *Professional Communication, IEEE Transactions on,* 46**,** 81-93.

Cadez, Igor V., Scott Gaffney and Padhraic Smyth (2000) A General Probabilistic Framework for Clustering Individuals and Objects. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.* Boston, Massachusetts, United States.

Cancedda, Nicola, Gaussier Eric, Goutte Cyril and Renders Jean Michel (2003) Word Sequence Kernels. *J. Mach. Learn. Res.,* 3**,** 1059-1082.

Carpineto, Claudio, Osi Stanislaw, ski, Romano Giovanni and Weiss Dawid (2009) A Survey of Web Clustering Engines. *ACM Comput. Surv.,* 41**,** 1-38.

Chen, Jisong, Yeh Chung-Hsing and Chau Rowena (2006) Identifying Multi-Word Terms by Text-Segments. *Proceedings of the Seventh International*

Conference on Web-Age Information Management Workshops. Hong Kong, China, IEEE Computer Society

Cherkassky, Vladimir and Filip Mulier (2007) Learning from Data - Concepts, Theory, and Methods. Chicester, GB, John Wiley and Sons Ltd.

Chi, M. T. H., R. Glaser and M. J. Farr (1988) *The Nature of Expertise,* New York, Hillsdale.

Chi, Michelene T. H., Paul J. Feltovich and Robert Glaser (1981) Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science,* 5**,** 121-152.

Cowie, Jim, Joe Guthrie and Louise Guthrie (1992) Lexical Disambiguation Using Simulated Annealing. *Proceedings of the workshop on Speech and Natural Language.* Harriman, New York, Association for Computational Linguistics.

Cutting, Douglass, R.,, R. Karger David, O. Pedersen Jan and W. Tukey John (1992) Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.* Copenhagen, Denmark.

Deerwester, Scott, Susan Dumais, Thomas Landauer, George Furnas and Richard Harshman (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science,* 41**,** 391-407.

Denaux, Ronald, Catherine Dolbear, Glen Hart, Vania Dimitrova and Anthony G. ohn (2011) Supporting Domain Experts to Construct Conceptual Ontologies: A Holistic Approach. *Web Semantics: Science, Services and Agents on the World Wide Web,* 9**,** 113-127.

Dhillon, Inderjit S. and Dharmendra S. Modha (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning,* 42**,** 143-175.

Dwivedi, S. K. and Rastogi Parul (2009) Critical Analysis of Wsd Algorithms. *Proceedings of the International Conference on Advances in Computing, Communication and Control.* Mumbai, India.

Eissen, Sven, Benno Stein and Martin Potthast (2005) The Suffix Tree Document Model Revisited. *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05).* Know-Center. ISSN 0948-695x.

Engelbrecht, Paul and aItiel Dror, E. Dror (2009) How Psychology and Cognition Can Inform the Creation of Ontologies in Semantic Technologies. *Proceeding of the 2009 conference on Information Modelling and Knowledge Bases XX.* IOS Press.

Estivill-Castro, Vladimir (2002) Why So Many Clustering Algorithms: A Position Paper. *SIGKDD Explorations Newsletter,* 4**,** 65-75.

Everitt, B., S. Landau, and and M. Leese (2001) *Cluster Analysis,* London, Edward Arnold, Ltd.

Facolta, D., S. Universita and I.R. Italie (2008) Multi-Class Categorization Based on Cluster Analysis and Tfidf. *Economia***,** 209-217.

Firth, John (1957) A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis***,** 1-32.

Fountain, Tony, Hussein Almuallim and Thomas G. Dietterich (1991) Learning with Many Irrelevant Features. *In Proceedings of the Ninth National Conference on Artificial Intelligence.* Anaheim, California, AAAI Press.

Frigui, H. and O. Nasraoui (2004) Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents. IN Berry, Michael (Ed.) *Survey of Text Mining.* Springer.

Fung, B., K. Wang and M. Ester (2003) Hierarchical Document Clustering Using Frequent Itemsets. *SIAM International conference of Data mining, SDM'03.* San Francisco, CA, USA.

Fung, B., K. Wang and M. Ester (2005) Hierarchical Document Clustering. *The Encyclopedia of Data Warehousing and Mining.* John Wang ed. NY, USA, Idea Group.

Gale, W. A., K. W. Church and D. Yarowsky (1992) One Sense Per Discourse. *Proceedings of the workshop on Speech and Natural Language.* Harriman, New York, Association for Computational Linguistics.

Glaser, R. (1991) Expertise and Assessment. IN Wittrock, M. C. and Baker, E. L. (Eds.) *Testing and Cognition.* New York, Prentice Hall.

Gliozzo, A., C. Strapparava and I. Dagan (2004) Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language,* 18**,** 275-299.

Goldsmith, Timothy E., Peder J. Johnson and William H. Acton (1991) Assessing Structural Knowledge. *Journal of Educational Psychology,* 83**,** 88-96.

Gordon, A. D. (1999) *Classification,* London,, Chapman and Hall.

Grefenstette, E. (2009) Analysing Document Similarity Measures. *Computing Laboratory.* Oxford, University of Oxford.

Gruber, Thomas (1993) A Translation Approach to Portable Ontology Specifications. *Knowl. Acquis.,* 5**,** 199-220.

Hammouda, Khaled M. and Mohamed S. Kamel (2004) Efficient Phrase-Based Document Indexing for Web Document Clustering. *IEEE Trans. on Knowl. and Data Eng.,* 16**,** 1279-1296.

Hearst, Marti A. (1993) Texttiling: A Quantitative Approach to Discourse. University of California at Berkeley.

Hearst, Marti A. (1997) Texttiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Comput. Linguist.,* 23**,** 33-64.

Hearst, Marti A. (1999) The Use of Categories and Clusters in Information Access Interfaces. *Natural Language Information Retrieval.* Kluwer.

Hotho, A., S. Staab and A. Maedche (2001) Ontology-Based Text Clustering. *Proc. IJCAI'01 Workshop ``Text Learning: Beyond Supervision''.*

Hotho, A., S. Staab and G. Stumme (2003a) Ontologies Improve Text Document Clustering. *Procd of the International Conference on Data Mining.* Los Alamitos, IEEE Press.

Hotho, A., S. Staab and G. Stumme (2003b) Ontologies Improve Text Document Clustering. *Third IEEE International conference on Data Mining, 2003. ICDM 2003. .*

Hotho, A., S. Staab and G. Stumme (2003c) Text Clustering Based on Background Knowledge. *Technical Report N425.* Institute of Applied Informatics and Formal Description Methods, AIFB, University of Karlsruhe, D–76128 Karlsruhe, Germany.

Hotho, Andreas and Steffen Staab (2003) Ontology-Based Text Document Clustering. *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'03 Conference held in Zakopane.*

Hsien-Hsun, Chen, Cheng Shein-Yung and Heh Jia-Sheng (2005) Assessing Users' Mental Knowledge by Using Structural Approach and Concept Map. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics.*

Huang, Anna (2008) Similarity Measures for Text Document Clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand***,** 49-56.

Huma, Lodhi, Saunders Craig, Shawe-Taylor John, Cristianini Nello and Watkins Chris (2002) Text Classification Using String Kernels. *J. Mach. Learn. Res.*, 2**,** 419-444.

Ide, Nancy and Jean Veronis (1998) Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Comput. Linguist.*, 24**,** 2-40.

Jain, A. K., R. P. W. Duin and Mao Jianchang (2000) Statistical Pattern Recognition: A Review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 22**,** 4-37.

Jain, A. K., M. N. Murty and P. J. Flynn (1999) Data Clustering: A Review. *ACM Computing Surveys*, 31**,** 264-323.

Jain, Anil K. and Richard Dubes (1988) *Algorithms for Clustering Data*, NJ, USA, Prentice Hall.

Jarmasz, M. and S. Szpakowicz (2003a) Not as Easy as It Seems: Automating the Construction of Lexical Chains Using Roget's Thesaurus. *Proceedings of the 16th Canadian Conference on Artificial Intelligence* Halifax, Canada.

Jarmasz, M. and S. Szpakowicz (2003b) Roget's Thesaurus and Semantic Similarity. *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003).*

Jing, Liping (2008) Survey of Text Clustering. *Department of Mathematics.* HongKong, China, the University of Hong Kong.

Jones, William P. and George W. Furnas (1987) Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science,* 38**,** 420-442.

Karatzoglou, Alexandros and Ingo Feinerer (2006) Text Clustering with String Kernels in R. Vienna, Department of Statistics and Mathematics, WU Vienna University of Economics and Business.

Karypis, G., Han Eui-Hong and V. Kumar (1999) Chameleon: Hierarchical Clustering Using Dynamic Modeling. *Computer,* 32**,** 68-75.

Katz, P. and P G. Pinkham (2006) Word Sense Disambiguation Using Latent Semantic Analysis.

Kaufman, Leonard and Peter Rousseeuw (2005) *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics),* Wiley-Interscience.

Kaufman, Leonard and Peter J. Rousseeuw (1990) *Finding Groups in Data: An Introduction to Cluster Analysis,* New York, Wiley.

Khan, L., F. Luo and I. Yen (2003) Automatic Ontology Derivation from Documents. *Proceedings of the 15th conference on Advanced Information Systems Engineering (CAISE '03).* Klagenfurt/Velden, Austria.

Khan, Latifur and Feng Luo (2002) Ontology Construction for Information Selection. *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence.* IEEE Computer Society.

Krishnapuram, R., A. Joshi and Yi Liyu (1999) A Fuzzy Relative of the K-Medoids Algorithm with Application to Web Document and Snippet Clustering. *Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99. 1999 IEEE International.*

Krovetz, Robert (1993) Viewing Morphology as an Inference Process. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* Pittsburgh, Pennsylvania, United States.

Lakkaraju, Praveen, Susan Gauch and Mirco Speretta (2008) Document Similarity Based on Concept Tree Distance. *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia.* Pittsburgh, PA, USA.

Leacock, C. and M. Chodorow (1998) Combining Local Context and Wordnet Similarity for Word Sense Identification. *Wordnet: An Electronic Lexical Database.* In C. Fellbaum (Ed.), MIT Press.

Lee, Chang-Shing, Yuan-Fang Kao, Yau-Hwang Kuo and Mei-Hui Wang (2007) Automated Ontology Construction for Unstructured Text Documents. *Data Knowl. Eng.,* 60**,** 547-566.

Lee, Lillian (1999) Measures of Distributional Similarity. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics.* College Park, Maryland, Association for Computational Linguistics.

Lee, M. D., B. Pincombe and M. Welsh (2005) A Comparison of Machine Measures of Text Document Similarity with Human Judgments. *27th Annual Meeting of the Cognitive Science Society CogSci2005,* 27**,** 1254–1259.

Lesk, Michael (1986) Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 5th annual international conference on Systems documentation.* Toronto, Ontario, Canada.

Lewis, David, D. (1992) Text Representation for Intelligent Text Retrieval: A Classification-Oriented View. *Text-Based Intelligent Systems.* L. Erlbaum Associates Inc.

Li, Tao, Sheng Ma and Mitsunori Ogihara (2004) Document Clustering Via Adaptive Subspace Iteration. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* Sheffield, United Kingdom.

Li, Yanjun, Soon M. Chung and John D. Holt (2008) Text Document Clustering Based on Frequent Word Meaning Sequences. *Data & Knowledge Engineering,* 64**,** 381-404.

Lin, Dekang (1998a) Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 17th international conference on Computational linguistics - Volume 2.* Montreal, Quebec, Canada, Association for Computational Linguistics.

Lin, Dekang (1998b) An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning.* Morgan Kaufmann Publishers Inc.

Lin, J. (1991) Divergence Measures Based on the Shannon Entropy. *Information Theory, IEEE Transactions on,* 37**,** 145-151.

Lipsman, Andrew, Carmela Aquino and Stephanie Flosi (2013) 2013 U.S. Digital Future in Focus. http://www.comscore.com/Insights/Presentations_and_Whitepapers/2013/2013_US_Digital_Future_in_Focus, comScore.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008) *Introduction to Information Retrieval,* Cambridge UP, Cambridge University Press.

Matveeva, Irina (2006) Document Representation and Multilevel Measures of Document Similarity. *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: doctoral consortium.* New York, New York, Association for Computational Linguistics.

McCarthy, Diana (2009) Word Sense Disambiguation: An Overview. *Language and Linguistics Compass,* 3**,** 537-558.

McDonald, Sharon, John Tait, Vladimir Dobrynin, David Patterson and Niall Rooney (2004) Contextual Document Clustering. *Advances in Information Retrieval.* Springer Berlin, Heidelberg.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller (1990) Introduction to Wordnet: An on-Line Lexical Database. *Journal of Lexicography,* 3**,** 235--244.

Mugunthadevi, MS. K., MRS. S.C. Punitha and Punithavalli Dr..M (2011) Survey on Feature Selection in Document Clustering. *International Journal on Computer Science and Engineering,* 3**,** 1240-1244.

Ng, A., M. Jordan and Y. Weiss (2001) On Spectral Clustering: Analysis and an Algorithm. IN Dietterich, T., Becker, S. and Ghahramani, Z. (Eds.) *Advances in Neural Information Processing Systems.* MIT Press.

Norvig, Peter (1987) Inference in Text Understanding. *In National Conference on Artificial Intelligence AAAI-87.*

Novak, Joseph D. (1990) Concept Mapping: A Useful Tool for Science Education. *Journal of Research in Science Teaching,* 27**,** 937-949.

Parsons, Lance, Ehtesham Haque and Huan Liu (2004) Subspace Clustering for High Dimensional Data: A Review. *SIGKDD Explor. Newsl.,* 6**,** 90-105.

Patwardhan, Siddharth (2003) Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. *Graduate School.* Duluth, University of Minesota.

Peng, Xiaogang and Ben Choi (2005) Document Classifications Based on Word Semantic Hierarchies. *In Proceedings of the International Conference on Artificial Intelligence and Applications***,** 362-367.

Porter, M. F. (1997) An Algorithm for Suffix Stripping. IN Karen Sparck Jones and Willett, Peter (Eds.) *Readings in Information Retrieval.* 137 ed. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

Punitha, S.C., K. Mugunthadevi and M. Punithavalli (2011) Impact of Ontology Based Approach on Document Clustering. *International Journal of Computer Applications,* 22**,** 22-26.

Resnik, Philip (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1.* Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc.

Robertson, Stephen (2004) Understanding Inverse Document Frequency: On Theoretical Arguments for Idf. *Journal of Documentation,* Volume 60**,** 503-520.

Rooney, Niall, David Patterson, Mykola Galushka and Vladimir Dobrynin (2006) A Scaleable Document Clustering Approach for Large Document Corpora. *Information Processing & Management,* 42**,** 1163-1175.

Rousseeuw, Peter (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics,* 20**,** 53-65.

Rubner, Yossi, Carlo Tomasi and Leonidas Guibas (2000) The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision,* 40**,** 99-121.

Salton, G. and C. Buckley (1998) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management,* 24**,** 513-523.

Salton, Gerard and Michael J. McGill (1986) *Introduction to Modern Information Retrieval,* New York, NY, USA, McGraw-Hill, Inc.

Sánchez, David, Montserrat Batet and David Isern (2011) Ontology-Based Information Content Computation. *Knowledge-Based Systems,* 24**,** 297-303.

Setchi, R and Q. Tang (2007) Semantic-Based Representation of Content Using Concept Indexing. *Proceedings of I*PROMS.* Cardiff, UK.

Setchi, Rossitza, Qiao Tang and Carole Bouchard (2009) Ontology-Based Concept Indexing of Images. *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part I;* . Santiago, Chile, Springer-Verlag.

Setchi, Rossitza, Qiao Tang and Ivan Stankov (2011) Semantic-Based Information Retrieval in Support of Concept Design. *Advanced Engineering Informatics,* 25**,** 131-146.

Shavelson, Richard J. (1974) Some Methods for Examining Content Structure and Cognitive Structure in Instruction 1. *Educational Psychologist,* 11**,** 110-122.

Slonim, Noam, Nir Friedman and Naftali Tishby (2002) Unsupervised Document Classification Using Sequential Information Maximization. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.* Tampere, Finland.

Slonim, Noam and Naftali Tishby (2000) Document Clustering Using Word Clusters Via the Information Bottleneck Method. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* Athens, Greece.

Steinbach, M., G. Karypis and V. Kumar (2000) A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining, Boston.*

Voorhees, Ellen M. (1998) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* Melbourne, Australia, ACM.

Wan, Stephen, CÃ©cile Paris and Robert Dale (2010) Supporting Browsing-Specific Information Needs: Introducing the Citation-Sensitive in-Browser Summariser. *Web Semantics: Science, Services and Agents on the World Wide Web,* 8**,** 196-202.

Wan, Xiaojun (2007) A Novel Document Similarity Measure Based on Earth Mover's Distance. *Inf. Sci.,* 177**,** 3718-3730.

Wan, Xiaojun and Yuxin Peng (2005a) The Earth Mover's Distance as a Semantic Measure for Document Similarity. *Proceedings of the 14th ACM international conference on Information and knowledge management.* Bremen, Germany.

Wan, Xiaojun and Yuxin Peng (2005b) A New Retrieval Model Based on Texttiling for Document Similarity Search. *Comput. Sci. Technol.,* 20**,** 552-558.

Wan, Xiaojun, Jianwu Yang and Jianguo Xiao (2007) Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* Association for Computational Linguistics.

Wandersee, James H. (1990) Concept Mapping and the Cartography of Cognition. *Journal of Research in Science Teaching,* 27**,** 923-936.

Wang, Ke, Xu Chu and Liu Bing (1999) Clustering Transactions Using Large Items. *Proceedings of the eighth international conference on Information and knowledge management.* Kansas City, Missouri, United States, ACM.

Wang, Lipo, Yaochu Jin, Jiangning Wu and Guangfei Yang (2005) An Ontology-Based Method for Project and Domain Expert Matching. *Fuzzy Systems and Knowledge Discovery.* Springer Berlin / Heidelberg.

Witten, Ian H. (2010) Wikipedia and How to Use It for Semantic Document Representation. *Proceedings of the 5th international conference on Rough set and knowledge technology.* Beijing, China, Springer-Verlag.

Wu, Zhibiao and Martha Palmer (1994) Verb Semantics and Lexical Selection. *32nd. Annual Meeting of the Association for Computational Linguistics.*

Xiao, Yu (2010) A Survey of Document Clustering Techniques & Comparison of Lda and Movmf. Stanford University, Computer Science Project, Class 229.

Xu, Jinxi and W. Bruce Croft (1998) Corpus-Based Stemming Using Cooccurrence of Word Variants. *ACM Trans. Inf. Syst.,* 16**,** 61-81.

Xu, Rui and D. Wunsch (2005) Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions on,* 16**,** 645-678.

Yang, XiQuan, DiNa Guo, XueYa Cao and JianYuan Zhou (2008) Research on Ontology-Based Text Clustering. *Semantic Media Adaptation and Personalization, 2008. SMAP '08. Third International Workshop on.*

Yarowsky, David (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of the 14th conference on Computational linguistics - Volume 2.* Nantes, France, Association for Computational Linguistics.

Yong, Wang and J. Hodges (2006) Document Clustering with Semantic Analysis. *System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on.*

Yun, Jiali, Liping Jing, Jian Yu and Houkuan Huang (2010) Semantics-Based Representation Model for Multi-Layer Text Classification. *Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II.* Cardiff, UK, Springer-Verlag.

Zadeh, L. (1965) Fuzzy Sets. *Information and Control,* 8**,** 338-353.

Ženko, Bernard (2007) Learning Predictive Clustering Rules. *Faculty of computer and information science.* Ljubljana, University of Ljubljana.

Zhang, W., Yoshida T., Ho T. B. and Tang X. J. (2009) Augmented Mutual Information for Multi-Word Extraction. *International Journal of Innovative Computing Information and Control,* 5**,** 543-554.

Zhang, Wen, Taketoshi Yoshida and Xijin Tang (2011) A Comparative Study of Tf*Idf, Lsi and Multi-Words for Text Classification. *Expert Systems with Applications,* 38**,** 2758-2765.

Zhang, Yongpeng, F. Mueller, Cui Xiaohui and T. Potok (2010) Large-Scale Multi-Dimensional Document Clustering on Gpu Clusters. *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on.*

Zhao, Y. and G. Karypis (2001) Criterion Functions for Document Clustering: Experiments and Analysis. *Technical Report, Department of Computer Science, University of Minnesota, Minneapolis, MN.*

Zhao, Ying and George Karypis (2002) Evaluation of Hierarchical Clustering Algorithms for Document Datasets. *Proceedings of the eleventh international conference on Information and knowledge management.* McLean, Virginia, USA, ACM.

Zheng, Hai-Tao, Bo-Yeong Kang and Hong-Gee Kim (2009) Exploiting Noun Phrases and Semantic Relationships for Text Document Clustering. *Information Sciences,* 179**,** 2249-2262.

Zobel, Justin and Alistair Moffat (1998) Exploring the Similarity Space. *SIGIR Forum,* 32**,** 18-34.

Appendix A - Data supporting the illustrative example in chapter 5

The documents below appear in the order presented in the illustrative example from chapter 5.

Content	Document Name
Muhammad Shaaban. Muhammad Shaaban (born on 13 June 1942) is a career Egyptian diplomat and currently serves as the United Nations Under-Secretary-General for General Assembly and Conference Management. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in February 2007. Shaaban has acquired extensive knowledge about the United Nations. From	AY-wiki00011(c2=90-

1984 to 1988, he served as Egypt's representative to the Second Committee of the United Nations Economic and Social Council, the United Nations Development Programme (UNDP), and several intergovernmental bodies and committees. Between 1985 and 1986, he acted as the Coordinator of the "Group of 77" developing countries and China. A seasoned diplomat, Shaaban served his government in various diplomatic capacities. From 1993 to 1997, he was Ambassador to Belgium and Luxembourg and Head of Egypt's Permanent Mission to the European Union in Brussels. In the following year, he served as Assistant Foreign Minister for African Affairs. Between 1998 and 2000, he was Ambassador to Denmark and Lithuania. From 2000 to 2001, he was Assistant Minister for Information, Research and Assessment and National Coordinator for Information, Research and Assessment. In the following three years, he served as Assistant Foreign Minister for European Affairs. Shaaban has been National Coordinator for Reform Initiatives in the Middle East since 2004. In this capacity, he maintained relations with his foreign partners and coordinated with various departments in the Egyptian government as well as political parties and civil society. He also advised the Foreign Minister on various issues. He obtained his Ph.D. in political science and his Master of Arts in international relations from Brussels University. He speaks fluently English and French and has fair knowledge of Portuguese and Spanish.	w2=156):
TafirÃ©. TafirÃ© is a town and commune of the Katiola department in the VallÃ©e du Bandama region of CÃ´te d'Ivoire. It is served by a station on the nation railway system.	AY-wiki74450(c2=58-w2=12):
Ambassador of India to Russia. The following people have served as Ambassadors of India to Russia and it predecessor state, the Soviet Union:	AY-wiki97072(c2=65-w2=10):
Wolfgang StÃ¶ckl. Wolfgang StÃ¶ckl (born on 11 June 1948) currently serves as the Vice Chairman of the United Nations International Civil Service Commission. Before becoming the Vice Chairman of ICSC, StÃ¶ckl was Ambassador and Special Coordinator for German Personnel in International Organizations. From 2000 to 2002, he served as Director of Economic and Development Affairs in the United Nations and Global Affairs Department of the Foreign Office of the Federal Republic of Germany. From 1997 to 2000, he was First Counsellor of the Permanent Mission of the Federal Republic of Germany to the Organization for Economic Cooperation and Development in Paris, in charge of the Public Management Committee and human resources management issues. He was a member of the ICSC from 1997 to 2002. In 1997, he served as Chairman of the Committee for Programme and Coordination of the United Nations. Between 1995 and 1997, he was member of the UN Advisory Committee on Administrative and Budgetary Questions(ACABQ). From 1991 to 1997, he was Counsellor of the Permanent Mission of the Federal Republic of Germany to the United Nations in New York. In this capacity, he was in charge of the UN reform, common system and human resources management issues. Earlier in his career, he served on a variety of capacities in the German foreign service, including as Special Adviser for Management and Personnel Questions to the Minister for Foreign Affairs of the German Democratic Republic in 1990, as Deputy Head of the Organization and Management Division of the Foreign Office from 1989 to 1991, as Head of the Headquarters Inspection Unit of the Foreign Office, Director of the German-Saudi Arabian Liaison Office for Economic Affairs in Riyadh, Saudi Arabia, as German Counsul in Cairo, Egypt and as Assistant Director in the Training Centre	AY-wiki01800(c2=90-w2=221):

of the Federal Foreign Service. He joined the German Foreign Office in 1977. Previously, he served the Ministry of Interior in Land Hesse, Germany, in 1976. He was Assistant judge and assistant prosecutor in the Ministry of Justice in Land Hesse from 1973 to 1974. He studied in the Training Centre of the Federal Foreign Office in Bonn from 1977 to 1979. He holds a Postgraduate Degree in Public Administration at the Postgraduate School of Public Administration in Speyer, German. He passed the Second State Examination in law(Bar Examination) in 1975 and the First State Examination in law(Masters Degree) in 1972. From 1967 to 1971, he studied law at the University of Marburg, Germany.	
Ali'ioaiga Feturi Elisaia. Ali'ioaiga Feturi Elisaia, born in 1954, is a Samoan diplomat. He is currently Samoa's Permanent Representative to the United Nations. He obtained a postgraduate certificate in diplomacy from Oxford University, and also holds a Bachelor of Arts degree in political science and administration from the University of the South Pacific. Elisaia first served in the Samoan Mission to the United Nations in 1979. From 1979 to 1981, he was Acting Division Head, and then as Division Head, at the Economic and Aid Division of the Samoan Ministry of Foreign Affairs. From 1981 to 1984, he served as First Secretary at Samoaâ€™s High Commission in New Zealand. He was Deputy Secretary for Foreign Affairs from 1984 to 1988, then co-director of the Hanns Seidel Foundation in Samoa from 1988 to 2001. From 2001 to 2003, he was Chief Executive Officer at the Samoan Ministry of Foreign Affairs. Elisaia was appointed Permanent Representative of Samoa to the United Nations in 2003.	AY-wiki56672(c2=90-w2=89):
Yohannes Mengesha. Yohannes Mengesha currently serves as United Nations Assistant Secretary-General for General Assembly and Conference Management. Previously, he was Director of the Office of the Deputy Secretary-General. During his career at the United Nations, Mengesha served in a variety of capacities. In 1976, he joined the UN as project officer for the Eastern Caribbean Office of the World Food Programme (WFP) in Trinidad and Tobago. From 1980 to 1992, he held various posts in the Southern Africa Bureau of WFP. From 1992 to 1994, he was senior adviser at the Department of Humanitarian Affairs. He served as Regional Manager in the Eastern and Southern Africa Bureau of WFP between 1994 and 1996. Mengesha held the position of Director for Iraq Programme at the Department of Humanitarian Affairs from 1996 to 1998. An Ethiopian national, Mengesha holds a Bachelor of Arts (Honours) and a Master of Arts in Law from Cambridge University, United Kingdom.	AY-wiki45194(c2=90-w2=79):
JosÃ© Victor da Silva Angelo. JosÃ© Victor da Silva Angelo (born in 1949) currently serves as the Special Representative and Head of the United Nations Mission in Central African Republic and Chad (MINURCAT). He was appointed to the position by UN Secretary-General Ban Ki-moon in January 2008. Angelo obtained a master's degree in sociology from Instituto Superior Economico e Social of the University of Evora, Portugal and studied for a Doctor of Philosophy in sociology at lâ€™UniversitÃ© Libre de Bruxelles, Belgium. He started his career as a university lecturer and served as Senior Statistician in the Portuguese National Institute of Statistics (INE). He was a member of the Electoral Commission of Portugal. Later on, he joined the United Nations, where he served in various capacities, including United Nations Development Programme (UNDP) Special Envoy for East Timor and Asia, Deputy Regional Director for Africa at UNDP in New York, Resident Coordinator/Resident Representative in the United Republic of Tanzania and the Gambia, and Deputy Resident Representative in the Central African Republic. He also served as United Nations Population Fund (UNFPA) Representative in Mozambique and United Nations Adviser in Sao Tome and Principe. His extensive experience brought him to serve on more senior positions at the United Nations. From 2000 to 2004, he served as UNDP Resident Representative in Zimbabwe. From 2005 to 2007, he was the Executive Representative of the Secretary-General for Sierra Leone, as well as the Resident Coordinator of the United Nations System in Freetown.	AY-wiki17349(c2=90-w2=136):

Catherine Bragg. Catherine Bragg (born in 1953 in Hong Kong) currently serves as Deputy Emergency Relief Coordinator in the Office for the Coordination of Humanitarian Affairs. She was appointed to this position by United Nations Secretary-General Ban Ki-moon in December 2007. Bragg obtained a PhD in Criminal Justice from the University at Albany, SUNY, a Master of Philosophy in Criminology from the University of Cambridge and a Bachelor of Science in Psychology from the University of Toronto. Throughout her career, she has served in various capacities in the Federal Public Service in the Government of Canada. In the Privy Council Office, she formulated policy advice to the Prime Minister and the Cabinet. In the Department of National Defence, she worked on human resource issues. In the Department of Justice, she focused on evaluation and strategic planning. Prior to joining the United Nations, Bragg served as the Director-General of the Humanitarian Assistance, Peace and Security Programme in the Canadian International Development Agency (CIDA) since 2004. She is the Chair of the Office for the Coordination of Humanitarian Affairs Donor Support Group and a member of the Advisory Group of the Central Emergency Response Fund.	AY-wiki37929(c2=90-w2=100):
Ike, Texas. Ike is an unincorporated community in Ellis County, Texas, USA. The community is served by Farm to Marked Road 878. The nearest city to Ike is Waxahachie.	AY-wiki35697(c2=62-w2=14):
Linnea Mellgren. Linnea Mellgren (born May 17, 1989 Sweden) is a Swedish figure skater. She is the 2006 and 2007 Swedish junior nationals silver medalist.	AY-wiki42562(c2=41-w2=14):
Peter van Walsum. Peter van Walsum was formerly the United Nations Secretary-General's Personal Envoy for Western Sahara. He was appointed to the position by United Nations Secretary-General Kofi Annan in July 2005 and left the position in September 2008 when his mandate expired. Previously, he was the Netherlands' Permanent Representative to the United Nations. He also served as his country's representative on the Security Council in 1999 and Chairman of the Iraq Sanctions Committee in 2000. He obtained his law degree from the University of Utrecht in 1959. He served in the military from 1960 to 1962 and then joined the Civil Emergency Planning in the Ministry of General Affairs from 1962 to 1963. Later on, he served at the Netherland's Ministry of Foreign Affairs, where he was posted to various positions in his nearly four decades career, including the Permanent Mission to the North Atlantic Treaty Organization (NATO) in Paris, the Embassy in Bucharest, the Permanent Mission to the United Nations in New York, the Embassy in both New Delhi and London, and the Permanent Mission to the European Commission in Brussels.	AY-wiki33493(c2=90-w2=103):
Maurice Gourdault-Montagne. Maurice Gourdault-Montagne (born on 16 November 1953) is a career diplomat and the current French Ambassador to the United Kingdom. Career. Mr Gourdault-Montagne joined the French Foreign Ministry in 1978. He served as First Secretary at the French Embassy in New Delhi (1981-83) and Deputy to the Minister to Foreign Affairs (1993-95). He became head of the Prime Minister's Office between 1995 and 1997. He served as the Ambassador to Japan in 1998.	AY-wiki49753(c2=90-w2=38):
Kaire Mbuende. Kaire Mbuende (born 28 November, 1953) is a Namibian politician and diplomat. Mbuenda has been the Namibian ambassador to the United Nations since his appointment in August 2006. An ethnic Herero, Mbuende has been a high-level member of the ruling South West Africa People's Organization (SWAPO) since 1974, when he became the Information officer in Lusaka, Zam-	AY-wiki53968(c2=75-

bia.	w2=33):
Lamberto Zannier. Lamberto Zannier is an Italian diplomat who currently serves as the United Nations Special Representative for Kosovo and Head of the United Nations Interim Administration Mission in Kosovo(UNMIK). He was appointed to this position by UN Secretary-General Ban Ki-moon in June 2008. Zannier has served for the foreign service of Italy for more than 30 years. Before his appointment as Special Representative for Kosovo, he played a leading role at the Italian Ministry of Foreign Affairs in its participation in European security and Defense Policy field operations. Between 2002 and 2006, he was Director of the Conflict Prevention Centre of the Organization for Security and Cooperation in Europe in Vienna. In this capacity, he managed more than 20 civilian field operations. From 2000 to 2002, he served as Permanent Representative of Italy to the Executive Council of the Organization for the Prohibition of Chemical Weapons in The Hague. From 1997 to 2000, he was chairperson of the negotiations on the adaptation of the Treaty on Conventional Armed Forces in Europe. From 1991 to 1997, he served as Head of Disarmament, Arms Control and Cooperative Security at the North Atlantic Treaty Organization. Zannier obtained a law degree from the University of Trieste.	AY-wiki17299(c2=90-w2=103):
Cornelius V. Clickener. Cornelius V. Clickener was the first major of Hoboken, New Jersey after the city was incorporated in 1855. He served from 1855 to 1857.	AY-wiki61621(c2=58-w2=16):
Jane Holl Lute. Jane Holl Lute currently serves as United Nations Assistant Secretary-General for Peacebuilding Support. Previously, she was Assistant Secretary-General for Mission Support in the Department of Peacekeeping Operations since August 2003. Lute holds a doctorate degree in Political Science from Stanford University and a J.D from Georgetown University. From 1991 to 1994, she served as director of European Affairs in the National Security Council staff at the White House. Between 1994 and 1999, Lute headed up the Carnegie Commission on Preventing Deadly Conflict and was a senior public policy fellow at the Woodrow Wilson Centre for International Scholars. Prior to joining the UN Secretariat, Lute served as Executive Vice President and Chief Operating Officer of the United Nations Foundation and the Better World Fund, which is established to administer Ted Turnerâ€™s $1 billion contribution to support the goals of the United Nations. Before that, she served as Executive Director of the Association of the United States Armyâ€™s project on the role of American Military Power in 2000. On January 23, 2009, President Barack Obama announced his intention to nominate Jane Holl Lute as Deputy Secretary of Homeland Security.	AY-wiki41362(c2=90-w2=109):
Harmoko. Harmoko'" (born 7 February 1939) is an Indonesian politician. He served as information minister in the New Order regime of President Suharto from 1983 until 1997 and chairman of the People's Consultative Assembly and People's Representative Council from 1997 until 1999.	AY-wiki49572(c2=90-w2=24):
Ahmed Shaaban. Ahmed Shaaban, (Arabic: Ø£ØÙ…Ø¯ Ø´Ø¹Ø¨Ø§Ù†) (born October 10, 1978) is an Egyptian footballer. He plays the defensive midfielder for Egyptian club Petrojet as well as Egypt national football team. He was a member of Egypt's squad in Ghana 2008 African Cup of Nations.	AY-wiki76183(c2=63-w2=24):
Akiko Yuge. Akiko Yuge currently serves as Assistant Administrator and Director of the Bureau of Management at the United Nations Development Programme. She was appointed to this position by United Nations Secretary-General	AY-

Kofi Annan in August 2006. Yuge began her career at UNDP as a Programme Officer in Thailand in 1976. She later served as Programme Officer in the Regional Bureau for Asia and the Pacific at UNDP Headquarters in New York and as Area Officer for China and the Philippines. From 1984 to 1987, she took special leave from UNDP and worked as a freelance development consultant and as Project Officer for the Japanese Engineering Consulting Firms Association (ECFA). In 1988, she rejoined UNDP and served in Thailand as Assistant Resident Representative until 1990. From 1990 to 1994, she was Deputy Resident Representative in Indonesia. From 1994 to 1998, she was UN Resident Coordinator and UNDP Resident Representative in Bhutan. She also taught Development Studies as a Professor at Ferris University in Yokohama in the meantime. Between 2001 and 2002, Yuge was a member of the Second Consultative Committee on ODA Reform, an advisory group to the Foreign Minister of Japan. She was also a member of the Advisory Group of International Cooperation for Peace organized by the countryâ€™s Chief Cabinet Secretary in 2002 and a member of the Eminent Personsâ€™ Group on UN Reform, an advisory group to the Foreign Minister. Since 2002, she served as Director of UNDPâ€™s Tokyo Liaison Office. Yuge obtained her Bachelorâ€™s Degree in Psychology from Barnard College at Columbia University in New York and her Masterâ€™s Degree in Development Economics from New York University.	wiki88595(c2=90-w2=142):
Operation Azure. Operation Azure'" is the name given to the Australian Defence Force's contribution to the United Nations Mission in Sudan (UNMIS). The '"United Nations Mission in the Sudan (UNMIS) was established by the United Nations under UN Security Council Resolution 1590 of the UN Security Council on March 24, in 2005, in response to the signing of the Comprehensive Peace Agreement between the government of the Sudan and the Sudan People's Liberation Movement on January 9, 2005 in Nairobi, Kenya.	AY-wiki31628(c2=90-w2=42):
Ernesto Benedettini. Ernesto Benedettini (born 5 March, 1948) is a politician of San Marino. He is Captain Regent of San Marino for the term from 1 October 2008 to April 2009 together with Assunta Meloni. He served as Captain Regent from April to October 1992. Benedettini is a member of the Sammarinese Christian Democratic Party.	AY-wiki55901(c2=77-w2=25):
Ora Namir. Ora Namir (, born 1 September 1930) is a former Israeli politician and diplomat who served as Minister of the Environment and Minister of Labour and Social Welfare during the 1990s, before becoming the country's ambassador to China and Mongolia. Biography. Namir was born in Hadera during the Mandate era. She served as an officer in the IDF during the 1948 Arab-Israeli War, before studying classics and English literature at Hunter College in New York City. She served as secretary of Mapai's parliamentary group and the coalition administration during the second Knesset (1951-55), before becoming secretary to the Israeli delegation at the United Nations. Between 1967 and 1974 she was secretary-general of the Na'amat organisation's Tel Aviv branch. In 1973 Namir was elected to the Knesset on the Alignment's list, and served as chairwoman of the Prime Minister's Committee for the Examination of the Status of Women in Israel from 1975 until 1978. Re-elected in 1977, 1981, 1984, 1988, Namir ran in the Labour Party leadership election in 1992, but came fourth. After retaining her seat in the 1992 elections she was appointed Minister of the Environment in Yitzhak Rabin's government, but was unpopular with staff in the ministry. In December that year she became Minister of Labour and Social Welfare (Rabin had kept the position free in the hope of attracting one of the ultra-orthodox parties to join the coalition), a role she retained when Shimon Peres formed a new government following Rabin's assassination. On 21 May 1996 she resigned from the Knesset and the cabinet to become ambassador to China and non-resident ambassador to Mongolia, role she held until 2000. Namir was married to Mordechai Namir, also a Minister of Labour, 33 years her senior.	AY-wiki22311(c2=90-w2=146):

Jean Arnault. Jean Arnault (born in 1951 in France) currently serves as United Nations Secretary-Generalâ€™s Special Representative for Georgia and Head of the United Nations Observer Mission in Georgia (UNOMIG). He was appointed to the position by UN Secretary-General Kofi Annan in July 2006. He has gained much experience in working for international organizations, especially in the UN peace operations. In 1991, he was Political Adviser to the Special Representative for Western Sahara and Senior Political Affairs Officer in Namibia and Afghanistan. Between January 1994 and December 1996, he served as Observer and then Mediator in the Guatemala peace negotiations. From 1997 to 2000, he was Special Representative for Guatemala. In the following year, he was appointed Representative of the Secretary-General in Burundi. From March 2004 to February 2006, he acted as the Special Representative for Afghanistan and Head of the United Nations Assistance Mission in Afghanistan, where he was also Deputy since March 2002. Arnault studied philosophy and graduated from the University of Sorbonne-Paris I. He holds postgraduate diploma in conference interpretation from the Polytechnic of Central London. In 2001, he was a visiting fellow at the Center for International Studies at Princeton University. He speaks English, French, Russian and Spanish.	AY-wiki88591(c2=90-w2=107):
Mari Simonen. Mari Simonen currently serves as Deputy Executive Director, External Relations, United Nations Affairs and Management of UNFPA, the United Nations Population Fund and focuses on United Nations reforms in particular. Her appointment was approved by UN Secretary-General Kofi Annan in March 2006. Ms. Simonen, of Finland, was most recently the Director of UNFPAâ€™s Technical Support Division, a post she had held since November 1999. In that capacity, she oversaw staff comprised of international technical experts in public health; reproductive health; HIV/AIDS; population studies; gender and human rights; and other specialized areas of work in support of population and development issues worldwide. Prior to that position, Ms. Simonen was the Chief of the Office of the Executive Director at UNFPA, a strategic position from which she helped the Executive Director carry out her functions as the secretary-general of the historic 1994 Cairo International Conference on Population and Development. Before joining the United Nations in 1980, Ms. Simonen worked at the University of California, Berkeley. She holds a doctorate degree from that university in Education. She has a masterâ€™s degree from Stanford University in Sociology of Education as well as a bachelorâ€™s degree in Sociology from the same institution.	AY-wiki47675(c2=90-w2=118):
Knut Eggum Johansen. Knut Eggum Johansen (born 25 September 1945) was a Norwegian civil servant. He was born in BodÃ¸, and graduated as cand.oecon. from the University of Oslo in 1979. He made a career in the Ministry of Finance, being promoted to deputy under-secretary of state in 1990. Since 1999 he serves as director of the Norwegian Competition Authority.	AY-wiki40937(c2=90-w2=29):
Pak Ui-chun. Pak Ui-chun () is a North Korean diplomat and politician. He is the current Minister for Foreign Affairs of the Democratic People's Republic of Korea. Pak began his diplomatic career in 1972, and went on to serve as ambassador of North Korea to Algeria, Syria and Lebanon. From 1989 to 2007, he served as ambassador to Russia, before being appointed Foreign Affairs Minister upon the death in office of his predecessor Paek Nam-sun.	AY-wiki61065(c2=90-w2=37):
Huber, Indiana. Huber is an unincorporated community in Fayette County, Indiana, USA. The community is served by Indiana State Road 1 and is near the airport Mettel Field.	AY-wiki29206(c2=71-w2=13):
Shamshad Ahmad Khan. Shamshad Ahmad Khan () was Pakistan Ambassador to the United Nations and Foreign Secretary of Pakistan.	AY-

	wiki06090(c2=41-w2=10):
Simon Mbatshi Batshia. Simon Mbatshi Batshia (born 24 May, 1949) is a politician from the Democratic Republic of the Congo. He has been the Governor of Kongo Central Province since 24 February, 2007. In 2007, Batshia ordered the opening of the border between DR Congo and the neighboring Republic of the Congo.	AY-wiki04783(c2=51-w2=24):
Funda, Angola. Funda is a town in Angola to the east of the capital, Luanda. Transport. It is served by a station on a branch railway of the Luanda Railway.	AY-wiki16648(c2=80-w2=12):
Scott Petri. Scott Petri is a member of the Pennsylvania House of Representatives from the 178th Legislative District. He currently serves on the House Appropriations, Liquor Control, Local Government and Urban Affairs Committees. Career. Prior to being elected to the House, Petri was a practicing attorney with his own firm, he served as council to Upper Makefield Township and New Britain. He also served on the Upper Makefield Township planning commission and as solicitor to the township. In 2002, Petri defeated Philadelphia sportscaster Carl Cherkin to succeed retiring Rep. Roy Reinard. He has been re-elected to each succeeding session of the House. Personal. Petri graduated from in 1985. He is a graduate of Washington and Jefferson College in Washington, Pennsylvania and the Villanova University School of Law. He also graduated from Downingtown High School. He resides in New Hope, Pennsylvania with his wife and son.	AY-wiki90818(c2=90-w2=72):
Felix Aboagye. Felix Ahmed Aboagye (born December 5, 1975) is a Ghanaian International footballer currently playing for Mumbai FC in the I-League. Trivia. He represented his homeland by the 1998 African Cup of Nations in Burkina Faso and 1996 African Cup of Nations in South Africa. He was member of the Ghana national football team at the 1996 Summer Olympics in Atlanta.	AY-wiki53698(c2=74-w2=31):
Kari SÃ¸rheim. Kari SÃ¸rheim (born 12 October 1948) is a Norwegian politician for the Christian Democratic Party. She served as a deputy representative to the Norwegian Parliament from Hordaland during the terms 1997â€“2001 and 2001â€“2005. On the local level she has been a member of Masfjorden municipal council.	AY-wiki36547(c2=90-w2=22):
Dick Stevenson. Richard R. "Dick" Stevenson'" (born February 11, 1945) is a member of the Pennsylvania House of Representatives, elected in 2000 to represent the 8th District. In the current legislative session, Stevenson serves on the House Appropriations, Judiciary and Professional Licensure Committees. Career. Stevenson served for eight years on the borough council of Grove City, Pennsylvania from 1985-1993, including five years as the council president. In 1996, Stevenson joined the Mercer County Board of Commissioners and was elected Chairman. Stevenson was first elected to the House in 2000 to replace Howard Fargo. That year, he defeated the Armstrong County district attorney, George Kepple, in the Republican primary election with 55% of the vote. In the general election, Stevenson defeated James Coulter, taking over 63%. Stevenson has won re-election to each succeeding session of the House. Since 2004, he has run unopposed in the primary and general elections. Personal. Stevenson served in the United States Air Force from 1968 to 1972. He served as Korean Language Specialist with the USAF Security Service. Stevenson received a Bachelor of Arts degree from Saint Francis College in New York and a Master of Business Ad-	AY-wiki38975(c2=90-w2=119):

ministration from Suffolk University in Massachusetts. He and his wife have two children, Sarah Hatfield and Emily Vallozzi, and three grandchildren.	
Asabot. Asabot is a town in eastern Ethiopia. It is located in the Mirab Hararghe Zone of the Oromia Region. Transport. It is served by a railway station on the Addis Ababa - Djibouti Railway.	AY-wiki32856(c2=62-w2=13):
Afelee F. Pita. Afelee F. Pita, born February 11, 1958, is a Tuvaluan diplomat. He is currently Tuvalu's Permanent Representative to the United Nations. Pita holds a Master's degree in public administration from the University of Canberra and a Bachelor of Arts degree in administration and accounting from the University of the South Pacific. He began his career as a senior official in government administration as assistant Secretary, and then Secretary, at the Tuvaluan Ministry of Commerce and Natural Resources, from 1987 to 1988. He was Assistant Secretary for Commerce from 1989 to 1993, then Acting Secretary at the Ministry of Trade, Commerce and Public Corporations in 1993. From 1994 to 1994, he served as Permanent Secretary in several successive ministries (Health and Sports, Labour and Communication, Resources and Environment, Finance). From 2001 to 2004, Pita was Adviser to the Executive Director of the Asian Development Bank in Manila, where he served as representative for Australia, Azerbaijan, Cambodia, Hong Kong, Kiribati, the Federated States of Micronesia, Nauru, the Solomon Islands and Tuvalu. Returning to Tuvalu, Pita served as Permanent Secretary to the Ministry of Natural Resources and Lands from 2004 and 2006, before being appointed as Permanent Representative to the United Nations. In April 2007, Pita addressed the Special Session of the United Nations Security Council on Energy, Climate and Security, and "beseech[ed] the Security Council to act urgently to address the threats to [Tuvalu]'s national security" - namely, climate change.	AY-wiki56708(c2=90-w2=139):
Quixinge, Angola. Quixinge is a town in Angola Transport. It is served by an extension of a branch railway of the northern railway.	AY-wiki61613(c2=70-w2=10):
Johan Hambro. Johan Randulf Bull Hambro'" (1915 - 1993) was a Norwegian journalist. He was the son of Norwegian politician C. J. Hambro, and brother of Carl and Edvard Hambro. From 1940 to 1945, during the German occupation of Norway, he was employed at the Norwegian consulate-general in New York City. After the war he worked as a journalist for "Aftenposten" from 1946 to 1948 and the Norwegian News Agency from 1949 to 1953. He was then press attachÃ© to the United Nations for two years. From 1955 to 1982 he was the secretary general of Nordmanns-Forbundet, a cultural association for Norwegian-Americans.	AY-wiki40850(c2=90-w2=52):
Luinha, Angola. Luinha is a town in northern Angola Transport. It is served by a station on the Luanda Railway. There is a junction to a branchline to the south.	AY-wiki24422(c2=56-w2=10):
George Alleyne. Sir George Alleyne'" (born in St. Philip, Barbados, on 7 October 1932) currently serves as United Nations Secretary-General's Special Envoy for AIDS in the Caribbean region. He was appointed to the position by UN Secretary-General Kofi Annan in February 2003. Alleyne studied medicine at the University of the West Indies (UWI) and graduated as the the gold medallist in 1957. Subsequently, he pursued his postgraduate training in internal medicine in the	AY-wiki49472(c2=90-w2=137):

United Kingdom and the United States. In 1972, he became Professor of Medicine at the UWI. In 1976, he was appointed Chairman of the Department of Medicine. In october 2003 George Alleyne was appointed the Chancellor of the University of West Indies. Besides his academic experience, Alleyne also gained much experience in working for international organizations. In February 1995, he became Director of the Pan American Health Organization (PAHO), Regional Office of the World Health Organization (WHO). He served two four-year terms in this position until the end of 2002 and was elected Director Emeritus. As Special Envoy for AIDS in the Caribbean region, he is responsible for ensuring follow-up to the United Nations General Assembly special session on HIV/AIDS and the Pan-Caribbean Partnership against HIV/AIDS, in the Caribbean region. Occasionally, he also represents the UN Secretary-General at events related to HIV/AIDS in the Caribbean region. He was made Knight Bachelor by Queen Elizabeth II in 1990, and awarded the Order of the Caribbean Community in 2001. In October 2008 Alleyne received the "Science of Peace Award" from the Inter American Heart Foundation.	
B'rov am hadrat melech. "B'rov am hadrat melech" (×'×¨×' ×¢×□ ×"×"×¨×ª ×ž×œ×š, "in multitudes there is glorification of the king") is a principle in Jewish law that recommends that mitzvot be performed as part of as large a gathering as possible, with the intention of providing greater honor to God. "Talmudic" examples of application. The "Talmud" provides many examples of the practical application of this principle. One such example is brought by a "Tosefta", which quotes a situation in which many individuals were gathered together and learning in a study hall when a candle arrived for use in the "havdalah" prayer that is receited at the end of "Shabbat". In such a case, either each individual could recite his own blessing on the fire, or one person can recite the blessing and all of the others can listen and respond "amen", thereby fulfilling their obligation to recite the blessing. Whereas the Academy of Shammai proposed that each person recite their own blessing, the Academy of Hillel proposed that one person should recite the blessing on behalf of everyone present in fulfillment of the principle of "b'rov am hadrat melech". The law follows the latter opinion. Another example is in reference to blowing the "shofar". The "Mishna" mandates that the "shofar" be blown during the "musaf" prayer service, and the "Gemara", ostensibly providing an explanation to why the "shofar" is not blown in the earlier "shacharit" prayer, provides the rationale that inclusion within the "musaf" prayer is because of the principle of "b'rov am hadrat melech", as more people are in the synagogue by the time the congregation has reached "musaf". This rationale is immediately debunked, as the "Gemara" continues to ask why "Hallel" (when recited) is included in "shacharit" if "b'rov am hadrat melech" is indeed governing into which prayers the additions are added.	AY-wiki29756 (c2=90-w2=74)
Soundmap. A soundmap'" is a form of locative media that links a place and its sonic representations. It is an example of the personalized map content described alternately as web mapping and neogeography. Soundmaps convey the soundscape of a place, often by organizing multiple soundmarks or "community sound[s] which is unique, or possesses qualities which make it specially regarded or noticed by the people in that community" in a web-based map. Our sense of hearing, which has until recently been underappreciated as a means of representing data, can be used to expand the representational repertoire of cartographic design...Sound, in other words, provides us with more choices for representing data and phenomena and thus more ways in which to explore and understand the complex physical and human worlds we inhabit. â€"'"John Krygier'", "Making Maps with Sound"	AY-wiki52756(c2=90-w2=74)

Appendix B ----------Data supporting the illustrative example in chapter 6

The documents below appear in the order presented in the illustrative example from chapter 6.

Content	Document name
Bo Asplund. Bo Asplund currently serves as United Nations Secretary-General's Deputy Special Representative for Afghanistan, the United Nations Resident Coordinator and the Humanitarian Coordinator in Afghanistan. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in August 2007. Asplund obtained his master's degree in international economics from Columbia University's School of International and Public Affairs and master's and bachelor's degrees in economics, political science and statistics from the University of Lund (Sweden). He also holds a Certificat d'Etudes Politiques from the Institut d'Etudes Politiques in Paris. In the beginning of his career, he worked for the Ministry for Foreign Affairs of the Swedish Government in Stockholm. He was also posted to Chile and the Swedish Mission to the United Nations. Besides his diplomatic experience, Asplund has worked in various capacities for international organizations. He has served as Deputy Assistant Administrator of UNDP's Regional Bureau for Arab States at the Organization's Headquarters in New York, United Nations Resident Coordinator and UNDP Resident Representative in Algeria, and UNDP Senior Deputy Resident Representative in the Sudan. From 2001 to 2007, he was the United Nations Resident and Humanitarian Coordinator and the United Nations Development Programme (UNDP) Resident Representative in Indonesia.	AY-wiki88537.html
José Victor da Silva Angelo. José Victor da Silva Angelo (born in 1949) currently serves as the Special Representative and Head of the United Nations Mission in Central African Republic and Chad (MINURCAT). He was appointed to the position by UN Secretary-General Ban Ki-moon in January 2008. Angelo obtained a master's degree in sociology from Instituto Superior Economico e Social of the University of Evora, Portugal and studied for a Doctor of Philosophy in sociology at l'Université Libre de Bruxelles, Belgium. He started his career as a university lecturer and served as Senior Statistician in the Portuguese National Institute of Statistics (INE). He was a member of the Electoral Commission of Portugal. Later on, he joined the United Nations, where he served in various capacities, including United Nations Development Programme (UNDP) Special Envoy for East Timor and Asia, Deputy Regional Director for Africa at UNDP in New York, Resident Coordinator/Resident Representative in the United Republic of Tanzania and the Gambia, and Deputy Resident Representative in the Central African Republic. He also served as United Nations Population Fund (UNFPA) Representative in Mozambique and United Nations Adviser in Sao Tome and Principe. His extensive experience brought him to serve on more senior positions at the United Nations. From 2000	AY-wiki17349.html

to 2004, he served as UNDP Resident Representative in Zimbabwe. From 2005 to 2007, he was the Executive Representative of the Secretary-General for Sierra Leone, as well as the Resident Coordinator of the United Nations System in Freetown.	
Ali'ioaiga Feturi Elisaia. Ali'ioaiga Feturi Elisaia, born in 1954, is a Samoan diplomat. He is currently Samoa's Permanent Representative to the United Nations. He obtained a postgraduate certificate in diplomacy from Oxford University, and also holds a Bachelor of Arts degree in political science and administration from the University of the South Pacific. Elisaia first served in the Samoan Mission to the United Nations in 1979. From 1979 to 1981, he was Acting Division Head, and then as Division Head, at the Economic and Aid Division of the Samoan Ministry of Foreign Affairs. From 1981 to 1984, he served as First Secretary at Samoa's High Commission in New Zealand. He was Deputy Secretary for Foreign Affairs from 1984 to 1988, then co-director of the Hanns Seidel Foundation in Samoa from 1988 to 2001. From 2001 to 2003, he was Chief Executive Officer at the Samoan Ministry of Foreign Affairs. Elisaia was appointed Permanent Representative of Samoa to the United Nations in 2003.	AY-wiki56672.html
Wolfgang Stöckl. Wolfgang Stöckl (born on 11 June 1948) currently serves as the Vice Chairman of the United Nations International Civil Service Commission. Before becoming the Vice Chairman of ICSC, Stöckl was Ambassador and Special Coordinator for German Personnel in International Organizations. From 2000 to 2002, he served as Director of Economic and Development Affairs in the United Nations and Global Affairs Department of the Foreign Office of the Federal Republic of Germany. From 1997 to 2000, he was First Counsellor of the Permanent Mission of the Federal Republic of Germany to the Organization for Economic Cooperation and Development in Paris, in charge of the Public Management Committee and human resources management issues. He was a member of the ICSC from 1997 to 2002. In 1997, he served as Chairman of the Committee for Programme and Coordination of the United Nations. Between 1995 and 1997, he was member of the UN Advisory Committee onwiki58368.html Administrative and Budgetary Questions(ACABQ). From 1991 to 1997, he was Counsellor of the Permanent Mission of the Federal Republic of Germany to the United Nations in New York. In this capacity, he was in charge of the UN reform, common system and human resources management issues. Earlier in his career, he served on a variety of capacities in the German foreign service, including as Special Adviser for Management and Personnel Questions to the Minister for Foreign Affairs of the German Democratic Republic in 1990, as Deputy Head of the Organization and Management Division of the Foreign Office from 1989 to 1991, as Head of the Headquarters Inspection Unit of the Foreign Office, Director of the German-Saudi Arabian Liaison Office for Economic Affairs in Riyadh, Saudi Arabia, as German Counsul in Cairo, Egypt and as Assistant Director in the Training Centre of the Federal Foreign Service. He joinedwiki58368.html the German Foreign Office in 1977. Previously, he served the Ministry of Interior in Land Hesse, Germany, in 1976. He was Assistant judge and assistant prosecutor in the Ministry of Justice in Land Hesse from 1973 to 1974. He studied	AY-wiki01800.html

in the Training Centre of the Federal Foreign Office in Bonn from 1977 to 1979. He holds a Postgraduate Degree in Public Administration at the Postgraduate School of Public Administration in Speyer, German. He passed the Second State Examination in law(Bar Examination) in 1975 and the First State Examination in law(Masters Degree) in 1972. From 1967 to 1971, he studied law at the University of Marburg, Germany.	
Catherine Bragg. Catherine Bragg (born in 1953 in Hong Kong) currently serves as Deputy Emergency Relief Coordinator in the Office for the Coordination of Humanitarian Affairs. She was appointed to this position by United Nations Secretary-General Ban Ki-moon in December 2007. Bragg obtained a PhD in Criminal Justice from the University at Albany, SUNY, a Master of Philosophy in Criminology from the University of Cambridge and a Bachelor of Science in Psychology from the University of Toronto. Throughout her career, she has served in various capacities in the Federal Public Service in the Government of Canada. In the Privy Council Office, she formulated policy advice to the Prime Minister and the Cabinet. In the Department of National Defence, she worked on human resource issues. In the Department of Justice, she focused on evaluation and strategic planning. Prior to joining the United Nations, Bragg served as the Director-General of the Humanitarian Assistance, Peace and Security Programme in the Canadian International Development Agency (CIDA) since 2004. She is the Chair of the Office for the Coordination of Humanitarian Affairs Donor Support Group and a member of the Advisory Group of the Central Emergency Response Fund.	AY-wiki37929.html
Alan Doss. Alan Doss (born on January 7, 1945) currently serves as United Nations Special Representative of the Secretary-General for the Democratic Republic of the Congo. He was appointed to the position by United Nations Secretary-General Ban Ki-moon. Doss has served the United Nations in a variety of capacities. His earlier work within the Organization include assignments in China, Kenya, Niger, Democratic Republic of the Congo (formerly Zaire) and Benin. Later on, he acted as United Nations Resident Coordinator and UNDP Regional Representative in Bangkok. In the meantime, he was Director of the United Nations Border Relief Operation, in charge of United Nations assistance to displaced Cambodians on the Thai-Cambodiawiki58368.html border. In addition, he also served as Director of the United Nations Development Group (UNDG). Before working for the UNDG, he was Director of the UNDP European Office in Geneva, in charge of strengthening the UNDP's outreach and liaison work in Western Europe. His recent appointments include as Principal Deputy Special Representative of the Secretary-General for Côte d'Ivoire in June 2004. Previously, he was the Deputy Special Representative of the Secretary-General for Sierra Leone, while concurrently serving as United Nations Resident Coordinator, Humanitarian Coordinator and United Nations Development Programme (UNDP) Resident Representative. Doss was born in Cardiff, Wales in the United Kingdom. He graduated from the London School of Economics.	AY-wiki04013.html
Choi Young-jin. Choi Young-jin (born on 29 March 1948 in Seoul, Republic of Korea) currently serves as United Nations Special Repre-	AY-wiki02748.html

sentative for Côte d'Ivoire. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in October 2007. A career diplomat, Choi has served in various capacities in the foreign service of the Republic of Korea. From 2005 to 2007, he was the Permanent Representative of the Republic of Korea to the United Nations. In 2004, he was the Vice Minister for Foreign Affairs and Trade. In 2003, he served as Chancellor of Institute of Foreign Affairs and National Security (IFANS). Previously, he was Ambassador to Austria and Slovenia, and Permanent Representative to all international organizations in Vienna, Austria. Between 2000 and 2001, he served as Deputy Minister for Policy Planning and International Organizations, in charge of foreign policy planning, North Korean affairs, the United Nations system, disarmament and non-proliferation, and democracy and human rights. From 1998 to 1999, Choi was Assistant Secretary-General for Peacekeeping Operations at the United Nations, responsible for overseeing planning and support for 17 peacekeeping operations, including those in Kosovo, Timor-Leste, Sierra Leone and the Democratic Republic of the Congo. In addition, he served on several other duties at the Ministry of Foreign Affairs, including as Director-General of the International Economic Affairs Bureau from 1994 to 1995, as First Senior Coordinator in the Ministry's Office of Policy Planning from 1991 to 1993, and as an Economic Counsellor at the Korean Embassy in Washington, D.C., from 1988 to 1990. He obtained his master's and doctorate degrees in international relations from the University of Paris I (Panthéon-Sorbonne), respectively, in 1980 and 1985 and his bachelor's degree with distinction from the Department of International Relations, Yonsei University in 1973. In 2007, he was a resident diplomat scholar at the Fletcher School of Law, Tufts University, Boston, Massachusetts. He speaks English and French fluently and knows 3,000 characters in Chinese.	
Jorge Urbina. Jorge Urbina, born on 2 May 1946, is the Permanent Representative to the United Nations for Costa Rica. He assumed the position in October 2006. He is married with two children. Education. Urbina received a master in law degree from the University of Costa Rica and a doctorate in law from the University of Bordeaux. Career. From 1982 to 1984, Urbina was the Deputy Permanent Representative to the United Nations for Costa Rica. After this appointment, he was Vice Minister for Foreign Affairs for two years from 1984 to 1986. In 1986, Urban then moved to the position of the Executive President of the National Institute for Municipal Counselling and Promotion until 1990; he also served as Costa Rica's Minister of Information from 1989 to 1990. From 1990 to 1993, Urban was an Associate Researcher at Centro de Investigaciones Económicas y Sociales in Montes de Oca, Costa Rica, and also was a professor at the International Affairs School of Universidad Nacional in Heredia, Costa Rica, for the same period. After that, he served as a permanent consultant at the Programme for Democratic Governance in Central America, United Nations Development from 1993 to 1998. Until he was appointed Permanent Representative to the United Nations, Mr. Urbina was a Programme Coordinator at the International Centre for Human Development, San José, Costa Rica, from 1998 to 2006.	AY-wiki40347.html

Jean Arnault. Jean Arnault (born in 1951 in France) currently serves as United Nations Secretary-General's Special Representative for Georgia and Head of the United Nations Observer Mission in Georgia (UNOMIG). He was appointed to the position by UN Secretary-General Kofi Annan in July 2006. He has gained much experience in working for international organizations, especially in the UN peace operations. In 1991, he was Political Adviser to the Special Representative for Western Sahara and Senior Political Affairs Officer in Namibia and Afghanistan. Between January 1994 and December 1996, he served as Observer and then Mediator in the Guatemala peace negotiations. From 1997 to 2000, he was Special Representative for Guatemala. In the following year, he was appointed Representative of the Secretary-General in Burundi. From March 2004 to February 2006, he acted as the Special Representative for Afghanistan and Head of the United Nations Assistance Mission in Afghanistan, where he was also Deputy since March 2002. Arnault studied philosophy and graduated from the University of Sorbonne-Paris I. He holds postgraduate diploma in conference interpretation from the Polytechnic of Central London. In 2001, he was a visiting fellow at the Center for International Studies at Princeton University. He speaks English, French, Russian and Spanish.	AY-wiki88591.html
Hans Engen. Hans Engen"' (1912 - 1966) was a Norwegian journalist, diplomat and politician for the Labour Party. He was born in Ringebu. During the German occupation of Norway from 1940 to 1945, he was a coordinator of the cooperation with the Norwegian government-in-exile and the Norwegian resistance movement. From 1946 to 1949 he worked as the foreign affairs editor of newspaper "Verdens Gang". From 1951 to 1952 he worked as Norway's counsellor of embassy to the United Nations; he was then ambassador to the UN to 1958. From 1958 to 1963, under the third cabinet Gerhardsen, Engen served as state secretary in the Ministry of Foreign Affairs. Finally, he was Norwegian ambassador to the United States from 1963 to 1966.	AY-wiki46929.html
Alan Doss. Alan Doss (born on January 7, 1945) currently serves as United Nations Special Representative of the Secretary-General for the Democratic Republic of the Congo. He was appointed to the position by United Nations Secretary-General Ban Ki-moon. Doss has served the United Nations in a variety of capacities. His earlier work within the Organization include assignments in China, Kenya, Niger, Democratic Republic of the Congo (formerly Zaire) and Benin. Later on, he acted as United Nations Resident Coordinator and UNDP Regional Representative in Bangkok. In the meantime, he was Director of the United Nations Border Relief Operation, in charge of United Nations assistance to displaced Cambodians on the Thai-Cambodia border. In addition, he also served as Director of the United Nations Development Group (UNDG). Before working for the UNDG, he was Director of the UNDP European Office in Geneva, in charge of strengthening the UNDP's outreach and liaison work in Western Europe. His recent appointments include as Principal Deputy Special Representative of the Secretary-General for Côte d'Ivoire in June 2004. Previously, he was the Deputy Special Representative of the Secretary-General for Sierra Leone, while concurrently serving as United Nations Resident Coordinator, Humanitarian Coordinator and United Nations Development	AY-wiki04013.html

Programme (UNDP) Resident Representative. Doss was born in Cardiff, Wales in the United Kingdom. He graduated from the London School of Economics.	
Choi Young-jin. Choi Young-jin (born on 29 March 1948 in Seoul, Republic of Korea) currently serves as United Nations Special Representative for Côte d'Ivoire. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in October 2007. A career diplomat, Choi has served in various capacities in the foreign service of the Republic of Korea. From 2005 to 2007, he was the Permanent Representative of the Republic of Korea to the United Nations. In 2004, he was the Vice Minister for Foreign Affairs and Trade. In 2003, he served as Chancellor of Institute of Foreign Affairs and National Security (IFANS). Previously, he was Ambassador to Austria and Slovenia, and Permanent Representative to all international organizations in Vienna, Austria. Between 2000 and 2001, he served as Deputy Minister for Policy Planning and International Organizations, in charge of foreign policy planning, North Korean affairs, the United Nations system, disarmament and non-proliferation, and democracy and human rights. From 1998 to 1999, Choi was Assistant Secretary-General for Peacekeeping Operations at the United Nations, responsible for overseeing planning and support for 17 peacekeeping operations, including those in Kosovo, Timor-Leste, Sierra Leone and the Democratic Republic of the Congo. In addition, he served on several other duties at the Ministry of Foreign Affairs, including as Director-General of the International Economic Affairs Bureau from 1994 to 1995, as First Senior Coordinator in the Ministry's Office of Policy Planning from 1991 to 1993, and as an Economic Counsellor at the Korean Embassy in Washington, D.C., from 1988 to 1990. He obtained his master's and doctorate degrees in international relations from the University of Paris I (Panthéon-Sorbonne), respectively, in 1980 and 1985 and his bachelor's degree with distinction from the Department of International Relations, Yonsei University in 1973. In 2007, he was a resident diplomat scholar at the Fletcher School of Law, Tufts University, Boston, Massachusetts. He speaks English and French fluently and knows 3,000 characters in Chinese.	AY-wiki02748.html
Jorge Urbina. Jorge Urbina, born on 2 May 1946, is the Permanent Representative to the United Nations for Costa Rica. He assumed the position in October 2006. He is married with two children. Education. Urbina received a master in law degree from the University of Costa Rica and a doctorate in law from the University of Bordeaux. Career. From 1982 to 1984, Urbina was the Deputy Permanent Representative to the United Nations for Costa Rica. After this appointment, he was Vice Minister for Foreign Affairs for two years from 1984 to 1986. In 1986, Urban then moved to the position of the Executive President of the National Institute for Municipal Counselling and Promotion until 1990; he also served as Costa Rica's Minister of Information from 1989 to 1990. From 1990 to 1993, Urban was an Associate Researcher at Centro de Investigaciones Económicas y Sociales in Montes de Oca, Costa Rica, and also was a professor at the International Affairs School of Universidad Nacional in Heredia, Costa Rica, for the same period. After that, he served as a permanent consultant at the Programme for	AY-wiki40347.html

Democratic Governance in Central America, United Nations Development from 1993 to 1998. Until he was appointed Permanent Representative to the United Nations, Mr. Urbina was a Programme Coordinator at the International Centre for Human Development, San José, Costa Rica, from 1998 to 2006.	
Jean Arnault. Jean Arnault (born in 1951 in France) currently serves as United Nations Secretary-General's Special Representative for Georgia and Head of the United Nations Observer Mission in Georgia (UNOMIG). He was appointed to the position by UN Secretary-General Kofi Annan in July 2006. He has gained much experience in working for international organizations, especially in the UN peace operations. In 1991, he was Political Adviser to the Special Representative for Western Sahara and Senior Political Affairs Officer in Namibia and Afghanistan. Between January 1994 and December 1996, he served as Observer and then Mediator in the Guatemala peace negotiations. From 1997 to 2000, he was Special Representative for Guatemala. In the following year, he was appointed Representative of the Secretary-General in Burundi. From March 2004 to February 2006, he acted as the Special Representative for Afghanistan and Head of the United Nations Assistance Mission in Afghanistan, where he was also Deputy since March 2002. Arnault studied philosophy and graduated from the University of Sorbonne-Paris I. He holds postgraduate diploma in conference interpretation from the Polytechnic of Central London. In 2001, he was a visiting fellow at the Center for International Studies at Princeton University. He speaks English, French, Russian and Spanish.	AY-wiki88591.html
Hans Engen. Hans Engen"' (1912 - 1966) was a Norwegian journalist, diplomat and politician for the Labour Party. He was born in Ringebu. During the German occupation of Norway from 1940 to 1945, he was a coordinator of the cooperation with the Norwegian government-in-exile and the Norwegian resistance movement. From 1946 to 1949 he worked as the foreign affairs editor of newspaper "Verdens Gang". From 1951 to 1952 he worked as Norway's counsellor of embassy to the United Nations; he was then ambassador to the UN to 1958. From 1958 to 1963, under the third cabinet Gerhardsen, Engen served as state secretary in the Ministry of Foreign Affairs. Finally, he was Norwegian ambassador to the United States from 1963 to 1966.	AY-wiki46929.html
Inga Björk-Klevby. Inga Björk-Klevby currently serves as United Nations Deputy Executive Director of the United Nations Human Settlements Programme (UN-HABITAT). She was appointed to this position by United Nations Secretary-General Kofi Annan in October 2005. She obtained her master's degree from the Stockholm School of Economics. Björk-Klevby has been a diplomat for many years. She has served as served as the Ambassador of Sweden to Kenya, Rwanda, Seychelles and the Comoros and as Permanent Representative to the United Nations Environment Programme (UNEP) and UN-HABITAT. Later on, she became the Assistant Undersecretary of the Ministry of Foreign Affairs of Sweden who is responsible for international development cooperation policies, programmes and budget. For over 20 years, she worked in international finance and senior management capacities with the Central Bank of Sweden, the International Monetary Fund (IMF), the World Bank, and the Asian Development Bank. She became Execu-	AY-wiki88514.html

tive Director for the African Development Bank where she represented Nordic countries, Switzerland and India. Prior to her appointment as Deputy Executive Director of the UN-HABITAT, she was Ambassador of Sweden to Côte d'Ivoire, Burkina Faso, Guinea, Liberia and Sierra Leone.	
Jun Yamazaki. Jun Yamazaki (born in 1956, in London) is a Japanese diplomat who currently serves as United Nations Assistant Secretary-General at the Office of Programme Planning, Budgets and Accounts, and Controller. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in August 2008. Yamazaki worked the Japanese Ministry of Foreign Affairs for a number of years. He was Counsellor in the Embassy of Japan in Indonesia. He also acted as Deputy Director in the United Nations Administration Division of the Multilateral Cooperation Department of the Ministry of Foreign Affairs. In this capacity, he was in charge of administrative and budgetary affairs of the United Nations. He later became Director of the International Peace Cooperation Division and Deputy Director-General for Global Issues in the International Cooperation Bureau of the Ministry of Foreign Affairs in Japan and served until August 2008. From 2003 to 2007, he was a member of the United Nations Advisory Committee on Administrative and Budgetary Questions (ACABQ). Yamazaki obtained his Bachelor of Arts in international relations from the University of Tokyo.	AY-wiki88603.html
Omar Abdi. Omar Abdi currently serves as Deputy Executive Director of the United Nations Children's Fund (UNICEF). He was appointed to the position by United Nations Secretary-General Ban Ki-moon in July 2007. Abdi obtained a bachelor's degree in civil engineering in 1982, a master's degree in regional planning in1988, and a doctorate in development economics in 1991. Before joining UNICEF, he served in the Government of Somalia as Senior Planning Officer in the National Refugee Commission from 1982 to 1986 and Director of the Planning Department from 1987 to 1989. In 1992, he became a Programme Officer in Liberia and was transferred to New York in 1994. From September 1996 to June 1998, he served as UNICEF Representative in Liberia. Subsequently, he became UNICEF Representative in Ghana in June 1998. Since June 2000, he served as Deputy Director of the Programme Division at United Nations Headquarters until 2003. In the subsequent years, Abdi held various positions within UNICEF, including as UNICEF Representative in Islamabad, Pakistan and as UNICEF Director of the Middle East and North Africa Region. Abdi was born in Somalia and is currently a Canadian citizen.	AY-wiki37567.html
Konrad Osterwalder. Konrad Osterwalder currently serves as the Rector of the United Nations University (UNU). He was appointed to the position by United Nations Secretary-General Ban Ki-moon in May 2007. He succeeded Prof. Hans van Ginkel from the Netherlands to be the fifth Rector of the United Nations University. A Swiss physicist, Prof. Osterwalder served as the Rector of ETH-Zürich since 1995. His research focused on the mathematical structure of relativistic quantum field theory, elementary particle physics and statistical mechanics. He obtained his doctorate degree in theoretical physics at ETH in 1970 and was appointed as full professor at ETH Zürich in 1977. He had held	AY-wiki33109.html

positions at New York University, Harvard University and the Alfred P. Sloan Foundation and had been a visiting fellow in many universities around the world.	
Roble Olhaye. Roble Olhaye (born 24 April 1944) is the Permanent Representative to the United Nations for the Republic of Djibouti. He has been Permanent Representative to the United Nations and Ambassador of Djibouti to the United States since 1988. In 1989, Olhaye was appointed as non-resident Ambassador to Canada. He is married and has five children. Education. Olhaye Intermediate Diploma in Commerce from the Addis Ababa Commercial College. After studying in the field of accounting, finance, taxation, law and management, he received qualification as a professional accountant in England. He is a Fellow of the Association of International Accountants (UK) and a member of the British Institute of Management. Career. After being appointed as the Permanent Representative, Olhaye has represented Djibouti in the Security Council, served as President of the Security Council and Chairman of the Sanctions Committee established by United Nations Security Council Resolution 841 on Haiti. While a member of the Security Council Mission to Mozambique, he helped in the process of democratic elections held in 1994. As Dean of the African diplomatic corps in Washington D.C., he received one of Djibouti's highest medals of honor awards for improving how the nation was viewed internationally Prior to his appoint to the Permanent Representative post, he served as Djibouti's Permanent Representative to the United Nations Environment Programme (UNEP) and United Nations Centre for Human Settlements (Habitat). He also served as Honorary Consul of Djibouti to Kenya and established ongoing diplomatic relations between the two nations. He has also worked in the private section with commerce, finance, and management.	AY-wiki69411.html
Minas Hadjimichael. Minas Hadjimichael (born in 1956) is the Permanent Representative to the United Nations for Cyprus. He presented his credentials to UN Secretary-General Ban Ki-moon on 25 August 2008. Education. Hadjimichael holds a bachelor of laws degree from the University of Athens and a masters of arts in political science and international relations, which he received at Georgia Southern University in the United States. He has also received instruction in European Union concerns from the Civil Service College of London, and was a participant in a program hosted by the United States Information Agency (USIA) on the United States Federal Government System. In addition to his academic credentials, Hadjimichael also speaks Greek, English and French. Career. Hadmichael was Director of the Cyprus Question and European Union-Turkey Affairs Division of the Cyprus Ministry of Foreign Affairs, and served as the Ministry's Acting Permanent Secretary, prior to his taking office at the United Nations. His long diplomatic service includes postings as Cyprus' Ambassador to France, Tunisia, Andorra, and Algeria. He has served in Cyprus' European Union Division as Deputy Director. He has also served as Deputy Chief of Mission at the Cyprus Embassy in Athens, Greece and Director of the Cypriot Foreign Minister's Cabinet.	AY-wiki63598.html
Akiko Yuge. Akiko Yuge currently serves as Assistant Administrator and Director of the Bureau of Management at the United Nations De-	AY-wiki88595.html

velopment Programme. She was appointed to this position by United Nations Secretary-General Kofi Annan in August 2006. Yuge began her career at UNDP as a Programme Officer in Thailand in 1976. She later served as Programme Officer in the Regional Bureau for Asia and the Pacific at UNDP Headquarters in New York and as Area Officer for China and the Philippines. From 1984 to 1987, she took special leave from UNDP and worked as a freelance development consultant and as Project Officer for the Japanese Engineering Consulting Firms Association (ECFA). In 1988, she rejoined UNDP and served in Thailand as Assistant Resident Representative until 1990. From 1990 to 1994, she was Deputy Resident Representative in Indonesia. From 1994 to 1998, she was UN Resident Coordinator and UNDP Resident Representative in Bhutan. She also taught Development Studies as a Professor at Ferris University in Yokohama in the meantime. Between 2001 and 2002, Yuge was a member of the Second Consultative Committee on ODA Reform, an advisory group to the Foreign Minister of Japan. She was also a member of the Advisory Group of International Cooperation for Peace organized by the country's Chief Cabinet Secretary in 2002 and a member of the Eminent Persons' Group on UN Reform, an advisory group to the Foreign Minister. Since 2002, she served as Director of UNDP's Tokyo Liaison Office. Yuge obtained her Bachelor's Degree in Psychology from Barnard College at Columbia University in New York and her Master's Degree in Development Economics from New York University.	
Chikadibia Isaac Obiakor. Lieutenant General Chikadibia Obiakor (born on 18 February 1951 in Nigeria) current serves as United Nations Military Adviser for Peacekeeping Operations. Previously, he was Force Commander of the United Nations Mission in Liberia (UNMIL), a position he held was appointed to by UN Secretary-General Kofi Annan in January 2006. Lieutenant General Obiakor started his military career by joining the Nigerian Army in 1973. He served the army in various capacities, including as the Commander of the Economic Community of West African States Monitoring Group (ECOMOG) Artillery Brigade in Liberia in 1996 and 1997, as the ECOMOG Chief Coordinator of the Liberian elections in July 1997, and as the General Officer Commanding, Second Mechanized Division of the Nigerian Army. He also held the position of Chief of Administration of the Nigerian Army, in charge of the welfare, discipline and medical services for all Nigerian military personnel. Lieutenant General Obiakor is graduated from the National War College in Abuja. He obtained a Master of Science degree in strategic studies from the University of Ibadan in Nigeria and has participated in numerous international military courses.	AY-wiki47225.html
David Tolbert. David Tolbert currently serves as United Nations Assistant Secretary-General, Expert on the United Nations Assistance to the Khmer Rouge Trials (UNAKRT). Previously, he was Deputy Prosecutor of the International Criminal Tribunal for the former Yugoslavia (ICTY). He was appointed to this position by UN Secretary-General Kofi Annan in August 2004. Before becoming the Deputy Prosecutor of ICTY, he was Deputy Registrar in the same institution. He has worked in the area of international law for many years. He has served as the Executive Director of the American Bar Association's Central European	AY-wiki88619.html

and Eurasian Law Initiative (ABA CEELI), an institution that manages rule of law development programmes throughout Eastern Europe and the former Soviet Union. Prior to that, he worked for four years as Chef de Cabinet to former President Gabrielle Kirk McDonald and as the Senior Legal Adviser of Registry at the ICTY. He also serves as Chief of General Legal Division of the United Nations Relief and Works Agency (UNRWA) in Vienna, Austria and Gaza. Tolbert obtained his B.A. magna cum laude from Furman University, his J.D. from the University of North Carolina and his LL. M. with distinction from the University of Nottingham. He has published widely regarding international criminal justice, the ICTY and the International Criminal Court (ICC) and has represented the ICTY in the discussions leading up to the creation of the ICC. He has also taught international law and human rights at the post-graduate level in the United Kingdom and practiced law for many years in the United States.	
Thomas Stelzer. Thomas Stelzer (born June 19, 1955) currently serves as United Nations Assistant Secretary-General for Policy Coordination and Inter-Agency Affairs, Department of Economic and Social Affairs. He was appointed to the position by UN Secretary-General Ban Ki-moon in February 2008. He holds a doctorate in law from Vienna University, a Master of Arts in Latin American Studies from Stanford University, and a diploma in International Relations from the Johns Hopkins University, School of Advanced International Studies, Bologna Center. Stelzer served in a variety of diplomatic and international positions in his early career. He was Deputy Director of the Austrian Cultural Institute in New York, Special Assistant to the Executive Secretary of the CTBTO Preparatory Commission, and Minister-Counsellor at the Permanent Mission of Austria to the United Nations in New York. He also serves as the Austrian Representative to the governing bodies of the United Nations Development Programme (UNDP), the United Nations Children's Fund (UNICEF) and the United Nations Population Fund (UNFPA) and Delegate to the Committee for Disarmament and International Security (First Committee) of the General Assembly. Ambassador Stelzer has been serving since August 2001 as Permanent Representative of Austria to the United Nations (Vienna), United Nations Industrial Development Organization (UNIDO), International Atomic Energy Agency (IAEA), and the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) Preparatory Commission. From 2002 to 2007, he was Facilitator and Chair of the Vienna Terrorism Symposiums. In 2003, he was Chair of the CTBTO Preparatory Commission. Between 2005 and 2006, he serves as President of the UNIDO Industrial Development Board. Most recently, he served as a Vice-Chair of the Second Conference of States Parties of the United Nations Convention against Corruption.	AY-wiki48318.html
Warren Sach. Warren Edward Sach (born in 1946 Essex, England) currently serves as United Nations Assistant Secretary-General for Central Support Services and as Representative of the Secretary-General for Investments. Prior to this appointment, he was United Nations Controller. Sach has served the United Nations in a variety of positions. Earlier in his career, from October 1968 to September 1970, he worked as a Junior Professional Officer with the United Nations Development Pro-	AY-wiki47608.html

gramme (UNDP) in Kenya. From May 1974, he became first as a Recruitment Officer and later as a Budget Officer at United Nations Environment Programme (UNEP) headquarters, Nairobi, Kenya. Between May 1979 and October 1988, Sach served in the Budget Division as a Budget Officer. Later on, he became Chief of the Data Analysis and Systems Control Unit. Sach was the Chief of the Salaries and Allowance Division of the International Civil Service Commission (ICSC) for seven years and since January 1996 he served as Deputy Director of the Programme Planning and Budget Division. He was Director of the same Division from 1997 to March 2005. In April 2005, he was appointed United Nations Controller at the level of Assistant Secretary-General. Sach was educated at University College, London, and Magdalene College, Cambridge, where he studied first economics and then development economics, respectively.	
Robert H. Serry. Robert H. Serry (born c. 1950 in Calcutta) currently serves as the United Nations Special Coordinator for the Middle East Peace Process and UN Secretary-General's Personal Representative to the Palestine Liberation Organization and the Palestinian Authority. A career diplomat, Serry has served in a variety of diplomatic positions for his country's foreign service. He was the Dutch Ambassador to Ireland and had served as the Deputy Assistant Secretary-General for Crisis Management and Operations at the North Atlantic Treaty Organization (NATO). He had been posted to Moscow, New York (United Nations) and Kyiv. While in the Netherlands, he led the Middle Eastern Affairs Division of the Dutch Foreign Ministry. He had participated in the events leading to the Middle East Peace Conference in Madrid in November 1991. He obtained his degree in political science from the University of Amsterdam. Following his Ukrainian posting Serry has written a book titled "Standplaats Kiev" (Podium, 1997) available in Dutch and Ukrainian languages.	AY-wiki33154.html
Peter van Walsum. Peter van Walsum was formerly the United Nations Secretary-General's Personal Envoy for Western Sahara. He was appointed to the position by United Nations Secretary-General Kofi Annan in July 2005 and left the position in September 2008 when his mandate expired. Previously, he was the Netherlands' Permanent Representative to the United Nations. He also served as his country's representative on the Security Council in 1999 and Chairman of the Iraq Sanctions Committee in 2000. He obtained his law degree from the University of Utrecht in 1959. He served in the military from 1960 to 1962 and then joined the Civil Emergency Planning in the Ministry of General Affairs from 1962 to 1963. Later on, he served at the Netherland's Ministry of Foreign Affairs, where he was posted to various positions in his nearly four decades career, including the Permanent Mission to the North Atlantic Treaty Organization (NATO) in Paris, the Embassy in Bucharest, the Permanent Mission to the United Nations in New York, the Embassy in both New Delhi and London, and the Permanent Mission to the European Commission in Brussels.	AY-wiki33493.html
George Alleyne. Sir George Alleyne"' (born in St. Philip, Barbados, on 7 October 1932) currently serves as United Nations Secretary-General's Special Envoy for AIDS in the Caribbean region. He was appointed to the position by UN Secretary-General Kofi Annan in February 2003.	AY-wiki49472.html

Alleyne studied medicine at the University of the West Indies (UWI) and graduated as the the gold medallist in 1957. Subsequently, he pursued his postgraduate training in internal medicine in the United Kingdom and the United States. In 1972, he became Professor of Medicine at the UWI. In 1976, he was appointed Chairman of the Department of Medicine. In october 2003 George Alleyne was appointed the Chancellor of the University of West Indies. Besides his academic experience, Alleyne also gained much experience in working for international organizations. In February 1995, he became Director of the Pan American Health Organization (PAHO), Regional Office of the World Health Organization (WHO). He served two four-year terms in this position until the end of 2002 and was elected Director Emeritus. As Special Envoy for AIDS in the Caribbean region, he is responsible for ensuring follow-up to the United Nations General Assembly special session on HIV/AIDS and the Pan-Caribbean Partnership against HIV/AIDS, in the Caribbean region. Occasionally, he also represents the UN Secretary-General at events related to HIV/AIDS in the Caribbean region. He was made Knight Bachelor by Queen Elizabeth II in 1990, and awarded the Order of the Caribbean Community in 2001. In October 2008 Alleyne received the "Science of Peace Award" from the Inter American Heart Foundation.	
Lino Sima Ekua Avomo. Lino Sima Ekua Avomo (born 4 April 1957, in Mongomo, Equatorial Guinea) is the Permanent Representative of Equatorial Guinea to the United Nations. He presented his credentials to Secretary-General Kofi Annan on 21 May 2003. Education. Sima Ekua Avomo attended the Diplomatic School of Madrid, Spain and the Carlos Lwanga National Institute of Secondary Education in Bata, Equatorial Guinea. Career. Sima Ekua Avomo served as Minister of State for International Cooperation and Francophone Affairs prior to his appointment to the United Nations. He has also served as Director General of International Cooperation in the Ministry of Foreign Affairs, International Cooperation and Francophone Affairs, Ambassador to France with jurisdiction in the United Kingdom, Portugal and Switzerland, and representative to the United Nations Office at Geneva. He has also been First Secretary in the Embassy of Equatorial Guinea in Addis Ababa, Ethiopia, and representative to the Organization of African Unity. From 1982-1984, he served as Second Secretary in Equatorial Guinea's Embassy in the Union of Soviet Socialist Republics.	AY-wiki86367.html
Michael Adlerstein. Michael Adlerstein currently serves as Assistant Secretary General of the United Nations and is the Executive Director of the United Nations Capital Master Plan, a five-year program to restore and renovate the historic United Nations' Headquarters in New York, NY. He was appointed to the position by United Nations Secretary-General Ban Ki-moon in July 2007. Adlerstein obtained his architectural degree from Rensselaer Polytechnic Institute and was a Loeb Fellow at Harvard University’s Graduate School of Design. He has extensive experience in restoration of historical sites. Most recently before joining the United Nations, he was the Vice-President and Chief Architect at the New York Botanical Gardens, where he headed a multi-year restoration and design initiative. He previously served in various positions throughout the National Park Service. In this capacity, he was in	AY-wiki37788.html

charge of the planning, design and construction program for the northeast region, including complex partnership projects at Gettysburg, Valley Forge, Acadia and Jamestown. Earlier in his career, he served as Project Director for the restoration of Ellis Island and the Statue of Liberty, which is considered the United States Department of the Interior's most ambitious historic restoration project ever. In this position, he managed and led the team of architects and engineers to plan, design, and construct the Ellis Island. The success of the project led to his promotion as Chief Historical Architect. He was recognized as the national expert in the field of historic preservation. He served as a Peace Corps Volunteer in Colombia, and has worked as a State Department consultant on preservation issues on numerous projects, including the preservation of the Taj Mahal. He has won numerous awards for his achievements and was made a Fellow of the American Institute of Architects.	
Catherine Pollard. Catherine Pollard (born in 1960 in Georgetown, Guyana) currently serves as United Nations Assistant Secretary-General for Human Resources Management. She was appointed to the position by UN Secretary-General Ban Ki-moon in April 2008. Prior to her current appointment, Pollard serves as Chief of Staff in the Department of Peacekeeping Operations. In this capacity, she was responsible for implementing the ongoing restructuring in the Departments of Peacekeeping Operations and of Field Support. She also helped the Secretary-General's task force to streamline management practices, including human resource management. Before this, she was Director of the Peacekeeping Financing Division in the Department of Management. In this capacity, she implemented the Secretary-General's reforms to enhance the strategic focus of budgets and to streamline the budget process. Pollard also held a variety of assignments in the area of financial and human resource management of the United Nations Volunteers Programme, the United Nations Development Programme, and the United Nations Protection Force in Croatia. Pollard obtained her master's degree in accounting from the University of the West Indies, Kingston, Jamaica.	AY-wiki47559.html
Atul Khare. Atul Khare (born in 1959) currently serves as the United Nations Special Representative for Timor-Leste and Head of the United Nations Integrated Mission in Timor-Leste(UNMIT). He was appointed to the position by UN Secretary-General Kofi Annan in December 2006. Khare started his career as an Indian diplomat in 1984. He has served in various capacities in the Indian foreign service, including as Deputy High Commissioner of India to Mauritius, Counsellor at the Permanent Mission of India to the United Nations in New York and Chargé d'affaires of the Indian Embassy in Senegal with concurrent accreditation to Mali, Mauritania, Gambia, Guinea Bissau and Cape Verde. He was Chef de Cabinet of the Foreign Secretary of India and of Director of the United Nations Division in the Ministry of External Affairs in New Delhi and served as Director of the Nehru Centre and Minister (Culture) of the High Commission of India in London since 2005. Besides his diplomatic career, Khare also acquired extensive experience in the United Nations. He served as Chief of Staff and later as Deputy Special Representative of the Secretary-General with the United Nations Mission of Support in East Timor (UNMISET) from June 2002	AY-wiki17486.html

until its completion in May 2005. Khare obtained a master's degrees in business administration and in leadership from the University of Southern Queensland, an advanced diploma (with distinction) in French from the Indian Defence School of Languages, a Bachelor of Medicine and a Bachelor of Surgery (with honours) from the All India Institute of Medical Sciences.	
Rebeca Grynspan. Rebeca Grynspan currently serves as Assistant Administrator of the United Nations Development Programme (UNDP) and Director of UNDP's Regional Bureau for Latin America and the Caribbean. She was appointed to the position by United Nations Secretary-General Kofi Annan in December 2005. Grynspan obtained a Bachelor of Science in economics from the University of Costa Rica and a Master of Science in economics from Sussex University. Prior to her appointment, she was Director of the Subregional Headquarters in Mexico of the Economic Commission for Latin America and the Caribbean (ECLAC). She was also a member of the UN Millennium Project's Task Force on Poverty and Economic Development and of the UN High-Level Panel on Financing for Development. Before joining the UN, she served as Vice-President of Costa Rica from 1994 to 1998 and concurrently as coordinating minister of the Government's social and economic sectors, housing and human settlements minister. She was a professor at the University of Costa Rica and researcher at the Instituto de Investigaciones de Ciencias Económicas de Costa Rica.	AY-wiki38614.html
Bader Al-Dafa. Bader Al-Dafa currently serves as the United Nations Executive Secretary of the Economic and Social Commission for Western Asia (ESCWA). He was appointed to the position by United Nations Secretary-General Ban Ki-moon in July 2007. A seasoned diplomat, Al-Dafa has acquired extensive experience in foreign affairs. He was the recent Ambassador of Qatar to the United States and Permanent Observer to the Organization of American States (OAS). In addition, he has served as Qatar's Ambassador to the Russian Federation, France, Egypt and Spain, and as non-resident Ambassador to Finland, Greece, Latvia, Lithuania, Estonia, Switzerland and Mexico. Besides his diplomatic experiences, he has also served in a variety of capacities in international non-governmental organizations. He supervised programmes of building housing for families with limited income in Africa and participated in landmine removal programmes in the Balkans. He also played an active role in fund-raising programmes with children's hospitals in Asia and North America and in supporting the empowerment of women in North Africa and Central Asia. He also organized conferences on free trade, democracy and inter-religious dialogue. Al-Dafa holds a master's degree in international public policy from Johns Hopkins University and a bachelor's degree in political science and economics from Western Michigan University. He obtained the award the Ordre du Merite from France. He is fluent in Arabic, English and Spanish.	AY-wiki02605.html
Afelee F. Pita. Afelee F. Pita, born February 11, 1958, is a Tuvaluan diplomat. He is currently Tuvalu's Permanent Representative to the United Nations. Pita holds a Master's degree in public administration from the University of Canberra and a Bachelor of Arts degree in administration and accounting from the University of the South Pacific. He began his career as a senior official in government administration as	AY-wiki56708.html

assistant Secretary, and then Secretary, at the Tuvaluan Ministry of Commerce and Natural Resources, from 1987 to 1988. He was Assistant Secretary for Commerce from 1989 to 1993, then Acting Secretary at the Ministry of Trade, Commerce and Public Corporations in 1993. From 1994 to 1994, he served as Permanent Secretary in several successive ministries (Health and Sports, Labour and Communication, Resources and Environment, Finance). From 2001 to 2004, Pita was Adviser to the Executive Director of the Asian Development Bank in Manila, where he served as representative for Australia, Azerbaijan, Cambodia, Hong Kong, Kiribati, the Federated States of Micronesia, Nauru, the Solomon Islands and Tuvalu. Returning to Tuvalu, Pita served as Permanent Secretary to the Ministry of Natural Resources and Lands from 2004 and 2006, before being appointed as Permanent Representative to the United Nations. In April 2007, Pita addressed the Special Session of the United Nations Security Council on Energy, Climate and Security, and "beseech[ed] the Security Council to act urgently to address the threats to [Tuvalu]'s national security" - namely, climate change.	
Choi Soon-Hong. Choi Soon-hong (born in Seoul, Republic of Korea, in 1950) currently serves as United Nations Chief Information Technology Officer at the level of Assistant Secretary-General. He was appointed to the position by UN Secretary-General Ban Ki-moon in July 2007. Choi has experience both in the public sector and the private sector. Starting his career as Quality Assurance Manager at Systems Automation Corporation in 1977, he also worked as Information Systems Analyst at TRW, Inc. between 1980 and 1981. In 1981, he joined the International Monetary Fund (IMF), where he served in numerous technical and business operations and team leadership positions until 1997. He was Division Chief of Technology Infrastructure in between 1997 and 1999 and Senior Budget Manager and Strategy Adviser from 1999 to 2004. Later on, he served as Head of Information Technology Services at IMF from 2004 to February 2007. He has been the IMF representative to the Information and Communications Technology Network of the United Nations Chief Executives Board for Coordination, and a member of the United Nations International Computing Centre's Management Committee since 2005. In his new position as UN Chief Information Technology Officer, he is in charge of all substantive and operational needs relating to information and communications technology of the United Nations, including developing, maintaining and monitoring the implementation of the effective information and communication strategy. Choi has conducted research and lectured on public policy, strategic management and innovation. His recent research interests are globalization, technology competition, digital society, knowledge sharing and information and communications technology for development. Choi holds a bachelor's degree in engineering from Sogang University, a master's degree in computer science from George Washington University, an MBA from the Wharton School of the University of Pennsylvania, and a Ph.D. in strategic management and public policy from George Washington University.	AY-wiki24175.html
Charles Thembani Ntwaagae. Charles Thembani Ntwaagae (born in 1953, Tutume, Botswana) is the Permanent Representative to the Unit-	AY-wiki52137.html

ed Nations for Botswana. He took office in July 2008. He is married with three children. Education. Ntwaagae holds a master's degree from Pennsylvania University, and a bachelor's degree from the University of Botswana and Swaziland. Career. Ntwaagae was Permanent Secretary in Botswana's Ministry of Foreign Affairs and International Cooperation prior to his taking office at the United Nation. He has also served as Permanent Representative to the United Nations Office at Geneva, in Austria, and in Greece, for Botswana. Other offices he has held include Deputy Permanent Secretary in the Foreign Ministry; Chief Executive at the National Secretariat of the National Conservation Strategy (Coordinating) Agency; and Deputy Permanent Secretary in the Ministry of Local Government, Lands and Housing.	
Sin Son Ho. Sin Son Ho (), born July 5, 1948, is a North Korean diplomat. He is currently North Korea's Permanent Representative to the United Nations. In his career as a diplomat, he has primarily represented North Korea in its relations with African countries. Sin graduated from Kim Il Sung University in 1972. That same year, he was appointed as Third Secretary to the North Korean embassy to Egypt, where he served until 1979. From 1979 to 1983, he was a senior officer in the Foreign Ministry, then, from 1983 to 1986, counsellor at the North Korean embassy to Lesotho. From 1986 to 1990, he was Division Chief at the Foreign Ministry, before serving as counsellor to the North Korean embassy in Zimbabwe from 1990 to 1995. From 1995 to 1999, he was Deputy-Director of the Foreign Ministry. Sin's first appointment to the United Nations came in 1999, when he worked at the North Korean Mission to the United Nations for four years before returning home upon being appointed Director-General of the Foreign Ministry. He served in the latter position until May 2008, when he was appointed Permanent Representative to the United Nations. In his latter role, Sin has been tasked with expressing North Korea's positions on the issue of his country's nuclear activities. In October 2008, Sin told a session of the General Assembly that North Korea had developed nuclear weapons in response to the threat it perceived from the United States, but that it had begun dismantling its nuclear apparatus and that it supported the "denuclearisation of the Korean peninsular". Sin added that his country wished for a peace agreement to replace to 55 year-old Korean War armistice, and that North Korea hoped for the eventual reunification of Korea.	AY-wiki56683.html
Rachel Mayanja. Rachel N. Mayanja currently serves as United Nations Secretary-General's Special Adviser on Gender Issues and Advancement of Women. She was appointed by the United Nations Secretary-General in 2004. Mayanja joined the United Nations by working for the Division for Equal Rights for Women, a division within the Centre for Social Development and Humanitarian Affairs. She served as the Special Assistant to the Assistant Secretary-General for Social Development and Humanitarian Affairs and actively participated in the development of policies and attended conferences at the intergovernmental and non-governmental levels on topics dealing with gender, the youth, the aged, the disabled and family. From 1989 to 1990, Mayanja went to work for the peacekeeping missions in Namibia (UNTAG), where she worked with the United Nations civilian police to oversee the elections	AY-wiki45167.html

leading to independence. From 1992 to 1994, she served in the UN Mission in Iraq/Kuwait (UNIKOM). Later on, Mayanja served on different senior positions in the Office of Human Resources Management, including as Chief, Common System and Specialist Service, dealing with policies regarding salaries and entitlements, as well as appeals and disciplinary cases. In 2000, she joined the United Nations Food and Agriculture Organization (FAO) as the Director, Human Resources Management Division. She played a vital role in the implementation of the reform of human resources management at FAO. Ms Mayanja, a national of Uganda, holds a law degree from Makarere University and a master's degree in law from the Harvard University Law School, United States.	
Sivert Andreas Nielsen (1916-2004). Sivert Andreas Nielsen (1916 - 2004) was a Norwegian civil servant, banker and politician for the Labour Party. He was born in Copenhagen as the son of Konrad Nielsen. A jurist by education, he spent the years 1941–1945 in Nazi German captivity. After the World War II he served as a diplomat to the United Nations from 1946 to 1948 and to the United States from 1948 to 1950. He then worked in the Ministry of Defence, as assistant secretary from 1950, deputy under-secretary of state from 1952 and state secretary from 1955 to 1958. From 1958 to 1966 he returned as Norwegian ambassador to the United Nations. From 1966 to 1976 he worked as a banker.	AY-wiki60119.html
Jan Beagle. Jan Margaret Beagle is a diplomat from New Zealand, Deputy Director-General of the United Nations Office at Geneva. She served as United Nations Assistant Secretary-General for Human Resources Management, appointed to the position by UN Secretary-General Kofi Annan in September 2005, and succeeded by Catherine Pollard of Guyana in April 2008. She started her career by serving the New Zealand Ministry of Foreign Affairs. From 1974 to 1979, she was a delegate of her country to the United Nations. In 1979, Beagle joined the United Nations, first in the Department of Political and Security Council Affairs, then as Political Affairs Officer from 1979 to 1989. She later became Senior Officer in the Office of the Under-Secretary-General for Management and Special Assistant to the Controller in the following year, Special Assistant to the Associate Administrator of the United Nations Development Programme from 1990 to 1992, and Principal Officer in the Executive Office of the Secretary-General from 1992 to 1996. Beagle became Director of the Division for Organizational Development in 1997 and has been serving in this capacity until her new appointment as ASG for Human Resources Management. She completed her Master of Arts in History and International Relations at the University of Auckland. Controversy. In January 2007, the United Nations Staff Council passed vote of no confidence in Beagle, saying that she prevented reforms to the UN's "archaic" system of employee rights.	AY-wiki88523.html
Eliyahu Sasson. Eliyahu Sasson (, born 1902, died 8 October 1978) was an Israeli politician and minister. Biography. Born in Damascus in Syria, Sasson studied at an Alliance School in his hometown and the Université Saint-Joseph in Beirut. He became a member of the Arab National Movement and edited a Jewish-Arab newspaper named "Al-	AY-wiki01669.html

Hayat". He made aliyah in 1927 and worked as an electrician, journalist and lecturer on Middle East affairs. He began working in the political department of the Jewish Agency, serving as head of the Arab department between 1933 and 1948. A member of the Jewish delegation to the United Nations between 1947 and 1948 and at the ceasefire negotiations in 1949, he worked as director of the Middle East department of the Foreign Affairs Ministry between 1948 and 1950, before heading an office in Paris for contacts with Arab nations. He also served as the Israeli envoy to Turkey (1950-1952), an envoy and ambassador to Italy (1953-1960) and ambassador to Switzerland (1960-1961). In 1961 he returned to Israel and was appointed Minister of Postal Services by David Ben-Gurion. He was elected to the Knesset in the 1965 elections, and retained his cabinet post until 2 January 1967, when he became Minister of Police. Although he was re-elected in 1969, he lost his ministerial post upon the formation of the new government. He lost his seat in the 1973 elections.	
Neven Jurica. Neven Jurica, (born 4 April 1952), is the Permanent Representative to the United Nations for Croatia. He took office in February 2008. Jurica is married with two children. Education. Jurica has a degree in comparative literature and philosophy, and a master of arts in literary theory from the University of Zagreb. Career. Jurica was the Croatia Ambassador to the United States prior to taking office at the United Nations. He has also served as Ambassador to Australia, New Zealand, Bulgaria, and Norway. He was a founding member of the Croatian Democratic Union and served as Political Secretary. Following the first democratic elections in Croatia in 1990, he was elected to the Parliament and served as Chairman of the Human Rights Committee (1990-1992), as well as, Chairman of the Foreign Affairs Committee (2003-2004). Jurica holds membership in the Croatian Writers' Association and P.E.N. (Poets, Essayists, and Novelists). From 1980 to 1989, he worked as a writer and published in excess of 16 books on literary theory and criticism. At the same time, he oversaw a literary forum, “Literary Friday”.	AY-wiki58368.html