Extracting Place Semantics From Geo-Folksonomies

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Ehab ElGindy

2013

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

Massive interest in geo-referencing of personal resources is evident on the web. People are collaboratively digitising maps and building place knowledge resources that document personal use and experiences in geographic places. Understanding and discovering these place semantics can potentially lead to the development of a different type of place gazetteer that holds not only standard information of place names and geographic location, but also activities practiced by people in a place and vernacular views of place characteristics.

The main contributions of this research are as follows. A novel framework is proposed for the analysis of geo-folksonomies and the automatic discovery of place-related semantics. The framework is based on a model of geographic place that extends the definition of place as defined in traditional gazetteers and geospatial ontologies to include the notion of place affordance. A method of clustering place resources to overcome the inaccuracy and redundancy inherent in the geo-folksonomy structure is developed and evaluated. Reference ontologies are created and used in a tag resolution stage to discover place-related concepts of interest. Folksonomy analysis techniques are then used to create a place ontology and its component type and activity ontologies.

The resulting concept ontologies are compared with an expert ontology of place type and activities and evaluated through a user questionnaire. To demonstrate the utility of the proposed framework, an application is developed to illustrate the possible enrichment of search experience by exposing the derived semantics to users of web mapping

applications. Finally, the value of using the discovered place semantics is also demonstrated by proposing two semantic based similarity approaches; user similarity and place similarity. The validity of the approaches was confirmed by the results of an experiment conducted on a realistic folksonomy dataset.

Acknowledgements

It is a great achievement to finish years of research and to write this thesis. However, I could not do any of this without the support and help I received from many people. In particular, I am thankful to my supervisor, Dr. Alia Abdelmoty, who was very supportive and guided me in each step to complete the research.

I am grateful to my parents, who have, as always, given me an indispensable moral support. There are really no words to describe how important was the continuous support my wife gave me since I started my PhD. I really would not achieve this without you. Kady, my little princess, your are the light of my life. Thank you all.

I would like to thank my friends and colleagues at Cardiff School of Computer Science and Informatics for their help and encouragement. Vitaly Teterin, Mona Ali, Ahmed Alazzawy, Sultan Alyahya, Hmood AlDossari, Haya AlMagwash, Shada Alsalamah, Abdelbaset Greede and Abdelhamid Alwaer. Thank you all.

Contents

List of Figures

List of Tables

List of Listings

Chapter 1

Introduction

1.1 Background and Motivation

Social bookmarking applications were introduced as part of the web 2.0 wave, where users are given the facility to publish and annotate contents/resources on the web. In such applications, users annotate web resources, e.g. web pages, using a set of keywords, namely *tags*, the annotation process is called *tagging* whilst the resulting structure of users, tags and resources is called *folksonomies*. The main purpose of the social bookmarking applications is to allow users to organise and index the resources with their own selection of tags. The tags may include keywords that cannot be extracted from the resources. The reason for that is some resources are not text-based such as images, or because users select different terms than the ones included in the resources based on their understanding of the document's topic.

The tagging process may not employ any sort of syntax validation, checking for spelling mistakes or controlled vocabulary restrictions to validate the user input. Such simple style of data acquisition requires no technical knowledge or special skills from the users, which is the main reason for the popularity of the tagging applications. On the other hand, this simplified user input approach introduces certain limitations which can affect the quality of the tags. For example, tags can be misspelled, vague or written in slang language.

Users with different backgrounds and expertise, which are reflected in their selection of

tags, may not access the resources annotated by each other unless the semantics of the tags are considered in the search and navigation tools. To a certain extent, dictionary resources may be employed to relate tags with linguistic relationships, such as polysemy and synonymy, to fill this gap. However, using formal data sources, including dictionaries, will fail to relate terms that have informal relationships known within a community of users, and will also fail to process new terms that are not included in the dictionary, such as the term "folksonomy".

As folksonomies directly reflect the vocabulary of users [67], they enable matching of users' real needs and language. On-going research efforts, such as in [93, 70, 19, 105, 78], realised the importance of the emergent semantics extracted from folksonomies as they capture the concepts and their relationships as understood by users.

A typical use of the emergent semantics extracted from folksonomies is to feedback to the social bookmarking application they are collected from to enhance the search and browsing experience. For example, the semantics can be used to enrich user queries with terms that other users think are semantically related to the terms used in the original queries.

Geo-tagging of resources on the web has become prevalent. Geographic referencing has evolved to become a natural method of organising and linking information with the aim of facilitating its discovery and use. Indeed, a significant portion of search queries include reference to geographic places [90]. GPS-enabled devices allow people to store their mobility tracks, tag photos, and events. In response, many applications on the web are enabling geo-tagging of resources, such as geo-locating photos on Flickr[1] and tweets on Twitter[2], and people are collaboratively building their own map resources and gazetteers (e.g. GeoNames[3] and OpenStreetMap[4]). While typical place name resources provided by mapping agencies, referred to as geographic thesauri, record name

[1] http://www.flickr.com
[2] http://www.twitter.com
[3] http://www.geonames.org
[4] http://www.openstreetmap.org

and map coordinates of a place, collaborative mapping on the social web provides an opportunity for people to create maps that document their social and personal experiences in a place. Thus university buildings may be a place of work and study for a group of people, a conference venue for another group, and a sports facility for a different group. Understanding and encoding this information in place name resources can eventually result in a different type of place gazetteer that documents not only where a place is, but also what happens at a place.

Some social bookmarking applications, such as Tagzania[5], are specialized in tagging geographic places using a map-based web interface. These applications generate a special kind of folksonomy, denoted **geo-folksonomy** in this thesis. Place resources in geo-folksonomies have some characteristics which do not exist in normal web resources: a) place resources are created to reference places in the real world, while normal web resources already exist in the web space and they are just referenced using unique URLs. Although it is possible to assign a unique URI for any resource (including place resources [7]), URIs are not used to locate places as people always refer to places by spatial and thematic attributes such as location and place name respectively; b) the values of spatial attributes, such as longitude and latitude, are acquired using a map-based applet. This method of acquiring data can be imprecise and is dependent on the user being able to identify and digitize a precise location on a map offered on the user interface of these applications. The accuracy is also related to the map scales offered to users and the difficulty in matching the precise location across map scales and c) the values of thematic attributes, such as place names, are acquired using a free-text input. Although they add valuable semantics to the place resources, they are associated with complexity, where people use non-standard, vernacular place names [28] and abbreviations. Hence, an immediate challenge is to analyse the quality of the place resources in geo-folksonomies.

Tags in folksonomies are created to describe general concepts in different topics, while

[5]http://www.tagzania.com

tags in geo-folksonomies are created mainly to describe places and place-related concepts. Hence, research has addressed the problem of extracting the place semantics embedded in geo-folksonomies, such as in [82, 79, 81, 22], where the place semantics are represented using lightweight ontologies that model the hierarchical gazetteer of place names, a set of place and events, or a set of clustered places that share common social aspect. Nevertheless, geo-folksonomies can be a potential source of information to build a more comprehensive place model that captures the social aspects of places including what activities people can do and how they realise the services provided by individual places. As a result, an additional challenge emerges to capture those types of semantics.

The aim of the research presented in this thesis is to provide an approach for extracting place semantics embedded in geo-folksonomies. Social/informal knowledge about places is targeted here, which are different to the semantics provided by formal place gazetteers or place ontologies. In particular, perceptions of users about place affordance and human activities related to places are captured to build place type and activity ontologies. The approach addresses the quality problems evident in the tags and place resources through a cleaning process; it also provides a place ontology model to capture the desired place semantics, and utilises external semantic resources and statistical co-occurrence methods to build the place ontology. The resulting ontology is evaluated and the applicability of the approach is also demonstrated.

1.2 Research Problem and Hypothesis

The research carried out in the scope of this thesis addresses the problem of extracting place semantics from geo-folksonomies. In particular, the main question investigated here is **How and to what extent the user tags and resources in geo-folksonomies can be utilised to build models that capture the social aspect of geographic places** and **How valuable are the new types of place semantics represented in these models?**

This problem can be further specified with the following research questions:

1. **How good is the quality of tags and place resources in geo-folksonomies?**
   In addition to the quality problems of the tags inherited from general folksonomies, place resources in geo-folksonomies introduce different quality problems such as the imprecise spatial locations and non-standard, vernacular names associated with the place resources. Answering this research question requires identifying and analysing the different quality problems in a realistic sample of a geo-folksonomy dataset. Additionally, it is also required to identify a method to quantitatively measure the quality of the dataset to evaluate any proposed cleaning approach.

2. **How different are the place semantics extracted from geo-folksonomies from the semantics represented by place ontologies and gazetteers?**
   The aim of the place semantics extracted from geo-folksonomies is to represent the way the users recognise and experience places. To answer this research question, concepts and semantic relationships embedded in geo-folksonomies need to be identified and extracted. A suitable representation model to capture these semantics needs to be designed and evaluated against existing models of place.

3. **How can the place semantics extracted from geo-folksonomies be evaluated?**
   Generally, evaluating semantics extracted from folksonomies is a challenging research task. Existing evaluation methods need to be considered and a suitable evaluation strategy needs to be identified to judge the successfulness of the approach.

4. **Can the place semantics extracted from geo-folksonomies be utilised to calculate user similarity based on their place perceptions?**
   A user profile can be constructed in social bookmarking applications from the tags used by that user which represent their topics of interest. The answer to this research question requires investigating the value of using the extracted place-

related semantics to enrich user profiles on the web as well as provide a dimension for evaluating users' similarity on the social web.

5. **Can the place semantics extracted from geo-folksonomies be used to derive a new measure of place similarity that complements traditional dimensions used in the literature?**
   Similarity of geographic places is normally a function of their spatial and thematic attributes. The geo-folksonomy tags can be employed to devise a place similarity measure based on the collaboration and interaction of the users who tag the places on the social web. Moreover, the semantics embedded in the tags can also be utilised as a place similarity application which is the focus of this research question.

Research Hypothesis

"User interaction on the social and collaborative mapping web can be used to deduce geographic and place-related concepts of relevance to the user. The deduced geo-semantic concepts are relevant to places and can be used to enhance people's understating of the places they live in."

Importance of Discovered Geo-Semantics

Users' interaction and collaboration on social and mapping web generate a new source of place information, where the information generated by users represent informal and social place semantics that reflect their experiences and sentiments about places. Such information can be beneficial to complement the formal place information provided by mapping agencies to build comprehensive place gazetteers. Moreover, this information can be utilised to enhance the user experience of using collaborative mapping applications and can also be used to deduce semantic similarity measures based on users' understanding of places.

1.3 Overview of the Thesis

The work carried out in the scope of the research is presented as follows:

Chapter 2: provides an overview of the literature related to the research discussed in the thesis. The chapter begins with an overview of concepts from library sciences, such as taxonomies and thesauri, to explain the origin of the resource organisation problem. The chapter then links these concepts to the web 2.0 social tagging and folksonomies, focusing on the research that addresses the problem of extracting the embedded semantics from user tags. Moreover, research addressing the geographical aspects of the folksonomies is discussed and the open issues on extracting place semantics from the folksonomies are identified to motivate the work in the thesis.

Chapter 3: presents a design of a place ontology model that captures the place semantics embedded in geo-folksonomies. Additionally, the chapter presents an overview of the framework proposed in this research to extract the place semantics from geo-folksonomies. The framework consists of three stages: pre-processing stage, tag resolution stage, and semantics association and ontology building stage. The details of the framework are discussed in Chapters 4 and 5.

Chapter 4: discusses the details of the pre-processing stage where several quality problems in the geo-folksonomies are identified and a cleaning approach is devised to address the identified problems. Also, this chapter discusses the evaluation strategy used to assess the quality of the output.

Chapter 5: discusses the details of the tag resolution stage where an approach is presented to identify the place-related concepts in the tag space via utilising external semantic data sources. Additionally, the chapter discusses the approaches used to infer the semantic relationships between the different concepts. Two approaches to evaluating the resulting semantics are used; a questionnaire is designed to validate the quality of the extracted semantics, and an automated semantic similarity service is also used to validate the inferred semantic relationships against the general semantics on the web.

Chapter 6: presents the details of the implementation of the work carried out in this research. A service-oriented application design is presented that contains several components to crawl the folksonomy from the web, analyse the collected folksonomy to extract place semantics, store and query the semantics encoded as OWL ontology. The chapter presents the details of the service layer which exposes a set of functions that can be called remotely to query the folksonomy and extracted semantics. Finally, an overview is provided on the implementation of a mapping-based application, SemTag, which utilises the induced semantics to enhance the user experience provided by the folksonomy-based applications.

Chapter 7: The aim of this chapter is to study whether the induced place ontology can be utilised as an application of user similarity. The chapter discusses building user profiles from folksonomies which are enriched using a statistical co-occurrence approach and using the induced place semantics. The user similarity is calculated using the different profile approaches and the output is presented and discussed.

Chapter 8: The aim of this chapter is to study whether the induced place ontology can be used to produce a semantic similarity measure for places. The chapter compares different approaches to calculating place similarity using folksonomies, that includes using the direct tags attached to each place, using the direct tags along with their similar tags, and using the direct tags along with their semantically similar tags retrieved from the induced ontology. The place similarity is calculated using the different approaches and the output is presented and discussed.

Chapter 9: concludes the thesis with an overview of the work carried out, the contributions of this study and an outlook for future research.

1.4 Contributions

The contributions of this thesis can be summarised as follows:

- Studying and identifying possible problems in the representation of geo-folksonomy datasets that can affect the quality of the data which do not exist in general folksonomies, particularly problems in the place resources, and introducing a pre-processing approach to limit the effects of the identified problems. The proposed approach was shown to improve the overall quality of the geo-folksonomy structure.

- Introducing a place ontology model to capture the social aspects of places including place affordance and the human activities. The model design is unlike other place ontologies and gazetteers which focus on the geographical aspects such as topological relationships.

- Extend existing place models to capture place-related semantics embedded in users' annotations and tags, particularly related to actions and activities associated with a place as well as categories for classifying place types.

- Suggesting a hybrid evaluation approach for ontologies extracted from folksonomies which consists of questionnaire and automatic web-based evaluations.

- Showing that the extracted place ontology can be utilised to produce user profiles that represent the place-related interests of users.

- Showing that the extracted place ontology can be utilised to produce semantic similarity measure for places.

1.5 Publications

- **ElGindy, E. & Abdelmoty, A. (2012), Enhancing the Quality of Place Resources in Geo-folksonomies, in Liwei Wang; Jingjue Jiang; Jiaheng Lu; Liang Hong & Bin Liu, ed., 'Web-Age Information Management', Springer Berlin / Heidelberg, , pp. 1-12.**

- **ElGindy, E. & Abdelmoty, A. (2012), Capturing Place Semantics From Users' Interaction on the GeoSocial Web, submitted to the semantic web journal**

- **ElGindy, E. & Abdelmoty, A. (2012), Enriching User Profiles using Geo-Social Place Semantics Induced from Geo-Folksonomies, submitted to the international journal of geographical information science**

Chapter 2

Background and Related Work

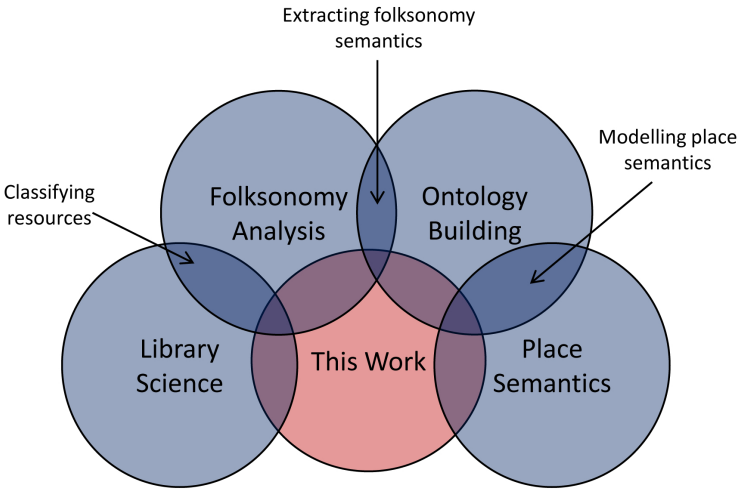


Figure 2.1: An illustration of the related research areas.

The research presented in this thesis is based on a variety of research areas and technologies including library and information sciences, folksonomy analysis, ontologies and semantic web, extracting semantics from user-generated content on web 2.0 and knowledge representation of geographic places. The chapter starts with an overview of using metadata to organise resources along with a presentation of the classification methods, originated in the library and information sciences, which are utilised by various approaches to extract semantics embedded in folksonomies. As folksonomies are the source of information to be analysed in this thesis, this chapter provides an overview of the definition and characteristics of folksonomies followed by the methods used in this thesis to calculate the similarity between folksonomy entities. The focus

is then switched to ontologies as they are employed in this thesis to represent the extracted semantics. Hence, this chapter provides a background of ontologies followed by a literature review on the approaches of extracting ontologies from folksonomies. The attention is then directed to the problem of extracting place semantics which is the main focus of this thesis. The limitations of the current approaches in the context of extracting place semantics are then presented. Finally, a summary of the chapter is given.

2.1 Organising Resources Using Metadata

Metadata is structured data that describes the characteristics of a dataset. The most straightforward definition of metadata is "data about data". In library and information sciences, library catalogues are good examples of metadata. The typical library catalogue contains information about each book in the library such as author, title, publishing date and the location of the book in the library [71]. In this case, the library catalogue is supplementary data used to describe the books (resources) in the library. Having an indexed library catalogue can ease the process of searching for and locating a specific book in the library. Similarly, pages on the web can expose metadata through special HTML elements "meta tags". For example, authors of web pages can provide a set of keywords as meta tags which can be indexed by search engines to allow finding these pages if the search query contained specific keywords that are referenced within those pages.

Authoring metadata to describe resources is traditionally carried out by dedicated professionals. For example, the metadata in library systems should be syntactically written in a standard format that facilitates machine processing, such as the Machine-Readable Cataloguing (MARC) standard. Additionally, a standard metadata vocabulary should be defined and followed by the authors. A well known vocabulary for metadata is Dublin Core [113] which defines a standard set of properties to describe documents.

Examples of these resources are 'title', 'creator', 'subject', 'description', 'date', and 'language'.

In most web-based systems, metadata creation is typically carried out by the authors of the resources - web pages, images and videos, for example, - to allow search engines to index these resources. Some web-based systems, such as corporates or news portals, publish the metadata through specialised content management systems (CMS) which facilitate the metadata authoring process to non-technical content editors. In web 2.0 collaborative and social applications, the metadata creation process is completely different. The metadata is typically provided in terms of single keywords (tags) entered by users and they could be stored in data stores separated from the resources being described, for example delicious[1] which allows users to index and organise their preferred web resources by annotating them with tags of their choice. The authoring process here is not carried out by professional or trained editors, and the metadata is provided by normal untrained web users, which of course has an impact on the quality and certainty of the provided data.

The process of organising a set of resources can be described by the terms "categorization" and "classification". Despite both terms seeming to be similar, these are different but overlapping processes. Categorization refers to the process of dividing the world into groups of entities whose members are in some way similar to each other, while classification refers to three distinct but related concepts: a system of classes, a group or class in the classification system, and the process of assigning entities to classes. The categorization process is an unsystematic process, and it does not depend on the features of the entity but it depends on similarity assessment which involves immediate context, personal sentiment or individual experience. On the other hand, the classification process involves systematic approaches for classifying entities based on their characteristics or features that define each class [52]. The following sections provide a discussion on the classification and the categorization processes with respect to the

[1] http://www.del.icio.us

research presented in this thesis.

2.2 Classification

Metadata of a resource is a set of attributes that describe what the resource is about in terms of discrete subjects. Several subject-based classification [34] techniques have been devised to group resources based on their subjects, these include controlled vocabularies, taxonomies, thesauri and faceted classification. However, it is important to clarify that there is a distinction between describing the resources being classified, and describing the metadata used to classify the resources. The subject-based classification approaches below are about classifying the metadata rather than classifying the resources. Such classification methods help connect the resources to the metadata and the subjects they are about.

2.2.1 Controlled Vocabularies

Controlled vocabulary, also known as "indexing language" in library science, is a predefined set of terms used to describe resources. Each term represents the name of a specific concept. A concept can have multiple names and each name refers to only one subject to avoid ambiguities [34]. Controlled vocabularies are closed sets of keywords that do not allow resources to be described using keywords not defined by the provided vocabulary. Such a controlled approach can be beneficial to avoid using keywords with problems such as being vague, too broad, too narrow or misspelled. Moreover, the problem of having multiple morphological forms of the same keyword can also be avoided.

Controlled vocabularies can also be beneficial in some cases where the resources need to be classified according to a specific domain. For example, controlled vocabulary of country names can be used to classify books in a library or in an online book store

based on the country of publishing. However, this classification approach can fail in other scenarios where there is no specific domain for classification. For example, there is no controlled vocabulary that can cover all the keywords used to describe images uploaded on Flickr[2].

2.2.2 Taxonomies

Taxonomy is a term that originated in life sciences when Carl Von Linnï¿½ [11] introduced a hierarchical classification system for life forms. Taxonomy is used in the 18th century to classify all the plants and animals on earth. Each animal or plant is represented by a node in a tree of hierarchical relationships between other nodes representing other species [34].

The term taxonomy is adopted in information sciences. However, having a term ported from a different domain can lead to having multiple definitions of this term in the new domain. Gilchrist [35] argued that the term taxonomy is a generic term and can have different meanings according to the type of the application it is used in. He classified the applications of taxonomies into: web directories, taxonomies to support automatic indexing, taxonomies created by automatic categorization, taxonomies to support searching and browsing, and corporate taxonomies.

Garshol [34] emphasized the hierarchical relations between terms and defined taxonomy as: "a subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy without doing anything further".

Hepp and de Bruijn [46] focused on the semantic aspect of the taxonomy and argued that a taxonomy represents a subsumption relationship between concepts. In other words, a "sub class of" relation in which any instance from a class is implicitly an instance of all the parent classes to that class. For example, in a taxonomy of place types, "Chinese Restaurant" is subsumed by "Asian Restaurant" which is also sub-

[2]A popular photo sharing for uploading and tagging images. `http://www.flickr.com`

sumed by the type “Restaurant”. It also implies that “Chinese Restaurant” is subsumed by “Restaurant”. However, if the relationship between the classes represents broader or narrower terms relationships, then it should be called “hierarchical classification” instead of taxonomy.

In this thesis, the term taxonomy will be considered to be referring to any hierarchical structure of concepts that has parent-child relationships regardless of the semantic meaning of the relations. Ontologies can be used to address semantic relationships and will be discussed later in this chapter.

2.2.3 Thesauri

Thesaurus can be considered as an extended version of taxonomies. Taxonomies classify terms in a hierarchical manner using parent-child relationships, while thesaurus allows more relationships to be used to classify terms. Thesaurus is described using two ISO standards; ISO 2788 which describes monolingual thesauri and ISO 5964 which describes multilingual thesauri. Basically, ISO 2788 defines several properties for thesauri such as:

- **BT**: stands for ‘broader than’, and is used to refer to a term which has wider or less specific meaning and it is always above in the hierarchy structure. ‘BT’ has an inverse relationship called ‘NT’ which stands for ‘narrower than’. The properties ‘BT’ and ‘NT’ allow thesauri to provide similar functionality provided by taxonomies, as they are the relationships responsible for defining the hierarchical structure of terms.
- **USE**: used to refer to another term that is preferred to be used instead of the current term.
- **RT**: stands for ‘related term’, and is used to link two terms that have related meanings which cannot be defined by ‘BT’, ‘NT’ or without being a synonym.

2.2.4 Faceted Classification

The term 'faceted classification' first originated in library sciences by S.R. Ranganathan [3]. The structure of the 'faceted classification' can be seen as a thesaurus-like structure where properties such as 'BT' and 'NT' can be used. However, each resource is classified using more than one perspective (facet), each facet contains a number of terms and each term cannot belong to more than one facet [101]. Resources to be classified are given one term from each facet, which gives a description for the resources from the different perspectives defined by the facets.

Ranganathan proposed the first faceted classification model to classify books in libraries by using the following (PMEST) facets:

- **Personality**: the main facet of the classification which describes what the resource is about.
- **Matter**: the material that the resource is about.
- **Energy**: the activities that take place in relation to the resource.
- **Space**: the location that the resource is about.
- **Time**: the time that the resource is about.

Although faceted classification originated in 1930s, it is still used in e-commerce application and auction web sites such as ebay[4]. For example, ebay users can narrow the scope of the item they are trying to find by specifying more than one facet such as (type, location, condition, buying format).

[3] http://www.boxesandarrows.com/view/ranganathan_for_ias

[4] A popular online auction website `http://www.ebay.com`

2.3 Tagging and Folksonomies

Web 2.0 has introduced a new type of application where users can assign keywords of their choice to web resources (such as web pages, photos or scholarly publications). In the web 2.0 world, these keywords are termed tags, and the process of assigning keywords to resources is termed tagging.

Tagging can be considered as a kind of assigning metadata to web resources. This can be mystifying if compared to the classification methods discussed earlier where the metadata creation process is carried out mostly by professionals rather than casual and untrained web users. Adam Mathes makes a distinction between three different metadata categories: professional, author and user-created metadata, and considered the tags to fall in the last category [67].

The main difference between the keywords created by professionals or authors on one side and the tags created by users on the web on the other side is that the tags are completely uncontrolled. The set of tags is managed by a number of users and each user is free to choose the tags he believes best describe the resource he wants to tag. Such a process can lead to a continuous creation of new tags as long as the tagging process is in place.

The tagging process became prevalent as a part of the web 2.0 wave, where users took an active role in publishing content on the web. There are four different parties/entities involved in the tagging process: actors (users), tags, resources and tagging systems [40, 108]. There exists a number of web sites built to publish contents that are fully created by users where tags are used to index and search the created contents. For example, the social bookmarking site Delicious, the publication sharing system Bibsonomy [5] and the image sharing site Flickr. Users of such systems can enter any tag of their choice to annotate resources. The aggregation of tags, users and resources is known as a Folksonomy.

[5] `http://www.bibsonomy.org`

The word “Folksonomy” is a concatenation of two words “folks” and “taxonomy”. The term was first coined by Thomas Vander Wal in July 2004 in a reply to a question posted in the Asylomar Institute for Information Architecture (AIFIA) closed list; the question was if there is a name for the informal social classifications generated in services such as Flickr and Del.icio.us.

Vandel Wal describes the folksonomy as [111]

> *"Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information."*

It is debatable that describing the folksonomy as taxonomy is rather inaccurate or incorrect, and some authors chose not to use the word taxonomy in their work at all such as in [37]; this is because the tagging process itself is considered as a categorization process [67, 37, 42] while the taxonomy is considered as a classification process. Despite the fact that both classification and categorization might be used synonymously, a clear distinction between both is provided in [52]. Classification assigns resources into distinct classes which have clear boundaries, that is opposite to the categorization where there are no clear boundaries defined. Folksonomies suffer from the lack of hierarchy, synonyms control and semantic precision but these reasons lead to a simpler tags authoring process which explains why folksonomy works [16]. Also, it is argued that folksonomies cannot be seen as a replacement or substitute for the professional classification approaches of librarians [80].

In this thesis, it is agreed that the term folksonomy can be misleading if considered as taxonomy replacement, firstly because the folksonomy on its own does not provide explicit hierarchical relationships between tags and secondly because it is more related to categorization because of the nature of assigning uncontrolled keywords to resources. However, the term folksonomy will be used in this thesis to refer to the well-established and defined data structure generated by users' interactions in tagging applications.

2.3.1 Folksonomy Representation

Folksonomies created in tagging applications via users' interaction on web 2.0 consist of three main entities: actors, tags and resources. Although the application used to create the folksonomy can be considered as a fourth entity (system), it is ignored in this thesis and it is assumed that only one system is dealt with.

A folksonomy can be modelled as a tripartite graph with hyper edges, which is also called a three-mode graph [69]. The vertices in this graph are classified into three disjoint sets $A = \{a_1, a_2, ..., a_k\}$, $C = \{c_1, c_2, ..., c_k\}$, $I = \{i_1, i_2, ..., i_k\}$ representing Actors (users), Concepts (tags) and Resources respectively. Each edge in this graph is a ternary association that connects a user, a tag and a resource, where no associations are allowed between elements in the same set. Accordingly, a folksonomy relation can be represented by a set of annotations $T \subseteq A \times C \times I$ that shows the relations between users, resources they create and the tags they use to annotate those resources.
The folksonomy tripartite graph is defined as follows:

$$H(T) = (V, E) \tag{2.1}$$

where $V = A \cup C \cup I, E = \{\{a, c, i\} | (a, c, i) \in T\}$
Although tripartite graphs can be easily used to describe folksonomies, the major problem with such representation is that they are not easy manipulated or analysed before being decomposed to bipartite (two-mode) graphs [112].

The bipartite graphs are similar to the tripartite graphs except that there are two sets of vertices instead of three. Moreover, the edges are regular in the sense that each edge connects two vertices. Any folksonomy tripartite graph can be decomposed to three bipartite graphs; Actor/Concept (AC graph), Concept/Objects (CO graph) and Actor/Resources (AI graph).

Decomposing tripartite graphs can be achieved using different methods; in the field of social network analysis (SNA), the 'Projection' method is one solution to the problem. Also, the 'aggregation' methods proposed in [66, 18], such as Distributional and Col-

laborative aggregation, can also be solutions to the problem. All those methods are based on the same idea of removing one of the 'modes' and modelling it as weights on the resulting two-mode graph. However, each method calculates the weight differently. For example, the SNA's 'Projection' method of building the AC graph uses the count of the resources annotated by the user and the tag represented by each edge as weights. However, in the 'Distributional' method, the weights are calculated differently so that the information content (entropy) associated with the set membership relationships between the two-modes are considered.

2.3.2 Similarity Measures

In general, the similarity between two entities is normally measured by comparing the values of their corresponding attributes. Hence, the similarity directly depends on the values represented by each attribute. On the other hand, folksonomies link entities of three different sets: users, tags and resources. Such links can be analysed to measure the similarity between entities in the same set based on their relationships to the entities in the other two sets. For example, similarity between two tags can be calculated based on the number of resources annotated using both tags, or based on the number of users who used both tags to annotate resources. The similarity calculated using folksonomy is independent on the attributes of the entities and it represents the similarity as a function of the tagging activities performed by the folksonomy users.

Several statistical methods exist in the literature to calculate the similarity between entities [66], mostly based on co-occurrence analysis and can be explained as follows. Assume that there exist a feature vector X that represents an entity (user or resource) x, such that each element in X represents a weighted relationship w_{xy} between the entity x and tag y. Assuming a binary representation, the value of $|X|$ is equivalent to the number of tags directly attached the entity x. For example, assume that the tag space contains only three tags t_1, t_2 and t_3. The vector $X = [1, 1, 0]$ of a place x indicates that the two tags t_1 and t_2 are associated with the place, and the total number of tags

used to describe that place $|X|$ is 2 tags. The similarity measure is represented by the symbol σ and can be calculated using methods such as Cosine, Dice or Overlap. More information about similarity measures can be found in [66].

In this thesis, the Cosine similarity is used to measure the similarity in several parts of the analysis. It measures the similarity between two vectors by calculating the Cosine of the angle between them and derived from the following Euclidean dot product formula:

$$X_1 \cdot X_2 = \|X_1\| \, \|X_2\| \, cos(\theta) \tag{2.2}$$

Hence, the Cosine similarity is calculated as follows:

$$\sigma(x_1, x_2) = cos(\theta) = \frac{X_1 \cdot X_2}{\|X_1\| \, \|X_2\|} = \frac{|X_1 \cap X_2|}{\sqrt{|X_1| \, . \, |X_2|}} \tag{2.3}$$

2.3.3 Broad versus Narrow Folksonomies

Folksonomies can be classified into two types according to the way they are used in the tagging applications: broad and narrow folksonomies [109]. The main difference between both types is the way the resources are linked to tags and users. In broad folksonomies, the same resource can be tagged by a big number of users (for example bookmarks on Del.icio.us), while in narrow folksonomies, each resource is tagged with a small number of users and in most cases by one user who created the resource (for example photos on Flickr).

Broad Folksonomies

Broad folksonomies exist when the same resource is tagged by many users, and every user can tag the resource using their own set of tags [109]. Figure 2.2 shows a visualisation of an example of the broad folksonomy. There are five groups of users (A,B,C,D,E

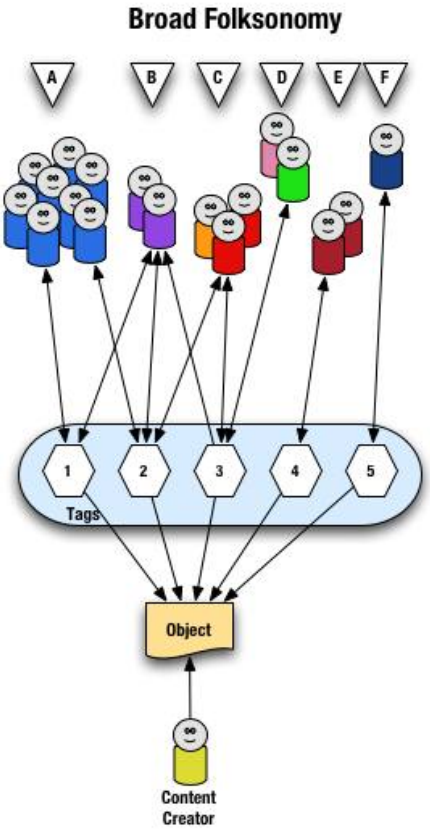


Figure 2.2: An example of a broad folksonomy [109].

and F), and each group is connected through an arrow to one or more tag; tags are represented by numbers from 1 to 5. Each group describes resources/objects using a different set of tags. This type of tagging usually leads to creating a folksonomy with power law distribution in which a few popular tags are frequently used while the rest of the tags are used only a few times. More details about the power law distribution are presented later in this chapter.

Narrow Folksonomies

Contrary to the broad folksonomies, narrow folksonomies exist when a resource is tagged by one or a small number of users. Usually, this happens in applications where the resources are not easily searchable or there is no other way to describe resources

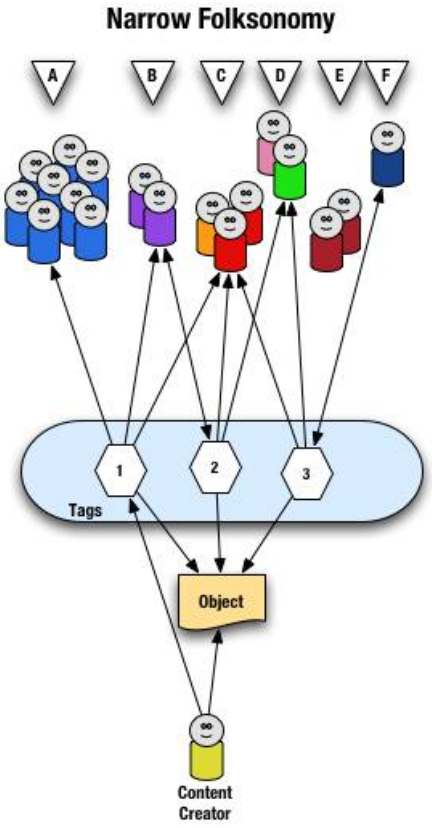


Figure 2.3: An example of a narrow folksonomy [109].

using text, for example Flickr, where photos are tagged only by their publishers. In Figure 2.3, only two groups of users (B and F) are annotating the resource while the rest of the groups (except group E) retrieve the resource by using the tags provided by the groups B and F. An example of this scenario is a blog post where the author provides tags for his article to be searchable by other users. Another example is Twitter[6] in which every tweet (a micro post of 140 character) can be annotated with hash tags to be searchable.

[6] `http://www.twitter.com`

2.3.4 Power Law Distribution

In tagging applications, where a broad folksonomy approach is followed, there are a small set of popular tags that are frequently used by all users while the rest of the tags are used a few times. Plotting the distribution of the tags' usage frequency shows a graph with a long tail known as a power law distribution graph [109, 67, 80, 42]. The tags' usage distribution in broad folksonomies has been shown by [42] to follow a power law evident on a data set from Del.icio.us that contains around 18,000 tags.

The power law distribution is defined by Newman [72] as being: "When the probability of measuring a particular value of some quantity varies inversely as a power of that value, the quantity is said to follow a power law". Examples of distributions that follow a power law are: the sizes of earthquakes, the frequencies of words in most languages and citation of papers. Power law distribution curves have a characteristic which, when plotted on logarithmic axes, the resulting graph shows as almost a straight line as shown below:

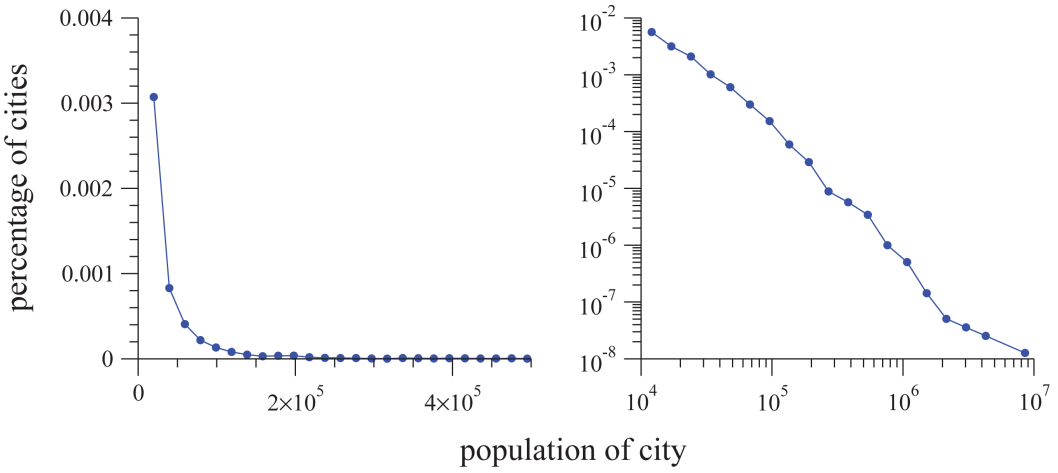


Figure 2.4: Power law distribution function [72].

2.4 Ontologies

The term Ontology means in Greek “being or existence”, but originally it comes from the Latin word ‘ontologia’. Ontologies became a popular research topic in the early 1990s. They have been the focus of several artificial intelligence (AI) research communities, such as knowledge engineering, natural-language processing and knowledge representation. More recently, ontologies have also been utilised in other fields, such as intelligent information integration, information retrieval and knowledge management [23].

The AI community was attracted to ontologies as they believed that ontologies could be used to represent formal knowledge needed to allow communication between knowledge based systems. In particular, knowledge based systems can communicate to answer the same question even if the knowledge concepts are modelled differently inside individual systems [41]. Similar usage of ontologies has been promised to the knowledge management community in general which can be described as “a shared and common understanding of a domain that can be communicated between people and application systems”.

As ontology is being used in different domains, different definitions exist describing the different aspects of using ontologies in each domain. Gruber [41] has defined ontology as “explicit specification of a conceptualization” and more recently defined ontology as a collection of concepts, relationships, and other elements that are critical to describe a domain [40].

Another definition offered by Jarrar and Meersman [38] is that it is “a branch of knowledge engineering, where agreed semantics of a certain domain is represented formally in a computer resource, which then enables sharing and inter-operation between information systems”. De Troyer et al. [25] defined ontologies as “concepts in a domain as well as relationships between these concepts and the terminology used”. A more comprehensive guide to ontologies can be found in [38].

According to Uschold [107], there are three different goals of using ontologies: communication between people and organizations, interoperability between machines, and improving systems engineering. The level of formality of an ontology is determined by its goal. For example, ontologies needed for communication between people can be informal while ontologies used by machines for interoperability need to be expressed in a formal approach. In this thesis, ontologies used by machines are the focus. Hence, an overview of ontology languages is provided in the next section.

2.4.1 Languages for Representing Ontologies

Different languages exist to support expressing ontologies in a formal way, such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). An ontology can be expressed via a set of assertions called statements or triples, where each statement is made up of three parts: subject, predicate, and object. A statement describes the subject using a relation to the object. For instance, the statement "John knows Rob" contains a subject "John", a predicate "knows", and an object "Rob" connected to the subject via the predicate. The RDF language defines a standard way of writing such statements in several formats. The three most popular formats are RDF/XML, the Terse RDF Triple Language (Turtle), and N-Triples. As the name suggests, RDF/XML format is based on the Extensible Markup Language (XML) as a standard supported by almost every platform. Hence, the RDF/XML is used in the interoperability scenarios. The Turtle format is not XML-based and is more human-friendly. The N-Triples format is a simplified version of Turtle but with fewer features. OWL is considered as an ontology standard by W3C. It can be seen as an extension to the RDF/XML with more expressiveness features and with vocabulary designed to model ontologies rather than a general triple/statement model supported by the RDF. OWL has three different versions: OWL Lite, OWL DL and OWL Full. More details about RDF and OWL can be found in [45].

2.5 Discovering Folksonomy Emergent Semantics

Folksonomies represent users' interaction on the web by capturing the links between tags, users and resources. Such a structure allows the semantics embedded in the folksonomy to emerge. The co-occurrence frequency of tags, resources and users is a vital characteristic of folksonomies [49, 100, 105, 43] which is utilised to discover embedded semantics, where entities are anticipated to be semantically related if they co-occurred together with a high frequency.

Peter Mika [69, 70] is one of the first researchers who addressed the problem of discovering folksonomy semantics. Mika represented the folksonomy as a tripartite graph with hyper edges, where nodes represent three distinct sets of tags, users and resources and each edge connects three nodes such that no nodes from the same set are allowed to be connected. He applied several Social Network Analysis (SNA) methods [112] in the folksonomy graph in order to build a lightweight ontology of tags (concepts) based on the co-occurrence with users and resources in the folksonomy. Other early research work was carried out by Begelman et al. [9], in which a weighted undirected graph is used to represent the tags. The weights represent the strength of the relation between tags and are calculated based on the co-occurrence frequency. Spectral clustering is used after that to induce clusters of related tags. Similar to Mika's work, the induced relationships among tags are general and do not represent specific semantic relations.

Schmitz [93] focused on building a taxonomy-like hierarchy of tags from folksonomies, where a probabilistic model for subsumption is used to discover the parent-child relationships. The hypothesis behind this method is that tag a subsumes tag b if the probability of appearance of a given b is above a certain threshold and the opposite is lower. However, considering the relationships induced by this method as a "sub-class-of" may be inaccurate as this method builds a hierarchical representation of tags based on the way they are used and this does not guarantee that every subsumption relation can semantically represent a "sub-class-of" relationship between two concepts. For example, the results of applying this method on Flickr tags [93] resulted in subsumption

relationships between tags e.g. (glass->blow, glass->stained), which obviously do not represent a "sub-class-of" relations.

As a common characteristic of broad folksonomies, tags follow a power law distribution. This was confirmed by Haplin et al. [42] in their study of the dynamics of tagging systems over a dataset from Del.icio.us. The study showed that high frequency tags that follow a stabilised power law distribution describe a general consensus on the topic of the resource. An empirical examination of concepts hierarchies built using a number of heuristics along with the information value of the tags, such as the number of resources linked to a tag, was presented in this study.

Heymann et al. [49] proposed an algorithm that utilises the SNA *betweenness* centrality measure to build concept hierarchies from tags. The idea behind the algorithm is that tags with higher centrality values represent more abstract concepts. Hence, those tags are moved to a higher level in the hierarchy.

Zhou et al. [115] employed an unsupervised model to automatically derive hierarchical concepts from tags. The deterministic annealing (DA) clustering is used to break down the tags into "effective clusters" whose semantics can be generalised by some specific tags, named as "leading tags". Hierarchical semantics was deduced through the leading tags.

A novel approach for learning tags hierarchies based on hybrid heuristic rules and a concept-relationship acquisition algorithm was presented in [105]. The evaluation of the proposed approach showed a high precision and recall rate. However, this cannot be generalised as the dataset used for evaluation was relatively small in size.

As a useful guideline for using the co-occurrence methods to extract folksonomy semantics, a survey study of several co-occurrence methods was presented by Cattuto et al. [19], where the methods were tested on a large-scale dataset from Del.icio.us and the induced semantics were compared to the hierarchy of Wordnet. The study suggested that the choice of the co-occurrence method should be based on the application, as

methods such as resource context similarity perform better in discovering synonyms while other methods such as FolkRank [50] are better in building concept hierarchies.

The above research exploits different approaches, mostly statistical-based, in order to build lightweight ontologies from folksonomies to represent emergent semantics. One possible problem in such approaches is that the popularity of a tag can be mistaken for generality which can produce inaccurate hierarchical relationships between concepts. Popular tags, with high frequency of usage, can represent concepts that are too generic.

Plangprasopchok et al. [78] tackled this problem by using additional information to induce global hierarchies from personal user-specific hierarchies on Flickr. Graph and lexical similarities were used to merge the individual users' hierarchies to build a taxonomy of concepts. This work was built around a feature offered only by Flickr, user-specific hierarchies, which limits the approach to work with folksonomies collected from other data sources. Also, as highlighted by the authors, a key issue with their approach is that only a small percentage of users apply such organisation to their content. A more generic approach of using additional information in the ontology building process was also carried out by Kim et al. [58, 57], where a folksonomy contextualisation method based on Formal Concept Analysis was proposed to build conceptual hierarchies from tags in the blogosphere. This approach showed that concepts hierarchies of context-centric shared collections of tags can be deduced by utilising the references among the blogs.

This section reviews different approaches used to extract the semantics embedded in folksonomies. However, the discussed approaches target only the domain-independent emergent semantics. The emergent semantics are represented by lightweight ontologies which, arguably, have two forms: a graph of concepts in which the degree of relatedness is represented by weights, or a taxonomy of concepts in which concept hierarchies are deduced from the folksonomy structure. The next section reviews the research on extracting domain-dependent place semantics from folksonomies.

2.6 Extracting Place Semantics from Folksonomies

Place semantics can be extracted from collaborative and social mapping applications. Semantics associated to place concepts are more specific. In particular, a geographic place is associated with spatial properties, representing its location, spatial extent and spatial relationships between other entities in space, and non-spatial properties, qualifying other properties, such as its type, name and purpose. Recently, collaborative mapping web applications have emerged where users are contributing to the development of web gazetteers as well as providing detailed descriptions of places and related information. A prominent example of a web gazetteer is *GeoNames*, currently containing around 10 million[7] geographic names. Also, some research has focussed on the problem of building gazetteers from user generated data on Web 2.0 [82, 79, 81].

2.6.1 Types of Place Semantics

On the semantic web, place name (or toponym) ontologies are employed to facilitate the utilisation of gazetteers to support geographic information retrieval tasks, such as disambiguation and expansion of terms in search engine queries [39, 56, 99]. An ontology of place names is defined as a model of terminology and structure of geographic space and named place entities [26, 2]. It extends the traditional notion of a gazetteer to encode semantically rich spatial and non-spatial entities, such as the historical and vernacular place names and events associated with a geographic place [76]. In addition to place qualification using place type categorisation, qualitative spatial relationships, commonly used in search queries, are also modelled to relate place instances.

Functional differentiation of geographical places, in terms of the possible human activities that may be performed in a place or place affordance, has been identified by Relph [84] as a fundamental dimension for the characterisation of geographical places. For Relph, the unique quality of a geographical place is its ability to order and focus human

[7] `http://www.geonames.org/about.html`

intentions, experiences, and actions spatially.

It has been argued that place affordance is a core constituent of a geographical place definition, and thus ontologies for the geographical domain should be designed with a focus on the human activities that take place in the geographic space [59, 29]. The term "action-driven ontologies" was first coined by Camara et al. [17] in categorising objects in geospatial ontologies. Affordance of geospatial entities refers to those properties of an entity that determine certain human activities. In the context of spatial information theory, research has attempted to study and formalise the notion of affordance [86, 60, 96, 94, 83, 92]. The assumption is that affordance-oriented place ontologies are needed to support the increasingly complex applications requiring semantically richer conceptualisation of the environment. Realising the value of the notion of affordance for building richer models of geographic information, the Ordnance Survey (the national mapping agency for the GB) proposed its utilisation as one of the ontological relations for representing their geographic information [44] and made an explicit use of a "has-purpose" relationship in building their ontology of buildings and places [8].

2.6.2 Extracting Place Semantics

Early research in this area was carried out by Rattenbury et al. [81], where the feasibility of automatically extracting event and place semantics from Flickr tags was tested. The research presented in this thesis exploited the geo-tagging feature of Flickr, where images are annotated with the spatial location of where they are taken. Burst-analysis and scale-structure identification techniques were used to recognise the spatial and temporal tagging patterns of event and place semantics. Although the results showed a successful extraction of places and events from the tags, there were no semantic relations deduced between the extracted concepts.

[8] `http://www.ordnancesurvey.co.uk/oswebsite/ontology`.

There is other research on automatic gazetteer building from folksonomies such as in [79], where an algorithm was proposed to analyse several online collaborative sites to extract a geographic gazetteer. Places in the extracted gazetteer were organised under categories which use a simple hierarchy structure.

Intagorn et al. [51] proposed an approach for learning geospatial concepts and relations from Flickr. The proposed approach identifies tags representing place names via consulting GeoNames[9]. This was followed by a data cleaning process to remove the noise and resolve disambiguation of place names. Finally, hierarchical relationships were induced using a probabilistic subsumption method.

Location Sharing Applications (LSAs) are becoming more popular every day due to the ubiquity of GPS-enabled smartphones. Examples of such applications are Twitter, Foursquare[10], Facebook Places[11] and Google Latitude[12]. LSA allow users to record activities such as check-ins in Foursquare, which generates highly dynamic and real-time data. Tang et al. [102] distinguished between two types of LSAs, social-driven and purpose-driven. The first is built to support location sharing within social networks, such as Twitter, while the latter is built for a special purpose such as collecting place data, for example OpenStreetMap. They showed that the type of LSA affects users' decisions about what information to share. In social-driven LSAs, which are more related to the focus of the research in this thesis, the motivation scenarios always emphasize the social aspects of location sharing. For example, Foursquare users share their check-ins to places to let their friends know where they are; they are not sharing the information, for example, to build a complete map of places. Social information, such as the user check-ins at places, is a valuable source of information to extract place semantics.

An interesting piece of research was carried out by Cranshaw et al. [22] to build a

[9] http://www.geonames.org
[10] http://www.foursquare.com
[11] http://www.facebook.com
[12] http://www.google.com/latitude

model of place that represents the character of life (livehoods) rather than the traditional municipal organizational units, such as neighbourhoods. An algorithm was presented to process a large-scale dataset downloaded from Foursquare. The algorithm utilised a spectral clustering approach to discover the local urban areas from the social check-in data. The authors presented a successful process of grouping places based on the pattern of users' movements.

Normally, the process of extracting semantics from folksonomies requires a pre-processing process to clean the tags. Quality problems, such as spelling mistakes, may exist in the tag space which is caused by the uncontrolled input approach provided by the social bookmarking applications, where no input validation is utilised. Hence, a pre-processing cleaning process is suggested by researchers, such as [108, 77, 51], which basically involves utilising stemming algorithms to identify the different forms of the same tag and using lexical resources such as online dictionaries to check the spelling. More details about the tags cleaning are provided in Chapter 4. On the other hand, the structure of the place resources in geo-folksonomies creates further complexity with respect to the pre-processing process. A basic place resource contains thematic attributes such as place name and type, and spatial attributes such as the location of this place. The thematic attributes inherit the same problems evident in the tags due to using the same uncontrolled input approach, while the spatial attributes are usually imprecise and inaccurate as they are acquired using a map-based interface which relies on the user being able to identify and digitise a precise location on a map.

The place semantics extraction approaches discussed in this section target simple place model representation. For example, the model represented in [81] produces a controlled vocabulary of place names and events, lacking the existence of any semantic relationships while in [79] a richer place model is used to capture the hierarchical relationships between place names in a taxonomy-like structure. An interesting model of place was represented in [4] which emphasized modelling place types and services offered by places. Although the model can be relevant to the work presented here,

the semantics extraction approach targeted a different structure of data collected using GPS devices. Another line of research which focuses on the LSAs utilises the social interaction data to understand the dynamics of places, such as [22], where the employed place model is still simple and represented by a graph structure connecting places with similar dynamics. Building a rich model of place which can capture both places and their related social information from geo-folksonomies will complement the work in this research area.

2.7 Limitations

The work presented in this thesis targets extracting place semantics from geo-folksonomies. Limitations of the approaches in the current literature are summarised as follows:

The need for specific geo-folksonomy cleaning approaches

Folksonomies are user-generated data created by users' interaction and collaboration using social bookmarking applications. Typically, such applications are designed to acquire the input from users in free-text format to simplify the user interface. As a result, the generated folksonomies contain an uncontrolled vocabulary of keywords (tags) with several problems such as polysemy (a word which has multiple related meanings) and synonymy (different words that have identical or very similar meanings) [37]. Geo-folksonomies contain place resources which are a specialised type of web resources that represent places in the real world through thematic and spatial attributes. The representation of the place resources, especially the spatial dimension, requires the folksonomy cleaning approaches to address the inaccuracy of the spatial data acquired from users along with the existing quality problems inherited from folksonomies. Thematic attributes such as place names are free text entered by users which, unlike tags, can be made up of multiple words. Moreover, the spatial attributes such as location of places, are acquired using a map-based user interface which is subject to

imprecision. Redundant place resources that refer to the same place in the real world are a problem that might affect the quality of geo-folksonomies.

The need to model user-generated place semantics

Semantics extracted from folksonomies are normally represented using a simple lightweight ontology model, where concepts of the ontology represent the frequently used tags, and a relation between two concepts is created if the tags representing those concepts co-occur frequently. However, in geo-folksonomies, the lightweight ontology model, which normally represents simple semantic relationships between instances of one concept, may not be sufficient to capture the domain-specific place semantics extracted from geo-folksonomies that requires a richer representation. The existing place models need to be investigated to check the possibility of being adopted or extended to model the required place semantics.

The need for devising an approach to capture the place semantics

The approaches used to extract general semantics from ontologies are generally based on co-occurrence analysis with the assumption that two tags or terms are semantically related if they frequently co-occurred together. However, extracting domain-specific place semantics requires further approaches to identify the place-related concepts, such as place affordance, as well as infer the different semantic relations linking the place concepts.

2.8 Summary

Enormous amounts of data are generated on the web due to the users' interaction and collaboration on web 2.0. Social and collaborative applications allow users to collaborate and provide information. Such applications allow users to describe their gener-

ated content using single keywords called “tags”. The aggregation of tags along with the users and the annotated resources create a user-generated index known as “Folksonomy”.

Folksonomies have been the focus of research as they contain embedded semantics and reflect users’ understandings about the annotated resources, which can be different to how these resources are formally described. There are two main ways to extract embedded semantics from folksonomies; the first is to extract general semantics that are not domain specific, and these are called “Emergent Semantics”. The second approach is to extract domain specific semantics such as place semantics.

The emergent semantics are characterized by a lightweight ontology of concepts and simple relationships, and each relationship can represent either related-to or subsumption relation between two concepts. Most of the approaches proposed to build emergent semantics from folksonomies are based on statistical co-occurrence methods, where identifying the concepts and relationships is based on the way the tags are co-occurred with users and resources.

Research has targeted extracting place semantics from folksonomies, where the place semantics are in the form of a hierarchical gazetteer of place names, a set of place and events, or a set of clustered places that have common social dimension. Folksonomies that contain geo-tagged resources (geo-folksonomies) can be a valuable source of information to build a more comprehensive place model that represents the semantic relations between concepts such as places, place affordance and user activities.

The next chapter presents the research conducted in this thesis to provide a framework and place ontology design to extract place semantics from geo-folksonomies, while the two chapters thereafter discuss the framework in detail. There are of course more specific links between existing research and the work in this thesis and these will be discussed throughout the thesis when and where they become relevant.

Chapter 3

Framework and Ontology Overview

This chapter provides an overview of the proposed framework for inducing place semantics from geo-folksonomies. The framework is based on a semantic model that captures particular aspects of place semantics related to types and activities. A discussion of the proposed framework is presented in Section 3.1. The design of the place ontology is provided in Section 3.2. Finally, a summary of the chapter is provided in Section 3.3.

3.1 A Framework for Inducing Place Semantics from Geo-Folksonomies

The type of semantics targeted to be extracted from the folksonomy determines the design of the ontology extraction process. The extracted semantics can be in the form of lightweight ontology or domain ontology. The process of extracting lightweight ontologies from folksonomies is addressed by several research works such as in [108] where an abstract 5-step process is provided as follows:

1. Cleansing and preparation of tags, where the problems caused by the uncontrolled user input are addressed, such as spelling mistakes and stop words.

2. Statistical analysis of folksonomies, where similar tags are grouped into clusters and concept hierarchies are induced from the co-occurrence relations between

the tags and users/resources.

3. Exploiting online lexical resources, where the concepts extracted from the previous step are validated using online lexical resources such as Google and Wikipedia. This approach is capable of validating new keywords such as 'folksonomies' which may not be included in normal dictionary resources.

4. Linking to ontologies and semantic web resources, where the concepts obtained in the previous step can be enriched by trying to establish mappings to elements in other ontologies.

5. Mapping and matching approaches, where it is suggested that the formal classification theory of [36] can be used to map the labels of existing classifications with the concepts obtained from the folksonomy.

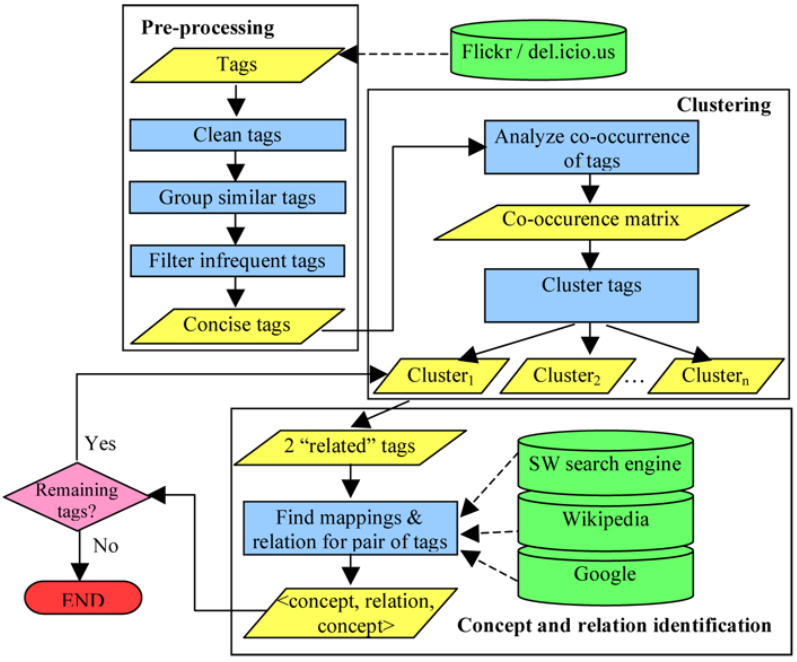


Figure 3.1: The process of building lightweight ontologies from folksonomies [100].

The abstract process above provides the essential steps to guide the design of extracting lightweight ontologies from folksonomies. This process is realised by the framework

provided by Specia et al. [100], aiming to extract a lightweight ontology from Flickr and Delicious tags. The design of the framework is shown in Figure 3.1.

The framework provides three stages of processing folksonomies: the pre-processing stage where the tags are cleaned to remove misspelled and unusual tags; The clustering stage where tags are clustered into groups of similar tags based on their co-occurrence with users and resources and finally, the concept and relations identification stage, where tags that represent concepts are identified and the relationships between the tags are discovered using external ontologies and online resources such as Google and Wikipedia.

In this thesis, a framework is provided that follows the same design principles of the works discussed above. The goal of the approach proposed here is to derive an understanding of implicit place semantics from geo-folksonomies. Starting with "raw" folksonomy resources, the framework involves three main stages: a) folksonomy pre-processing, b) tag resolution, and c) semantics association and ontology building.

A particular characteristic of geo-folksonomies is the possible redundancy in place resource creation and the resulting fragmentation of folksonomy relationships that can affect the quality of the analysis. The first stage in the proposed approach thus involves two main tasks: a) cleaning the tags to filter out noise such as stop words, and b) clustering of place resources and the reconstruction of the folksonomy structure.

The tag resolution stage involves domain-dependent analysis tasks for resolving and isolating tags that refer to domain concepts. The approach proposed here is to utilise existing domain ontologies for matching domain concepts. The process involves identification and building place type and human activity ontology bases and using these as reference sources for matching against the tag collection.

The final stage is the semantics association and ontology building stage, where the individual identified domain-dependent tag collections are first analysed to derive relationships and create ontologies using the folksonomy structure. A place type sub-ontology

and a place activity sub-ontology are created to represent a folksonomy-specific view of these concepts. A tag integration process is then applied to link the tags from both sub-ontologies using the inherent folksonomy relationships. The resulting structures are associated with the clustered place resources from the first stage and used to populate the place ontology. Further semantic analysis can be applied to the tag collection. Here, a sentiment analyser is developed to estimate a sentiment score for each place resource.

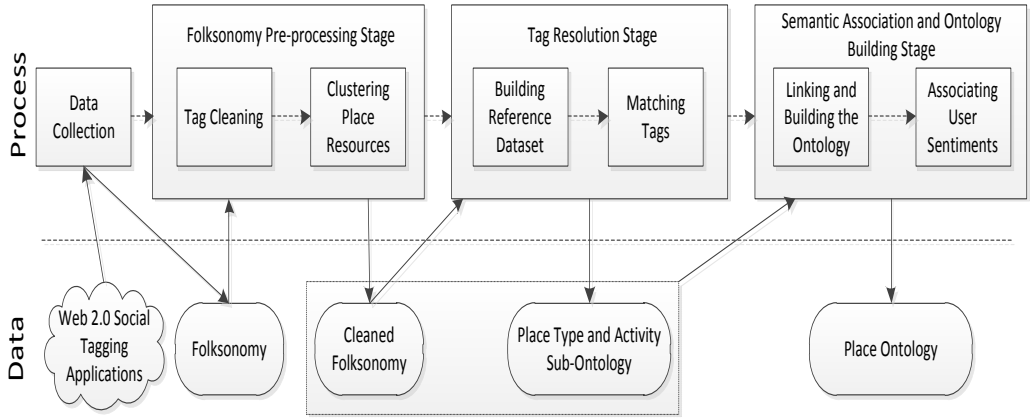


Figure 3.2: The process of building place ontology from folksonomies.

An outline of the framework is shown in Figure 3.2 and the different stages are described in more detail in Chapters 4 and 5. The following section describes in detail the model of the place semantics used in this work.

3.2 Modelling Place Semantics

Places, whether natural or man-made, can normally be associated with specific functions, services, economic activities or other human activities that they provide to individuals. This dimension of a geographical place definition is typically evident in

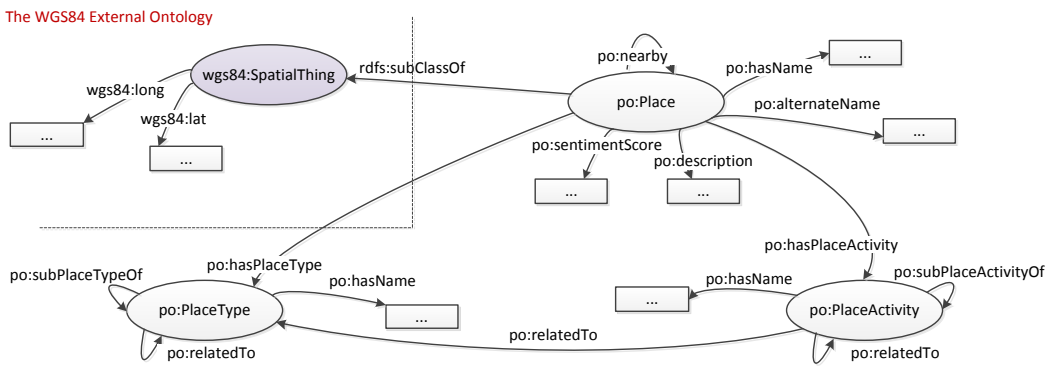


Figure 3.3: Place ontology represents the place semantics captured from folksonomies.

catalogues of place type specifications produced by national mapping and other geographical data collection agencies, and are used for the purpose of classification of place entities. For example, the following descriptions are parts of the definitions associated with place types in the Ordnance Survey Mastermap specification[1]: *Amusement park; a permanent site providing entertainment for the public in the form of amusement arcades, water rides and other facilities*, and a *Comprehensive school; a state school for teenagers, which provides free education*. Classification of economic activities of business establishments is often used for place type categorisation. For example, national bodies such as the Office of National Statistics of the UK (ONSUK) [2] and Eurostat (the statistical office of the European Commission), produce classifications and definitions of economic activities for classifying business establishments by the type of economic activity in which they are engaged[3]. Notably, a business place can be associated with a number of services, where some of these are principal activities that determine its primary classification while others are ancillary activities (such as

[1] `http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap`

[2] `http://http://www.statistics.gov.uk`

[3] See The Standard Industrial Classification of all Economic Activities (SIC), `http://www.statistics.gov.uk/methods_quality/sic/downloads/sic2007explanatorynotes.pdf`

accounting, transportation, purchasing, and repair and maintenance) that exist solely to support the principal ones.

Whereas these formal classifications of place types and services are useful and required for many contexts, they are general and are not intended to capture any specific experiences of users in a place. There is an emergent need for recognising and sharing the experiences of people in geographic places, evident from the ever-growing volumes of data and applications that allow users to check-in and tag places [21, 91]. Such experiences are associated with particular instances of geographic place and may not be generalised.

Hence, in this work a model of place is adopted where a geographic place can be associated with possibly multiple place types and place activities. Place types and place activities may themselves form individual subsumption hierarchies. A place type may be associated with more than one type or activity and vice versa. A distinguishing characteristic in this model is that it allows for a specific place instance to be associated with an activity that may not be derived from its association with a specific place type.

Hence, for example, a specific instance of a school may be associated with several place types, such as primary school, public school and nursery, from which it can derive activities, such as learning and teaching, but can also be associated with activities, such as dancing, weight training, and adult education, where it offers external services to the community after school hours. The former list is derived from the association with a particular place type, but the latter list may come from direct annotation by users of the place.

The proposed place ontology is shown in Figure 3.3. The model contains three concepts: *Place*, *Place Type* and *Place Activity* as well as properties and inter-relationships between them. The spatial location of a place is modelled by extending the WGS84 *SpatialThing* concept to inherit the spatial properties *lat, long*. A *Place* has a *name* and possibly 0 or more *alternate names* and may be involved with different types of spatial relationships with other place instances. Explicit modelling of qualitative spatial

relationships are adopted in various proposals of place ontologies such in SPIRIT [56], TRIPOD [1] and Geonames. One example of such relationships, namely, proximity or *near by*, is shown in Figure 3.3.

The model extends previous proposals, for example, that of the Ordnance Survey Building and Place ontology (OSBP) [4], where a similar notion of place activity is explicitly modelled and associated with a place type through a relationship "has-purpose". The difference in the research presented in this thesis is that a place concept is introduced which also exhibits separate relationships between types and activities. In addition, inter-relationships between place types and place activities were not modelled in the OSBP ontology.

The design of the place ontology is implemented using OWL. All classes and properties are qualified with the prefix **po**[5]. Note that, in general, the associations in this model are dynamic as a result of the accumulation of users' experiences and annotations. Hence, the relationships $po : hasPlaceType$, $po : hasPlaceActivity$ and $po : relatedTo$ would be time-stamped. However, the time dimension is out of the scope of the current work and is the subject of future research.

3.3 Summary

A framework is proposed in this chapter to induce place semantics from geo-folksonomies. The framework involves three main stages of processing geo-folksonomies: a) folksonomy pre-processing stage where the tags and place resources are cleaned to enhance the quality of the data; b) tag resolution stage where external resources are consulted to identify place-related concepts represented by the tags and c) semantic association and ontology building stage where the semantic relations between the identified concepts are inferred. Moreover, a semantic model of place was also proposed in this chapter,

[4] http://www.ordnancesurvey.co.uk/oswebsite/ontology/

[5] http://cs.cardiff.ac.uk/2010/place-ontology#

where particular aspects of place semantics related to types and activities are captured.

The proposed framework is discussed in detail in the next two chapters; the geo-folksonomy pre-processing stage is discussed in Chapter 4 while the tag resolution stage, and the semantic association and ontology building stage are discussed in Chapter 5.

Chapter 4

Folksonomy Pre-processing

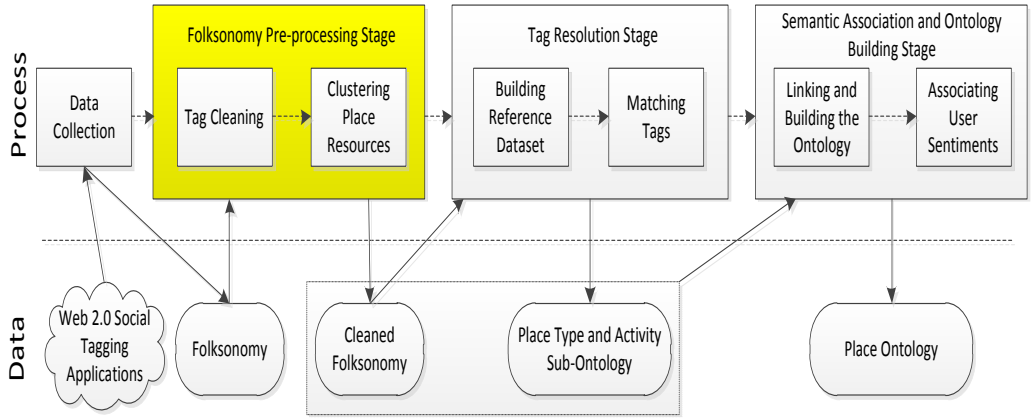


Figure 4.1: The process of building place ontology from geo-folksonomies.

Geo-folksonomies contain tags and place resources created by users. The uncontrolled data acquisition approach provided to users by the collaborative mapping applications can affect the quality of tags and the accuracy of place resources. In this chapter, a sample of geo-folksonomy tags is studied to identify the potential problems and a tag cleaning process is designed and discussed in Section 4.1 that addresses the identified problems.

Moreover, a place resources clustering process is discussed in Section 4.2 that addresses the imprecision problems in place resources which are evident in: a) the im-

precise place locations due to the digitization of the map-based interfaces provided by the collaborative mapping applications and b) the imprecise and vernacular place names used by users. Such problems lead to misclassification and duplication of place resources in geo-folksonomies.

The methods proposed in this chapter are tested on a geo-folksonomy data set collected from Tagzania and the results are discussed in Section 4.3. An evaluation of the provided work is presented in Section 4.4. Finally, a summary of the chapter is presented in Section 4.5.

4.1 Tag Cleaning

A set of arbitrary queries is used to explore the tags in the dataset in order to identify the problems that might exist in the tags. Table 4.1 lists the identified problems along with example tags of each problem. Generally, social tagging applications do not util-

Problem	**Example Tags**
Stop words such as articles and pronouns	a, an, the, we
Dialect	center, centre
Morphological forms of the same word	shop, shops, shopping
Numbers	20, 505, 2007
Synonyms	chair, seat
Homonyms	mean
Abbreviations	UK, EU
Concatenated terms	CardiffUniversity, London_Eye
Non-alpha-numeric letters	"ball
URLs	www.google.co.uk

Table 4.1: Sample of possible problems in the tag collection.

ize any kind of input validation on the tags provided by users. Such uncontrolled

user input can explain why tags are associated with problems, such as having stop words and sometimes being misspelled. Such problems can be avoided if the user interface is implemented differently, for example, a dictionary can be used to check the spelling before saving the tags. However, the user interface validations in social tagging applications are abandoned to encourage users to supply tags with minimum interaction. Other problems, such as abbreviations, synonyms and homonyms, require special methods for linguistic and semantic analysis.

Another problem identified is that some users try to use tags which consist of more than one word. Normally, users are aware that a tag by definition is a single word, thus they either use special characters to concatenate multiple words into one tag (e.g. London_Eye), or they concatenate the words directly by using naming conventions, e.g. Pascal casing such in (LondonEye). Other users wrap a whole sentence in double quotes, possibly assuming that the social bookmarking application will use it as one tag. For example, a place tagged with the following sentence "this is my house" will be split into the following tags *'"this'*, *'is'*, *'my'*, *'house"'*. The resulting set of tags include the stop words (is, my), term with non-alpha-numeric letter (house"), and a complex problem of non-alpha-numeric letters and stop word in the same term ("this). Hence, a process of tag cleaning is needed to isolate such problems and prepare the tags for processing.

In this thesis, a process for cleaning tags is proposed. The following section discusses two popular methods from the literature used in the context of folksonomy analysis; *Stemming and Lemmatization* and *Text Similarity*, and then the proposed cleaning process is discussed in Section 4.1.2.

4.1.1 Approaches to Tag Cleaning

In the literature on folksonomy analysis, part of tag cleaning process involves identifying redundant tags. *Stemming and Lemmatization* and *Text Similarity* are two ap-

proaches that are commonly used in cleaning tags [100, 14, 108, 5]. These are discussed below:

Stemming and Lemmatization

Stemming and Lemmatization are different techniques used to reduce inflected and derived words to their base or root form [65]. Stemming algorithm works by removing suffixes. For example, the words "Fisher" and "Fishing" are stemmed to the same word "Fish" [74]. Stemming algorithms are language dependent, as each language has its own suffixes [65]. The Porter stemming algorithm is one of the most widely used English language stemming algorithms and is utilized in the presented research work as discussed later in this chapter.

Although stemming can help identifying a tag that has different morphological forms, it is important that the semantic meaning of the tag is not lost in the process. There are two common problems related to stemming: *under-stemming* and *over-stemming* problems. Under-stemming happens when stemming lets two words referring to the same concept have different stems, for example *divide* and *division* are stemmed to *divid* and *divis* respectively. Over-stemming takes place when two words with different meanings are stemmed to the same root, for example the words *new* and *news* are stemmed to *new*.

On the contrary, lemmatization algorithms do not remove the suffixes. Instead, the word is transformed to its lemma, usually using a dictionary. For example, the word *good* is the lemma of the word *better*. Some words can have more than one lemma depending on how they are used in a sentence. Hence, lemmatization algorithms involve more complex tasks than stemming algorithms, such as understanding context and determining the part of speech of a word in a sentence. Examples of available

lemmatization tools are Collatinus[1], Lemmatizer.org[2] and MorphAdomre[3].

Problems: One limitation of both approaches is that they are language-dependant; if the dataset contains tags written in a different language, the stemming and lemmatization approaches will fail. Stemming works in a systematic way to remove suffixes and does not provide any semantic analysis. On the other hand, lemmatization takes the semantics into consideration by processing the containing sentence. However, tags are single words and they are not attached to a context, hence the advantage of the semantic processing offered by lemmatization cannot be utilized.

Text Similarity

Unlike exact matching, text similarity methods are fuzzy matching approaches that can measure how similar two strings are [65]. The Levenshtein edit distance and SoundEx are examples of text similarity algorithms. The Levenshtein edit distance algorithm calculates the minimum number of steps needed to transform one string into another, where the allowed steps are removing, adding and replacing a letter. The higher the Levenshtein distance, the less similar the two words are. If two words are exactly the same, the Levenshtein distance would be equal to zero.

SoundEx is a phonetic algorithm; it compares two words based on how they are pronounced, hence it can be used to match homophones, where two words have the same pronunciation but are spelled differently. SoundEx is implemented in popular databases such as Microsoft SQL and Oracle.

Problems: Text similarity is a useful tool to relate similar terms. However, it is not utilized in the tags cleaning process as it is found to be risky to consider similar tags, even with a high similarity threshold, as they are referring to the same concept. For

[1] an open-source lemmatizer for the Latin language

[2] an open-source lemmatizer for the English and Russian languages

[3] a Java open-source lemmatiser for the English language

example, the Levenshtein Distance between *New* and *News* is 1, implying they are very similar while they are semantically not. On the other hand, the distance between *Run* and *Running* is 4, implying that they are less similar while they are semantically similar. Also, SoundEx can help in specific cases, such as in dialect, for example the words *Center* and *Centre* will be found identical but it can fail in other cases such as *knows* and *nose*. Although text similarity is not used as a part of the tag cleaning process, it is utilised in this thesis to identify redundant place resources by matching similar place names.

4.1.2 The Tag Cleaning Process

Extracting place-related semantics modelled in Section 3.2 is the focus of this work. The proposed cleaning process involves the following steps:

1. Removal of special characters. All non alphanumeric characters are removed from tags. For instance, the tag *Cardiff&* is changed to *Cardiff*.
2. Filtering of all tags that are just one character in length.
3. Filtering of tags that represent URLs.
4. Filtering of stop-words. A list of 116 stop words, published by Microsoft [4] is used.
5. Stemming the tags. The Porter stemming algorithm[5] is applied such that each tag is transformed to its stem.
6. Removal of duplicate tags. Duplicates are removed in such a way as to preserve the relations between place resources and users.

[4] `http://msdn.microsoft.com/en-us/library/bb164590(v=vs.80).aspx`
[5] `http://tartarus.org/~martin/PorterStemmer/`

4.2 Clustering Place Resources

Most of the applications that generate geo-folksonomies aim to collect as much information as possible about places, which can be one of the reasons why such applications do not allow users to share place resources and why they require a new place resource to be created each time a user wants to tag a place. This results in having multiple place resources that reference the same place in the real world. This redundancy in the geo-folksonomy structure can produce inaccurate results when analysing folksonomies or computing tag-similarity.

Each tuple in the folksonomy represents a relation between a user, a resource and a tag. A simple query on such data can answer questions such as: what the most used tags for annotating resources are or who the most active user is. These are typical data retrieval questions that can be answered by simple database queries. However, questions such as what the most related tags to the tag "Cardiff" are, are more complicated where the answer requires co-occurrence analysis of tags to calculate tag similarity.

Web resources, such as documents, can be easily located and identified using URIs[6], where each document has a unique address on the World Wide Web. In social bookmarking applications, two users are considered to be tagging the same web resource only if the resources they are tagging have the same URI.

Unlike web resources, place resources in geo-social bookmarking applications may not be easily identified and located on the web, as such resources are not represented as web documents and consequently do not have URIs. Typically, place resources are associated with spatial attributes for representing the place location and thematic attributes, such as a place name and a place type, encoded as free text. Hence, two users can be considered to be tagging the same place resource only if the resources they are tagging are ‘spatially close’ and have similar names.

In a typical folksonomy application, the spatial location of place resources is acquired

[6]Unique Resource Identifier

via a map-based user interface. Users click on the location of the place they want to tag and the cursor location on the applet is translated to the corresponding longitude and latitude. While tagging a new place, the map interface does not reveal any places created by other users in the same area and thus a place resource can be created and tagged multiple times by different users. The same place may be given different names. For example, both "Cardiff University" and "Cardiff uni." refer to the same place by different users. Also, both instances may not be digitized at the exact same spatial location.

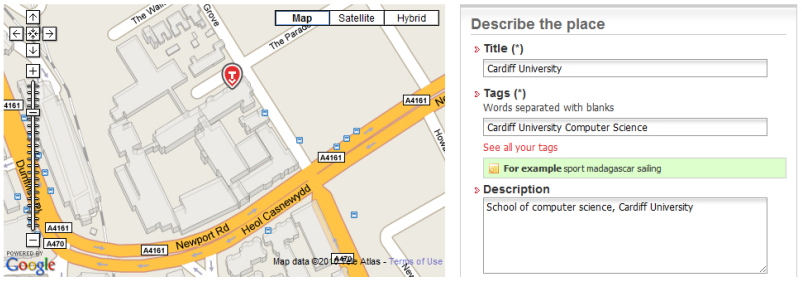


Figure 4.2: User interface for creating a new place resource in Tagzania.

Figure 4.2 shows the map-based user interface of Tagzania.com used for tagging new place resources. The map-based interface allows the current user to click on the map to locate the place and add required attributes, such as the place name, tags and description in free-text from.

As discussed above, a real-world place entity can be referred to using more than one place resource/instance in the geo-folksonomy. These redundant place resources are not linked and can thus lead to an increased uncertainty in the information content of the folksonomy and will adversely affect the result of any co-occurrence analysis applied to it. Hence, a process of clustering similar place resources is needed to enhance the certainty of the contained information in the folksonomy. A two-step clustering process based on the analysis of assigned spatial location and place names is used as follows:

1. First, a spatial clustering process is applied using a spatial similarity measure to

group place resources based on their relative proximity.

2. This is followed by a textual clustering process to isolate resources from the identified groups above based on similarity of given place names.

4.2.1 Spatial Clustering

The assumption behind spatial clustering is that close place instances may refer to the same geographic entity. The main objective of using a spatial similarity measure is to find place instances that are in close proximity to each other. Finding close instances can be achieved by using a cluster analysis method that groups place instances based on absolute distance between places, or by using a relative clustering approach that groups related places based on their belonging to predefined geographic zones. Both methods are described below.

The Quality Threshold (QT) clustering algorithm defined in [47] is used here. It has the advantage of not requiring the number of clusters to be defined apriori. In general, the QT algorithm assigns a set of objects into groups (or clusters), where objects in the same cluster satisfy a pre-defined threshold function. Here, place resources are added to a cluster if they are located within 500 meters, a reasonable threshold for the experiment, from the centre of that cluster which is determined by the QT algorithm.

Two methods are considered for reverse geo-coding the point locations of place resources (i.e. to identify a place given its spatial location); the Yahoo Where on Earth ID (WOEID) service and a postcode reverse geo-coding service. The *WOEID* web service provides a unique identifier for any geographic location based upon the closest street to that location. Hence, place resources with the same *WOEID* can be considered close, as they all have a common closest street. The postcode reverse geo-coding service, published by Geonames[7], provides a method that returns the postcode of any given spatial location. Both methods were tested and evaluated.

[7] `http://www.geonames.org/export/web-services.html`

ID	WOEID	Unit Level PC	District Level PC	QT ID
31758	44417	SW1A 0AA	SW1A	ID0
31759	44417	SW1A 0AA	SW1A	ID0
31760	44417	SW1A 2JR	SW1A	ID0
31761	44417	SW1A 2JR	SW1A	ID0
31762	44417	SW1A 0AA	SW1A	ID0
49775	44417	SW1A 2JR	SW1A	ID0
49776	44417	SW1A 0AA	SW1A	ID0
49777	44417	SW1A 0AA	SW1A	ID0

Table 4.2: Place resources referring to *Big Ben* in London, with their corresponding derived WOEIDs, postcodes and quality threshold identifiers.

An example is shown in Table 4.2, where place resources are shown that all refer to one place "Big Ben", located in the Palace of Westminster in London. Each resource is shown with its derived WOEID, postcode and its calculated QT cluster ID. As shown in the table, all instances are grouped into one WOEID, while the postcode divides the resources into two groups, with a common district-level code (*SW1A*), but separate unit-level codes. The unit-level postcode divisions are too restrictive in this context. Also, the district-level postcodes are much too broad and are likely to produce wrong clusters. In addition, postcode systems vary from one country to another, whereas the WOEID system of identification is more universal. Further experimentation with the data set confirmed that both the qualitative clustering using the WOEID and the QT clustering method are both highly successful in producing valid clusters. The QT method is however, computationally expensive with time complexity of $O(knt_{dist})$ where k is the number of clusters, n is the number of place resources and t_{dist} is the time needed to calculate the distance between the place resources.

4.2.2 Textual Clustering

After an initial clustering of place resources using their spatial location, a second step of filtering the clusters is applied based on place name similarity. The Levenshtein distance [61] is a method used for measuring text similarity. Unlike folksonomy tags, a place name can be made up of multiple words, for example "Cardiff University" and in some cases the words are used in different order, for example "University of Cardiff". The traditional Levenshtein distance between these two names will be high and they will not be detected as similar. An improved version of the Levenshtein distance [30] that is based on the word level matching as opposed to character level matching is used here and is defined as follows.

$$\sigma_t(n(r_1), n(r_2)) = 1 - \frac{LD(n(r_1), n(r_2))}{Max((n(r_1), n(r_2)))} \tag{4.1}$$

Where σ_t is the text similarity to be calculated, n is the place name of the resource r_i, LD is the Levenshtein Distance function and Max is the maximum length of place names of the instances compared.

4.3 Application and Results

4.3.1 Description of the Dataset

A data collection process is first used to build a local geo-folksonomy repository. A crawler software is developed to process pages from Tagzania[8]. Tagzania is a geo-social tagging application where users are able to collaboratively create, annotate and index geographic places on a background map. The crawler is used to extract the geo-folksonomy generated by user interaction on this application. For our experiments, the collected geo-folksonomy data set included 22,126 place instances in the UK and USA,

[8] http://www.tagzania.com

2,930 users and 12,808 distinct tags. The total number of collected geo-folksonomy tuples is 68,437.

4.3.2 Tag Cleaning

The tags cleaning process is applied on the collected folksonomy dataset. 741 tags were identified to contain special characters; those tags had 911 relations to users and 1,414 relations to place resources. 35 tags were identified to be one-character tags and they had 557 relations to users and 813 relations to place resources. 65 tags were identified to be representing URLs and they had 149 relations to users and 149 relations to place resources.

Although the stop word list contains 116 entries, there were 67 tags that matched the stop words in the list; those tags had 686 relations to users and 1,261 relations to place resources. Finally, 1,933 tags were found to have the same stems, they had 9,690 relations to users and 22,436 relations to places. Figures 4.3, 4.4 and 4.5 illustrate the results using pie chart representation.

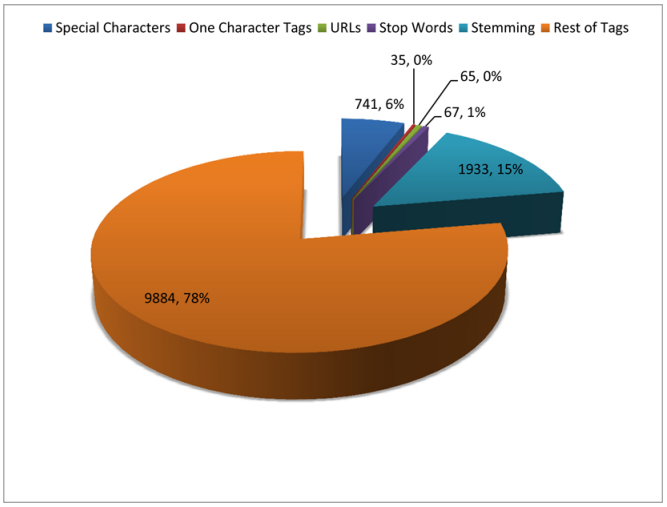


Figure 4.3: Results of the cleaning process showing the number of affected tags.

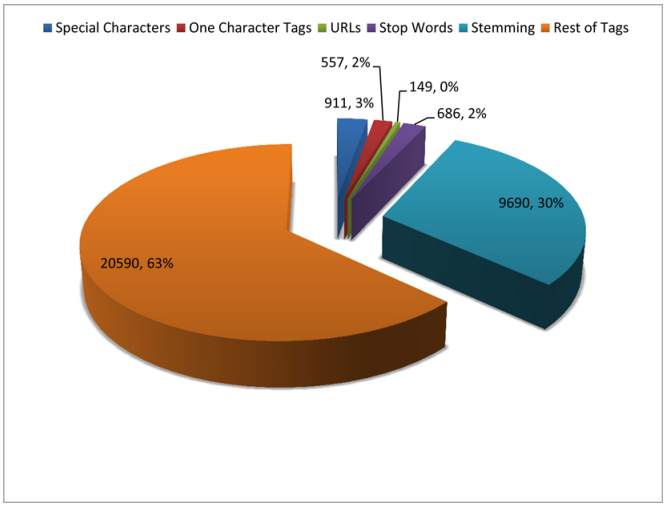


Figure 4.4: Results of the cleaning process showing the number of affected user-tags relations.

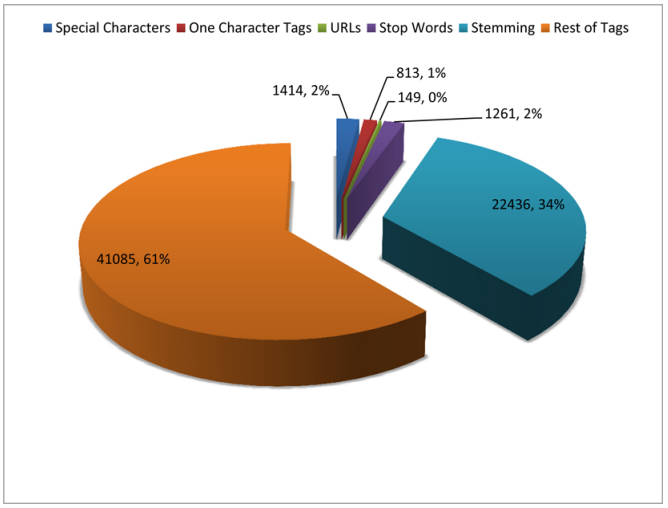


Figure 4.5: Results of the cleaning process showing the number of affected place-tags relations.

4.3.3 Place Clustering

10,119 unique WOEIDs were obtained covering all the place resources in the folksonomy; the average number of place resources sharing the same WOEID is two places. To understand the density of the spatial groups (where one WOEID is a group), it is worth considering how the place resources are distributed over the WOEIDs.

Figure 4.6 shows a histogram of the number of place resources over WOEIDs; the WOEIDs that group only two place resources are 1653 groups and this number drops to 627 (less than half) for the WOEIDs that group only three place resources. Again, this number drops to 350 (around half) for the WOEIDs that group only four places and so it continues.

The text similarity is applied with a threshold value set to 0.8 which was empirically

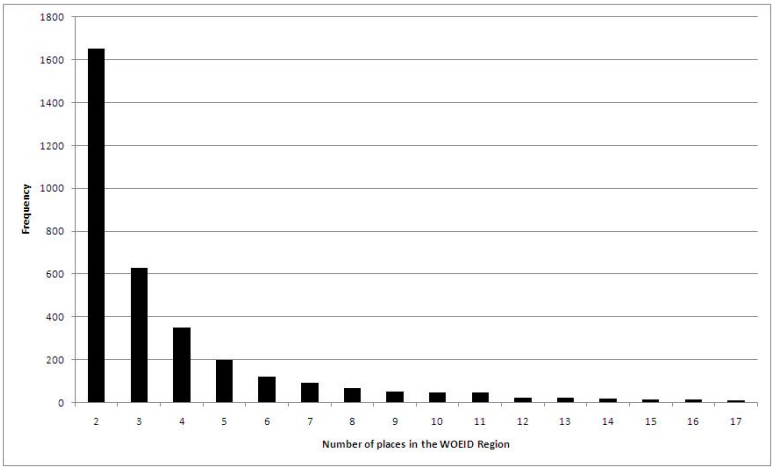


Figure 4.6: Histogram of the number of places grouped by WOEIDs.

found to be sufficient for the purpose of the work. Figure 4.7 shows the distribution of the created place clusters. The distribution of clusters follows the same distribution of WOEID groups shown in Figure 4.6. However, the magnitude is lower as the place resources in each cluster are a subset of the place resources in the container WOEID group. This distribution gives an idea about the density of the clusters. The density appears to be low in general except in certain regions (such as point of interests). This reflects the annotation behaviour of the users; relatively, a small number of places are annotated by too many users while the majority of places are annotated by a smaller number of users.

Figures 4.8 and 4.9 show two views of an area around the place **Big Ben** in London. Figure 4.8 shows the place resources, grouped in colour-coded clusters, after applying the spatial clustering method. Figure 4.9 shows the same place resources in different

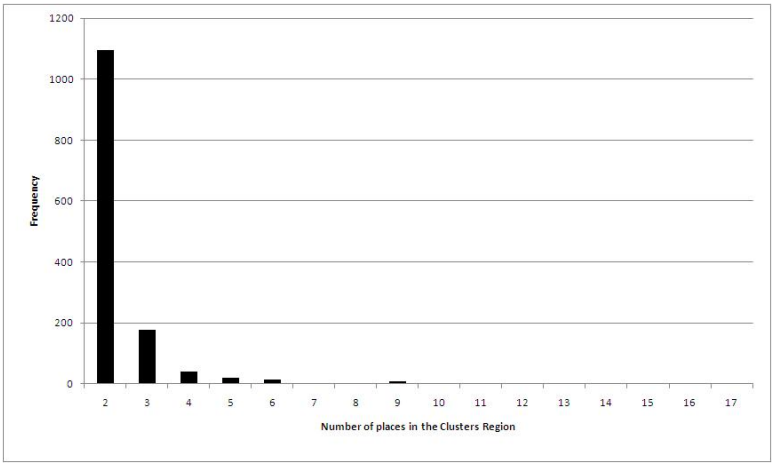


Figure 4.7: Histogram of the number of places grouped by clusters.

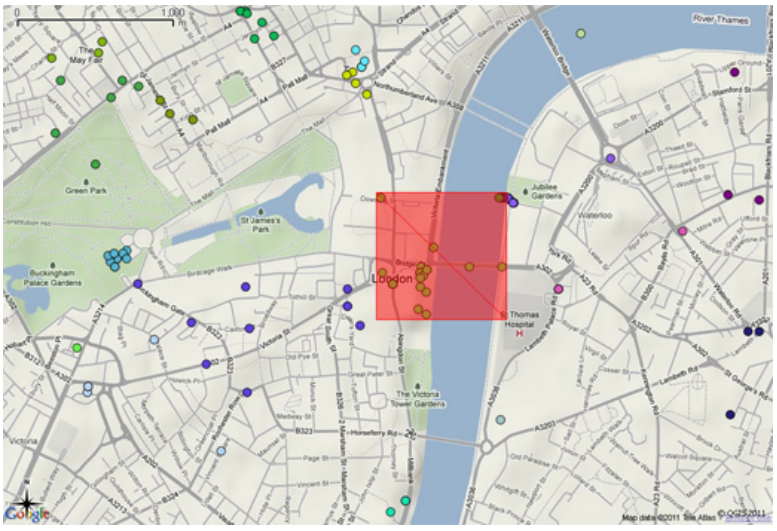


Figure 4.8: Place resources spatially clustered using WOEID.

clusters after identifying similar resources using both the spatial and textual clustering methods. The box in Figure 4.8 bounds the place resources with a unique WOEID including the place Big Ben in the first view. In Figure 4.9 the smaller box identifies the place resources which all refer to the Big Ben. The first box spans an area of 750 m. across its diagonal, whereas in second box, the area shrinks to around a 1/3 of this size. This demonstrates the quality and accuracy of the location of these place resources.

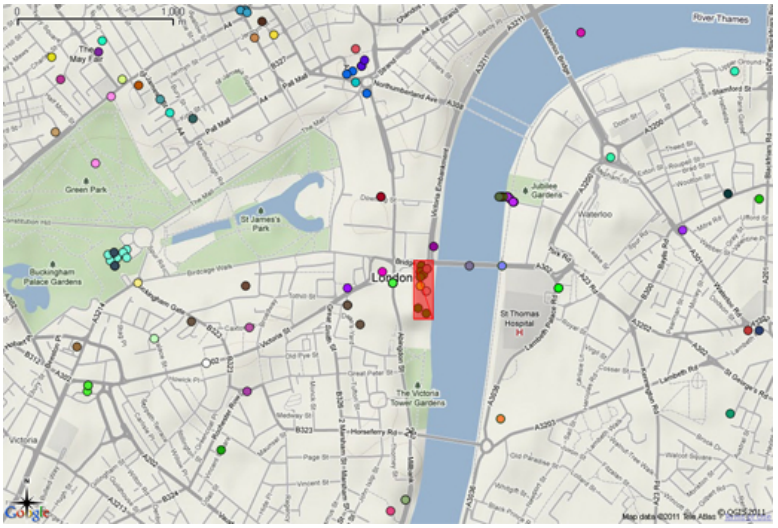


Figure 4.9: Place clusters after applying spatial and textual clustering.

4.4 Evaluation

The process of folksonomy preparation has changed the structure of the folksonomy. The tags have been cleaned and their total number has reduced as a result of removing the duplicate tags after applying the cleaning process. The place resources have been clustered into groups to identify the redundant place resources that represent the same place in real world. The tags cleaning and place resources clustering not only reduced the total number of tags and places in the resulting cleaned folksonomy, but also changed the associations between the tags, places and users. In this section, a quantitative evaluation approach is presented to compare the uncertainty in the folksonomy structure before and after the cleaning process.

4.4.1 Approach

In this experiment, Shannon's information gain [97] is used to measure the uncertainty in the folksonomy structure as follows:

$$I(t) = -\sum_{i=1}^{m} \log_2 p(x_i) \tag{4.2}$$

Where t is any given tag. m is the number of places annotated by the tag t and $p(x_i)$ defined by:

$$p(x) = \frac{w_{t,x}}{\sum_{j=1}^{m} w_{t,x_j}} \tag{4.3}$$

Where w is equal to the weight of the link between t and place x. The value of $p(x)$ will increase if the number of user votes increases and vice versa, high values of $p(x)$ indicates a high degree of certainty (lower information gain) of using tag t with place x.

Numerical Example

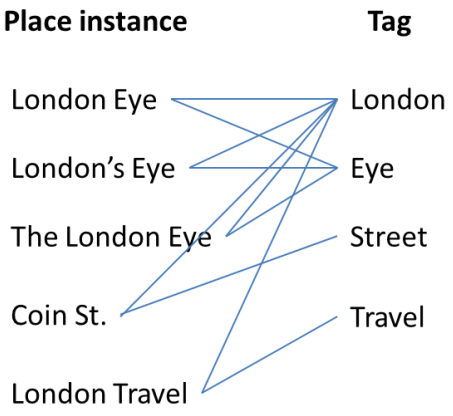


Figure 4.10: Example of un-clustered place instances.

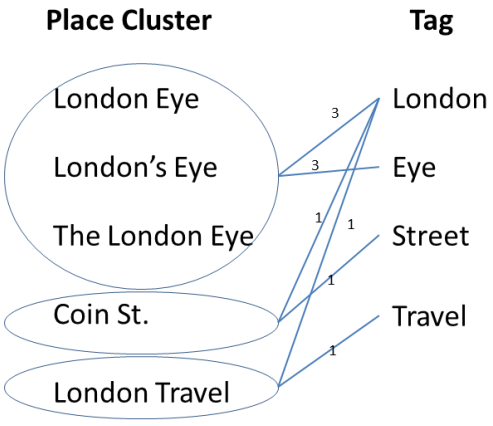


Figure 4.11: Example of clustered place instances.

In this section, an example is given to calculate the total information gain for the example folksonomy shown in Figures 4.10 and 4.11. The information gain values are calculated to measure the uncertainty in the folksonomy before and after the clustering process. First, the information gain before clustering is calculated as follows:
$I(London) = -\log_2 1/5 - \log_2 1/5 - \log_2 1/5 - \log_2 1/5 - \log_2 1/5 = 11.6096$

As there are no weights (all equal to one) and the tag 'London' is attached to all five places in the folksonomy, all the places have the same probability of $1/5$. Similarly, the remainder can be calculated as follows:

$I(Eye) = -\log_2 1/3 - \log_2 1/3 - \log_2 1/3 = 4.7549$
$I(Street) = -\log_2 1 = 0$
$I(Travel) = -\log_2 1 = 0$

Hence, the total information gain (uncertainty) is 16.3645 bits.
The information gain after clustering is calculated as follows:

$I(London) = -\log_2 3/5 - \log_2 1/5 - \log_2 1/5 = 5.379$

$I(Eye) = -\log_2 3/3 = 0$

$I(Street) = -\log_2 1 = 0$

$I(Travel) = -\log_2 1 = 0$

Hence, the total information gain (uncertainty) is 5.379 bits. This example shows that the uncertainty is reduced from 16.3645 bits to 5.379 bits by using the enriched Geo-Folksonomy instead of the original one.

4.4.2 Results

To evaluate the effect of identifying the place instances of the same place concept and build a richer geo-folksonomy, the information gain is calculated for the geo-folksonomy before and after using the proposed cleaning approach. The results show that the information gain reduced from 4011.54 to 3442.716 bits, which is around a 14% reduction in the uncertainty.

The uncertainty reduction is caused by the regions that have increased place annotation activities, in which there is likely to be multiple users annotating the same place using similar names. Table 4.3 shows a sample of WOEID regions, the number of places in each region and the information content before and after applying our clustering algorithm.

WOEID	Instances	(I) Before	(I) After	Reduction %
2441564	106	126	115	8.7%
2491521	86	11.7	6.9	41%
2441564	83	129	119	7.8%
2377112	80	23.6	18.8	20.3%
2480201	68	24.6	21.6	12.2%

Table 4.3: Information content (Uncertainty) for a sample of places identified by their WOEID code.

4.5 Summary

A geo-folksonomy pre-processing stage is introduced in this chapter that includes two processes: tag cleaning and clustering of place resources. The tag cleaning process is a multi-step process that employs different methods, such as removing stop words. However, a major part of this process is merging tags that have the same stem using the Porter stemming algorithm. Part of the embedded semantics might be lost as a result of the stemming approach. The other alternative is to use lemmatization tools. However, lemmatization is helpful when the input term is part of a sentence, while tags in this work are independent on each other and do not have any attached context.

The proposed methods used for tag cleaning and clustering of place instances were shown to be successful in filtering a significant percentage of un-cleaned tags and redundant place instances. Analysing the cleaned geo-folksonomy to build an ontology of place is the next step discussed in the following chapter.

Chapter 5

Ontology Population

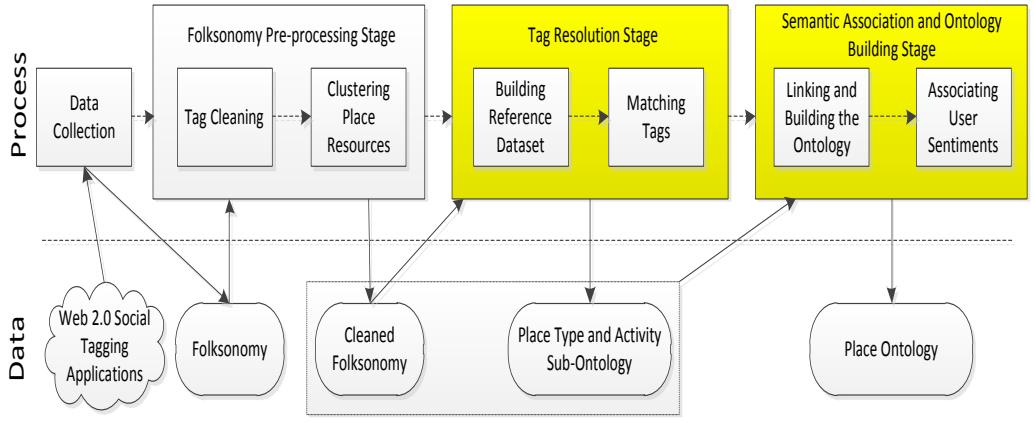


Figure 5.1: The process of building place ontology from folksonomies.

The work presented in this chapter builds on the output - the cleaned folksonomy - of the *pre-processing* stage discussed in Chapter 4. This chapter presents our approach to constructing place ontology from geo-folksonomies and this is achieved via two stages of processing; the *Tag Resolution* stage and *Semantic Association and Ontology Building* stage which are highlighted in Figure 5.1.

The *Tag Resolution* stage is designed to identify the tags that represent place types or activities by consulting external semantic data sources. The details of this stage are provided in Section 5.1. The *Semantic Association and Ontology Building* stage is

designed to construct the place ontology by creating ontology instances and inferring the semantic relationships between the concepts. The details of this stage are provided in Section 5.2. Results are presented in Section 5.3. Evaluation experiments of the proposed approach are discussed in Section 5.4. Finally, a summary of the chapter is presented in Section 5.5.

5.1 The Tag Resolution Stage

The tag resolution stage involves a process of tag classification and filtering of tag collections. In particular, the process is guided by pre-defined assumptions of possible semantics associated with the resources. In the case of geo-folksonomies, the place semantics, as defined in the model proposed earlier in Section 3.3, capture how users associate place types and activities to reflect their experiences in a place. Hence, the tag resolution stage involves first identifying and collecting place type and place activity reference dictionaries and using those as bases for matching and classification of the tag collection.

5.1.1 Building Reference Datasets

A place type is a basic concept used for classification purposes in any place gazetteer. Here, two different sources are used for collecting place type information, 1) an official data source, produced by the Ordnance Survey (OS), the national mapping agency of Great Britain, and b) the Geonames web gazetteer, built collaboratively by users and containing over 10 million place names. The OS provides an ontology of places, called the Buildings and Places ontology (OSBP)[1] that is used to describe building features and place types surveyed with the intention of improving use and enabling semi-automatic processing of this data. OSBP provides over 200 place types such

[1] `http://www.ordnancesurvey.co.uk/oswebsite/ontology`

as: (University, Hotel, Market and Stadium). Geonames also has a place ontology that associates places with a hierarchy of place types represented as feature codes. Geonames provides over 600 unique feature codes corresponding to place types such as: (Store, School and University).

Identifying possible human activities associated with a place is a not a simple task. Some research work has addressed this issue previously [3], where an approach was shown to automatically extract possible types of services and activities from definitions of place types. Here, two resources are also used for identifying possible human activities that can be associated with geographic places: a) the OSBP ontology includes a property *os:purpose* that is defined by experts to represent the possible service(s) associated with the place types, and b) the OpenCyc ontology[2], an open source version of the Cyc project that assembles a comprehensive ontology of everyday common sense knowledge. Each place type in the OSBP ontology is attached with one or more *purpose*. Table 5.1 shows example records of the place type and purpose associations. The OpenCyc ontology contains human activity concepts and offers a classification of

Place Type	**Purpose(s)**
University	Education
Hotel	Accommodation
Market	Trading
Stadium	Racing, Playing

Table 5.1: Example place types and corresponding purposes from OSBP.

different possible activities as follows:
(*cyc:HumanActivity, cyc:CommercialActivity,*
cyc:OutdoorActivity, cyc:RecreationalActivity,
cyc:CulturalActivity). Both ontologies are extracted and stored in a local RDF store. Listing 5.1 shows a sample of the SPARQL queries used to retrieve the activity types from both ontologies. Approximately 400 distinct activities are retrieved from both

[2] `http://www.opencyc.org/`

ontologies. An online implementation of this SPARQL endpoint can be found at [3]. Examples of the extracted place activities are: (Boating, Eating, Fishing, Travelling, Working, Walking).

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  PREFIX os: <http://www.ordnancesurvey.co.uk/ontology/
       BuildingsAndPlaces/v1.1/BuildingsAndPlaces.owl#>
3  PREFIX cyc: <http://sw.opencyc.org/2010/08/15/concept/en/>
4
5  SELECT ?placeActivity WHERE {
6  { ?placeActivity rdfs:subClassOf os:Purpose. }
7  UNION
8  { ?placeActivity rdfs:subClassOf cyc:HumanActivity. }
9  UNION
10 { ?placeActivity rdfs:subClassOf cyc:CommercialActivity. }
11 UNION
12 { ?placeActivity rdfs:subClassOf cyc:OutdoorActivity. }
13 UNION
14 { ?placeActivity rdfs:subClassOf cyc:RecreationalActivity. }
15 UNION
16 { ?placeActivity rdfs:subClassOf cyc:CulturalActivity. } }
```

Listing 5.1: The SPARQL query used to retrieve activities from the RDF store.

Another possibility to identify tags representing place activities is by matching against "action" verbs from a dictionary resource such as WordNet. However, activities or services offered by a place are more commonly expressed as verb phrases, composed of a combination of a verb and one or more nouns. The place types and activities extracted from the external data sources are stored in a local database and are used to classify the folksonomy tags through a matching process, as described below.

[3] http://hobzy.cs.cf.ac.uk/sparql/

5.1.2 Matching Tags

To match the tags in the folksonomy to the extracted lists of place types and place activities, these lists are first prepared as follows. Types and activities composed of multiple words are concatenated and added to the list. For example, the place type "Coffee Shop" is transformed to "CoffeeShop". Matching is carried out on stemmed tags against the list of stemmed types and activities, using the Porters stemming algorithm. The corresponding type or activity or both are then added to the ontology. For example, a tag "shop" can match a place type "shop" and a place activity "shopping" and hence both instances are created in the corresponding type and activity ontologies. The matching process resulted in 325 place type instances and 161 place activity instances.

5.2 Semantics Association and Ontology Building Stage

In this stage, the identified tag collections are structured in two steps. Firstly, subsumption relationships within individual tag collections of place types and activities are extracted and used to populate their respective sub-ontologies, and secondly, inter-relationships between types and activities are derived using the folksonomy structure. The place ontology is then populated with the resources and their associated tags from both the type and activity ontologies. Thus, the resulting place ontology reflects the associations between tags, resources and users in the folksonomy. The final step in this stage is enriching the place instances with the user sentiments.

5.2.1 Inferring Subsumption Relationships

This process infers the subclass hierarchical relationships between place type ontology instances and between place activity ontology instances represented by the properties *po:subPlaceTypeOf* and *po:SubPlaceActivityOf*. A probabilistic model of subsump-

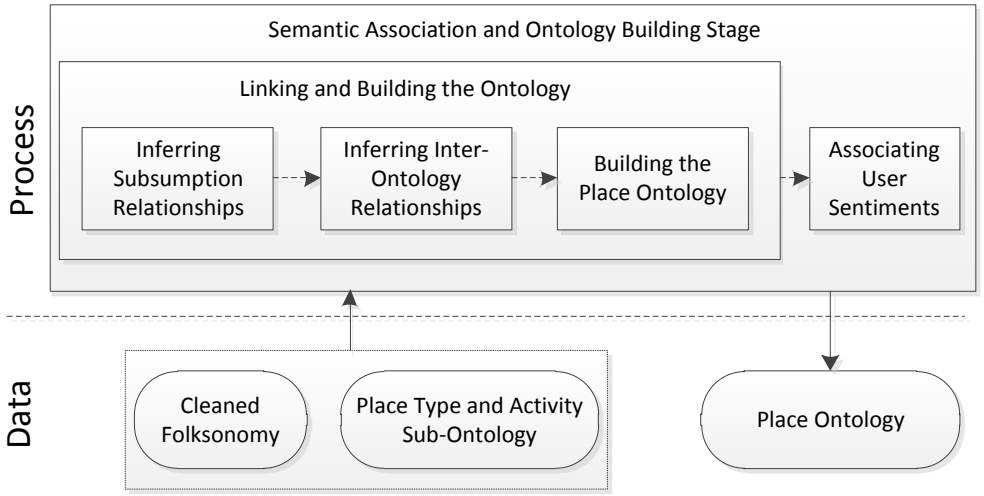


Figure 5.2: The semantics association and ontology building stage of the framework..

tion, originally introduced by Sanderson and Croft [89], can be used to derive concept hierarchies from text documents where for any given concepts/tags x and y: x subsumes y if

$$P(x|y) \geq 0.8 \; and \; P(y|x) < 1 \tag{5.1}$$

In other words x subsumes y if all the documents which contain y are a subset of the documents that contain x.

This model was extended for folksonomies [93] by including users and resources in the subsumption equation as follows: x subsumes y if

$$\begin{aligned} P(x|y) >= t \; and \; P(y|x) < t, \\ R_x \geq R_{min} \; , \; R_y \geq R_{min} \\ U_x \geq U_{min} \; , \; R_y \geq U_{min} \end{aligned} \tag{5.2}$$

Where t is the co-occurrence threshold, R_x is the number of resources tagged using x, and U_x is the number of users who used tag x. In [93], it was proposed to set R_{min} to a

value between 5 and 40, U_{min} to a value between 5 and 20, and the threshold t to 0.8, similar to values determined empirically in [89] where the same model was applied on a folksonomy dataset extracted from Flickr. The model was applied on the identified type and activity collections, resulting in the creation of 162 subsumption relationships, of which 143 were for the place types and 19 were for the place activities.

5.2.2 Inferring Inter-Ontology Relationships

Relating two tags in a folksonomy can be achieved by measuring the similarity between them, in the sense that the higher the similarity value between two tags, the more related they are. Tag similarity methods were developed to measure the similarity between tags based on their co-occurrence with users and resources in the folksonomy [66]. One of the commonly used methods to measure tag similarity is Cosine similarity [66], where similarity between two tags is defined by the following equation:

$$\sigma(t_1, t_2) = \frac{|T_1 \cap T_2|}{\sqrt{|T_1| \cdot |T_2|}} \tag{5.3}$$

Where t_i represents a tag and T_i represents the resources associated with the tag t_i in the folksonomy. A *po:relatedTo* relation is created in the place ontology between a place type and activity instance if the Cosine similarity between their corresponding tags was found to be equal or above 0.8, a threshold found empirically to be sufficient in this work. A total of 393 relationships were created, linking instances between the place type and the place activity sub-ontologies.

5.2.3 Building the Place Ontology

The process of building the place ontology involves linking the results from all the previous sub-processes and populating a place ontology with the identified semantics. A place instance of type (*po:Place*) is created for every place cluster in the restructured folksonomy. A total of 19,641 ontology place instances are created and their properties are populated as follows:

- **po:hasName**: is the most commonly used place name among the folksonomy place resources in the cluster.
- **po:alternateName**: each distinct name of the folksonomy place resources in the cluster other than the most commonly used name is represented by this property.
- **po:description**: is a concatenation of the comments attached to folksonomy place resources in the cluster.
- **wgs84:long** and **wgs84:lat**: is calculated by finding the centre location of the folksonomy place resources represented by the cluster.
- **po:nearby**: linking place instances that are spatially located within 1 km distance.

The inter-instances relations in the proposed ontology model are represented by the following properties:

- **po:hasPlaceType**: relating a place instance to a place type instance.
- **po:hasPlaceActivty**: relating a place instance to a place activity instance.
- **po:relatedTo**: relating place types, place activities, and type-activity instances.

A *po:hasPlaceType* relation is created in the place ontology between a place instance and a place type instance, if the place type is one of the tags associated with the place instance or its cluster. A *po:hasPlaceActivity* relation is created in a similar way between a place instance and an activity instance. A total of 12,736 explicit ontology relationships are created.

5.2.4 Associating User Sentiments

Folksonomy tags can reflect the opinions of users about places. The aim of sentiment analysis in this step is to calculate the sentiment score for each place resource in the folksonomy. The sentiment score for a place resource measures the positive, negat-

ive or neutral users' opinions about this place. Sentiment analysis has been used in similar research works to capture users' opinions from the interaction and collaboration activities on Web 2.0. Research works on microblogs [103, 54, 53, 75, 13], more specifically Twitter, target the problem of capturing users' opinions from posts of similar structure. In contrast to previous work, the sentiment analysis method developed here considers the influence of users and their tagging behaviour in the equations as described below.

A semantic classifier based on the Naïve Bayes classifier [87] is used here. It assumes conditional independence among features (tags in this context), which is fitting with the nature of folksonomies. Unlike other classifiers (such as Support Vector Machines), it requires a small amount of training data. The classifier is based on Bayes' theorem as follows:

$$P(S|T_1, ...T_n) = P(S) \prod_{i=1}^{n} P(T_i|S) \tag{5.4}$$

where S is a sentiment, T_i is a tag and n is number of tags associated with the place resource. Assuming an equal probability of positive, negative and neutral opinions, the equation can be simplified as follows:

$$P(S|T_1, ...T_n) = \prod_{i=1}^{n} P(T_i|S) \tag{5.5}$$

The output of the classifier depends on the way the features are selected. Here, a simple class feature model is used. However, considering different feature models such as N-Grams can be tested in the future. The data used to train the classifier is the AFINN wordlist [4] which contains 2477 words and phrases with valence between -5 and +5. The classes are defined as follows; a positive class includes words with valence between +5 and +1, a neutral class with valence of 0 and a negative class with valence between -1 and -5. An example of the classified word list is shown in Table 5.2.

[4] http://fnielsen.posterous.com/afinn-a-new-word-list-for-sentiment-analysis

Word	Valence	Classification
Perfect	+3	positive
Safe	+1	positive
Some Kind	0	neutral
Spam	-2	negative
Winner	+4	positive
Worried	-3	negative
Worst	-3	negative
WOW	+4	positive

Table 5.2: AFINN wordlist example.

After training the classifier, the algorithm in Listing 5.2 is applied to calculate the sentiment score for place clusters using the tags assigned to each place cluster.

```
1  places = GetPlaceResources();
2  for (p in places)
3  {
4    users = GetUsersOfPlace(p);
5    usersCount = 0;
6    sentimentScore = 0;
7    for (u in users)
8    {
9      usersCount++;
10     tagSet = GetTagSet(p,u);
11     sentimentScore += GetSentimentScore(tagSet);
12   }
13   SentimentScore = SentimentScore/usersCount;
14   SaveSentimentScore(p,SentimentScore);
15 }
```

Listing 5.2: Calculating the sentiment score for each place resource.

The algorithm starts by retrieving all the place resources in the dataset and finding the

associated users for each place resource. For each place-user pair, the associated tags are retrieved and stored in a set *tagSet*. The *tagSet* is used to calculate the sentiment scores for each place-user pair using the trained classifier, and then the average score is assigned to the place resource to neutralise the influence of individual user's scores. The sentiment score is a real value representing the overall users' sentiment about a place. The value ranges from -1 to +1, where -1 indicates that all the tags attached to a place are classified as negative sentiments, while +1 indicates that all the tags attached to a place are classified as positive sentiments. The sentiment score is the sum of the classifier output averaged by the number of users who annotated a given place. For example, a sentiment score with value 0.8 indicates a strong positive sentiment value while the value -0.2 indicates a weak negative sentiment value. An evaluation of the sentiment analysis process is presented in the following section.

5.3 Results

The data cleaning process resulted in identifying 19,614 clusters, corresponding to unique places instances. Hence, 2,512 place instances are merged (around 11% of the total number of place resources). Figure 5.3 shows the results of classifying the tags using the proposed framework. 32% of the tags are place names. 18% of the tags were classified as user's opinions and are processed by the sentiment analysis process. 2% of the tags correspond to place types and 3% correspond to place activities. The rest of the tags (45%) do not fit in any of the above categories.

The distribution of the tags in the geo-folksonomy dataset follow a power law distribution as shown in Figure 5.4. This is similar to the results reported by other empirical studies [20, 24]. It is noted that although the percentages of place type and activity tags are low, these tags are used more frequently than unclassified tags as shown in Figure 5.5, which plots the frequency distribution of the 10 most used tags in each category. Table 5.3 lists the top 10 frequently used tags in each category. 79% of the

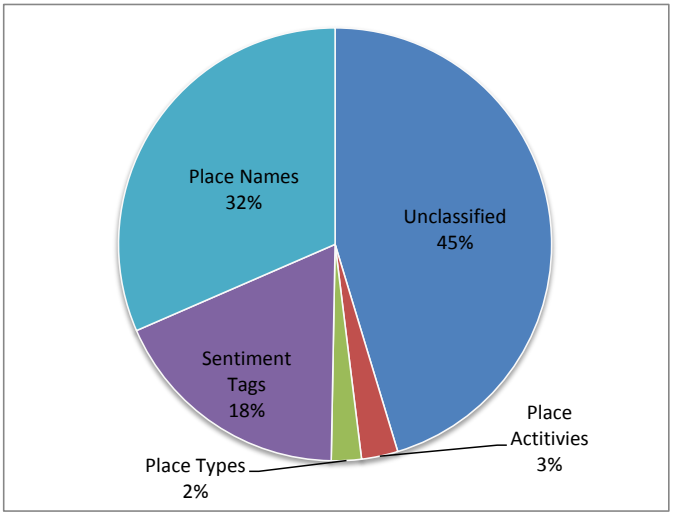


Figure 5.3: Tag classification chart.

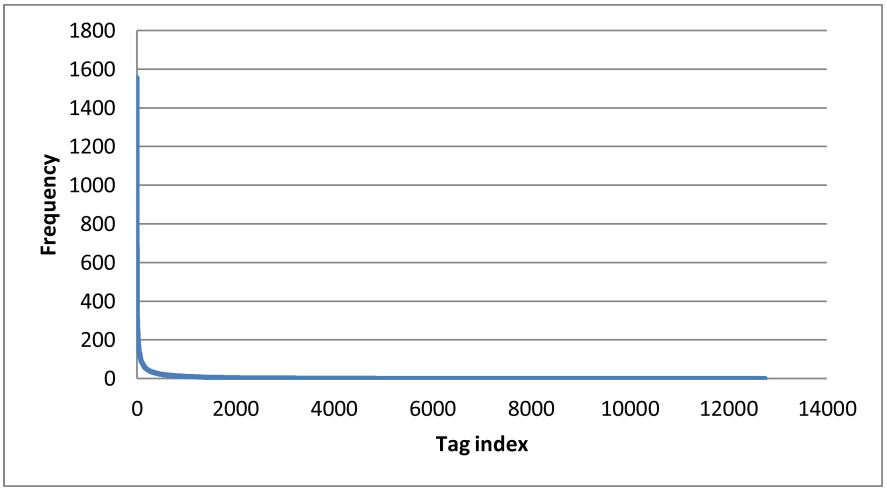


Figure 5.4: Frequency of tag usage over the entire geo-folksonomy dataset.

unclassified tags contribute to the long tail of the Zipf frequency graph as they were found to be used only once or twice. The unclassified tags include possible reference to temporal concepts, such as 2008 and $summer$, possible abbreviations (e.g. st. for street), or noise (e.g. two letter words: nv, vc, xy). The tag resolution stage resulted in identifying 346 activity types in the folksonomy, using a set of approximately 400

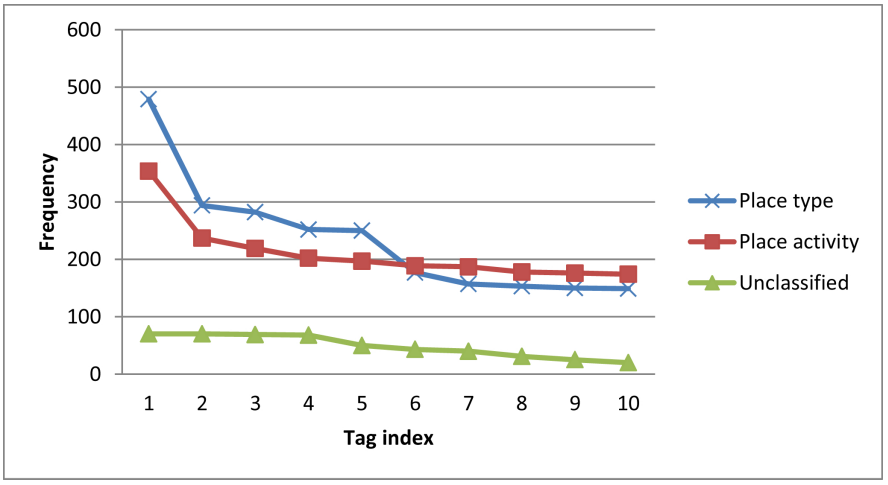


Figure 5.5: Detailed tag usage frequency of the 10 most used tags.

Rank	Place Type	Place Activity	Unclassified
1	food	housing	north
2	restaurant	travelling	clock
3	school	marketing	new
4	store	sale	one
5	hotel	visiting	family
6	university	servicing	TimeForPublicSpace
7	park	camping	apple_store
8	airport	socializing	high
9	museum	buying	2008
10	shop	business	recitation

Table 5.3: Most frequently used tags classified as place types, activities and other in the sample geo-folksonomy.

activity types in the reference data sets. It is interesting to observe that although 927 tags are identified as verbs using WordNet, only 107 of those corresponded to possible activities and types from the compiled list using the external ontology resources. Some examples of the unclassified verb tags include, *arm*, *arrest*, *assign*, *back* and *coin*.

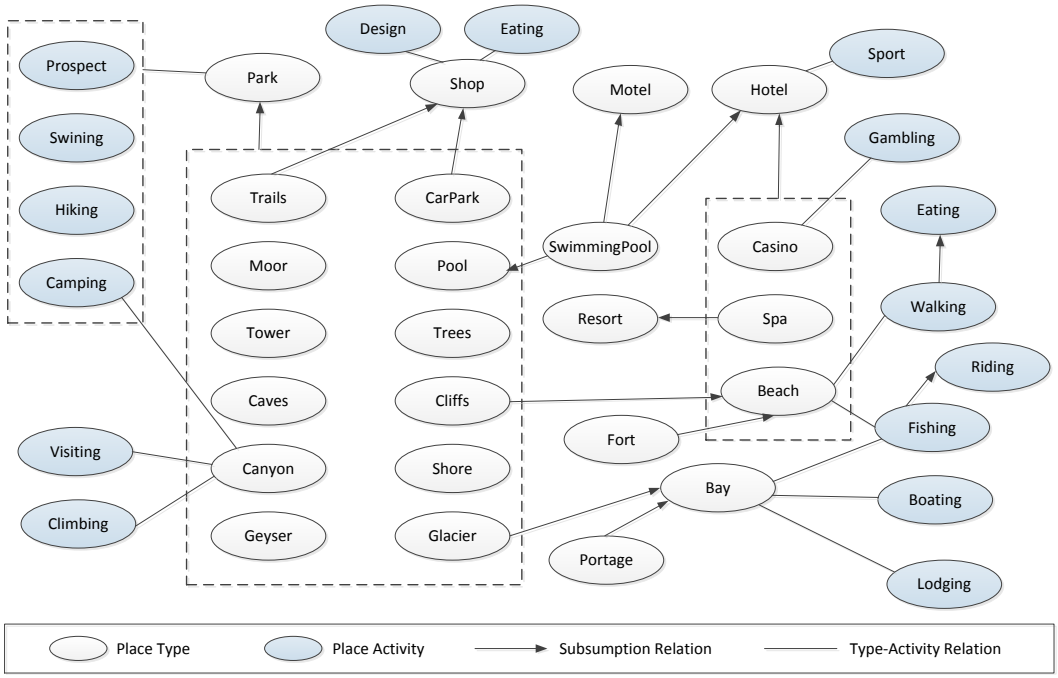


Figure 5.6: A snapshot of the derived ontology showing a number of place types, their related place activities and subsumption relationships.

Figure 5.6 shows a subset of the derived place semantics, in which 24 place types and 16 place activities are presented with their corresponding association and subsumption relationships.

Finally, the following table shows statistics about the induced ontology:

5.4 Evaluation

5.4.1 User-based Evaluation

A possible approach to ontology evaluation is to compare it to a "golden standard" which itself can be an ontology [64]. The OS Building and Place ontology is used

Entity	Count
Place instances	19,641
Place Type instances	211
Place Activity instances	346
Subsumption relations	136
Inter-Instance relations	12,736
Spatial relations (near by)	254,888

Table 5.4: Instances and relationships in the induced place ontology.

here for demonstration. Figure 5.7 compares the semantics related to the place type "Tourism Attraction" as defined in OSBP ontology to those related to the place type "Tourism" in the derived place ontology. As can be seen in the Figure, only one "purpose" (Entertainment) is associated with the "Tourism Attraction" place type in the OSBP ontology, whereas a much richer set of relationships is identified in the place ontology reflecting the usage of the concept in the specific folksonomy dataset ("Tourism" is related to 6 other place types and 4 place activities). However, it should be noted that an absolute comparison is not realistic as the ontologies represent different views and purposes and, as suggested previously, the ontology derived from the folksonomy is dynamic and its structure is likely to change with time.

To further evaluate the derived ontology, a questionnaire was designed to assess the quality of the derived concepts and their relationships. Five different places in London, UK, corresponding to different possible place types, were chosen, namely Hyde Park, Marriot Hotel, Tesco, Wagamama and the Imperial War Museum. The geographic region was chosen primarily because of popularity and, as such, more users were likely to be aware of the place names and secondly because of the density of the associated tags in the folksonomy. The questionnaire was issued to university students over a period of four weeks. 53 students participated in the survey, of which 76% were male users, approximately 90% were under 29 years old, 96% of users have a degree above high school, 65.9% were familiar with London and 80.4% were native English

Figure 5.7: An example of a place type concept "Tourism" as defined in the Ordnance Survey ontology and its computed definition in the derived place ontology.

speakers.

Two types of questions were asked for each place. The first type of questions aimed at evaluating the quality of the relationships between concepts. Figure 5.8 shows the responses of participants on questions about place-type relationships. The second type of questions were aimed at evaluating misclassified tags by asking the user to suggest a classification for tags co-occurring with the place resource, as either a place type, a place activity, a related concept or a non-related concept. Figure 5.9 shows the results of the second type of questions for the place "Hyde Park". Users' responses were used to calculate the recall, precision and F1 measure for evaluation. Table 5.5 lists the number of true positives, false positives, true negatives and false negatives used to calculate the precision (0.8), recall (0.5) and F1 (0.615). The experiment suggests a correlation between the derived ontology and users' perception of places and related

semantics. Finally, the survey also questioned the users' experiences, or impressions (if they did not visit the places), of the five places. The responses again correlated with the output of the sentiment classifier. Though the experiment is limited, the results provide an indication of the validity of the methods.

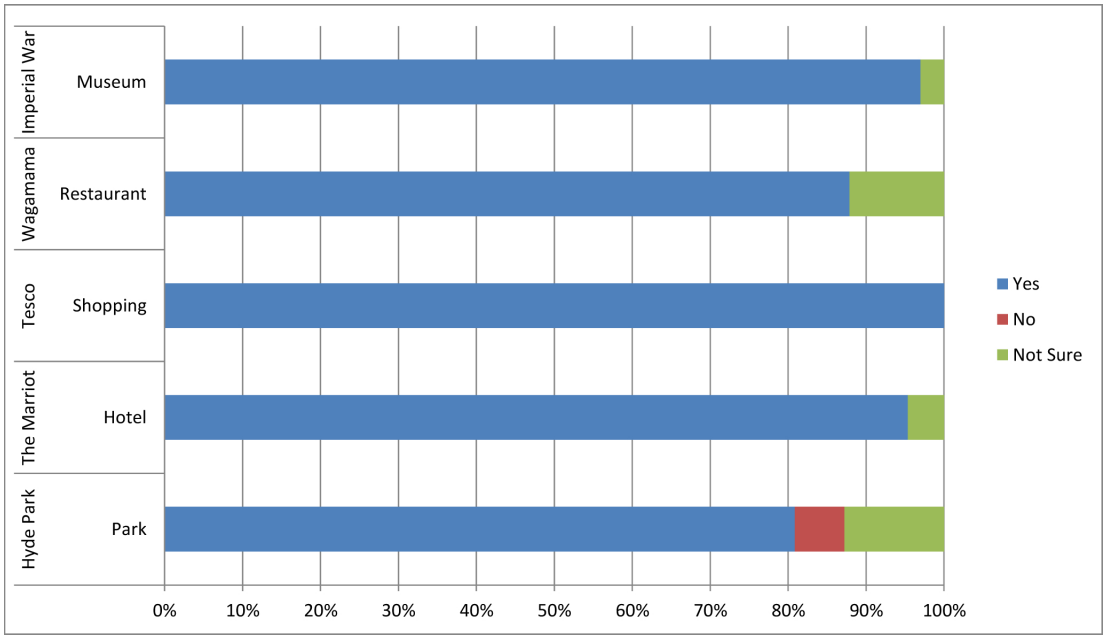


Figure 5.8: Level of agreement in the questionnaire with the derived relationships between concepts for the chosen place resources..

Place	TP	FP	TN	FN
Hyde Park	4	2	3	12
Marriot	4	0	10	5
Tesco	4	1	12	3
Wagamama	4	2	12	0
Imperial War	4	0	15	0
Total	20	5	52	20

Table 5.5: Evaluating the tag classification results with the questionnaire responses.

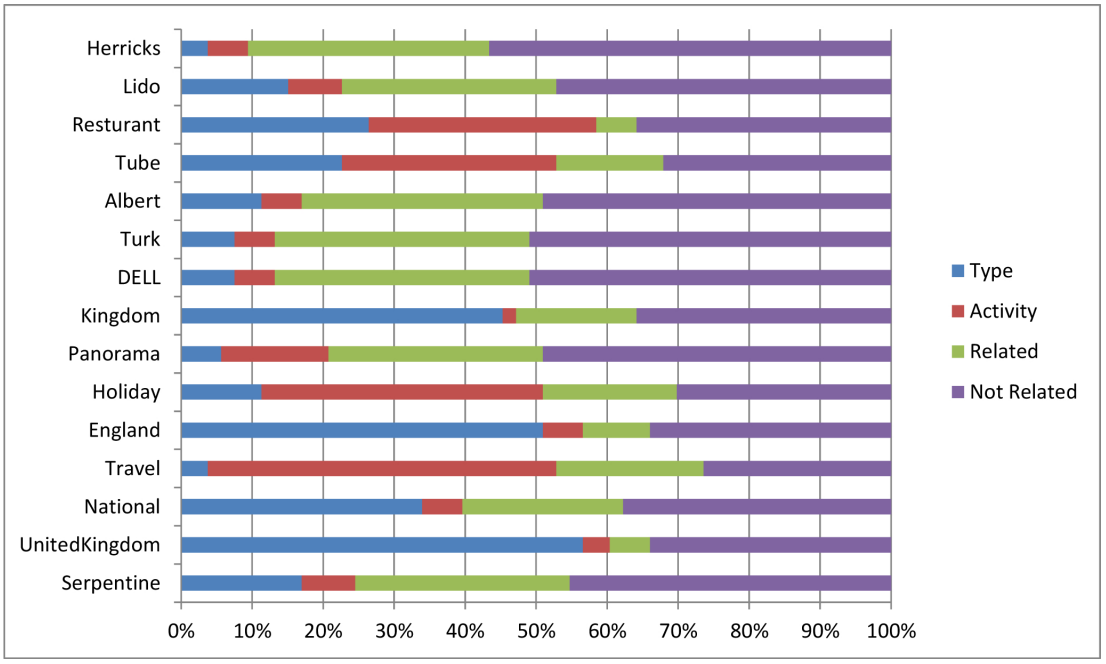


Figure 5.9: A sample of the users' responses classifying tags co-occurring with the place "Hyde Park".

5.4.2 Quantitative Evaluation Using Semantic Similarity

A quantitative evaluation experiment was designed here to measure the level of agreement between the semantics represented by the place type and place activity sub-ontologies on one side and the general semantics on the web on the other side. The Measure of Semantic Relatedness (MSR) web service [110] provides a set of methods through a web-based API interface to calculate the semantic similarity between two terms[5]. Although the MSR provides different methods of calculating the semantic similarity, all of them are based on the same theory. The MSR assumes that the strength of the relation between two terms is proportional to the number of times the two terms co-occurred together in the same documents on the web. Even though the MSR does not employ any semantic analysis approaches and it is based only on co-occurrence of the terms, it assumes that the co-occurrence of two terms in the same document

[5] http://cwl-projects.cogsci.rpi.edu/msr/

implies that they are in the same context. The more frequently they appear, the more semantically related they are. The performance of the different MSR methods in terms of quality and accuracy is found to be dependent on the size and type of the input data [63]. More details and comparisons about the different MSR methods can be found in [27]. In this experiment, the Point-wise Mutual Information (PMI) [106] and the Normalised Search Similarity (NSS) [68] methods are chosen to calculate the similarity. Both methods can measure the semantic relatedness among terms in large datasets.

Relations in the induced ontology that link place types, place activities or both are evaluated using the PMI and the NSS methods. First, a set of SPARQL queries are used to retrieve the relations along with the concepts they connect. The appropriate MSR API functions are passed the two concepts of each relation to calculate the semantic similarity between them using the Google's search engine. The PMI and NSS similarity are measured for about 500 relations. Figure 5.10 shows a graph of the output of both measures along with the trend lines. As can be seen in the Figure, corresponding trend lines indicate a correlation between the two measures.

The strength of the similarity measured by the PMI-G method over the whole set has an average of 86% while the average strength of the similarity measured by the NSS-G is 78%. Table 5.6 illustrates the results of the experiment by showing a sample of the measures of PMI-G and NSS-G for 10 relationships. This experiment demonstrates the validity of the place semantics automatically extracted from the geo-folksonomies; the extracted semantics are found to be close to semantics embedded in web documents.

5.5 Summary

This chapter introduced the approach to extract the embedded place semantics in geo-folksonomies. The approach introduced here builds on the pre-processing steps introduced in Chapter 4, in which the geo-folksonomy tags and place resources are cleaned to enhance the quality before extracting the semantics. The cleaned folksonomy is ana-

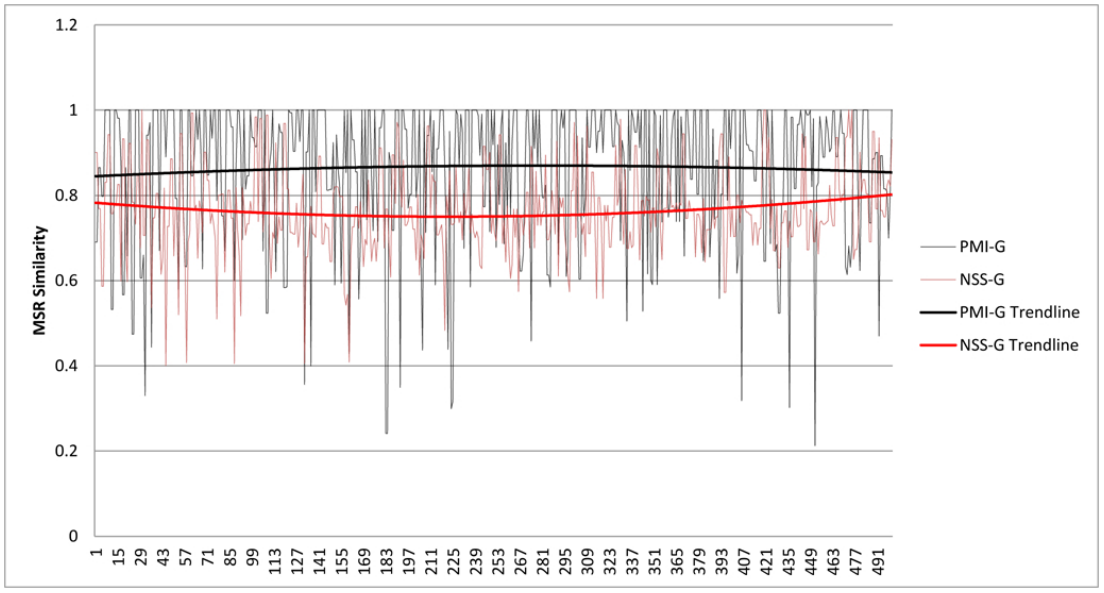


Figure 5.10: A graph showing the PMI-G and the NSS-G measures for a set of 500 ontology relationships.

Concept 1	Concept 2	PMI-G	NSS-G
Sale(A)	Flat(T)	69%	90%
Buy(A)	Sale(A)	100%	83%
Hotel(T)	Reservation(A)	97%	79%
University(T)	College(T)	100%	89%
Spa(T)	Hotel(T)	96%	91%
Boating(A)	Fishing(A)	100%	78%
Rock(T)	Climbing(A)	63%	65%
Casino(T)	Gambling(A)	93%	76%
Museum(T)	Park(T)	75%	80%
Rock(T)	Mountain(T)	86%	82%

Table 5.6: A sample of the MSR measures calculated using PMI-G and NSS-G applied on the ontology relations between places types (T) and activities (A).

lysed through a two-stage process; a tag resolution stage, in which external semantic data sources are used to classify the tags into place types and human activities, and a semantic association and ontology building stage, in which the relationships among ontology concepts are inferred and user sentiments are calculated for each place resource.

The proposed approach is based on folksonomy co-occurrence analysis as well as statistical analysis to build the infer relationships among concepts. The induced place ontology contains 19,641 places, 211 place types and 346 place activities and over 12,700 semantic relationships.

Two experiments were introduced in this chapter for the purpose of evaluating the induced place ontology; a user-based evaluation through a survey and a validation of the semantic relationships through an external semantic measurement web service. The results of the two evaluation experiments suggest that place semantics extracted from geo-folksonomies correlate with users' views and expectations generated on web 2.0.

Chapter 6

Implementation

6.1 System Overview

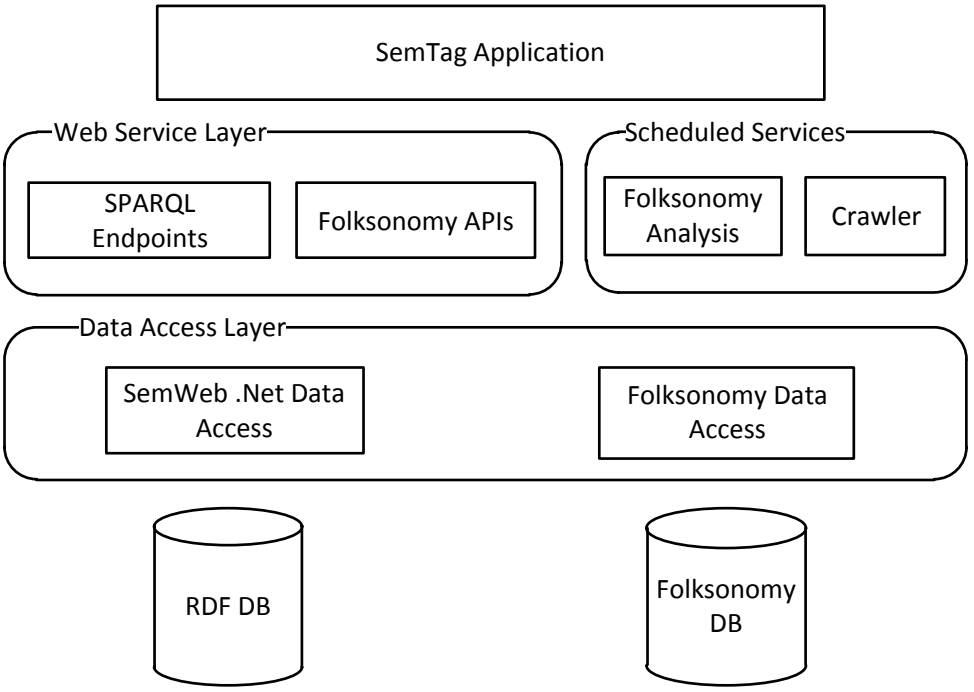


Figure 6.1: The components of the implemented system.

Testing the hypothesis of this research required a considerable amount of effort ded-

icated to designing and implementing a system of various software components. The aim of the system is to collect geo-folksonomies from the web, extract the embedded place semantics and present the geo-folksonomies along with extracted semantics on a mapping application.

This chapter describes the the overall architecture and the implemented system components. A system is designed following a typical three-tier service oriented architecture that consists of a data access layer, a service layer and an application layer. The data access layer contains the components responsible for the database operations such as adding and updating records. The service layer contains the components and methods that implement the approaches used to analyse, process and query the data. The application layer contains the application *SemTag* which provides a web-based user interface that allows users to search for a place and view its tags along with the induced place semantics attached to that place.

The system relies on two databases: the Folksonomy DB, which stores the folksonomy records collected from the web, and the RDF DB which stores RDF-encoded ontologies. The databases are strictly accessed only from the data access layer which manipulates the data in both databases. The system exposes a set of Application Programming Interface (API) functions to access both the folksonomy and the place semantics through an XML and Simple Object Access Protocol (SOAP) service layer. Also, there are two separate programs designed to run in the background of the hosting server to crawl the data from the web and to extract the place semantics from the collected geo-folksonomies. A detailed discussion of each component of the system is presented in the remainder of this chapter.

6.2 Database Design

The database engine used in this research is Microsoft SQL Server 2008. It has been selected for various reasons: a) the seamless integration and support with other Mi-

crosoft development tools used in this research such as Visual Studio 2008; b) the compatibility with semantic web tools as SemWeb.Net which will be discussed later in this chapter; c) the support of full text indexes and text similarity functions such as SoundEX which are already built in the database, unlike other databases such as PostgreSQL that require additional plug-ins such as Lucene[1] or Solr[2] to perform adequate text indexing.

Two database instances are created to support this research, *Folksonomy DB* and *RDF DB*. The *Folksonomy DB* is designed to support storing and searching of the collected folksonomy datasets as well as the output of the folksonomy co-occurrence analysis methods implemented. The data model of the database is shown in Figure 6.2.

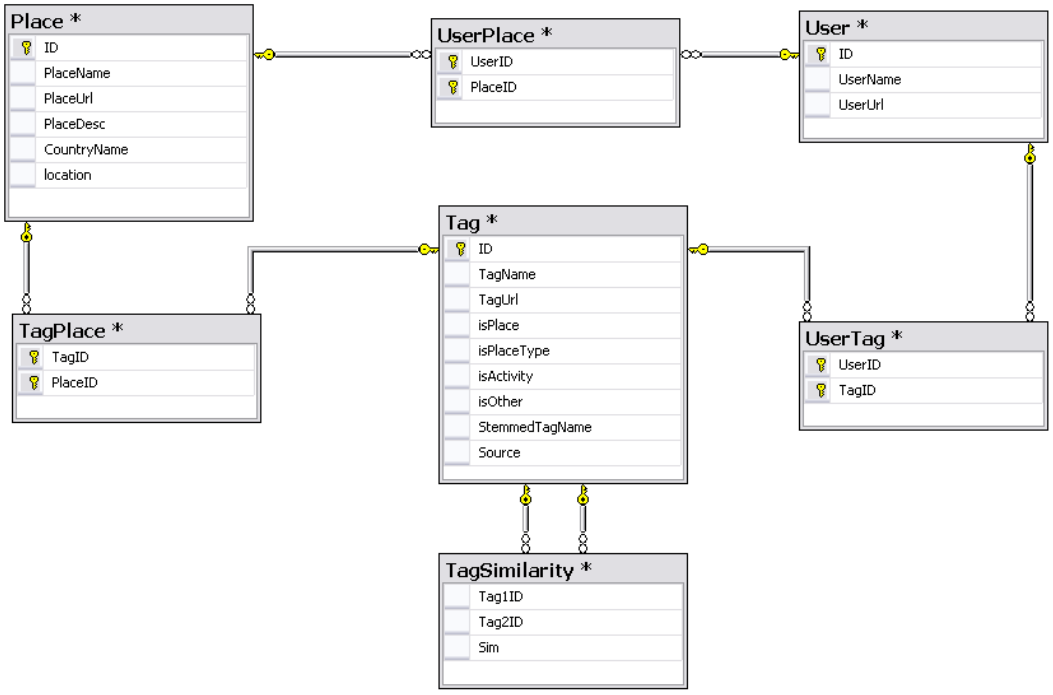


Figure 6.2: The main tables in the Folksonomy DB.

[1] http://lucene.apache.org/core/

[2] http://lucene.apache.org/solr/

The three distinct components of the geo-folksonomy are modelled using the *Place* table representing folksonomy place resources, the *Tag* table representing folksonomy tags, and the *User* table representing folksonomy users. Each table has a many-to-many relation to the other two tables represented by the *UserTag*, *TagPlace* and *UserPlace* tables.

A spatial index of type geography [3] is applied to the *Location* column in the *Place* table, where each place is represented by a single spatial point. Also, text indexes are applied to *PlaceName* and *CountryName* columns in the *Place* table as well as the *TagName* in the Tag table. The following are examples of the queries that can be applied to the database:

```
1 Select Distinct t.TagName
2 From Tag t
3 Join TagPlace tp on t.ID = tp.TagID
4 Join Place p on p.ID = tp.PlaceID
5 Where P.PlaceName = 'London␣Eye'
```

Listing 6.1: Retrieve all tags attached to place resources named 'London Eye'.

```
1 Select top 100 t.TagName, COUNT(tp.PlaceID)
2 From Tag t
3 Join TagPlace tp on t.ID = tp.TagID
4 Group By t.TagName
5 Order By COUNT(tp.PlaceID) Desc
```

Listing 6.2: Retrieve top 100 most used tags.

The database also contains several tables for storing the output of the folksonomy analysis such as tags similarity. The database table *TagSimilarity* shown in Figure 6.2 is a template for the similarity output tables, where each record contains the identifiers of the similar tags along with the calculated similarity value. This template is instantiated multiple times in the database, one time for each analysis method.

[3] http://msdn.microsoft.com/en-us/library/bb964711

The *RDF DB* is automatically created by the SemWeb.Net semantic web tool and is used to store RDF triples in a relational database instead of file system. More details about the SemWeb.Net library is presented in Section 6.4.2.

6.3 Semantic Web Tools and SemWeb

There are various tools and application libraries already developed to manipulate the RDF data. An investigation of existing tools has been carried out to evaluate their suitability for building the system, and a summary is presented in Table 6.1.

One disadvantage observed in some of the tools, such as in 4Suite and OWL API, is that the RDF file has to be fully loaded into the memory to be processed or queried. This usually causes an out of memory exception when dealing with large RDF files. The choices are narrowed down to either LinqToRDF or SemWeb.Net as being biased to Microsoft .NET as a rapid development platform. LinqToRDF is an extension for the .NET Language Integrated Query (LINQ) designed to support querying RDF files. The SemWeb.NET is a complete semantic web framework with a SPARQL query engine and inference engine. It also supports persisting RDF data in relational databases such as Microsoft SQL to address the memory problems. Hence, the SemWeb.NET is the tool chosen here to store and query the place semantics.

6.4 Data Access Layer

6.4.1 Folksonomy Data Access

This component provides simplified access to the data stored in the folksonomy database through a set of static functions listed in Table 6.2. The connection string of the database is configured through an XML configuration file named ‘app.config’; this design allows users to change the database connection without needing to recompile the

Tool	Description
4Suite	Python-based toolkit for XML application development, it features a library of integrated tools for XML processing, implementing open technologies. More information can be found at `http://pypi.python.org/pypi/4Suite-XML`.
Jena	A commonly used semantic web framework for Java. Provides a SPARQL interface, RDF and OWL APIs, and inference support. More information can be found at `http://jena.sourceforge.net`.
Sesame	Another commonly used semantic web framework for Java. Provides a SPARQL interface and an HTTP server interface. More information can be found at `http://www.openrdf.org`.
OWL API	An implementation for Java. Provides OWL APIs and contains a common interface for many reasoners. More information can be found at `http://owlapi.sourceforge.net`.
RAP RDF API	An open-source RDF API and software suite for storing, querying and manipulating RDF in PHP. More information can be found at `http://sourceforge.net/projects/rdfapi-php`.
Redland	An implementation for C. Provides a collection of RDF libraries for parsing and querying. More information can be found at `http://librdf.org`.
LinqToRDF	A semantic web framework for .NET built on the LINQ. More information can be found at `http://code.google.com/p/linqtordf`.
SemWeb.NET	A semantic web framework for .NET. Provides APIs to keep RDF in persistent storage (MS SQL, MySQL, etc.). Also provides SPARQL query engine and inferencing functionality. More information can be found at `http://razor.occams.info/code/semweb`.

Table 6.1: Tools for manipulating RDF data.

source code of the application. The *CreateConnection* factory function is responsible for reading the value of the connection string from the configuration file and returns a ready-to-use connection object. The connection object is used by the other functions in the component to perform the database operations.

Function	**Description**
CreateConnection	A factory function returns a SQLConnection object configured to connect to the Folksonomy DB
ExecuteScalar	Executes an SQL query that returns a single value.
ExecuteNonQuery	Executes an SQL query that does not return values such as insert or delete statements.
ExecuteReader	Returns a connected and read only SqlDataReader used to iterate over the results of an SQL query.

Table 6.2: The APIs provided by the Folksonomy data access component.

Listing 6.3 shows the source code of the function *ExecuteNonQuery* where the SQL query and the database connection object (created using the *CreateConnection* function) are passed as input parameters. The function creates an *SqlCommand* object and configures its timeout, connection and command text properties before executing the command. This is an example of the encapsulated database access logic where users do not have to write the same logic every time they need to execute a query on the database.

```
1 public static void ExecuteNonQuery(string sql, SqlConnection con
      )
2 {
3       SqlCommand cmd = new SqlCommand();
4       cmd.CommandTimeout = 500;
5       cmd.Connection = con == null?CreateConnection():con;
6       cmd.CommandText = sql;
7       if (cmd.Connection.State == ConnectionState.Closed)
8           cmd.Connection.Open();
```

```
 9        cmd.ExecuteNonQuery();
10        if (con == null)
11        {
12            con.Close();
13            con.Dispose();
14        }
15    }
```

Listing 6.3: The source code of the ExecuteNonQuery function of the folksonomy data access component.

6.4.2 SemWeb.Net Data Access

SemWeb.Net is an open-source library developed to read and write RDF, keeping RDF in persistent storage (memory, Microsoft SQL, etc.), querying persistent storage via SPARQL, and executing SPARQL queries over remote endpoints. The version of the SemWeb.Net library used here is v1.0.7.

The library contains a set of classes providing different functionalities, and part of the provided classes is utilised in this research as follows. The *Store* class is used to specify the RDF persistent storage used by the library. Here, it is configured to use Microsoft SQL server. The *SparqlEngine* class provides the functionality to parse and execute SPARQL queries, it is passed a string object which contains the query to be executed over the *Store* object. The UML Sequence diagram in Figure 6.3 illustrates the logic of executing SPARQL queries using the SemWeb.Net.

SemWeb.Net also provides the functionality to import RDF files into the supported persistent storages via a standalone command line tool (rdfstorage.exe). The tool receives two command line parameters; the path of the input RDF file and the connection string of the database that the file will be imported to. The following example shows how to use the tool from the command line:

rdfstorage.exe PlacesData.rdf

–out “sqlserver:rdf:Database=SemWebDB;user id=xx;password=xx”

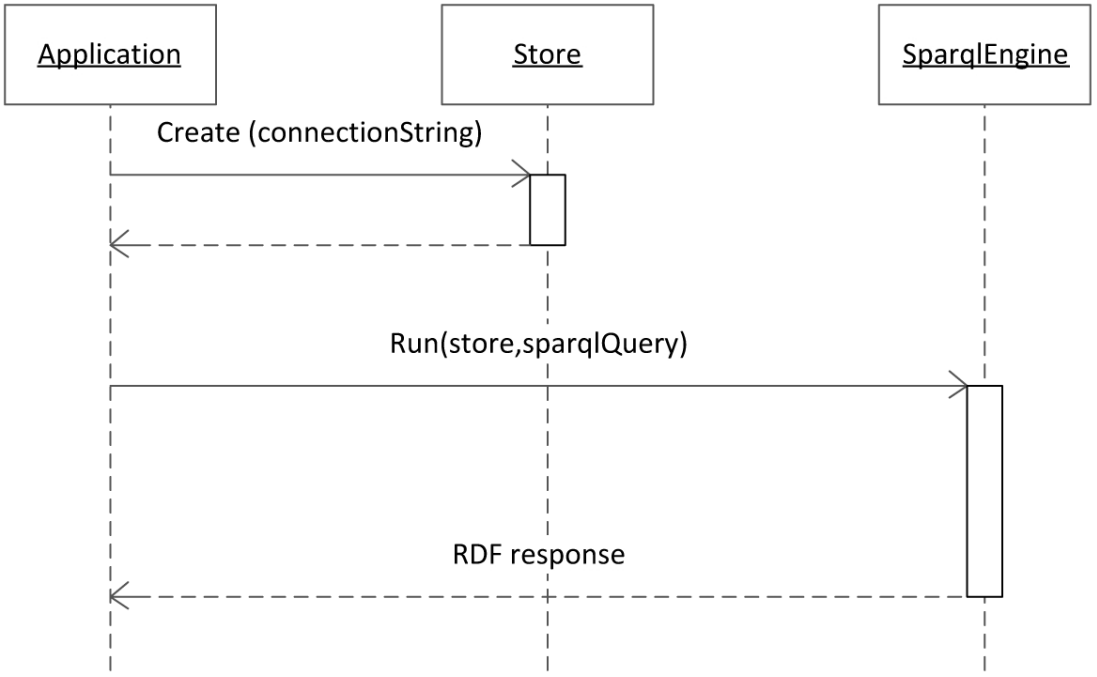


Figure 6.3: UML Sequence diagram showing how the SemWeb.Net components are used to execute SPARQL queries.

6.5 Web Service Layer

6.5.1 SPARQL Endponits

The query engine provided by the SemWeb.Net library is responsible for parsing and executing the SPARQL queries. SemWeb.Net exposes functionality though a set of APIs which cannot be called remotely. As the system is designed to be service oriented to allow the integration of the induced place semantics with external applications, a SPARQL endpoint is developed for this purpose. The SPARQL endpoint is

implemented as a web page with a server side code to receive SPARQL queries, validate their syntax, and send the queries to the SemWeb.Net component to execute if no syntax errors are present. Figure 6.4 shows a snapshot of the SPARQL endpoint used to query the extracted place semantics. There are two other SPARQL endpoints exposed by the system for Open Cyc and OSBP ontologies, and all are accessible from the following address: `http://hobzy.cs.cf.ac.uk/sparql`.

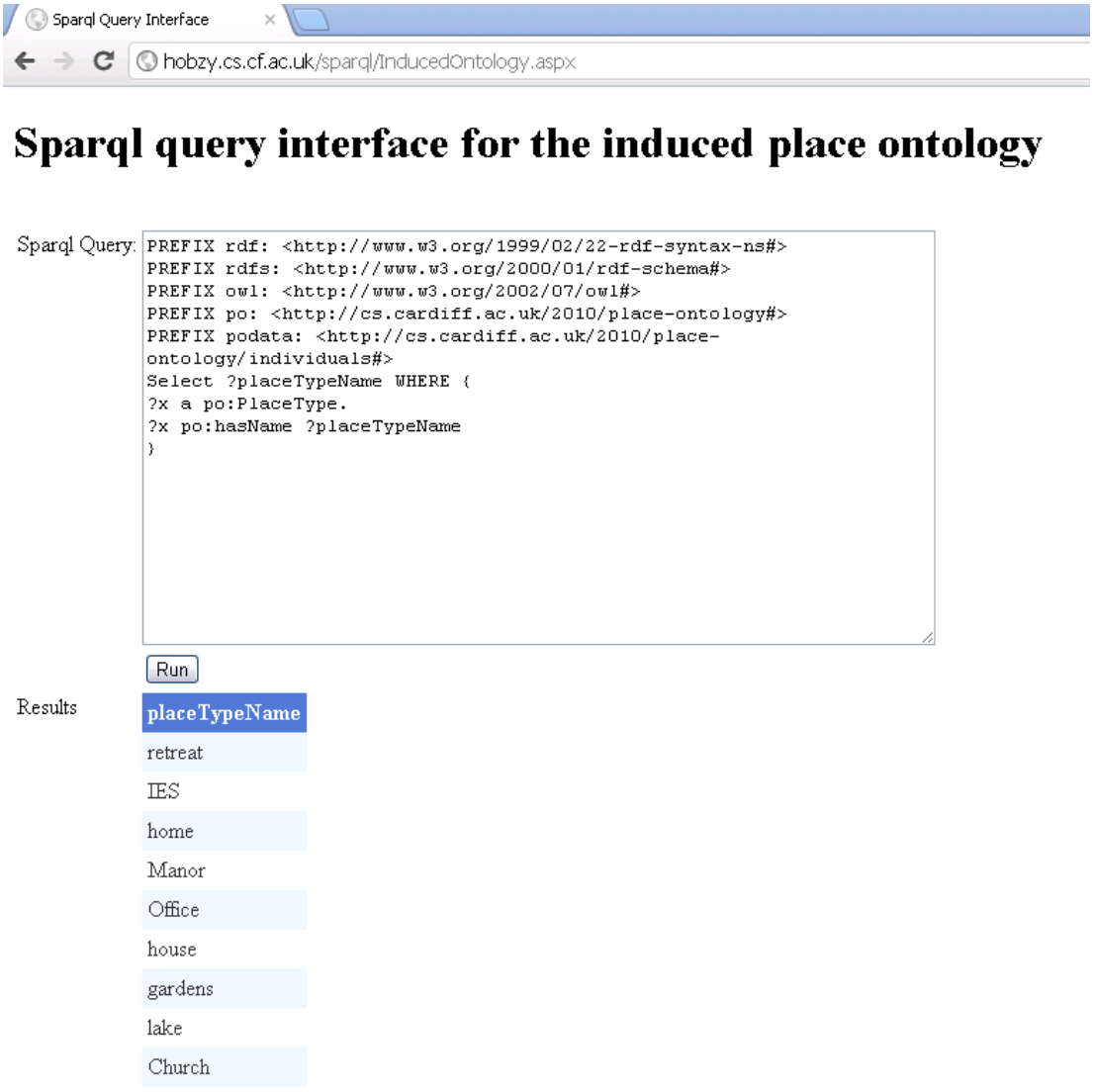


Figure 6.4: A snapshot of the SPARQL endpoint used to query the extracted place ontology.

6.5.2 Folksonomy APIs

This component exposes a set of remote API functions that allow external applications to query the collected geo-folksonomies as well as the output of the folksonomy analysis methods such as tags similarity. The APIs are implemented as XML/SOAP web service, which is the W3C standard for remote methods invocation, so third-party applications can use the exposed APIs regardless of the programming language used or the platform they are deployed on. Figure 6.5 shows a snapshot of the web service that exposes the tag and place similarity functions. For instance, the *CalcTagTagRelatednessCosine* function calculates the Cosine similarity for any given two tags. The function receives the tags as input parameters and returns an XML response which contains the calculated similarity measure.

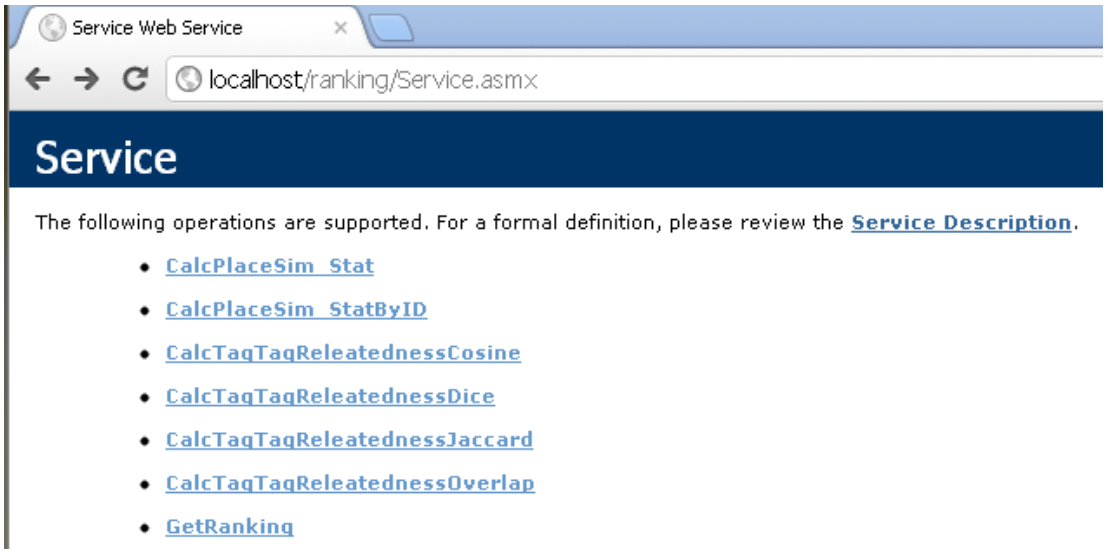


Figure 6.5: A snapshot of the XML/SOAP web service that exposes the geo-folksonomy APIs.

6.6 Scheduled Services

6.6.1 Folksonomy Analysis Application

All the folksonomy analysis work provided in this thesis is developed in a standalone console-based windows application. The application provides several analysis and data manipulation functions via two modes of operation: command line mode and menu mode. The command line mode allows the application to run as a scheduled service where no user input is required. The menu mode is designed to allow users to interact with the application; a menu of all the provided functions is printed on the screen and users are prompted to select the option they want to run.

Figure 6.6 shows a snapshot of the application. Similarity analysis using different measures, such as Cosine and Dice similarity, is provided through options 1 to 11. The input and output of the application is stored in the folksonomy DB described in Section 6.2. The tag classification and subsumption analysis are provided through options 12 to 17, where external ontologies are used to classify the tags into place types and activities, and the hierarchical relationships are inferred. Building the ontology and generating its RDF output are provided through options 18 to 22, and the generated RDF files are then imported using SemWeb.Net as described in Section 6.4.2. Finally, the methods used for evaluating the induced folksonomy using the Measure of Semantic Similarity (MSR) are provided through options 23 to 29.

6.6.2 Web Crawler

Custom crawler software was developed to collect geo-folksonomies from collaborative mapping applications on the social web. The design goal of the crawler is to be reusable and hence the implementation avoided hard-coding site-specific HTML/patterns in the code. For any geo-folksonomy application, the crawler assumes that there are separate pages to view places, tags and user information, and those pages are linked

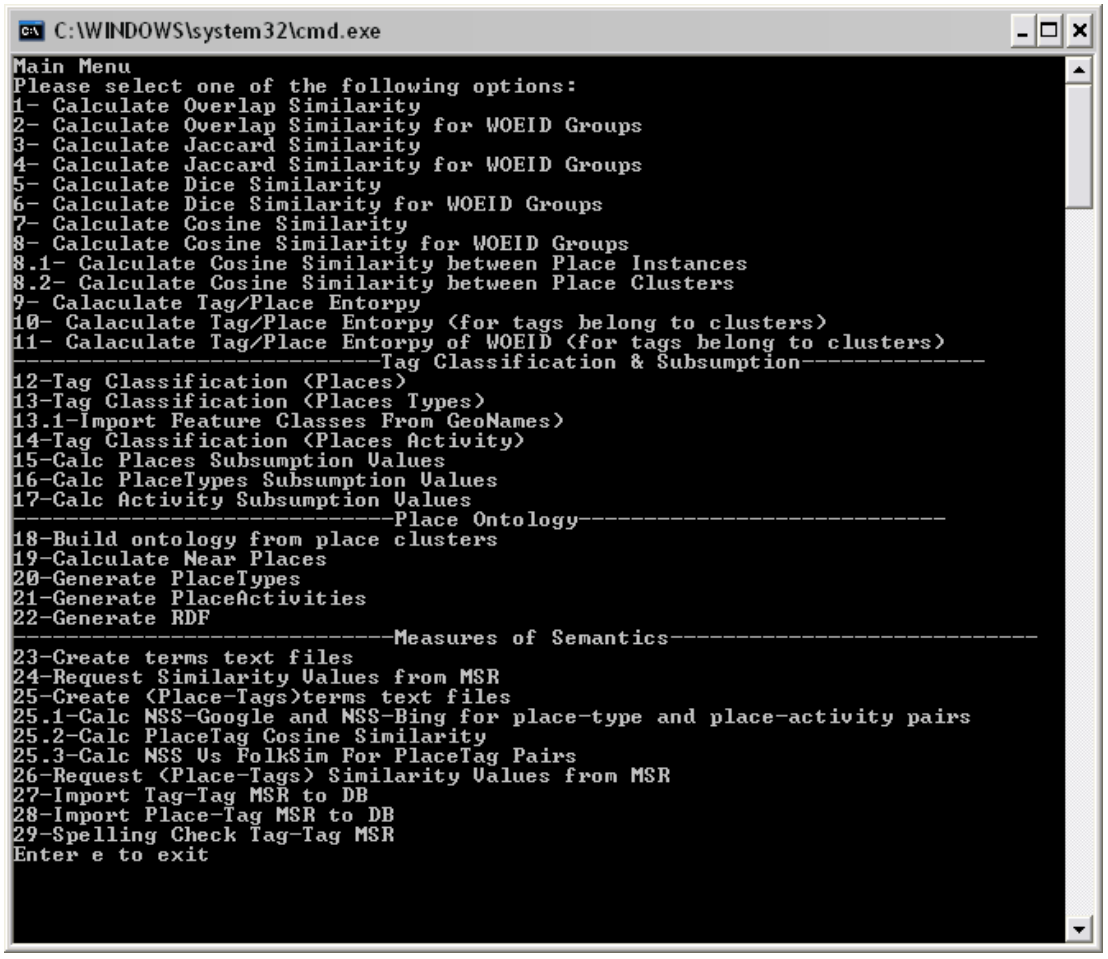
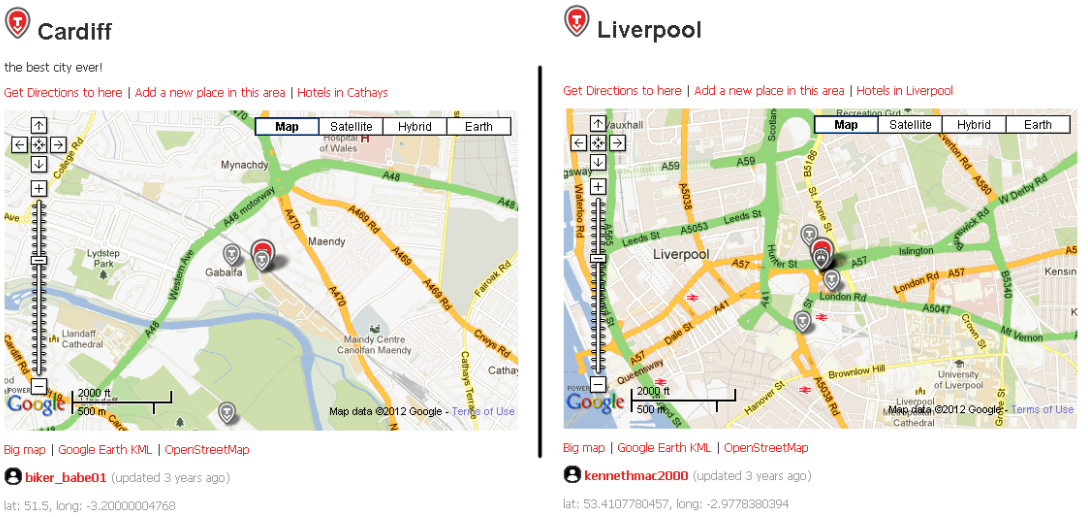


Figure 6.6: A snapshot of the folksonomy analysis application.

to each other using an HTML anchor element. Within a single geo-folksonomy application, all the pages that represent a single entity, i.e. a place, should have a consistent HTML pattern but different content. The crawler is designed to read the HTML patterns of the different entities from a separate configuration file, so the application can be configured to crawl different web sites without needing to recompile or rebuild the source code. In this research, the application is configured to process geo-folksonomies from Tagzania.com. Figures 6.7 (a) and (b) present a part of the place pages for *Cardiff* and *Liverpool* cities.

By having a closer look at the two snapshots, it is obvious that the pages follow the



(a) (b)

Figure 6.7: A snapshot of the place page for a) Cardiff and b) Liverpool on Tagzaina.com.

same pattern; place name is located on top, while the place description is located underneath the place name and is not mandatory. Google Maps applet is used to show the location of the place. The user who annotated the place along with the latitude and longitude of the place is rendered below the map applet. With such a consistent interface, a regular expression is an ideal solution to extract the required information from the place pages. Listing 6.4 shows an example regular expression that is used to extract the place location from the HTML.

```
1 <div class="\"geo\">
2 \n␣lat:␣<span␣class="\"latitude\">(?<lat>.+?)</span>, \n long: <
      span class="\"longitude\">(?<lon>.+?)</span>\n␣\n
3 </div>
```

Listing 6.4: Regular expression used to extract the location from the HTML page representing place information.

6.7 The SemTag Application

To demonstrate the utility of the proposed framework, an application, called SemTag, was developed [4] to display the derived place semantics. For comparison, these were displayed alongside the tag cloud for any given place resource. A tag cloud is used on social applications to display the most popular tags associated with a resource, directly based on co-occurrence analysis.

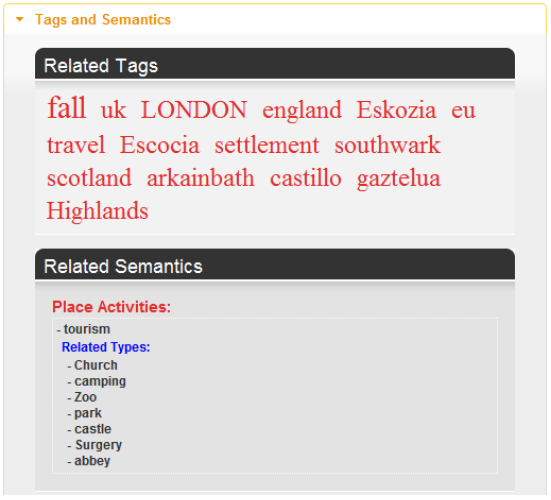


Figure 6.8: Screenshot of the SemTag application showing the derived place semantics for the place "London Eye".

The snapshot in Figure 6.8 shows part of the user interface displaying the tag cloud and the derived place types and activities for the place "London Eye". Note how the place type "tourism" is identified with this point of interest, but are not included in the tag cloud.

Figure 6.9 shows another snapshot for the place "London South Bank University". This is an interesting example of how different users can provide different semantics for the same place; the same place is tagged as "work" and "university"; both classified as place types. Also, some limitation of the tag resolution process is evident as shown,

[4] http://hobzy.cs.cf.ac.uk/SemTag

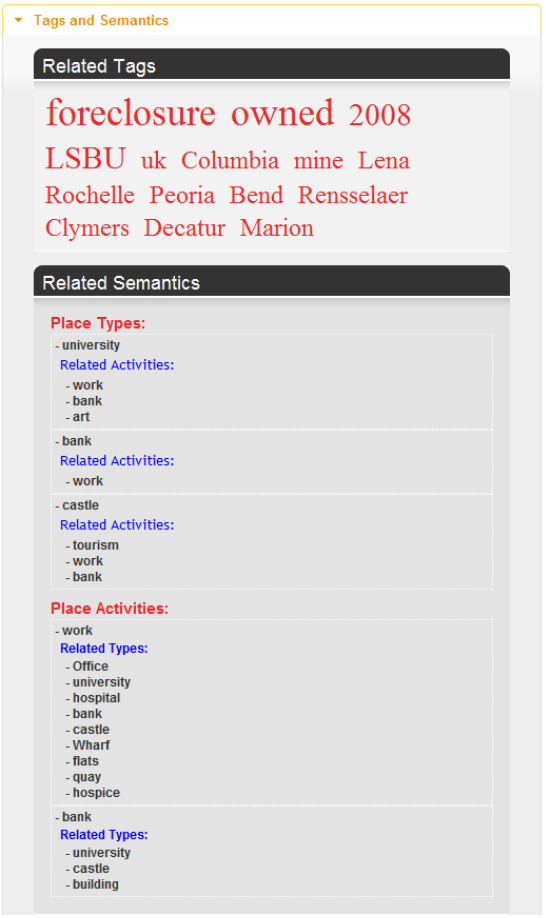


Figure 6.9: Snapshot of SemTag user interface showing the derived place semantics for the place "London South Bank University".

where "bank" was identified as an associated place type, whereas it is part of the place name. Further refinement of the tag resolution process and development of a more flexible place name recognition procedures can overcome this problem.

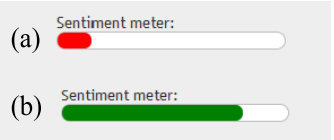


Figure 6.10: The sentiment score gadget showing a low score sentiment score.

A sentiment meter gadget is also implemented and presented on the interface to visualise the sentiment score of a place. The meter gadget is a 'progress bar'-like component

where colour is used to distinguish the score level: red colour for a low sentiment score, as in Figure 6.10(a), and a green colour for a high sentiment score as in Figure 6.10(b).

The application demonstrates the possible utility of the proposed framework, where it can be envisaged that the derived place semantics may be used to refine search queries and, when combined with the sentiment score, may be used to rank the retrieved search results.

6.8 Summary

This chapter covered the technical aspects of this research, where a multi-tier service oriented architecture was adopted to implement a system for extracting place semantics from geo-folksonomies. The system relies on two databases hosted in Microsoft SQL database server: a) Folksonomy DB for storing the collected geo-folksonomy along with the output of the analysis, b) RDF DB for storing the induced ontology. All the data access logic is implemented in the data access layer. The service layer contains web-based and windows-based components that encapsulate all the analysis and querying logic. The web-based components are designed to expose remote APIs to query the data, while the windows-based components are designed to collect and process the data. Components belong to the service layer access the databases though the methods implemented in the data access layer. A mapping application, SemTag, was presented which utilises the developed web services and APIs exposed by the service layer in order to demonstrate the utility of using place semantics to enhance the user experience on the web.

Chapter 7

Using Place Semantics to Enrich User Profiles

The collaborative and social interaction on web 2.0 allows users to create and annotate resources using tags. The tags created by individual users reflect their interest and can be used to build user profiles to support social network applications.

The methods used to create user profiles from social tags utilise the folksonomy co-occurrence analysis methods. Three different forms of user profiles built from folksonomies are discussed in this chapter. The simplest form of a user profile contains the tags that are directly used by that user in the folksonomy. A more complex form is to enrich user profiles with similar tags retrieved by co-occurrence similarity methods such as Cosine similarity. The co-occurrence methods used to enrich profiles are not capable of finding tags that are semantically related, more specifically, tags that represent related place concepts.

The work presented in this chapter builds on the discovered place semantics from Chapter 5. User profiles are enriched with concepts that are semantically related to the tags directly used by each user. The proposed user profile enrichment approach is demonstrated using a sample of geo-folksonomy dataset that covers an area in the City of London. In addition, user similarity is calculated using the enriched profiles approaches and the results are analysed and discussed.

7.1 Related Approaches to Extracting User Profiling Based on Folksonomies

Social tags generated by users' interaction on web 2.0 social bookmarking applications became the focus of much research in recent years. Social tags are uncontrolled vocabulary generated by users which represent their explicit topic interests. Moreover, tags may carry embedded semantics that reflect the user understanding of concepts and their relations. Analysing social tags can be beneficial to different research areas such as improving web search [8, 9, 10] and recommendation systems [73, 98, 55, 114].

Research on social tags can help improve functionalities of web applications, such as improving the current collaborative tagging systems [37], enhancing the navigation and the organization of web site content [10], extracting and modelling semantics embedded in social tags to enhance recommendation systems [33], and personalizing web search [8].

Social tags can be used to build user profiles. Sen et al. [95] argue that social tagging activities can be considered as an implicit rating behaviour, in other words, social tags can represent the interests and express the preferences of individual users. A user profile built from folksonomies is denoted by the set of tags representing the user interests with corresponding weights. The weight of a tag in the user profile represents the strength of the relationship between the user and that tag. Weights can be simplified by using a binary weighting approach such as in [12], or they can be calculated using methods such as TF-IDF [88], which is borrowed from text mining and is commonly used to assign weights to tags.

There are different approaches to build user profiles from social tags. Profiles can be built using users' own tags. For instance, Tso-Sutter et al. [104] proposed a user profiling approach that relates users to tags after converting the three-dimensional folksonomy relations into an extended user-tag rating matrix. Other approaches have been proposed to extend the process of building user profiles to use tags not directly used by

the user. For example, Niwa et al. [73] proposed an approach to build clusters of tags that are highly related based on tag similarity, then the clusters are used to extend user profiles. Au Yeung et al. [6] proposed a method called 'personomy', in which a cluster of all popular tags of the resources annotated by a user is used to profile topics of interest of that user. Other methods, such as association rules, as used in data mining, were used to find the related tags to tags in the user profile [48].

Although most of the user profiling approaches require decomposing the folksonomy tripartite graph into bipartite two-dimensional graphs, it is proposed by [85, 114] that user profiles can be built directly from the folksonomy graph. Rendle et al. [85] proposed the use of a three-dimensional tensor to profile users. Zhang et al. [114] suggested approaches to rank the weights of tags in the tripartite graphs to represent users' tagging behaviour. However, the conventional method of using the bipartite graph is followed in this chapter as it was found to be more convenient to illustrate and explain work on the user profile enrichment.

As discussed above, user profiles built from folksonomies are either basic, containing tags directly used by users, or enriched by including tags that are similar to the ones directly used by the user. The approaches used to find similar/related tags to enrich user profiles are based on the co-occurrence of the tags with users and resources. Such methods ignore the semantics that might be embedded in the tags. In this chapter, a user profile semantic enrichment approach is proposed based on the place semantics presented earlier in this thesis.

7.2 Constructing User Profiles from Folksonomies

A user profile built from a folksonomy can be represented by a vector Pf_u as follows:

$$Pf_u = (pf_{u,1}, pf_{u,2}, ..., pf_{u,|T|}) \tag{7.1}$$

Where $pf_{u,i}$ represents the strength of the association between the user u and the tag $t_i \in T$.

In this chapter, user profiles constructed from folksonomies are compared using the following approaches:

Direct Tags

Profiles constructed using this approach represent the interests of each user through the tags they used to annotate resources. The bipartite AC folksonomy graph is used to construct the profiles. The AC graph is defined as follows:
$AC = \langle A \times C, E_{ac} \rangle$
Where
$E_{ac} = \{(a, c) | \exists m \in M : (a, c, m) \in E\}$ and
$w : E \to \mathbb{N}, \forall e = (a, c) \in E_{ac}, w(e) := |\{m : (a, c, m) \in E)\}|$
Where M is the set of the values of the weight $w(e)$. The AC bipartite graph links users to tags that they have used to annotate resources. Each link is weighted by the number of times the user has used that tag to annotate resources. Hence, the user profiles can be calculated directly from the AC graph as $pf_{u,i} = ac_{u,i}$.
Where the AC matrix is denoted as $AC = \{ac_{u,i}\}$.

Similar Tags

A basic user profile is first constructed similar to the *Direct Tags* approach presented above. However, $pf_{u,i}$ is set to the number of times a tag t_i is used by a user u. The basic profile is enriched by using a tag similarity method, such as Cosine similarity, to find the tags similar to the ones in the basic profile. In this case, $pf_{u,i}$ is set to the value calculated by the tag similarity method. The enriched user profile $\hat{pf}_{u,i}$ is constructed using the following equation:

$$\hat{pf}_{u,i} = \begin{cases} \alpha & \text{if the tag is directly used by user} \\ \beta Max(Sim(t_i, t_j)) & \forall t_j \in T | pf_{u,j} > 0 \end{cases}$$

Where α and $\beta \in (0, 1]$ and can be used to facilitate building user profiles with different representations of direct and similar tags.

Semantically-Related Tags

A basic user profile is also constructed first. However, the place ontology introduced earlier in this thesis is utilised to enrich user profiles with tags that are semantically similar to the tags in the basic profile. Each tag in the basic profile is used to query the place ontology; if a tag is identified as a place type or place activity, all related concepts to this tag, within a specified semantic distance, are retrieved and used to enrich the basic profile. The enriched user profile $\hat{pf}_{u,i}$ is constructed using the following equation:

$$\hat{pf}_{u,i} = \begin{cases} \alpha & \text{if the tag is directly used by user} \\ \beta / Min(SemDist(t_i, t_j)) & \forall t_j \in T | pf_{u,j} > 0 \end{cases}$$

Where $SemDist$ is the semantic distance between the two tags t_i, t_j and α and $\beta \in (0, 1]$ and can be used to facilitate building user profiles with different representations of direct and similar tags.

7.2.1 Example of Enriching Basic User Profiles Using Place Semantics

Figure 7.1 illustrates an example of folksonomy consisting of four users, five tags and six place resources. The tagging activity of each user is represented by a line connecting user-tag-place. Basic user profiles can be constructed from the folksonomy graph where the place resources are removed and replaced by weights on the edges

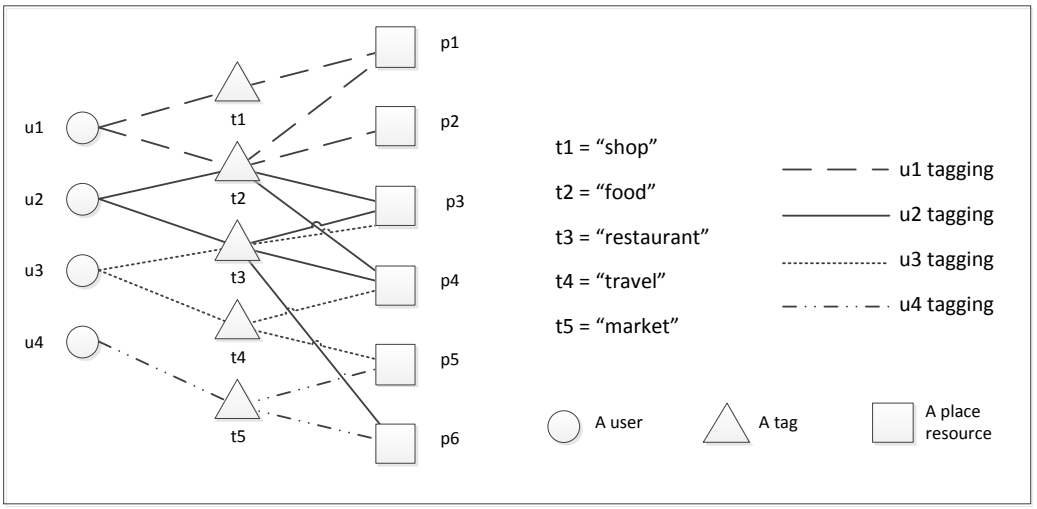


Figure 7.1: An example folksonomy.

between users and tags. Table 7.1 shows the matrix representation of the user profiles constructed from the folksonomy.

User/Tag	t1 (Shop)	t2 (Food)	t3 (Restaurant)	t4 (Travel)	t5 (Market)
U1	1	2	0	0	0
U2	0	2	3	0	0
U3	0	0	1	2	0
U4	0	0	0	0	2

Table 7.1: Basic user profiles constructed from the folksonomy.

Each row in Table 7.1 represents a user profile. The values in each cell are the weights/strengths of the relation between a user/tag pair. The weights in this example represent the number of place resources annotated by a user/tag pair.

For the purpose of illustrating the profile enrichment approach, assume that the folksonomy dataset does not contain any other tags and the semantic threshold is set to one step. For each tag, the place ontology is consulted to find the semantically related con-

cepts. Figure 7.2 shows a snapshot of the place type and place activity sub-ontologies, where concepts representing user profile tags in this example are highlighted. For example, the profile of user (U1) will be enriched with the tag "travel" because the profile already contains the tag "food" which has a one-step semantic distance to the tag "travel".

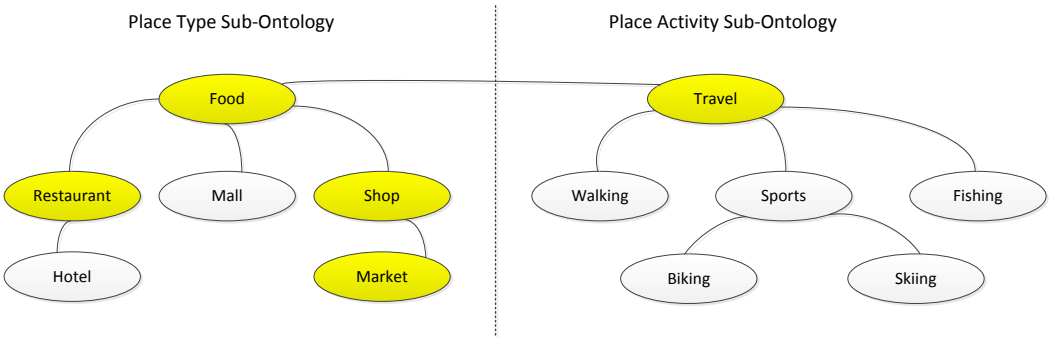


Figure 7.2: A snapshot of the place ontology illustrating the relations between the concepts in user profiles.

Using the above ontology for profile enrichment would change the profiles as shown in Table 7.2.

User/Tag	t1 (Shop)	t2 (Food)	t3 (Restaurant)	t4 (Travel)	t5 (Market)
U1	1	2	0.5	0.5	0.5
U2	0.5	2	3	0.5	0
U3	0	0.5	1	2	0
U4	0.5	0	0	0	2

Table 7.2: Enriched AC graph - User profiles constructed using $\alpha = 1$ and $\beta = 0.5$ for demonstration.

In the following section the proposed profile enrichment approach is applied to a real dataset and it is shown how the enriched profiles can be used to allow users to be

associated to relevant places (to their profiles). Moreover, the chapter studies how the enriched profiles affect user similarity calculation.

7.3 Description of the Dataset

A geographic region is chosen that covers places annotated within the City of London and is used in this experiment. The geo-folksonomy contains 299 users, 7810 tags and 9142 places. The average number of tags per place is 28, while the average number of tags per user is 52 tags. Also, the average number of users per place is four users.

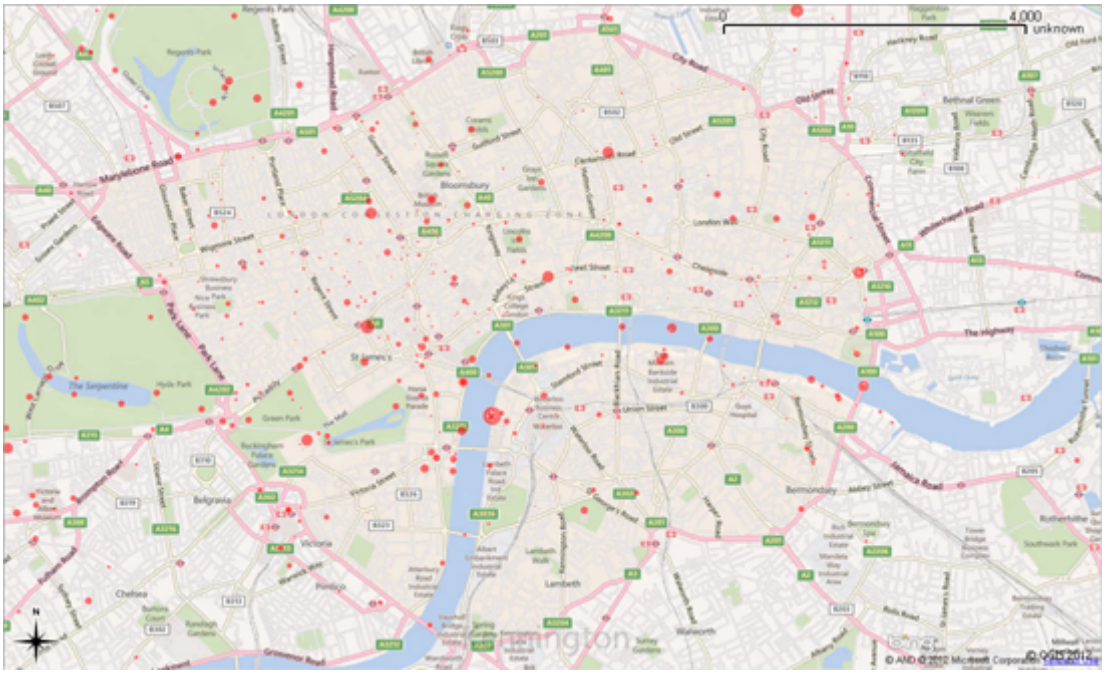


Figure 7.3: Place-Tag heat map.

Figures 7.3 and 7.4 show heat maps, covering the studied geographic area, presenting the density of relationships between places/tags and places/users respectively. The bigger the circle the larger the number of associations between a place and the users in the folksonomy.

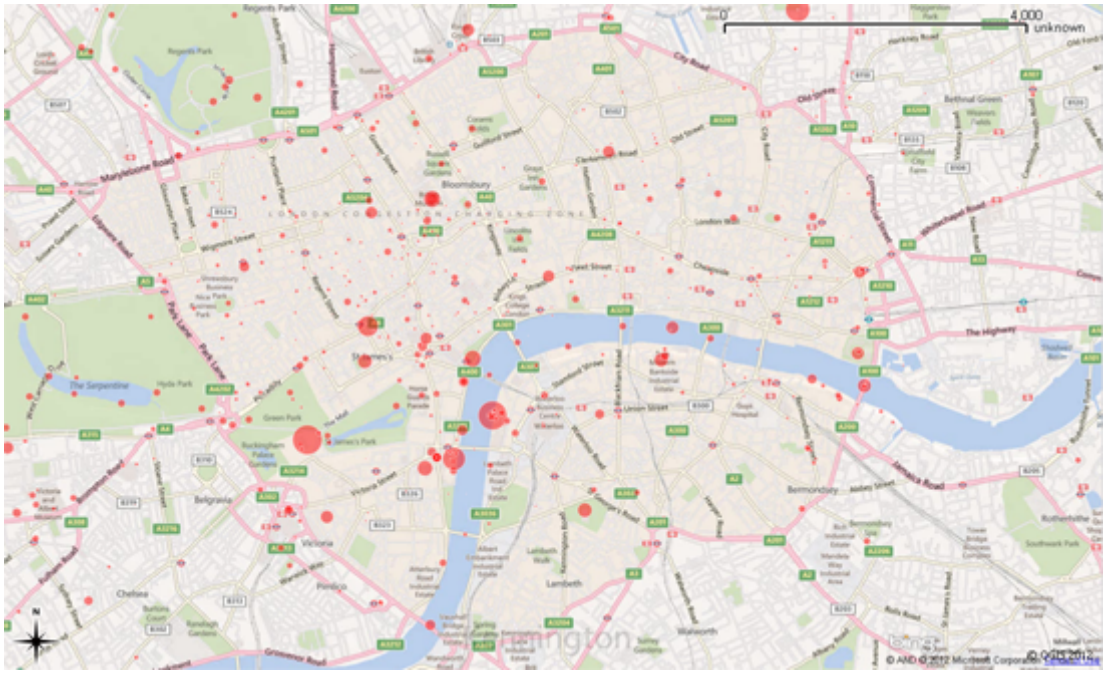


Figure 7.4: Place-User heat map.

7.4 Analysis and Results

7.4.1 User Profile

To analyse the place semantics generated using the different profile construction approaches, four user profile versions were built from the folksonomy: basic profile where the direct tags are used to construct the profiles; profile enriched with similar tags using Cosine similarity, and two profiles enriched with semantic-related tags using the place ontology with one and two-step semantic distance. Table 7.3 illustrates the output of the profiles in terms of the total number of place types and place activities against the total number of distinct tags used by the constructed profiles.

Enriching the basic user profiles using Cosine similarity with tags that are 80% or more similar to the tags directly used by users resulted in an increase of the total number of tags used in the profiles by 3252 tags, of which 41 are place types and 34 are place

Method/Count	Place types	Place activities	Distinct tags
Direct tags	191	63	3639
Cosine similarity	232	97	6891
Semantic similarity (1-step)	221	94	3700
Semantic similarity (2-step)	382	140	3907

Table 7.3: Total number of place types and activities in user profiles.

activities. Although a high threshold value is used, the number of the retrieved place semantics is small compared to the total number of tags retrieved.

Utilising the place ontology to enrich the basic user profiles by retrieving concepts with one-step semantic distance from the tags in the profile resulted in retrieving only 61 tags, of which 30 are place types and 31 are place activities. Also, there were 268 tags retrieved by increasing the threshold to two-step semantic distance, of which 191 are place types and 77 are place activities.

Enriching user profiles can also allow place resources in geo-folksonomies to be searchable and discoverable by more users. To illustrate this, the enriched user profiles were used to draw a heat map showing places and users who are related to this place. Two experiments were conducted to enrich the user profiles; in the first experiment, the related concepts were retrieved from the place ontology having the semantic distance set to one-step while in the second experiment, the semantic distance was set to two-steps.

The heat map shown in Figure 7.5 illustrates the relation between users and places after using the one-step semantic profile enrichment. The size of the circle representing a place increases if more users can be related to that place. A place and user are related if there is at least one common tag between the user profile and the tags of that place. Figure 7.6 shows the heat map after using the two-steps semantic profile enrichment. It is obvious that increasing the semantic distance in the profile enrichment process enables users to discover more resources.

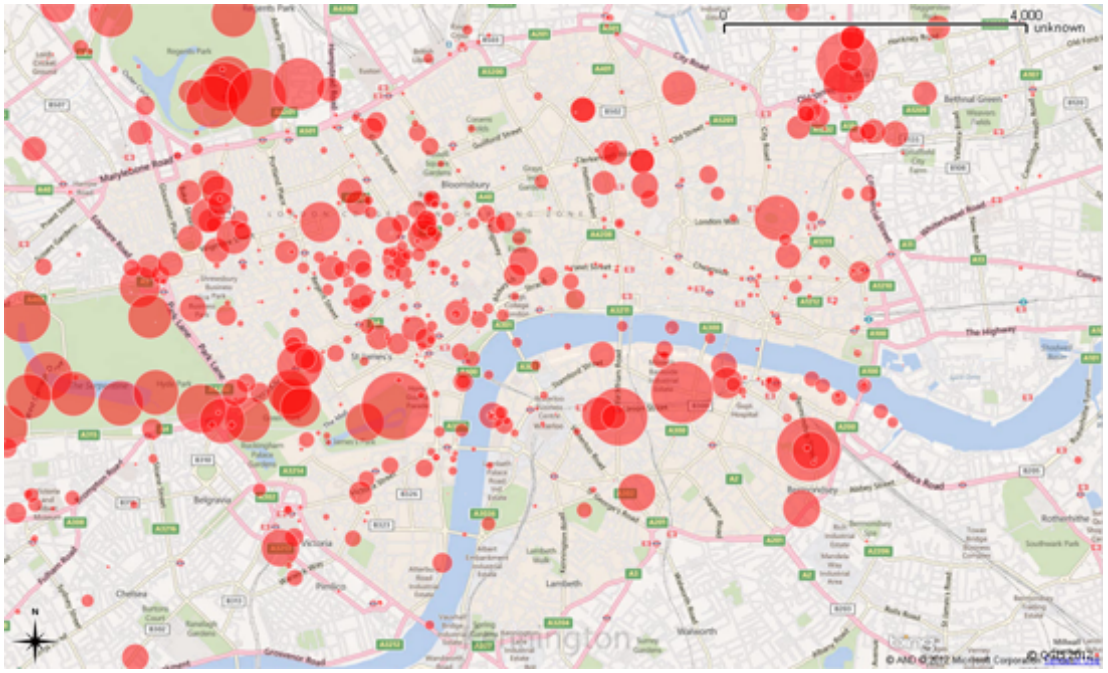


Figure 7.5: Place-User heat map with 1-step semantic distance - Places are associated with a larger number of users compared to 7.4.

7.4.2 User Similarity

Another experiment was conducted to analyse the effect of enriching user profiles on users' similarity. User to user similarity was calculated using Cosine similarity for the three versions of user profiles: direct tags, one-step and two-steps semantically enriched profiles. Table 7.4 shows statistics for the user similarity based on the three profile versions.

Profiles	Min	Max	Avg
Direct tags	0.0025	0.34	0.009
1-step semantic similarity	0.0025	0.4375	0.038
2-step semantic similarity	0.0025	0.56	0.191

Table 7.4: Statistics for user similarity using basic and semantically enriched profiles.

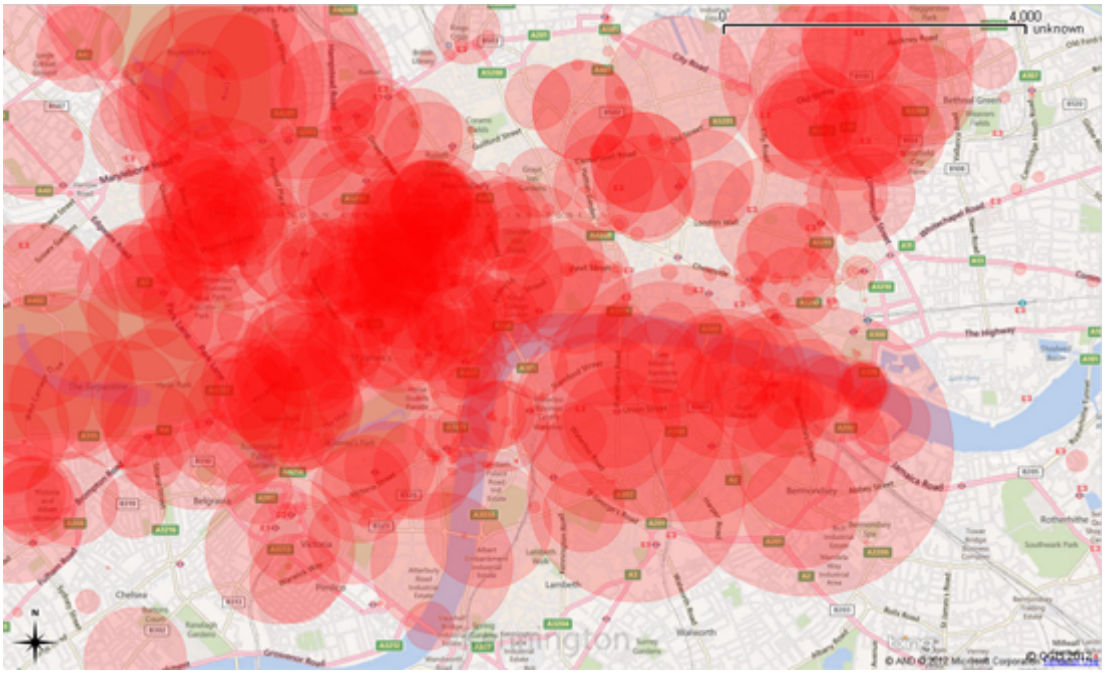


Figure 7.6: Place-User heat map with 2-steps semantic distance.

Figure 7.7 shows the complementary cumulative distribution function (CCDF) of user similarity using the three user profile versions. Here, the CCDF function describes the probability that a similarity value will be found at a value higher than or equal to x. It is noted that the enriched user profiles increase the probability of similarity matching.

For instance, the probability of having user similarity more than 0.1 is about 0.5 using the original profiles and about 0.55 using the enriched profiles (one-step) while it is about 0.7 using the enriched profiles (two-steps).

Another important factor to analyse is measuring the information content after enriching profiles. For example, enriching user profiles so that all place ontology concepts are used to enrich user profiles can lead to having all users to be almost 100% similar. However, such a scenario can result in having a very low information gain. However, measuring the amount of information (entropy) retrieved using all profile versions can be useful to understand the trade-off.

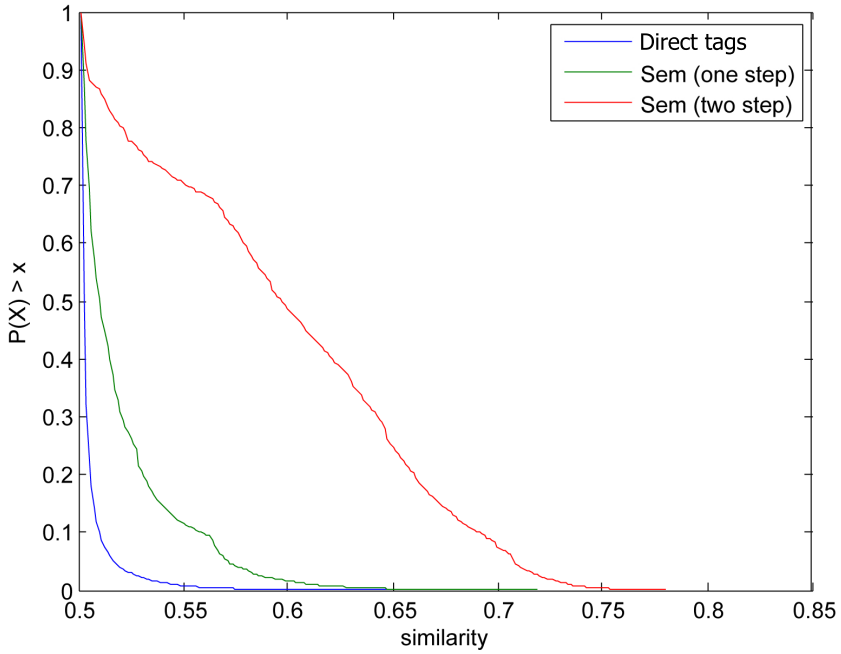


Figure 7.7: CCDF of user similarity using the three user profile versions.

Shannon's information gain [97] can be used to measure the amount of information produced in each experiment. The Mutual Information produced by each user u_x can be defined as:

$$I(u_x) = -\sum_{i=1}^{m} \log_2 k\left(u_{i,x}\right) \tag{7.2}$$

Where m is the number of users in the dataset and $k\left(u_{i,x}\right)$ defined by:

$$k\left(u_{i,x}\right) = \frac{s_{i,x}}{\sum_{j=1}^{n} s_{i,j}} \tag{7.3}$$

Where s is the user similarity value, and n is the number of users similar to user i. The information gain results are shown in table 7.5

It is clear that the information gain increases as the user profile gets richer. Although the uncertainty increases while information gain increases, it can be assumed that the maximum certainty in this case can exist if all users in the dataset are found to be

Method	Information gain
Direct tags	1.3669
1-step	3.5980
2-step	5.6198

Table 7.5: Information gain of the three versions of user profiles.

similar to each other, therefore the k will be equal to $1/m$ which implies that the maximum information gain can be calculated using the following equation:

$$I = m \log_2 m \tag{7.4}$$

Given that the number of users is 299, then the maximum entropy according to Equation 7.4 is 2458.98. Hence, the increase in the information gain (uncertainty) can be acceptable as it is a small fraction of the maximum information gain.

7.5 Summary

This chapter builds on the place ontology constructed from geo-folksonomy presented earlier in this thesis in Chapter 5. The possibility of using induced ontology to build user profiles was analysed here. Three approaches of building user profiles were discussed: basic profiles built with tags directly used by users to annotate resources; enriched profiles built with direct tags along with their similar tags using Cosine similarity, and semantically enriched profiles built with direct tags and their semantically related tags using the derived place ontology. The semantically enriched profiles were found to contain more place-related semantics when compared to the profiles enriched using Cosine similarity. Also, the semantically enriched profiles were used as different user similarity measures compared to the profiles with direct tags only, where users having interests that are similar, as derived from their associations with place semantics, could be related together.

Chapter 8

Using Place Semantics to Calculate Place Similarity

A place is normally represented using a set of attributes reflecting different facets, namely spatial and thematic attributes. Such attributes can be utilised to quantitatively measure place similarity. For instance, place location can be used to measure the spatial similarity between two places based on the distance between them such that the closer two places are the more similar they would be. Other place attributes can also be utilised to measure place similarity, such as place names and place types. In web 2.0 applications, places created using collaborative mapping applications are annotated with tags that are not place attributes; but these reflect users' views and experiences and hence can can be utilised to produce different place similarity views.

In this chapter, a folksonomy-based place similarity approach is presented, in which place profiles are constructed using the social tags of users who annotated those places. The created profiles are then used to measure the similarity between the places. Three types of place profiles are presented in this chapter: basic profiles built using the tags directly attached to the place; profiles enriched with similar tags retrieved by co-occurrence similarity methods and semantically-enriched profiles where place semantics, derived and encoded in place ontologies, are utilised to enrich place profiles. The place profile construction and enrichment approaches are demonstrated using a sample of the geo-folksonomy dataset and the results of the place similarity application are demonstrated and discussed.

8.1 Place Similarity Overview

A geographic place is normally represented by a set of properties that describe that place. Properties can capture spatial or geometric aspects such as location or boundary of a place, or they can capture thematic aspects such as place names and types. Moreover, properties can also capture relationships between multiple places, such as topological and directional relationships.

Modelling geographic places, in terms of what properties are used, is an important factor that affects how the place similarity is calculated. For example, in systems where place locations are modelled using a point representation, i.e. WGS84, together with a place name, a combined approach of spatial distance and string similarity can be used to measure place similarity [31].

Similarity between spatial scenes is a more general problem where a spatial scene contains multiple place objects along with their inter-relationships. In this case, measuring the similarity involves the assessment of the number of spatial operations needed to transform one scene to another using different spatial relationships such as topological, directional and metric [62, 15].

Place resources used in geo-folksonomies and geo-tagged web applications, such as Flickr photos and Tagzaina, are represented by simple objects that contains spatial and thematic properties. The spatial similarity approaches here are quantitative. Spatial similarity is a function of distance such that closer places are considered more similar, while the thematic similarity is calculated according to each thematic attribute. For instance, text similarity such as SoundEx or Levenshtein distance can be used to assist the similarity of place names.

In the GIR field, research has addressed the problem of improving spatial searches for geographic places by using thematic properties. For example, a method of assessing similarity is introduced in [32], where a combined similarity measure of place footprint, place name, place type and place hierarchy is used.

In this chapter, different methods of calculating place similarity are tested; a co-occurrence similarity approach is used to calculate the place similarity using the folksonomy structure. Also, the induced place semantics are used to calculate the semantic similarity between the places.

8.2 Constructing Place Profiles from Folksonomies

Similar to the approach proposed in Chapter 7 of building user profiles, a place profile built from a folksonomy can be represented by a vector Pf_o as follows:

$$Pf_o = (pf_{o,1}, pf_{o,2}, ..., pf_{o,|T|}) \tag{8.1}$$

Where $pf_{o,i}$ represents the strength of the association between the place resource o and the tag $t_i \in T$.

In this chapter, we compare place profiles constructed from folksonomies using the following approaches:

Direct Tags

Profiles constructed using this approach represent the keywords attached to each place through the tags used to annotate resources. The bipartite Concepts and Objects (CO) folksonomy graph, which links tags and places, is used to construct the profiles. The CO graph is defined as follows:

$CO = \langle C \times O, E_{co} \rangle$
Where
$E_{co} = \{(c, o) | \exists m \in M : (c, o, m) \in E\}$ and
$w : E \to \mathbb{N}, \forall e = (c, o) \in E_{co}, w(e) := |\{m : (c, o, m) \in E)\}|$

Where M is the set of the values of the weight $w(e)$. The CO bipartite graph links place resources to tags used by users to annotate resources. Each link is weighted by the number of times the user has used that tag to annotate resources. Hence, the user profiles can be calculated directly from the CO graph as $pf_{o,i} = co_{o,i}$.
Where the CO matrix is denoted as $CO = \{co_{o,i}\}$.

Similar Tags

A basic place profile is first constructed similar to the *Direct Tags* approach presented above. However, $pf_{o,i}$ is set to the number of times a tag t_i is used to annotate place o. The basic profile is enriched by using a tag similarity method, such as Cosine similarity, to find the tags similar to the ones in the basic profile. In this case, $pf_{o,i}$ is set to the value calculated by the tag similarity method. The enriched place profile $\hat{pf}_{o,i}$ is constructed using the following equation:

$$\hat{p}_{o,i} = \begin{cases} \alpha & \text{if the tag is directly used to annotate the place} \\ \beta Max(Sim(t_i, t_j)) & \forall t_j \in T | pf_{o,j} > 0 \end{cases}$$

Where α and $\beta \in (0, 1]$ and can be used to facilitate building place profiles with different representations of direct and similar tags.

Semantically-Related Tags

A basic place profile is also constructed first. However, the place ontology introduced earlier in this work is utilised to enrich place profiles with tags that are semantically similar to the tags in the basic profile. Each tag in the basic profile is used to query the place ontology; if a tag is identified as a place type or place activity, all related concepts to this tag, within a specified semantic distance, are retrieved and used to enrich the basic profile. The enriched place profile $\hat{pf}_{o,i}$ is constructed using the following equation:

$$\hat{p}_{o,i} = \begin{cases} \alpha & \text{if the tag is directly used to annotate the place} \\ \beta / Min(SemDist(t_i, t_j)) & \forall t_j \in T | pf_{o,j} > 0 \end{cases}$$

Where $SemDist$ is the semantic distance between the two tags t_i, t_j and α and $\beta \in (0, 1]$ and can be used to facilitate building place profiles with different representations of direct and similar tags.

8.3 Description of the Dataset

A popular area in central London, England, has been chosen for this demonstration. The size of the chosen area is about 16 km^2 and has the *British Museum* at its centre. The place dataset used here is the cleaned version of the geo-folksonomy built earlier in this work. The map in Figure 8.1 shows the *British Museum* place instances at the centre of the map represented by (yellow) stars. Each (green) circle represents a place cluster from the cleaned geo-folksonomy; a total of 283 unique places are shown in this map representing different kinds of places, for example: Wagamama, Design Museum, National Gallery and Madame Tussauds.

8.4 Analysis and Results

8.4.1 Place Profiles

To analyse the place semantics linked to each place through the user tags, four place profile versions are built from the folksonomy: a basic profile where the direct tags are used to construct the profiles, profile enriched with similar tags using Cosine similarity, and two profiles enriched with semantic-related tags using the place ontology with one and two-step semantic distance. Table 8.1 illustrates the output of the profiles in

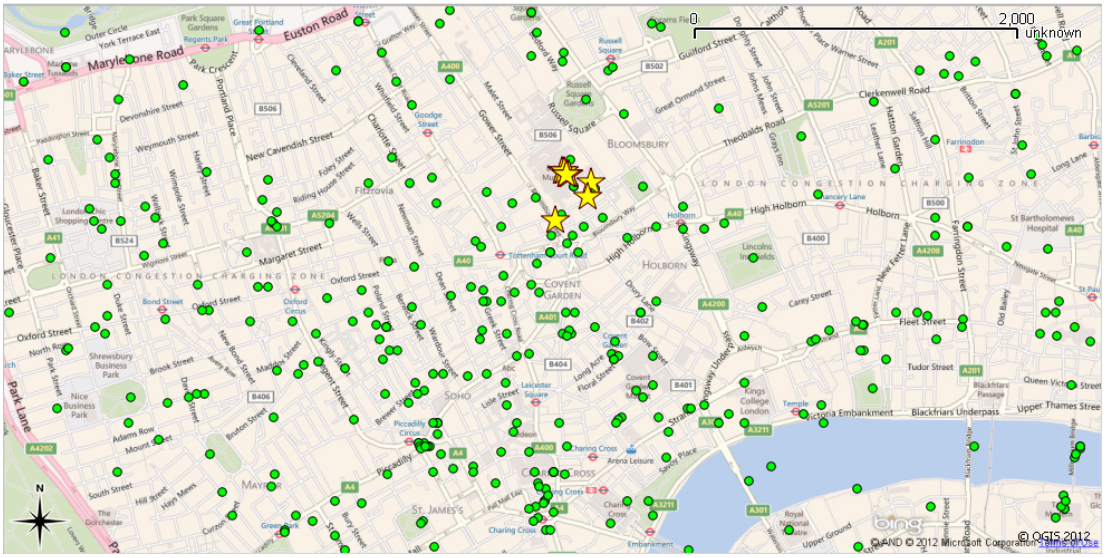


Figure 8.1: Places located around the British Museum in Central London.

terms of the total number of place types and place activities against the total number of distinct tags used in the constructed profiles for the 283 places in the dataset.

Method/Count	Place types	Place activities	Distinct tags
Direct tags	40	7	385
Cosine similarity	101	52	4462
Semantic similarity (1-step)	216	62	616
Semantic similarity (2-step)	328	87	721

Table 8.1: Total number of place types and activities in place profiles.

Enriching the basic place profiles using Cosine similarity with tags that are 80% or more similar to the tags directly used to annotate places resulted in an increase in the total number of tags used in the profiles by 4077 tags, from which 61 are place types and 45 are place activities. Although a high threshold value is used, the number of retrieved place semantics is small compared to the total number of tags retrieved.

Utilising the place ontology to enrich the basic place profiles by retrieving concepts with one-step semantic distance from the tags in the profile resulted in retrieving 231

tags, from which 176 are place types and 55 are place activities. Also, 366 tags were retrieved by increasing the threshold to two-step semantic distance, from which 286 are place types and 80 are place activities.

Enriching place profiles can also allow place resources in geo-folksonomies to be searchable and discoverable by more users. To illustrate this, the enriched place profiles are used to draw a heat map showing places and users who are related to this place. Two experiments are conducted to enrich the place profiles: in the first experiment, the related concepts are retrieved from the place ontology having the semantic distance set to one-step while in the second experiment, the semantic distance is set to two-steps.

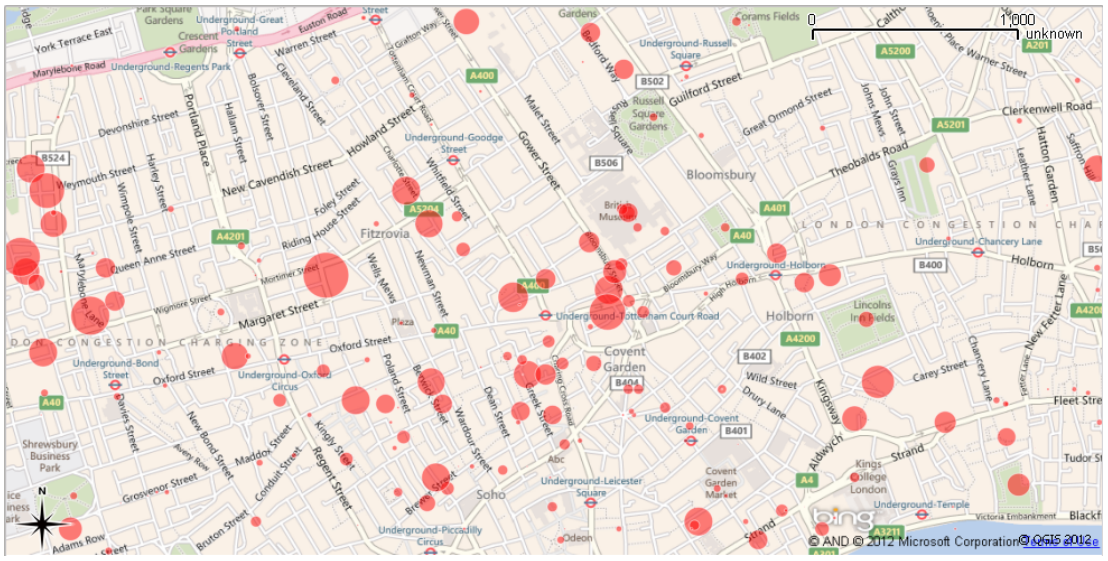


Figure 8.2: Place semantics heat map with 1-step semantic distance.

The heat maps shown in Figures 8.2 and 8.3 illustrate the amount of place semantics attached to each place using one and two-step semantically enriched profiles respectively. The size of the circle representing a place increases if more place semantics are attached to this place.

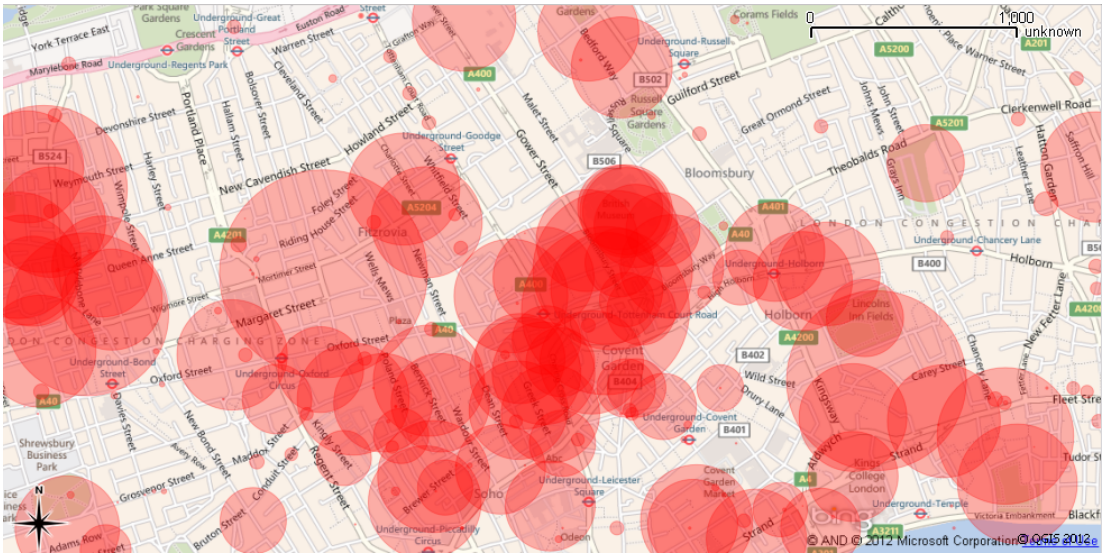


Figure 8.3: Place semantics heat map with 2-steps semantic distance.

8.4.2 Place Similarity

Using Folksonomy Co-Occurrence Analysis

Cosine similarity is used to calculate the similarity between the *British Museum* and the 283 place instances shown in Figure 8.1. The resulting similarity values range from 8.4% to 48.9% with a mean of 27.5% and standard deviation equal to 9.3%. A map-based representation of the Cosine similarity results is shown in Figure 8.4, where each place is represented by a circle and the size of the circle is directly proportional to the similarity value between the place represented by the circle and the *British Museum* instances represented by stars.

Figures 8.5 and 8.6 are different views of the Cosine similarity results showing the location of the places instances with similarity values less than (138 places) and more than (145 places) the average similarity value. Similar places are represented by triangles while circles are used to represent the rest of the places.

The top five places in each category along with their associated tags are listed in Table

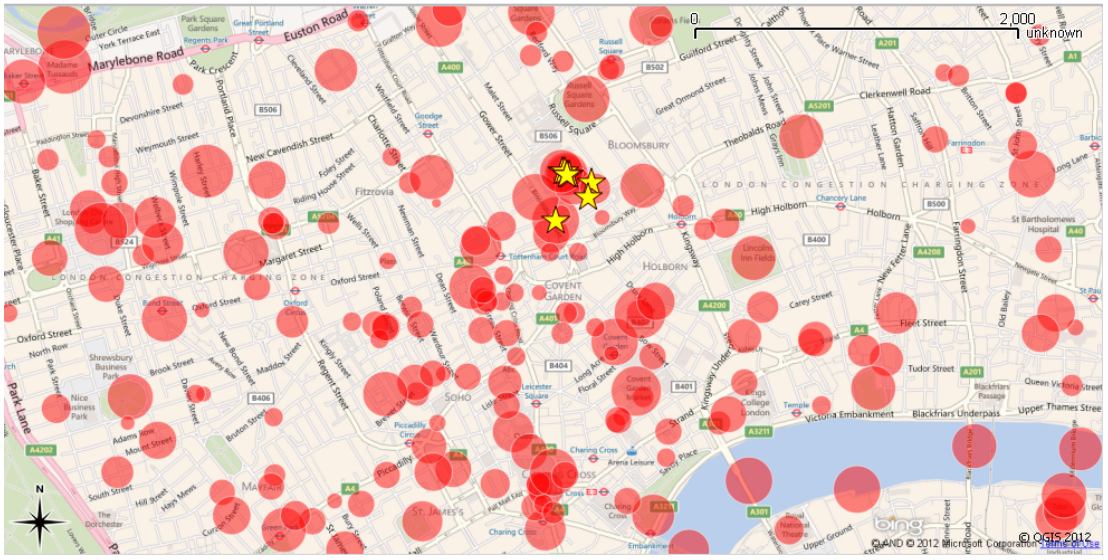


Figure 8.4: Heat map of places similar to British Museum using Cosine similarity.

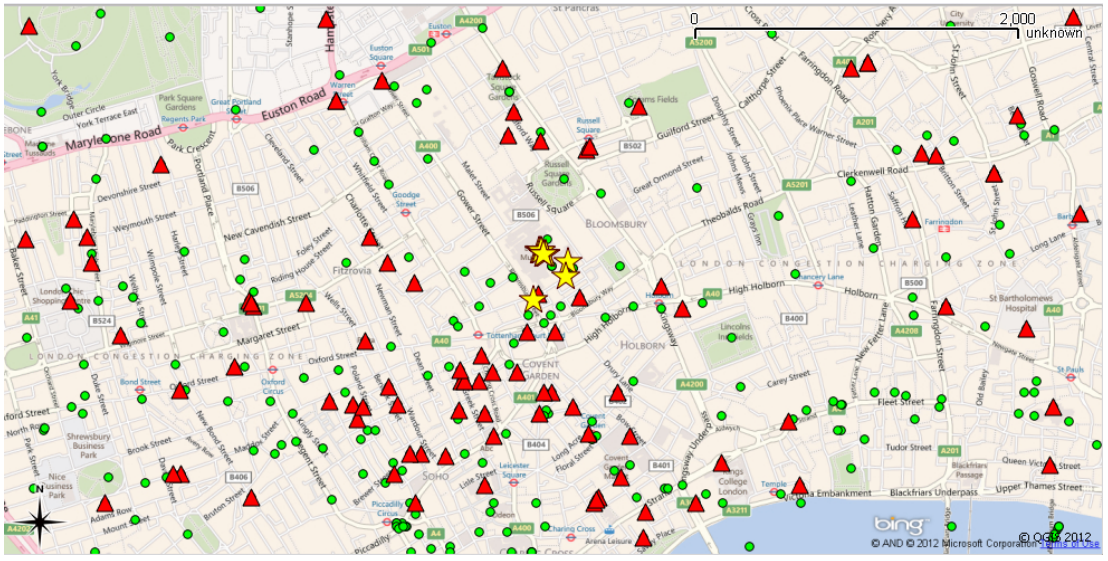


Figure 8.5: Location of the places similar to British Museum with similarity values < avg(sim).

8.2. The similarity maps show that there is no correlation between the spatial distribution of the place instances and the similarity value. In other words, similar place

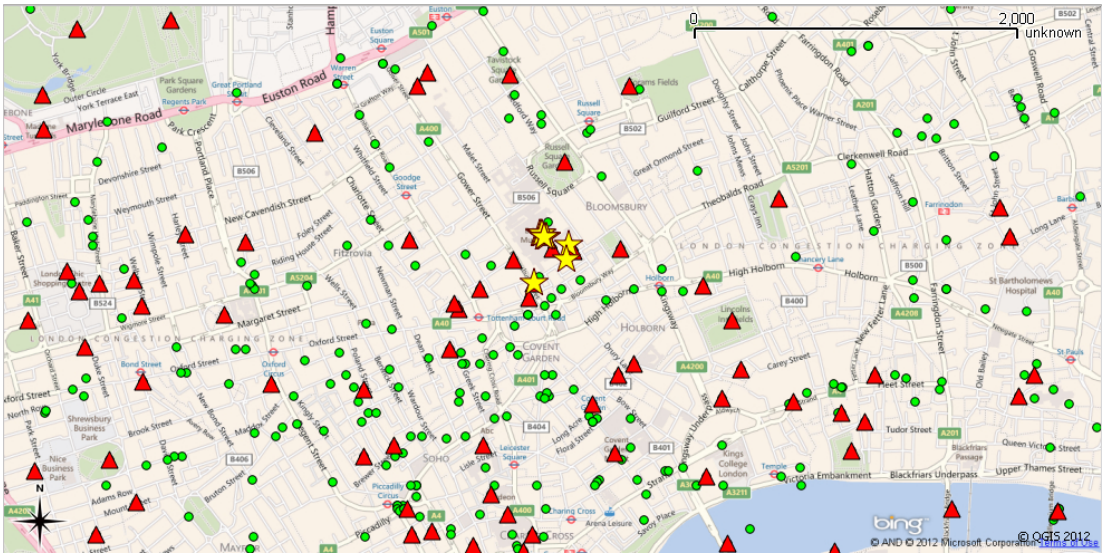


Figure 8.6: Location of the places similar to British Museum with similarity values $\geq$ avg(sim).

instances are not located spatially closer to the *British Museum*. This can be explained by the way Cosine similarity works; place instances are considered more similar if they share more tags in common and the method does not consider the spatial dimension while calculating the similarity.

Using The Induced Place Semantics

An interesting aspect of the research presented in this thesis is to be able to assess how semantically similar the places are. Semantic similarity can be guided by the place type and activity ontology introduced earlier in this work. Here, ontology is used to produce two different views of the places around the *British Museum*; a view that shows places that share the same semantics (of types and activities) attached to the *British Museum*, and another view of places with one-step semantic similarity distance. Place ontology is used for identifying tags that represent semantics as well as finding related semantics within a specified semantic distance. This process is carried out by running SPARQL

Sim	Place	Tags
sim < avg(sim)	The Green Park	park, bidaia, ikasketa, green, green_park_tag
	Trafalgar Square	ikasketa, square, bidaia, trafalgar, ikas
	Milk & Honey	bar, london
	Milroy's of Soho	whisky, london
	No. 6	restaurant, london
	London Bridge	bridge, london
sim ≥ avg(sim)	Old Operating Theatre Museum	southwark, museum, london, uk
	Madame Tussauds	travel, museum, waxworks, tussauds, london, uk
	Boating Lake - Regents Park	travel, panorama, united, kingdom, england, london, uk
	Harley Street	travel, panorama, united, kingdom, england, london, uk
	The Wallace Collection	travel, museum, art, wallace, collection, united, kingdom, england, london, uk

Table 8.2: Sample of similar to British Museum using Cosine similarity.

queries over the RDF store where the induced ontology is stored.

To produce the first view of the places that have the same semantics attached to the *British Museum*, the following SPARQL query is used:

```
1 SELECT ?concept WHERE {
2 {
3 ?concept a po:placeType.
4 ?concept po:hasName <tag>
5 }
6 UNION
7 {
```

```
?concept a po:placeActivity.
?concept po:hasName <tag>
}}
```

Listing 8.1: The SPARQL query used to check whether a tag represents a place type or activity.

Two tags are identified as carrying place semantics; *Museum* is identified as a place type and *Travel* is identified as a place activity.

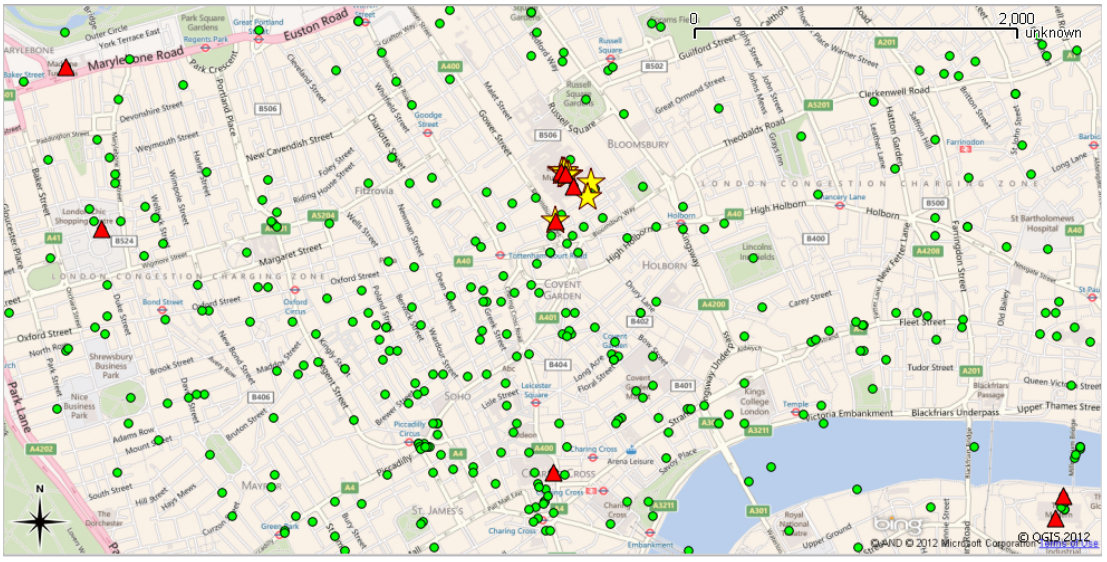


Figure 8.7: Places that have exact semantics as the British Museum.

Within the same area of central London, nine places were found to be semantically similar to the place *British Museum*, being annotated using the *Museum* or *Travel* tags. The locations of those places are shown in the map in Figure 8.7. The identified places include the following: Imperial War Museum, Design Museum, Science Museum, Natural History Museum, Madame Tussauds, The National Gallery and The London Dungeon. The place instances retrieved so far have strong semantic relations to the *British Museum*. However, the induced ontology can be used to find place instances that are semantically related to *British Museum* but with weaker semantic relationships, this can

be achieved by using the induced ontology to find instances that have semantics within n-steps distance from the source concepts. To illustrate this approach, a SPARQL query is executed over the induced ontology to retrieve all the concepts that are directly related to the concepts *Travel* and *Museum*. The general template of the query can be simplified as follows:

```
1 SELECT ?concept WHERE
2 {
3 ?x <relationType> ?concept.
4 ?x po:hasName <tag>
5 }
```

Listing 8.2: The SPARQL query used to retrieve concepts with specific relationships.

Where *relationType* is replaced with every relation linking place types and place activities in the ontology such as *po:subPlaceTypeOf* and *po:relatedPlaceType* while the *tag* is replaced by *Travel* and *Museum*. The above SPARQL query resulted in retrieving 82 concepts and a sample of the results is shown in Table 8.3.

Concept	Related Concept	Type/Activity
Museum	Art	A
Museum	Gallery	T
Museum	Design	A
Travel	Garden	T
Travel	Picnic	A
Travel	Park	T
Travel	Beach	T

Table 8.3: A Sample of the semantics that are one-step away from 'Travel' and 'Museum' concepts.

Places in the same area in central London that are annotated with any of the tags retrieved by the above SPARQL query are shown in triangles in the map in Figure 8.8. A

total of 140 places are identified to be semantically similar to the *British Museum* with a semantic distance of one step. Those places include the following: Piccadilly Circus, Waterloo Mainline Station, Houses of Parliament, Hyde Park, London Eye, Waterloo Bridge, Oxford Street and Marble Arch.

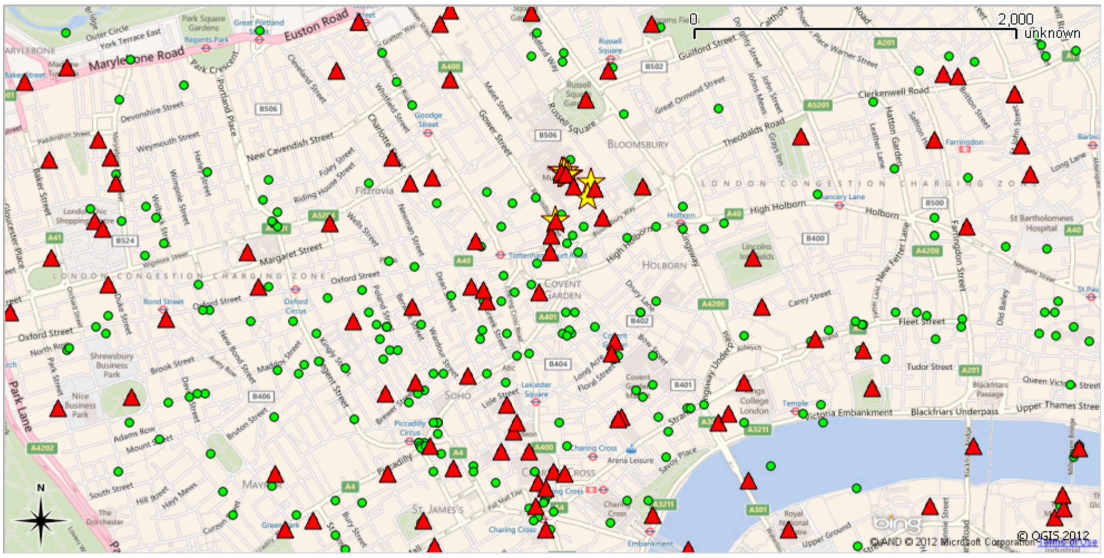


Figure 8.8: Places that have similar semantics (1-step) with the British Museum, shown as triangles.

8.4.3 Discussion

The geo-folksonomy created by the interactions of users on web 2.0 mapping applications can be used directly to assist the similarity of place instances using the co-occurrence analysis methods, in which the way the users annotate the places, reflected in the common tags between places, defines the place similarity. Also, the place type and activity ontology, which was originally induced from the geo-folksonomy dataset, can be used to assess the similarity of the place instances. It is important in this discussion to see the level of agreement between the two similarity approaches. Table 8.4 lists the top 10 semantically similar places along with the rank of each place in the

output of the cosine similarity:

Place	Ranking (Cosine)
Old Operating Theatre Museum	1
Imperial War Museum	–
Design Museum	–
The Wallace Collection	4
Science Museum	–
Natural History Museum	–
Earth Science Galleries	–
Madame Tussauds	2
The London Dungeon	10
Shakespeare's Globe	29

Table 8.4: The top 10 places that are semantically similar to the British Museum along with their ranking using the Cosine similarity.

The table shows that four places out of ten are found to be within the top ten Cosine similarity results while one place, Shakespeare's Globe, had a ranking of 29 in the results retrieved by the Cosine similarity. Half of the results could not be retrieved at all by the Cosine similarity, and almost all of the missed places are museums.

The Cosine similarity approach retrieves places that are annotated with common tags regardless of their associated semantics. However, co-occurrence analysis approaches are commonly used in web 2.0 applications for finding related tags represented usually as a Tag Cloud. There is almost a general consensus about the Tag Clouds which is that usually a small part of the tags is truly related while the rest of the tags are completely not related, either because they are too general or meaningless. For example, the tags uk, London, united and kingdom are used with most of the place resources in the folksonomy. Such tags are too general for a user searching for places located already in the UK. In this thesis, such tags are the main reason that un-related places are given high similarity values.

The semantic similarity approach overcomes the problems of the co-occurrence analysis approaches as the tags used in the similarity matched concepts in an ontology derived from the folksonomy. Another advantage of the semantic similarity is that different places can be considered similar even if they do not share any common tags. Grounding the tags to the place ontology facilitates finding related tags represented by concepts in the ontology, which is achieved by traversing the ontology relationships.

Chapter 9

Conclusion

Users' interactions and collaborations on web 2.0 mapping applications generate geo-folksonomies, in which geographic places are annotated with different kinds of place semantics, including vernacular place names, place types and activities people participate in, events, as well as personal opinions. Much interest has emerged in the geographic information retrieval community in the creation and population of place name resources to facilitate and enhance the search and retrieval of geographically-referenced information. Such research focuses primarily on finding place names and geographic locations of place instances. Geo-folksonomies embed rich user-oriented place semantics, which, if discovered, can potentially lead to much richer place knowledge resources and more personalized search and retrieval of web information content.

In this thesis, a framework is proposed for extracting some fundamental types of place semantics from tags in geo-folksonomies. In particular, a model of place, in which place types are associated with activities and services afforded, is used as a base to encode information derived from the folksonomies. Multiple web ontological resources are used to identify and match place type and activity concepts and statistical analysis is used to relate both types of concepts as presented in the folksonomy. A significant proportion of the tags associated with places can be analysed using sentiment analysis methods to discover general user opinions and feelings. Of the classified tags, place types and activities were more frequently used by users.

An application was developed to demonstrate how the discovered place semantics can

be employed to enhance the user experience in mapping applications, where, for each place instance, the related place semantics are displayed alongside the current method used in social applications of presenting tags as tag clouds. Moreover, the value of the discovered semantics is further revealed by deducing two semantic-based similarity approaches; user similarity and place similarity, where the results show that different similarity views can be produced by the proposed approaches which interestingly can represent different place and user dynamics based on the user tagging activities.

9.1 Evaluating Research Hypothesis

The research hypothesis for this thesis was presented in Chapter 1. To remind the reader, the core part of the hypothesis is reiterated below:

"User interaction on the social and collaborative mapping web can be used to deduce geographic and geo-semantic concepts of relevance to the user. Such relevant information can enhance their experience on the web in general."

The research documented in this thesis, particularly in Chapters 3, 4 and 5 tested this hypothesis to the point where it is possible to say that it does indeed hold true. The strategy followed to achieve this conclusion was to build a framework to a) collect realistic geo-folksonomy from the web that captures the users' interactions and collaboration on collaborative mapping applications; b) analyse the collected geo-folksonomy to extract the place semantics embedded in its structure and c) evaluate the extracted place semantics and explore their applications to enhance the user experience. Prior to analysing the geo-folksonomy, several quality problems in tags and place resources, which could affect the results of the analysis, are identified and addressed as discussed in Chapter 4. The analyses carried out to discover the semantics utilise external semantic data sources to identify the place-related concepts; the semantic relationships linking the identified concepts are discovered by employing several statistical co-occurrence methods as discussed in Chapter 5. The discovered semantics represent users' under-

standing of the places they are tagging which are found to be dissimilar to the place semantics provided by formal geographical data collection agencies such as Ordnance Survey. The evaluation of the framework is carried out manually via a survey study and automatically via validating the discovered semantic relationships using online semantic similarity services. The discovered place semantics are shown to be useful when utilised to improve the user interface of the collaborative mapping application as described in Section 6.7, and also shown to be beneficial to deduce semantic user similarity and semantic place similarity measures as described in Chapters 7 and 8 respectively.

9.2 Answers to the Research Questions and Problems

In this section, the research questions previously identified in Section 1.2 will be discussed in relation to the research undertaken in this thesis. Each research question will be repeated and the relevant research will be discussed including any related analysis, evaluation approaches and new knowledge that has been acquired.

1. **How good is the quality of tags and place resources in geo-folksonomies?**
   A folksonomy is a data structure generated from the users' interaction on social tagging applications that links tags, resources and users. Social tagging applications typically adopt an uncontrolled input approach which causes several problems to occur such as spelling mistakes. Such problems can affect the quality of folksonomy tags. Moreover, geo-folksonomies generated in social mapping applications introduce additional quality problems evident in place resources such as imprecise spatial locations and non-standard, vernacular place names. The combination of the problems in the tags and place resources can decrease the overall quality of the geo-folksonomy and can affect the results of any further analysis.

   A sample of a realistic geo-folksonomy dataset was explored to identify the

quality-related problems in tags and place resources. Several problems were identified which affected around 22% of the geo-folksonomy tags. A cleaning process targeting the tags collection was introduced in Section 4.1 which involves six steps, each of which targets a specific problem such as removing special characters, filtering stop words and removing duplicate tags. An additional cleaning process targeting the place resources was introduced in Section 4.2, where the redundant place resources referring to the same place in the real world were identified and merged using a hybrid textual and spatial clustering approach.

In order to quantify the quality improvement produced by the proposed cleaning approach, an evaluation method based on the Shannon's information gain is used to measure the uncertainty in the geo-folksonomy structure before and after the cleaning. The experiment described in Section 4.4 showed an improvement of the quality by around a 14% reduction in the uncertainty.

2. **How different are the place semantics extracted from geo-folksonomies from the semantics represented by place ontologies and gazetteers?**
   National mapping and geographical data collection agencies typically deliver place gazetteers and place type catalogues that capture the geographical dimension of places and are used for the purpose of classification of place entities. A place in general, can be associated with functions, services and activities that it provides to individuals. For example, national agencies such as the Office of National Statistics of the UK (ONSUK) provide classifications and definitions of economic activities for classifying business establishments by the type of economic activity in which they are engaged. Additionally, services afforded by a place are also modelled where a place can be associated with one or more service, some of which can be classified as primary services provided by that place while others can be classified as ancillary services that exist solely to support the principal ones.

The problem in such formal classifications of place types and services is that they are not intended to capture any specific experiences of users in a place. Since the main target of the research in this thesis is to capture users' understandings and experiences of places they tag in geo-folksonomies, a model of place is adopted where a geographic place can be associated with possible multiple place types and place activities. Place types and activities may themselves form individual subsumption hierarchies. Also, a place type can be associated with more than one type or activity and vice versa. This model of place allows to infer semantic relationships between the different entities (places, types, and activities) derived from the indirect associations as discussed in Section 3.2.

3. **How can the place semantics extracted from geo-folksonomies be evaluated?** The place ontology extracted from geo-folksonomies captures users' experiences and understandings of the places they are tagging. The most straightforward evaluation approach, comparing to a "golden standard" such as a formal place ontology, is not realistic here as the existing place ontologies are designed to capture geographical aspects of places.

   A questionnaire was designed to assess the quality of the extracted semantics which included five places in London, UK. The questions were designed to validate the concepts and relationships associated with the five places as discussed in Section 5.4.1 via two types of questions; the first type aimed at evaluating the quality of the relationships while the second type aimed at evaluating misclassified tags. Although the evaluation experiment was limited to a small number of places, the results suggested a correlation between the derived place ontology and users' perception of places and related semantics.

   Another evaluation experiment was conducted on a larger scale to measure the level of agreement between the derived semantics and the general semantics on the web. The Measure of Semantic Relatedness (MSR) web service is used which provides a set of methods to calculate the semantics similarity between

two terms. A total of 500 relationships from the induced place ontology were validated using the MSR service and the results demonstrated the validity of the place semantics which were found to be close to the semantics embedded in the web in general. The details of the experiment are provided in Section 5.4.2.

4. **Can the place semantics extracted from geo-folksonomies be utilised to calculate user similarity based on their place perceptions?**
   The work presented in Chapter 7 studied the feasibility of feeding back the discovered place semantics into the folksonomy to relate users who share similar understandings and experiences of places. User similarity calculated from folksonomies in general requires two steps of processing; firstly, constructing a profile for each user and secondly, calculating the similarity using the constructed profiles. In folksonomies, user profiles can be constructed straight away from the tags directly used by each user, so that a user profile is represented by a vector of dimensions equal to the total number of tags in the folksonomy. Vector-based similarity methods such as Cosine similarity can then be used to measure the similarity between any two profiles.

   In Chapter 7, two different profile enrichment approaches were presented; the first approach used tags that statistically related to the tags in each user profile retrieved using co-occurrence similarity. The second approach utilised the induced place ontology to build semantically-enriched user profiles. Two semantically enriched user profile versions are tested in this thesis; profiles enriched with one-step semantic distance and profiles enriched with two-step semantics distance from the tags directly associated with the user.

   A comparison of user profiles constructed using the mentioned approaches showed that the semantically enriched profiles contain more place-related concepts evident in the increased number of place types and place activities over the total number of tags. Such enriched profiles led to producing semantic similarity views of users based on their place interests.

5. **Can the place semantics extracted from geo-folksonomies be used to derive a new measure of place similarity that complements traditional dimensions used in the literature?**

   Place similarity is normally calculated using the spatial and thematic attributes of the place. The research presented in Chapter 8 studied the possibility of using the semantics extracted from geo-folksonomies to relate places based on the way people recognise the services provided by places and their related activities.

   The place similarity was calculated using the geo-folksonomy, where each place is characterised by the tags it was annotated with. The similarity value between two places was calculated as a function of those tags. The tags directly associated to each place were enriched by the similar tags using co-occurrence similarity. Also, a semantic-enrichment approach was employed to associate the places with tags that are semantically related to the tags directly attached to them. Two experiments were carried out for tags enrichment using the induced place ontology; using one-step semantic distance and two-steps semantic distance as described in Section 8.2.

   The place similarity calculated using the semantically enriched place profiles were compared to the place similarity calculated using the co-occurrence similarity approaches using the same experimental dataset to study the differences between both approaches. An experiment was conducted to test the overlap in the top 10 places that are similar to the British Museum, and it showed a weak overlapping between the outputs of both approaches. Moreover, the top 10 places retrieved by the semantic similarity approach are found (empirically) to be more related to the British Museum than the top 10 results retrieved by the co-occurrence similarity. The results strongly support the validity of the proposed approach of devising a place similarity approach based on the place semantics extracted from geo-folksonomies.

9.3 Utilising the Output of this Research

The research undertaken in this thesis has provided an approach for extracting place semantics from geo-folksonomies, where the extracted place semantics capture the social aspect of the places. Such semantics can be utilised in several ways to improve the existing state of the art. This thesis has highlighted how the existing web 2.0 mapping applications can improve the user experience by providing focused place-related information which could not be provided without this research. The details of this use case were given in Section 6.7. Moreover, this thesis has also highlighted how the extracted place semantics can be utilised in research. An approach of enriching the user/place profiles to produce place semantic similarity measures was discussed, in which the similarity measures were shown to be able to relate users and places based on the place affordance and user activities. These applications were described in detail in Chapters 7 and 8.

9.4 Future Work

This section describes the research not yet conducted, but that would be a valuable contribution to this research in the future.

9.4.1 Linking the Induced Ontology to other online Place Ontologies

The induced place ontology contains three different types of concepts: places, place types and place activities. Instances of those types can be linked to external ontologies, such as DBPedia and GeoNames, using the **rdf:seeAlso** or **owl:sameAs** properties. By linking the local concepts to external ontologies, users of the semantic web can benefit from integrating the knowledge produced by different providers about the

same concept. However, constructing those relationships to external entities requires research effort to choose the appropriate approaches to match the internal concepts to the external ones.

9.4.2 Extending the Framework to Use Multiple Folksonomy Data sources

The place ontology framework proposed in this research is designed to process geo-folksonomy data collected from one source. Building a richer geo-folksonomy data store collected from different sources can help in extracting richer place semantics. Integrating multiple data sources can lead to the problem of having redundant place instances of the same place. Identifying those place instances can be a challenging task especially because place resources, unlike normal web resources, cannot be uniquely identified using URLs. The place resources clustering approach proposed in Section 4.2 can be utilised to address this problem. However, further research may be needed to construct an integrated geo-folksonomy such as identifying same users across the different data sources.

9.4.3 Analysing the Unclassified Tags

More research is also needed to further analyse the unclassified tags, where there is a potential of extracting temporal information on events associated with a place; the tagging activities in particular are usually associated with the date and time when a user tagged a place. The unclassified tags can also be analysed to identify the homonyms and synonyms which can be represented by semantic relations in the induced ontology. Also, more resources can be used to enhance the process of place name identification and for handling abbreviations and vernacular names.

9.4.4 Improving the Sentiment Analysis Approach

The sentiment analysis used to calculate the sentiment score for a place is independent on the semantics attached to that place. This can be improved by calculating different sentiment scores based on the activities or types attached to a place. For example, a place such as “London Eye” can be given a positive sentiment score as a tourism place, but might be given negative score as a work place.

Appendix A

The OWL of the Place Ontology

A.1 Introduction

The following is the OWL of the induced place ontology presented earlier in this thesis in Section 3.2.

A.2 The OWL Source of the Ontology

```
1  <?xml version="1.0"?>
2  <rdf:RDF xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:po="
       http://cs.cardiff.ac.uk/2010/place-ontology#"
3  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
4    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
       -ontology#PlaceType">
5      <rdf:type>
6        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
           Class"/>
7      </rdf:type>
8    </rdf:Description>
9    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
       -ontology#PlaceActivity">
10     <rdf:type>
```

```
11        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              Class"/>
12      </rdf:type>
13    </rdf:Description>
14    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#Place">
15      <rdf:type>
16        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              Class"/>
17      </rdf:type>
18    </rdf:Description>
19    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#subPlaceTypeOf">
20      <rdf:type>
21        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
22      </rdf:type>
23      <rdfs:domain>
24        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
25      </rdfs:domain>
26      <rdfs:range>
27        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
28      </rdfs:range>
29      <owl:inverseOf>
30        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#superPlaceTypeOf"/>
31      </owl:inverseOf>
32    </rdf:Description>
33    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#superPlaceTypeOf">
34      <rdf:type>
35        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
```

```
36      </rdf:type>
37      <rdfs:domain>
38        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
39      </rdfs:domain>
40      <rdfs:range>
41        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
42      </rdfs:range>
43    </rdf:Description>
44    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#hasName">
45      <rdf:type>
46        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              DatatypeProperty"/>
47      </rdf:type>
48      <rdfs:domain>
49        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#Place"/>
50      </rdfs:domain>
51      <rdfs:domain>
52        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
53      </rdfs:domain>
54      <rdfs:domain>
55        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
56      </rdfs:domain>
57    </rdf:Description>
58    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#subPlaceActivityOf">
59      <rdf:type>
60        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
61      </rdf:type>
```

```
62      <rdfs:domain>
63        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
64      </rdfs:domain>
65      <rdfs:range>
66        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
67      </rdfs:range>
68      <owl:inverseOf>
69        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#superPlaceActivityOf"/>
70      </owl:inverseOf>
71    </rdf:Description>
72    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#superPlaceActivityOf">
73      <rdf:type>
74        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
75      </rdf:type>
76      <rdfs:domain>
77        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
78      </rdfs:domain>
79      <rdfs:range>
80        <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
81      </rdfs:range>
82    </rdf:Description>
83    <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#alternateName">
84      <rdf:type>
85        <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              DatatypeProperty"/>
86      </rdf:type>
87      <rdfs:domain>
```

```
88          <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
89        </rdfs:domain>
90      </rdf:Description>
91      <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#longitude">
92        <rdf:type>
93          <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                DatatypeProperty"/>
94        </rdf:type>
95        <rdfs:domain>
96          <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
97        </rdfs:domain>
98      </rdf:Description>
99      <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#hasPlaceType">
100       <rdf:type>
101         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                ObjectProperty"/>
102       </rdf:type>
103       <rdfs:domain>
104         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
105       </rdfs:domain>
106       <rdfs:range>
107         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#PlaceType"/>
108       </rdfs:range>
109     </rdf:Description>
110     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#hasPlaceActivity">
111       <rdf:type>
112         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                ObjectProperty"/>
```

```
113       </rdf:type>
114       <rdfs:domain>
115         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#Place"/>
116       </rdfs:domain>
117       <rdfs:range>
118         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
119       </rdfs:range>
120     </rdf:Description>
121     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
          -ontology#nearTo">
122       <rdf:type>
123         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
124       </rdf:type>
125       <rdfs:domain>
126         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#Place"/>
127       </rdfs:domain>
128       <rdfs:range>
129         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#Place"/>
130       </rdfs:range>
131       <rdf:type>
132         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              SymmetricProperty"/>
133       </rdf:type>
134     </rdf:Description>
135     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
          -ontology#relatedPlaceActivity">
136       <rdf:type>
137         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
138       </rdf:type>
```

```
139     <rdfs:domain>
140       <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
141     </rdfs:domain>
142     <rdfs:range>
143       <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
144     </rdfs:range>
145   </rdf:Description>
146   <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#relatedPlaceType">
147     <rdf:type>
148       <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              ObjectProperty"/>
149     </rdf:type>
150     <rdfs:domain>
151       <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceActivity"/>
152     </rdfs:domain>
153     <rdfs:range>
154       <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#PlaceType"/>
155     </rdfs:range>
156     <owl:inverseOf>
157       <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
              place-ontology#relatedPlaceActivity"/>
158     </owl:inverseOf>
159   </rdf:Description>
160   <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
        -ontology#hasDescription">
161     <rdf:type>
162       <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
              DatatypeProperty"/>
163     </rdf:type>
164     <rdfs:domain>
```

```
165         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
166       </rdfs:domain>
167     </rdf:Description>
168     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#hasID">
169       <rdf:type>
170         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                DatatypeProperty"/>
171       </rdf:type>
172     </rdf:Description>
173     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#latitude">
174       <rdf:type>
175         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                DatatypeProperty"/>
176       </rdf:type>
177       <rdfs:domain>
178         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
179       </rdfs:domain>
180     </rdf:Description>
181     <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/place
            -ontology#instancesCount">
182       <rdf:type>
183         <rdf:Description rdf:about="http://www.w3.org/2002/07/owl#
                DatatypeProperty"/>
184       </rdf:type>
185       <rdfs:domain>
186         <rdf:Description rdf:about="http://cs.cardiff.ac.uk/2010/
                place-ontology#Place"/>
187       </rdfs:domain>
188     </rdf:Description>
189   </rdf:RDF>
```

Listing A.1: The OWL of the induced place ontology.

Appendix B

Place Ontology Evaluation Survey

B.1 Introduction

The following is the summary of the responses of the survey used to evaluate the induced place ontology.

B.2 Summary of the Survey Responses

Place Information Survey

SurveyMonkey

1. Are you male or female?

	Response Percent	Response Count
Male	**76.5%**	**39**
Female	23.5%	12
	answered question	**51**
	skipped question	**2**

2. Which category below includes your age?

	Response Percent	Response Count
17 or younger	0.0%	0
18-20	**68.6%**	**35**
21-29	23.5%	12
30-39	3.9%	2
40-49	3.9%	2
50-59	0.0%	0
60 or older	0.0%	0
	answered question	**51**
	skipped question	**2**

3. What is the highest level of school you have completed or the highest degree you have received?

	Response Percent	Response Count
Less than high school degree	4.0%	2
High school degree or equivalent (e.g., GED)	**56.0%**	**28**
Some college but no degree	26.0%	13
Associate degree	2.0%	1
Bachelor degree	4.0%	2
Graduate degree	8.0%	4
answered question		**50**
skipped question		**3**

4. How familiar are you with city of London?

	Response Percent	Response Count
Very familiar	11.8%	6
A bit familiar	**45.1%**	**23**
Not familiar at all	43.1%	22
answered question		**51**
skipped question		**2**

5. Are you a native English speaker?

	Response Percent	Response Count
Yes	**80.4%**	**41**
No	19.6%	10
	answered question	**51**
	skipped question	**2**

6. Would you describe the place "Hyde Park" as a "Park"?

	Response Percent	Response Count
Yes	**79.2%**	**38**
No	6.3%	3
Not Sure	12.5%	6
Other (please specify)	2.1%	1
	answered question	**48**
	skipped question	**5**

7. Do you think "Parks" can be related to the activity of "Tourism"?

	Response Percent	Response Count
Related	41.7%	20
Maybe Related	**47.9%**	**23**
Not Related	10.4%	5
Other (please specify)	0.0%	0
	answered question	**48**
	skipped question	**5**

8. Do you think "Parks" can be related to "Water Activities" such as "Sliding" or "Swimming"?

	Response Percent	Response Count
Related	18.8%	9
Maybe Related	29.2%	14
Not Related	**52.1%**	**25**
Other (please specify)	0.0%	0
	answered question	**48**
	skipped question	**5**

9. Do you think "Parks" can be related to "Market Activities" such as "Buying" or "Selling"?

	Response Percent	Response Count
Related	14.6%	7
Maybe Related	29.2%	14
Not Related	**56.3%**	**27**
Other (please specify)	0.0%	0
	answered question	**48**
	skipped question	**5**

10. How would you describe the relation between "Parks" and "Walks"?

	Response Percent	Response Count
"Walks" can be part of "Parks"	**64.6%**	**31**
Related	31.3%	15
Not Related	4.2%	2
Not Sure	0.0%	0
Other (please specify)	0.0%	0
	answered question	**48**
	skipped question	**5**

11. How would you describe the relation between "Parks" and "Heritage Places"?

	Response Percent	Response Count
"Parks" can contain "Heritage Places"	**43.8%**	**21**
Related	25.0%	12
Not Related	10.4%	5
Not Sure	18.8%	9
Other (please specify)	2.1%	1
	answered question	**48**
	skipped question	**5**

12. In this question, each row represents a term that people use to describe places on the web, some of these terms might be strange or irrelevant. For each term, please choose the categories that you think are valid. You can select more than one category for each term.

	Place type	Activity you can do in the place	Other concept, but related	Response Count
serpentine	35.5% (11)	12.9% (4)	**51.6% (16)**	31
unitedkingdom	**75.6% (31)**	4.9% (2)	9.8% (4)	41
National	**57.6% (19)**	9.1% (3)	36.4% (12)	33
travel	5.1% (2)	**66.7% (26)**	33.3% (13)	39
england	**70.0% (28)**	7.5% (3)	15.0% (6)	40
Holiday	15.4% (6)	**56.4% (22)**	28.2% (11)	39
panorama	10.7% (3)	28.6% (8)	**64.3% (18)**	28
kingdom	**69.4% (25)**	2.8% (1)	27.8% (10)	36
DELL	15.4% (4)	11.5% (3)	**76.9% (20)**	26
Turk	14.8% (4)	11.1% (3)	**74.1% (20)**	27
albert	25.0% (6)	12.5% (3)	**79.2% (19)**	24
tube	38.2% (13)	**47.1% (16)**	26.5% (9)	34
resturant	42.4% (14)	**54.5% (18)**	12.1% (4)	33
lido	28.6% (8)	14.3% (4)	**60.7% (17)**	28
Herricks	8.3% (2)	12.5% (3)	**79.2% (19)**	24
			answered question	**43**
			skipped question	**10**

13. Would you describe "The Marriott" as a "Hotel"?

	Response Percent	Response Count
Yes	**95.3%**	**41**
No	0.0%	0
Not Sure	4.7%	2
Other (please specify)	0.0%	0
	answered question	**43**
	skipped question	**10**

14. How would you describe the relation between "Hotels" and "Casinos"?

	Response Percent	Response Count
"Hotels" can contain "Casinos"	**83.7%**	**36**
Related	9.3%	4
Not Related	7.0%	3
Not Sure	0.0%	0
Other (please specify)	0.0%	0
	answered question	**43**
	skipped question	**10**

15. How would you describe the relation between "Hotels" and "Swimming Pools"?

	Response Percent	Response Count
"Hotels" can contain "Swimming Pools"	**83.7%**	**36**
Related	11.6%	5
Not Related	4.7%	2
Not Sure	0.0%	0
Other (please specify)	0.0%	0
answered question		**43**
skipped question		**10**

16. How would you describe the relation between "Hotels" and "Venues"?

	Response Percent	Response Count
"Hotels" can contain "Venues"	**65.1%**	**28**
Related	23.3%	10
Not Related	9.3%	4
Not Sure	2.3%	1
Other (please specify)	0.0%	0
answered question		**43**
skipped question		**10**

17. In this question, each row represents a term that people use to describe places on the web, some of these terms might be strange or irrelevant. For each term, please choose the categories that you think are valid. You can select more than one category for each term.

	Place type	Activity you can do in the place	Other concept, but related	Response Count
TX	34.8% (8)	8.7% (2)	**56.5% (13)**	23
Courtyard	**90.0% (27)**	6.7% (2)	3.3% (1)	30
bbq	6.7% (2)	**93.3% (28)**	6.7% (2)	30
texasmonthly	5.0% (1)	30.0% (6)	**65.0% (13)**	20
rangers	5.0% (1)	20.0% (4)	**75.0% (15)**	20
ballpark	**48.1% (13)**	37.0% (10)	14.8% (4)	27
austin	**68.0% (17)**	4.0% (1)	28.0% (7)	25
houston	**77.8% (21)**	3.7% (1)	18.5% (5)	27
high	16.7% (3)	22.2% (4)	**61.1% (11)**	18
manassas	38.1% (8)	4.8% (1)	**57.1% (12)**	21
Fairfield	**56.5% (13)**	4.3% (1)	39.1% (9)	23
Syracuse	**60.0% (12)**	5.0% (1)	35.0% (7)	20
21	15.8% (3)	5.3% (1)	**78.9% (15)**	19
20	10.5% (2)	10.5% (2)	**78.9% (15)**	19
dallas	**77.8% (21)**	3.7% (1)	18.5% (5)	27
			answered question	**33**
			skipped question	**20**

18. Would you describe "Tesco" as a "Shopping Place"?

	Response Percent	Response Count
Yes	**97.5%**	**39**
No	0.0%	0
Not Sure	0.0%	0
Other (please specify)	2.5%	1
	answered question	**40**
	skipped question	**13**

19. Do you think "Shooping Places" can be related to "Market Activities" such as "Buying" or "Selling"?

	Response Percent	Response Count
Related	**92.5%**	**37**
Maybe Related	5.0%	2
Not Related	2.5%	1
Other (please specify)	0.0%	0
	answered question	**40**
	skipped question	**13**

20. How would you describe the relation between "Shopping Places" and "Sightseeing Places"?

	Response Percent	Response Count
"Shopping Places" can include "Sightseeing Places"	12.5%	5
Related	12.5%	5
Not Related	**72.5%**	**29**
Not Sure	0.0%	0
Other (please specify)	2.5%	1
	answered question	**40**
	skipped question	**13**

21. How would you describe the relation between "Shopping Places" and "Car Parks"?

	Response Percent	Response Count
"Shopping Places" can contain "Car Parks"	**82.5%**	**33**
Related	17.5%	7
Not Related	0.0%	0
Not Sure	0.0%	0
Other (please specify)	0.0%	0
	answered question	**40**
	skipped question	**13**

22. How would you describe the relation between "Shopping Places" and "Eating Places"?

	Response Percent	Response Count
"Shopping Places" can contain "Eating Places"	**77.5%**	**31**
Related	17.5%	7
Not Related	5.0%	2
Not Sure	0.0%	0
Other (please specify)		0
answered question		**40**
skipped question		**13**

23. In this question, each row represents a term that people use to describe places on the web, some of these terms might be strange or irrelevant. For each term, please choose the categories that you think are valid. You can select more than one category for each term.

	Place type	Activity you can do in the place	Other concept, but related	Response Count
for:jenna	6.7% (1)	13.3% (2)	**80.0% (12)**	15
Fashion	7.7% (2)	**42.3% (11)**	34.6% (9)	26
dresses	0.0% (0)	32.0% (8)	**56.0% (14)**	25
fun	4.0% (1)	**68.0% (17)**	20.0% (5)	25
NYC	**72.0% (18)**	4.0% (1)	24.0% (6)	25
mall	**76.7% (23)**	13.3% (4)	6.7% (2)	30
sports	0.0% (0)	**72.0% (18)**	24.0% (6)	25
clothing	0.0% (0)	33.3% (8)	**54.2% (13)**	24
friends	0.0% (0)	50.0% (10)	**55.0% (11)**	20
EBAFF	6.7% (1)	6.7% (1)	**86.7% (13)**	15
kitchen	**63.3% (19)**	13.3% (4)	16.7% (5)	30
youthmap	11.1% (2)	11.1% (2)	**77.8% (14)**	18
Bay	**80.8% (21)**	3.8% (1)	15.4% (4)	26
Outlet	**61.3% (19)**	16.1% (5)	16.1% (5)	31
sightseeing	4.0% (1)	**88.0% (22)**	8.0% (2)	25
			answered question	**32**
			skipped question	**21**

24. Would you describe "Wagamama" as a "Restaurant"?

	Response Percent	Response Count
Yes	**85.3%**	**29**
No	0.0%	0
Not Sure	11.8%	4
Other (please specify)	2.9%	1
	answered question	**34**
	skipped question	**19**

25. Would you describe "Wagamama" as a "Food Place"?

	Response Percent	Response Count
Yes	**91.2%**	**31**
No	0.0%	0
Not Sure	8.8%	3
Other (please specify)	0.0%	0
	answered question	**34**
	skipped question	**19**

26. Do you think "Food Places" can be related to "Clubs"?

	Response Percent	Response Count
Related	20.6%	7
Maybe Related	**44.1%**	**15**
Not Related	35.3%	12
Other (please specify)	0.0%	0
	answered question	**34**
	skipped question	**19**

27. Do you think "Food Places" can be related to "Market Activities" such as "Buying" or "Selling"?

	Response Percent	Response Count
Related	**61.8%**	**21**
Maybe Related	20.6%	7
Not Related	17.6%	6
Other (please specify)	0.0%	0
	answered question	**34**
	skipped question	**19**

28. How would you describe the relation between "Food Places" and "Restaurants"?

	Response Percent	Response Count
"Restaurants" can be classified as "Food Places"	**85.3%**	**29**
Related	11.8%	4
Not Related	0.0%	0
Not Sure	2.9%	1
Other (please specify)	0.0%	0
	answered question	**34**
	skipped question	**19**

29. Do you think "Restaurants" can be related to "Clubs"?

	Response Percent	Response Count
Strongly Related	5.9%	2
Related	35.3%	12
Not Related	**47.1%**	**16**
Not Sure	11.8%	4
Other (please specify)	0.0%	0
	answered question	**34**
	skipped question	**19**

30. In this question, each row represents a term that people use to describe places on the web, some of these terms might be strange or irrelevant. For each term, please choose the categories that you think are valid. You can select more than one category for each term.

	Place type	Activity you can do in the place	Other concept, but related	Related	Response Count
cleveland	**86.4% (19)**	0.0% (0)	13.6% (3)	0.0% (0)	22
voicebony2006	9.1% (1)	0.0% (0)	**81.8% (9)**	9.1% (1)	11
ramen	6.3% (1)	**37.5% (6)**	**37.5% (6)**	18.8% (3)	16
bostontrainmap	23.5% (4)	23.5% (4)	**41.2% (7)**	17.6% (3)	17
boston	**91.3% (21)**	0.0% (0)	8.7% (2)	0.0% (0)	23
station	**95.7% (22)**	0.0% (0)	4.3% (1)	4.3% (1)	23
train	**35.0% (7)**	**35.0% (7)**	30.0% (6)	10.0% (2)	20
ma	0.0% (0)	0.0% (0)	**66.7% (8)**	33.3% (4)	12
red+line	7.7% (1)	15.4% (2)	**38.5% (5)**	**38.5% (5)**	13
indonesian	**38.1% (8)**	0.0% (0)	**38.1% (8)**	23.8% (5)	21
dumpling	0.0% (0)	23.5% (4)	**52.9% (9)**	29.4% (5)	17
jakartan	33.3% (5)	13.3% (2)	**40.0% (6)**	13.3% (2)	15
				answered question	**26**
				skipped question	**27**

31. Would you describe "Imperial War Museum" as a "Museum"?

	Response Percent	Response Count
Yes	**97.0%**	**32**
No	0.0%	0
I don't know	3.0%	1
Other (please specify)	0.0%	0
	answered question	**33**
	skipped question	**20**

32. Do you think the place "Imperial War Museum" can be related to the the activity of "Travelling"?

	Response Percent	Response Count
Related	39.4%	13
Maybe Related	**48.5%**	**16**
Not Related	12.1%	4
Other (please specify)	0.0%	0
	answered question	**33**
	skipped question	**20**

33. How would you describe the relation between "Natural Places" and "Museums"?

	Response Percent	Response Count
"Natural Places" can be classified as "Museums"	30.3%	10
Related	21.2%	7
Not Related	**33.3%**	**11**
Not Sure	15.2%	5
Other (please specify)	0.0%	0
	answered question	**33**
	skipped question	**20**

34. How would you describe the relation between "Attraction Places" and "Museums"?

	Response Percent	Response Count
"Museums" can contain "Attraction Places"	**57.6%**	**19**
Related	39.4%	13
Not Related	0.0%	0
Not Sure	3.0%	1
Other (please specify)	0.0%	0
	answered question	**33**
	skipped question	**20**

35. In this question, each row represents a term that people use to describe places on the web, some of these terms might be strange or irrelevant. For each term, please choose the categories that you think are valid. You can select more than one category for each term.

	Place type	Activity you can do in the place	Other concept, but related	Response Count
kingdom	**75.0% (18)**	0.0% (0)	16.7% (4)	24
disaster	6.7% (1)	20.0% (3)	**73.3% (11)**	15
travel	9.1% (2)	**77.3% (17)**	4.5% (1)	22
england	**83.3% (20)**	0.0% (0)	8.3% (2)	24
panorama	6.3% (1)	18.8% (3)	**75.0% (12)**	16
politic	7.1% (1)	21.4% (3)	**71.4% (10)**	14
united	29.4% (5)	0.0% (0)	**70.6% (12)**	17
uk	**83.3% (20)**	0.0% (0)	8.3% (2)	24
LONDON	**79.2% (19)**	0.0% (0)	12.5% (3)	24
oxford	**86.4% (19)**	0.0% (0)	13.6% (3)	22
unitedkingdom	**78.3% (18)**	0.0% (0)	17.4% (4)	23
eu	**81.8% (18)**	0.0% (0)	18.2% (4)	22
settlement	**68.4% (13)**	10.5% (2)	26.3% (5)	19
scotland	**86.4% (19)**	0.0% (0)	13.6% (3)	22
Hyde	**85.0% (17)**	5.0% (1)	20.0% (4)	20
			answered question	**27**
			skipped question	**26**

36. If you have been to the following places before. How would you describe your experience?

	Positive experience	Neutral	Negative experience	N/A	Rating Average	Response Count
Hyde Park	37.5% (12)	6.3% (2)	0.0% (0)	**56.3% (18)**	0.86	32
The Marriott Hotel	28.1% (9)	0.0% (0)	3.1% (1)	**68.8% (22)**	0.80	32
Tesco	**56.3% (18)**	40.6% (13)	0.0% (0)	3.1% (1)	0.58	32
Wagamama	40.6% (13)	3.1% (1)	3.1% (1)	**53.1% (17)**	0.80	32
Imperial War Museum	31.3% (10)	0.0% (0)	0.0% (0)	**68.8% (22)**	1.00	32
				answered question		**32**
				skipped question		**21**

37. If you have any comments, feedback or suggestion, please feel free to let us know.

	Response Count
	2
answered question	**2**
skipped question	**51**

Glossary

API an Application Programming Interface (API) is a particular set of rules and specifications that a software program can follow to access and make use of the services and resources provided by another particular software program that implements that API. 90

Controlled vocabulary a list of predetermined terms that describe a specific domain. 14

Folksonomy is a type of categorization that consists of the aggregation of user-created keywords or tags used to describe a resource. 18

Geo-Folksonomy is a specialised type of folksonomy which contains only place resources instead of general web resources. 34

Lemmatization is a technique that transforms words to their base or dictionary forms. 50

Levenshtein edit distance a text similarity metric which calculates the distance between two words. More specifically, it counts how many letters have to be replaced, deleted, or inserted to transform on word into the other. The higher the Levenshtein edit distance, the more different two words are. 51

Ontology describes all concepts, instances and relations from a specific domain mostly

expressed in a formal format that is machine-interpretable. 26

Stemming is a technique that transforms words into their stems or roots. 50

Taxonomy belongs to the group of subject-based classification. It Puts all the terms in the controlled vocabulary into a hierarchy. 15

Thesaurus an extension of a taxonomy where different relations are included such as equivalence, hierarchical and associative relationships. 16

Acronyms

LINQ Microsoft Language Integrated Query. 94

LSAs Location Sharing Applications. 33

MSR Measure of Semantic Relatedness. 84

NSS Normalised Search Similarity. 85

ONSUK Office of National Statistics of the UK. 43

OS Ordnance Survey. 68

OSBP Ordnance Survey Building and Place ontology. 45

OWL Web Ontology Language. 27

PMI Point-wise Mutual Information. 85

QT Quality Threshold. 55

RDF Resource Description Framework. 27

SNA Social Network Analysis. 28

SOAP Simple Object Access Protocol. 90

WOEID Yahoo Where on Earth ID. 55

Bibliography

[1] AI Abdelmoty, P. Smart, and CB Jones. Building place ontologies for the semantic web:: issues and approaches. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 7–12. ACM, 2007.

[2] A.I. Abdelmoty, P.D. Smart, B.A. El-Geresy, and C.B. Jones. Supporting frameworks for the geospatial semantic web. In *SSTD '09: Proceedings of the 11th International Symposium on Advances in Spatial and Temporal Databases*, volume Lecture Notes in Computer Science 5644, pages 335–372. Springer-Verlag, 2009.

[3] Ahmed N. Alazzawi, Alia I. Abdelmoty, and Christopher B. Jones. An ontology of place and service types to facilitate place-affordance geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 4:1–4:2, New York, NY, USA, 2010. ACM.

[4] Ahmed N. Alazzawi, Alia I. Abdelmoty, and Christopher B. Jones. What can i do there? towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 26(2):345–364, 2012.

[5] S. Angeletou, M. Sabou, and E. Motta. Improving folksonomies using formal knowledge: A case study on search. *The Semantic Web*, pages 276–290, 2009.

[6] Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Contextualising

tags in collaborative tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 251–260, New York, NY, USA, 2009. ACM.

[7] S. Auer, J. Lehmann, and S. Hellmann. LinkedGeoData: Adding a spatial dimension to the Web of Data. *The Semantic Web-ISWC 2009*, pages 731–746, 2009.

[8] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, pages 501–510. ACM, 2007.

[9] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pages 22–26, 2006.

[10] S. Bindelli, C. Criscione, C. Curino, M. Drago, D. Eynard, and G. Orsi. Improving search and navigation by combining ontologies and social tags. In *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, pages 76–85. Springer, 2008.

[11] W. Blunt and W.T. Stearn. *Linnaeus: the compleat naturalist*. Princeton Univ Pr, 2002.

[12] T. Bogers and A. Van den Bosch. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290. ACM, 2008.

[13] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proc. of WWW 2009 Conference*, 2009.

[14] S. Braun, A. Schmidt, A. Walter, G. Nagypal, and V. Zacharias. Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In *Proceedings*

of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference (WWW 07), Banff, Canada, 2007.

[15] T. Bruns and M. Egenhofer. Similarity of spatial scenes. In *Proceedings of the 7th International Symposium on Spatial Data Handling*, pages 31–42, 1996.

[16] S. Butterfield. I am sharing this with you, August 2004. accessed on 22nd of June 2012.

[17] Gilberto Câmara, Antonio Miguel, Vieira Monteiro, Argemiro Paiva, Ricardo Cartaxo, and Modesto De Souza. Action-driven ontologies of the geographical space: Beyond the field-object debate. In *Proceedings 1st International Conference on Geographical Information Science, GIScience*, pages 52–54, 2000.

[18] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. *arXiv*, 805, 2008.

[19] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. *The Semantic Web-ISWC 2008*, pages 615–631, 2008.

[20] C. Cattuto, C. Schmitz, A. Baldassarri, V.D.P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *Ai Communications*, 20(4):245–262, 2007.

[21] H. Cramer, M. Rost, and L.E. Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 57–66. ACM, 2011.

[22] J. Cranshaw, R. Schwartz, J.I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM*,

volume 12, 2012.

[23] J. Davies, D. Fensel, and F. Van Harmelen. *Towards the semantic web*. Wiley Online Library, 2003.

[24] P. De Meo, G. Quattrone, and D. Ursino. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Information Systems*, 34(6):511–535, 2009.

[25] O. De Troyer, P. Plessers, and S. Casteleyn. Solving semantic conflicts in adience driven web design. In *Proceedings of the WWW/Internet 2003 Conference, Algarve Portugal*, 2003.

[26] Max J. Egenhofer. Toward the semantic geospatial web. In *Proceedings of the tenth ACM international symposium on Advances in geographic information systems*, pages 1–4. ACM Press, 2002.

[27] E. Emadzadeh, A. Nikfarjam, and S. Muthaiyah. A comparative study on measure of semantic relatedness function. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 1, pages 94–97. Ieee, 2010.

[28] F.A.Twaroch, C.B Jones, and A.I. Abdelmoty. Acquisition of vernacular place footprints from web sources. In Ricardo Baeza-Yates and Irwin King, editors, *Weaving Services and People on the World Wide Web*, pages 195–214. Springer Berlin Heidelberg, 2009.

[29] A. Frank. Ontology for Spatio-temporal Databases. volume Lecture Notes in Computer Science 2520, pages 9–77, 2003.

[30] J.C. French, A.L. Powell, and E. Schulman. Applications of approximate word matching in information retrieval. In *Proceedings of the sixth international conference on Information and knowledge management*, pages 9–15. ACM, 1997.

[31] G. Fu, C.B. Jones, and A.I. Abdelmoty. Building a geographical ontology for

intelligent spatial search on the web. In *Proceedings of IASTED International Conference on Databases and Applications*, pages 167–172. Citeseer, 2005.

[32] G. Fu, C.B. Jones, and A.I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, pages 1466–1482, 2005.

[33] W.T. Fu, T. Kannampallil, R. Kang, and J. He. Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(3):12, 2010.

[34] L.M. Garshol. Metadata? thesauri? taxonomies? topic maps! making sense of it all. *Journal of Information Science*, 30(4):378, 2004.

[35] A. Gilchrist. Thesauri, taxonomies and ontologies–an etymological note. *Journal of documentation*, 59(1):7–18, 2003.

[36] Fausto Giunchiglia, Fausto Giunchiglia, Maurizio Marchese, Maurizio Marchese, Ilya Zaihrayeu, and Ilya Zaihrayeu. Towards a theory of formal classification. In *Proceedings of the AAAI-05 Workshop on Contexts and Ontologies: Theory, Practice and Applications (C&O-2005*, pages 1–8. AAAI Press. ISBN, 2005.

[37] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[38] A. Gomez-Perez, M. Fernández-López, and O. Corcho. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Verlag, 2004.

[39] I.N. Gregory, C. Bennett, V.L. Gilham, and H.R. Southall. The Great Britain Historical GIS Project: from maps to changing human geography. *The Cartographic Journal*, 39(1):37–49, 2002.

[40] T. Gruber. Ontology of folksonomy: A mash-up of apples and oranges. *Interna-*

tional Journal on Semantic Web and Information Systems (IJSWIS), 3(1):1–11, 2007.

[41] T.R. Gruber et al. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 43(5):907–928, 1995.

[42] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220. ACM, 2007.

[43] Wen hao Chen, Yi Cai, Ho fung Leung, and Qing Li. Generating ontologies with basic level concepts from folksonomies. *Procedia Computer Science*, 1(1):573 – 581, 2010. ICCS 2010.

[44] G. Hart, S. Temple, and H. Mizen. Tales of the river bank: first thoughts in the development of a topographic ontology. In F. Toppen and P. Prastacos, editors, *Proceedings of the 7th AGILE Conference*, pages 165–168, Heraklion, 2004. Crete University Press.

[45] J. Hebeler, M. Fisher, R. Blace, and A. Perez-Lopez. *Semantic web programming*. Wiley, 2011.

[46] M. Hepp and J. de Bruijn. Gentax: A generic methodology for deriving owl and rdf-s ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. *The Semantic Web: Research and Applications*, pages 129–144, 2007.

[47] L.J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, 9(11):1106, 1999.

[48] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538. ACM, 2008.

[49] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford InfoLab, April 2006.

[50] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Folkrank: A ranking algorithm for folksonomies. pages 111–114, 2006.

[51] S. Intagorn, A. Plangprasopchok, and K. Lerman. Harvesting geospatial knowledge from social metadata. *Knowledge Creation Diffusion Utilization*, 2010.

[52] E.K. Jacob. Classification and categorization: a difference that makes a difference. *Library trends*, 52(3):515–540, 2004.

[53] B.J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3859–3864. ACM, 2009.

[54] B.J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[55] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *PKDD 2007: Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514, Berlin, Heidelberg, 2007. Springer-Verlag.

[56] C.B. Jones, A.I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The spirit spatial search engine:architecture, ontologies and spatial indexing. Lecture Notes in Computer Science 3234:125–139, 2004.

[57] H.G. Kim, S.H. Hwang, Y.K. Kang, H.L. Kim, and H.S. Yang. An agent environment for contextualizing folksonomies in a triadic context. *Agent and Multi-Agent Systems: Technologies and Applications*, pages 728–737, 2007.

[58] H.L. Kim, S.H. Hwang, and H.G. Kim. Fca-based approach for mining contextualized folksonomy. In *Proceedings of the 2007 ACM symposium on Applied computing*, pages 1340–1345. ACM, 2007.

[59] Werner Kuhn. Ontologies in Support of Activities in Geographical Space. *International Journal of Geographical Information Science*, 15(7):613–631, 2001.

[60] Werner Kuhn. An Image-Schematic Account of Spatial Categories. In *Spatial Information Theory*, volume Lecture Notes in Computer Science 4736, pages 152–168. Springer-Verlag, 2007.

[61] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions. *and reversals. Soviet Physics Doklady*, 10:707–710, 1966.

[62] B. Li and F. Fonseca. Tdd: A comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition & Computation*, 6(1):31–62, 2006.

[63] R. Lindsey, V.D. Veksler, A. Grintsvayg, and W.D. Gray. Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. In *8th International Conference of Cognitive Modeling, ICCM*, 2007.

[64] A. Maedche and S. Staab. Measuring similarity between ontologies. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 15–21, 2002.

[65] C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[66] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World Wide Web*, pages 641–650. ACM New York, NY, USA, 2009.

[67] A. Mathes. Folksonomies-cooperative classification and communication

through shared metadata. In *Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign*, 2004.

[68] I. Matveeva. *Generalized latent semantic analysis for document representation*. ProQuest, 2008.

[69] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5, 2007.

[70] P. Mika. *Social Networks and the Semantic Web (Semantic Web and Beyond)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2007.

[71] P. Morville. *Ambient findability*. O'Reilly Media, Inc., 2005.

[72] M.E.J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[73] Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Web page recommender system based on folksonomy mining for itng '06 submissions. In *Proceedings of the Third International Conference on Information Technology: New Generations*, pages 388–393. IEEE Computer Society, 2006.

[74] C.D. Paice. An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–50. Springer-Verlag New York, Inc., 1994.

[75] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), May 2010.

[76] Matthew Perry, Farshad Hakimpour, and Amit Sheth. Analyzing the space, and time: an ontology-based approach. In *Proceedings of the 14th annual ACM*

international symposium on Advances in geographic information systems, GIS '06, pages 147–154. ACM, 2006.

[77] A. Plangprasopchok and K. Lerman. Constructing folksonomies from user-specified relations on flickr. In *Proceedings of the 18th international conference on World Wide Web*, pages 781–790. ACM, 2009.

[78] A. Plangprasopchok, K. Lerman, and L. Getoor. Growing a tree in the forest: Constructing folksonomies by integrating structured metadata. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 949–958. ACM, 2010.

[79] A. Popescu, G. Grefenstette, and P.A. Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93. ACM, 2008.

[80] E. Quintarelli. Folksonomies: power to the people. *ISKO Italy-UniMIB meeting*, 24, 2005.

[81] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM New York, NY, USA, 2007.

[82] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1):1–30, 2009.

[83] M. Raubal and W. Kuhn. Ontology-based task simulation. *Spatial Cognition and Computation*, 4(1):15–37, 2004.

[84] E.C. Relph. *Place and placelessness*. Pion Ltd, 1976.

[85] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discov-*

ery and data mining, pages 727–736. ACM, 2009.

[86] R.D. Rugg, M.J. Egenhofer, and W. Kuhn. Formalizing behavior of geographic feature types. *Geographical Systems*, 4:159–180, 1997.

[87] S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice hall, 2010.

[88] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[89] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213. ACM, 1999.

[90] M. Sanderson and J. Kohler. Analyzing geographic queries. In *SIGIR Workshop on Geographic Information Retrieval*, 2004.

[91] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11:329–336, 2011.

[92] Simon Scheider and Werner Kuhn. Affordance-based categorization of road network data using a grounded theory of channel networks. *International Journal of Geographical Information Science*, 24(8):1249–1267, 2010.

[93] P. Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at World Wide Web, Edinburgh, Scotland*, 2006.

[94] S. Sen. Two types of hierarchies in geospatial ontologies. In *GeoSpatial Semantics*, volume Lecture Notes in Computer Science 4853, pages 1–19, 2007.

[95] S. Sen, J. Vig, and J. Riedl. Learning to recognize valuable tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 87–96. ACM, 2009.

[96] Sumit Sen. Use of affordances in geospatial ontologies. In *Proceedings of the 2006 international conference on Towards affordance-based robot control*, pages 122–139, Berlin, Heidelberg, 2008. Springer Verlag.

[97] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):55, 2001.

[98] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266. ACM, 2008.

[99] P.D. Smart, C.B. Jones, and F.A. Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Sixth international conference on Geographic Information Science, GIScience 2010*, volume Lecture Notes in Computer Science 6292, pages 234–248, 2010.

[100] L. Specia and E. Motta. Integrating folksonomies with the semantic web. *The semantic web: research and applications*, pages 624–639, 2007.

[101] E. Svenonius. *The intellectual foundation of information organization*. The MIT Press, 2000.

[102] K.P. Tang, J. Lin, J.I. Hong, D.P. Siewiorek, and N. Sadeh. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 85–94. ACM, 2010.

[103] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 2011.

[104] K.H.L. Tso-Sutter, L.B. Marinho, and L. Schmidt-Thieme. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1995–1999. ACM,

2008.

[105] Eric Tsui, W. M. Wang, C. F. Cheung, and Adela S. M. Lau. A concept-relationship acquisition and inference approach for hierarchical taxonomy construction from tags. *Inf. Process. Manage.*, 46(1):44–57, 2010.

[106] Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag.

[107] M. Uschold. Building ontologies: Towards a unified methodology. *TECHNICAL REPORT-UNIVERSITY OF EDINBURGH ARTIFICIAL INTELLIGENCE APPLICATIONS INSTITUTE AIAI TR*, 1996.

[108] C. Van Damme, M. Hepp, and K. Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. *Bridging the Gap between Semantic Web and Web*, 2:57–70, 2007.

[109] T. Vander Wal. Explaining and showing broad and narrow folksonomies, 2005. Accessed 22dn of June 2012.

[110] V.D. Veksler, A. Grintsvayg, R. Lindsey, and W.D. Gray. A proxy for all your semantic needs. In *29th Annual Meeting of the Cognitive Science Society*, 2007.

[111] Thomas Vander Wal. Folksonomy, url: http://www.vanderwal.net/folksonomy.html. 2007.

[112] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[113] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. 1998.

[114] Z.K. Zhang, T. Zhou, and Y.C. Zhang. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs. *Physica A: Statistical Mechanics and its Applications*, 389(1):179–186, 2010.

[115] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pages 680–693. Springer-Verlag, 2007.