Int J Digit Libr
DOI 10.1007/s00799-017-0227-5

Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents

Hélène de Ribaupierre[1] · **Gilles Falquet**[2]

Received: 30 September 2016 / Revised: 11 July 2017 / Accepted: 1 August 2017

Abstract When scientists are searching for information, they generally have a precise objective in mind. Instead of looking for documents "about a topic T", they try to answer specific questions such as finding the definition of a concept, finding results for a particular problem, checking whether an idea has already been tested, or comparing the scientific conclusions of two articles. Answering these precise or complex queries on a corpus of scientific documents requires precise modelling of the full content of the documents. In particular, each document element must be characterised by its discourse type (hypothesis, definition, result, method, etc.). In this paper, we present a scientific document model (SciAnnotDoc ontology), developed from an empirical study conducted with scientists, that models the discourse types. We developed an automated process that analyses documents effectively identifying the discourse types of each element. Using syntactic rules (patterns), we evaluated the process output in terms of precision and recall using a previously annotated corpus in Gender Studies. We chose to annotate documents in Humanities, as these documents are well known to be less formalised than those in "hard science". The process output has been used to create a SciAnnotDoc representation of the corpus on top of which we built a faceted search interface. Experiments with users show that searches using with this interface clearly outperform standard keyword searches for precise or complex queries.

✉ Hélène de Ribaupierre
deribaupierreH@cardiff.ac.uk

Gilles Falquet
gilles.falquet@unige.ch

[1] School of Computer Science and Informatics, Cardiff University, Cardiff, UK

[2] CUI, University of Geneva, Geneva, Switzerland

Keywords SciAnnotDoc model · Information retrieval · Knowledge management · Ontologies · Semantic publishing

1 Introduction

In their work, scientists need to gather very specific information with different goals and tasks in mind. For example, when they write a paper, they must access information about recent findings in their domain and compare them with their own work, or they need to find a definition of a concept and compare it with alternative definitions. When reading documents with these purposes in mind, they need to be exhaustive (they want to retrieve all the new findings in their domain, or all the definitions for this concept), but they are not necessarily interested in reading other parts of the documents (background, hypothesis, etc.). They face many challenges to accomplish these kinds of tasks. The number of publications increases every year (e.g. Medline has an annual growth rate of 0.5 million items [22]). Consequently, the time needed to search and read all available information tends to increase. Therefore, the researchers need tools to help them search, organise and read scientific papers, avoiding time spent on irrelevant articles. Recently, systems supporting electronic scientific publications have introduced some improvements compared with simple searches of printed edition indices, but the potential benefits of electronic publishing, for example, systems that could easily find a bit of information, or summarise scientific documents, or combine them, have not been realised yet.

The aim of this research is to improve the tools that search information in scientific documents to provide more accurate information retrieval for scientists. Currently, information retrieval systems (IRs) for scientific documents work with metadata (e.g. author's name, title) or with the full text

Published online: 09 August 2017

indices. This way of querying documents only works when scientists want a very specific document and know some metadata like the title or the name of the author; however, this method is not efficient for more complex queries. Typical examples of such complex queries are: "Find all the definitions of the term gender" or "Look for changes in the meaning of the term knowledge base over time" or "find all the findings that analyse why the number of women in academia falls more sharply than the number of men after their first child, using qualitative and quantitative methodologies". These examples were all extracted from real cases.[1] In this context, scientists need to find specific parts of the documents (we will call them fragments) and not the whole document, to find precise answers to be able to help them in their tasks. The information they are looking for can be classified as different types of discourse elements such as "definition", "hypothesis", "methodology", "findings".

The problem of complex queries of scientific documents has previously been addressed from three different perspectives. First, the scientific community has proposed machine-readable scientific document formats. However, those formats are not based, to the best of our knowledge, on empirical studies conducted with real users. These formats or models will omit some of the fragments or discourse elements needed for answering a precise or complex query. Second, the community of Natural Language Processing (NLP) has developed a variety of methods for extracting and disambiguating information from scientific documents and their metadata. However, we still lack effective methods to extract and describe semantically a number of more structured and fine-grained entities. Third, the community of Information Science has developed a variety of models of users' behaviour based on empirical studies. However, few of these works focus on the seeking process of scientists and if seeking patterns change depending on the task scientists want to accomplish.

The aim of this work is to bring these three perspectives together. By starting the creation of the annotation model with a user-centric perspective, the model should comprehend the users seeking behaviour based on empirical studies and could be used to automatically annotate scientific documents using some of the NLP methods and tools. The annotations will be evaluated with "standard" methodologies such as precision and recall. The effectiveness of the system will be evaluated with scientists. This work is a step towards the creation of new information retrieval systems that instead of working on the metadata of the scientific documents will work on the semantic contents. These annotations will also permit the creation of new visualisations of the knowledge contained in the scientific corpus of a research area and create new ways for cross-referencing documents.

[1] From interviews with scientists, see below.

The rest of the paper is organised as follows. In Sect. 2, we present the user studies made with scientists and the user model. In Sect. 3, we present the annotation model obtained from the previous user model. In Sect. 4, we detail the annotation process and present the results we have obtained with this process in terms of precision and recall. This model and the annotation process have already been published in [7–9]; however, the extended version of the user model and the extended version of the user evaluation are new contributions. In Sect. 5, we present the user evaluation with the faceted search interface and the keywords search interface, and finally, in Sect. 6, we discuss and summarise the main conclusions and outline future directions of research.

2 User studies and user model

In order to correctly annotate the documents, one needs to know what scientists consider as important in the text. Of the studies that examine scientists' behaviour, only a few spend some time examining how scientists read articles. Currently, most published studies examine why scientists are reading an article, how much time they spend reading it, or in which database they found it [17,35]. Fewer studies show findings on users reading behaviour and utilisation of scientific documents. For example, Bishop [3], showed that scientific writings are "disaggregated and reaggregated" in the process of constructing new knowledge in science and engineering. Bishop showed that some components that have been indexed in a digital library (figures, conclusion, references, article titles, title, heading, caption, authors, full text, etc.) can be used for this process, but Bishop focussed on the physical and logical structure of the documents more than on the semantic content of the documents. This study showed that users can utilise search engines retrieving specific components, but the study was restricted to researchers in science and engineering, as well as the structural elements such as tables and figures and did not consider the semantic content. Renear et al. [24] showed that scientific readers extract and accumulate bits of specific information such as findings, equations, protocols and data. They also found that the literature is usually scanned to find results and to monitor the progress of peers and competitors, and to extract facts and evidence to build a database. Tenopir et al. [30] showed that scientists are using the literature for a range of purposes, including teaching, writing articles, proposals, reports, continuous education, advising others and research. Tenopir et al. showed that reading scientific articles can inspire new ways of thinking, create new ideas, improve results, save time or other resources, resolve technical problems, etc. In [33], authors analysed the information needs of scientists in life science in relation to citation. Their analysis revealed two tasks often performed by participants: the appraisal task and

the citation-focused task. They implemented a prototype for Citation-Sensitive In-Browser Summariser (CSIBS) according to the user requirements, allowing scientists to determine whether or not a cited document is worth reading further. However, this work was done only in one field, life science, and both the interviews and summary were related to this area of research.

Understanding the behaviour of scientists is essential to building a user model that will be at the core of an efficient information retrieval system for scientific documents. To improve our understanding of the information needs of scientists, the current work examined which fragments of scientific documents are read in what priority by scientists. We also examined the relationship of the fragment to the task the scientists have to accomplish. These fragments can then be used to build a better indexing model for scientific documents.

In some cases, the noun of the structural fragment reflects the semantic nature of its content (e.g. methods describing the methodology of the research). However, we found numerous exceptions, and these exceptions could lead to annotate a fragment with the wrong semantic type if only the structural part is taken into account. For example, scientific documents using a structure such as IMRaD (Introduction, Methodology, Results and Discussion) should be written in a way that the reader could not find any methods description outside of the methods section; however, it is often the case that authors describe a complement of their methods in the results or conclusion sections. Another example is that authors may adopt IMRaD in the construction of their narrative, but the headings will not necessarily reflect that structure (e.g. in computer science, the part corresponding to the methodology or the results could be called implementation), or that the structure describes rhetorical parts or discourse elements but neglects others (e.g. background, aim, definition). In [15], the author argues that the headings of sections are too generic and are not strictly delimited. The same heading can convey information of different natures. The author argues that precise information retrieval requires the use of lower-level indices, such as paragraphs or sentences.

In addition, even if scientific documents began to be more standardised, and the IMRaD structure was adopted in various domains, especially in medicine [29], this structure is not adopted in all scientific documents and by all domains of research. To provide an example, we examine two journals in the domain of gender studies and sociology.

Out of 24 scientific documents extracted from the journal Feminist Studies in 2007, only 9 contain titles for paragraphs or sections, and the titles do not always correspond to the IMRaD model. Out of 18 articles extracted from the American Journal of Sociology in 2013,[2] 10 articles have a complete IMRaD structure, but 8 articles present only a partial or no IMRaD structure. These articles also contain other names for headings, such as Theory and Theoretical Framework, as well as more discursive titles (see Table 1).

Additionally, depending on the date of publication or the area of research, documents using structural models such as IMRaD do not even exist, as the documents pre-date the structural models or the models were never used.

This research allows us to build a model that should fit with the scientists' needs and all types of scientific documents. During the interview, we made a very clear distinction between the structural and the semantic, and we kept only the semantic definition of the fragments.

In the first step of this research, published in [6,25], we conducted a survey based on three hypotheses; (1) scientists seek and read scientific documents for different reasons; (2) scientists focus on specific parts of a document rather than the whole document; and (3) scientists do not seek information from the same parts of documents depending on the field of research. The results were promising and showed different behaviours and tendencies depending on the area of research. Analyses showed that scientists are more likely to find their answers in certain types of sentences or paragraphs than in others. However, the analyses were not completely conclusive as the responses were given by closed answers.

In the second step, the research hypotheses were similar, but the methodology differed as we conducted semi-structured interviews with scientists from different areas of research. We chose a qualitative approach instead of a

Table 1 Example of the logical structure of an article in Feminist Studies and in American Journal of Sociology

	Feminist studies
Section	The dual marginality ...
Section	Women's struggle In ...
Section	The Study
Section	Coping strategies...
Section	Discussion
Section	Notes
	American Journal of Sociology
Section	Trends in U.S...
Section	Theoretical background
Section	Data and methods
Section	Results
Section	Selection into migration ...
Section	Discussion and conclusions
Section	Appendix
Section	References

[2] A journal with more empirical studies than Feminist Studies.

Table 2 Sample population

Subject	Year of birth	Sex	Field of research	Highest education degree if Ph.D. year	No of article written as first author	Total no of article published	No of article read by month
P1	1983	M	Computer Science	Master	6	6	10
P2	1978	F	Computer Science	2009	15	20	10
P3	1976	F	Linguistic/Computer Science	2008	17	29	25
P4	1961	M	Computer Science/Health	Master	10	20	15
P5	1976	F	Information System	Master	1	10	3
P6	1946	F	Psychology	1975	68	141	6
P7	1957	M	Physics/Science History	1987	15	60	50
P8	1985	F	Translation and Communication	Master	2	3	20
P9	1983	F	Social Science, Gender Studies	Master	1	1	15
P10	1974	F	Medicine	2000	20	28	25

quantitative approach allowing us to spend more time with the scientists and examine some of our hypotheses more deeply. The sample consisted of 10 subjects (see Table 2). An important part of each interview was devoted to finding which fragments the scientists were focusing on, and if they associated a specific task to a specific fragment. The methodology also allowed the identification of the difficulties scientists have distinguishing between structural and semantic fragment types.

Our interview sheet comprised 33 questions which were the basis of an interview and was written in French.[3] Interviews time was between 1 and 2 h. The first part sought individuals' socio-demographic data. The second part examined individuals' search patterns, including which information retrieval systems (IRs) they would use, what type of information they would be looking for, and what kind of questions they would be planning to ask before querying an IRs. The final part examined individuals' reading behaviour relating to the search results, such as which part of the document (fragment) they would be focusing on when they are not reading the whole article, and if these parts were different depending on their reason for search. This article focuses only on the reading behaviour results and the association with the task they have to accomplish and on commonalities of the responses from the participants.

When asked if they were using scientific documents to answer a precise question, our survey participants all answered yes. When they developed their answer, the majority said that they would use their reading to develop a methodology, followed by looking at the findings, searching for a definition and looking at the state of the art.

Figure 1 shows which fragments the scientists were reading and demonstrates that not all scientists were reading all fragments of the documents. This pattern was common across

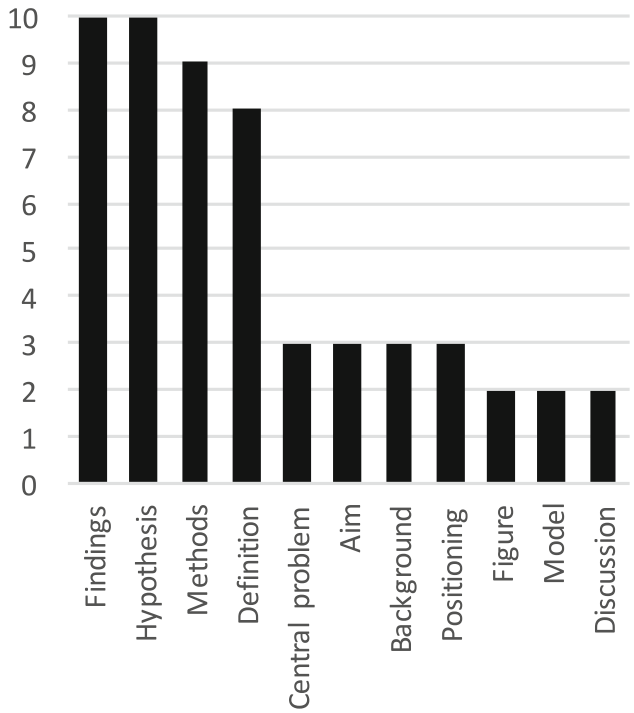

Fig. 1 Fragments read by person (interview)

all areas of scientific research examined. Most scientists read the findings and hypothesis, methodology, definition. We found that other fragments of the documents were consulted less frequently, e.g. aim, background, positioning, figure, model and discussion. To confirm the answers, we asked the participants if it was the section or the content of the fragment that they were interested in, and they all answered that the section was not always meaningful, and they were looking for these fragments all over the document, even the one in Medicine, where IMRaD is widely adopted [29].

We established a list of aims why scientists seek and read scientific documents. During the interviews, we discussed

[3] The interview sheet can be found in http://users.cs.cf.ac.uk/DeRibaupierreH/.

this list with the interviewees and added new aims the scientists pursue when seeking and reading scientific documents or reviewed the existing ones. The aims that the interviewees agreed most strongly with were (in descending order):

- to enhance their knowledge.
- to write papers.
- to find a new idea/inspiration.
- to discover another point of view.
- look at a new idea.
- to stay informed.
- because they were advised by someone.
- to compare documents.
- to write a state-of-the-art review.
- to find a specific definition.
- curiosity.
- to provide a new point of view.
- to compare methodologies.
- to provide expertise/review.
- to summarise their knowledge.
- to be able to take a position in a document.
- because it was referenced by another document.
- to resolve a problem.
- to build a model.
- to determine whether an idea was new or whether someone else had already worked on it.

Figure 2 shows which fragments the scientists read depending on the task they want to accomplish (we show only the most meaningful tasks cited by the interviewees). When reading to keep track of other studies, researchers focus primarily on findings. When they set out to discover a new area of research, the interviewees indicated that their focus was more widely distributed between fragments types. To write an article, researchers are more likely to read whole documents. To find new ideas, the interviewees' answers varied between reading fragments with a preference towards findings and whole documents. To acquire new knowledge (learning), the researchers indicated again that their focus was more widely distributed between fragments with a importance to the whole document. Furthermore, when planning to start a new project, they focus on the whole document, the findings and the background.

To have a stronger user model, we combined some of the results obtained during the first stage (survey) and the second stage (interview) to have a better understanding of which fragments scientists focus on. As the fragments indicated by the respondents were not exactly the same between the two studies (survey and interview) and the manner in which answers were provided, open (interview) vs. closed (survey), these factors may have affected the results. Figure 3 shows the fragments read by person for the interviews and the survey. The results of both studies confirm that scientists do not focus on the entire documents. This behaviour is a factor to consider in the construction of document models concerning indexing, queries, the display of information and evaluation.

Although the sample is not necessarily representative of a general academic population, the results of the two studies provide a sufficiently solid foundation for developing an empirical annotation model.

3 Annotation model

Various models have been developed to manually or automatically annotate scientific writing, with the aim to improve summarisation or information retrieval using, for example, rhetorical structure or discursive categories ([5,11–13,19,23,28,31,32]). These works focus mainly on the "hard" sciences like biology or biomedical where there is relatively little variation in describing the results, hypothesis, conclusions, etc. The sentences are generally short and more formalised than in social science and humanities document. In [2], the author showed that humanities writing scored

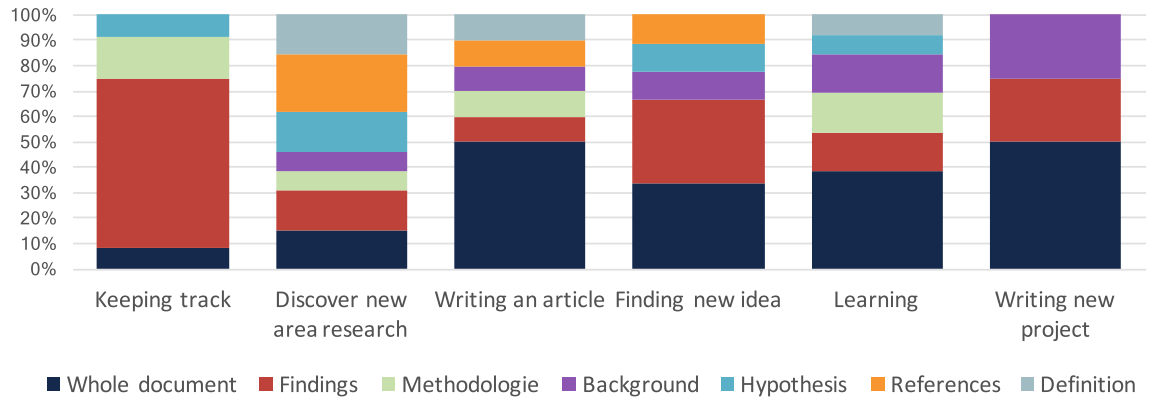

Fig. 2 Fragment by tasks

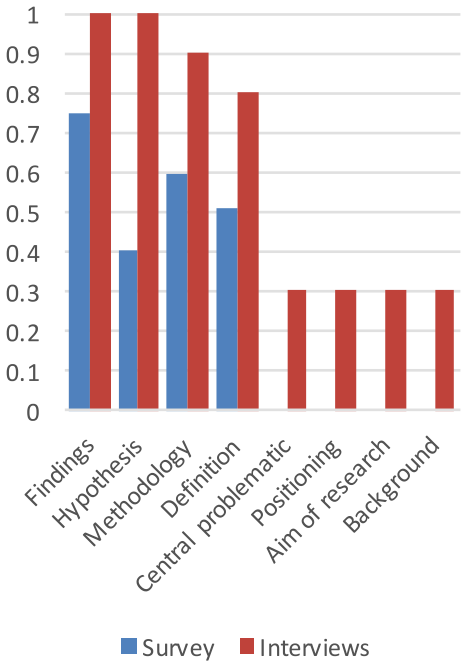

Fig. 3 Fragments read by person (interview and survey)

higher on the narrative dimension than engineering and technology. Our analysis of a sample of 1500 abstracts extracted of journal in gender studies and 1500 abstracts from PubMed showed that a sentence on average in a journal in gender studies is composed of 18 words (126,469 words, 6944 sentences) against 12 words (157,197 words, 12,912 sentences) in PubMed.

In this research, we are proposing a methodology and a model that describes the different areas of scientific knowledge including the "soft" sciences like sociology, psychology or gender studies. In addition, to the best of our knowledge, none of the existing models is based on empirical studies that aim to understand what type of information users want to retrieve or to use in selection criteria. The previous models defined some of the discourse elements types revealed by our empirical studies; however, they are not complete, hence the need for a new model. The model presented here is based on a combination of elements of previous studies, whilst also suggesting some new concepts, and using the potential of web semantic techniques. The resulting model is SciAnnotDoc (see Fig. 4).

The model contains four main dimensions:

3.1 Metadata

This consists of current metadata in the field of scientific documents (bibliographic data), such as the authors, title of the article, the journal name, publisher, date of publication.

3.2 Textual content

This consists of the representation of the terms contained in the document. The content of each discourse element is semantically indexed by means of concepts from three auxiliary ontologies: an ontology of the studied domain, an ontology of scientific objects (equations, models, algorithms, theorem, assertions, principle, etc.) and an ontology of methods (types of methods, types of variables, tools, material, etc.). The ontology of methods and the ontology of scientific objects are generic, whereas the domain ontology is domain dependent (e.g. an ontology of gender studies, an ontology of particle physics). Different domain ontologies can be added to the model to provide the most precise annotations. Using ontologies to annotate these terms can help the resolution of problems such as the ambiguity of a word, homonymy and synonymy.

3.3 Discourse elements

This dimension is associated with discourse elements and is the core of SciAnnotDoc model. A document is decomposed into structural fragments (usually paragraphs), and each fragment is composed of elements, generally one or a few sentences, that have specific roles in the scientific discourse. The discourse element types are *findings*, *definition*, *methodology*, *hypothesis* and *related work*. They correspond to the types that were most often mentioned by the scientists in our study. In addition, these types are not exclusive; a sentence that describes a definition could also describe a finding as shown in the following example.

> We find, for example, that when we use a definition of time poverty that relies in part on the fact that an individual belongs to a household that is consumption poor, time poverty affects women even more, and is especially prevalent in rural areas, where infrastructure needs are highest [1].

Additionally, if a sentence summarises a result presented in another article, this sentence is a *finding* and a *related work*. The reason for this kind of annotation derives from the results of the analyses we performed using the interview findings. Scientists are sometimes looking for a finding, a definition, a methodology or a hypothesis, but the attribution of the document to an author is not in their priority; it is only later that they might be interested to know the author(s) of the document or the referenced sentences. For example, the following sentence is a finding and also a related work, as this sentence refers to other works.

> The results of the companion study (Correll 2001), and less directly the results of Eccles (1994; Eccles et al. 1999), provide evidence that is consistent with the main

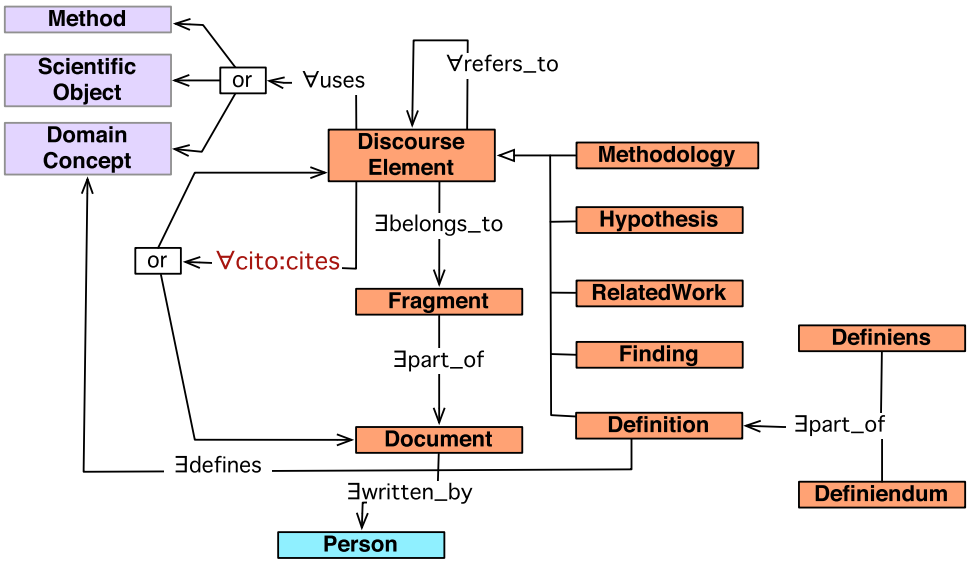

Fig. 4 SciAnnotDoc model: user-centric annotation model

causal hypothesis that cultural beliefs about gender differentially bias men and women's self-assessments of task competence [4].

3.4 Relational elements

This dimension consists of all explicit references from a document or discourse element to another document or discourse element. The relation between documents is a very important part of all scientific documents. Most of the models and research in this subject aim to compute bibliometric measures and take into account the document-to-document citation. More recent research such as [34] looks at the fragment level with a system that extracts fragments from the cited document which are related to the citing fragment. The authors in [16] are detecting the boundaries of citations in the full text of research papers, to improve paper summarisation. In this research, the aim is to help scientists to search documents with citation as a criteria. The annotation of the type of relationships that documents share is important, as CiTO ontology [27] proposes (such as agree, disagree with, confirm), but it should also be done at the discourse element level and not only at the document level. We reused the CiTO ontology and extended it to represent citations between documents and also between discourse elements since a large proportion of the citations do not refer to a whole document but to some, often very restricted, part (discourse element). This is necessary to answer precise queries such as "Find all the documents containing an outcome ('finding') about the difference between girls and boys in school and referencing a result of Zazzo." With this model, it becomes possible to perform more detailed analysis of the network of citations, depending on the types of citing or cited elements, and extract, for example, the lineage of idea of a discourse element.

The structure of the discourse elements is formally defined in the SciAnnotDoc OWL ontology.[4] The ontologies used to annotate the textual content, the metadata and the relational elements are kept separate from the discourse elements ontology so they can be easily interchanged, and there is a clear distinction between the categorisation of discourse elements on the one hand and their content on the other.

4 Annotation process

To assess the relevance of the SciAnnotDoc model, we annotated different gender studies and sociology documents. We chose this research area because the documents are largely heterogeneous, ranging from very empirical studies to more "philosophical" documents, and most of the documents are not using a structural model such as IMRaD.

To simplify automatic annotation, we have restricted a discourse element to one sentence even if some discourse elements could run over several sentences, and a fragment to one paragraph. We automatically annotated 1410 documents. We used a rule-based system to generate annotations based on the syntactic patterns detected in each sentences.[5] We used GATE,[6] a text engineering platform, ANNIE,[7] a component that forms a pipeline composed of a tokeniser, a gazetteer, a sentence splitter and a part-of-speech tagger, JAPE rules (Java Annotation Patterns Engine), a grammar

[4] Available at http://users.cs.cf.ac.uk/DeRibaupierreH/.

[5] All the different materials are available on http://users.cs.cf.ac.uk/DeRibaupierreH/.

[6] http://gate.ac.uk.

[7] http://gate.ac.uk/ie/annie.html.

Table 3 ANNIE tag sequence

On	This	Usage	,	Gender	Is	Typically	Thought	To	Refer	To	Personality	Traits	And
IN	DT	NN	,	NN	VBZ	RB	VBN	TO	VB	TO	NN	NNS	CC

language for operating over annotations based on regular expressions, and Ontology OWLIM2, a module for importing ontologies. These different components were used to detect the different kinds of discourse elements in documents as well as the domain concepts, the different methods of research and the scientific objects. The first step was to manually create the JAPE rules (20 rules to recognise findings, 34 for definitions, 11 for hypothesis and 19 for methodologies). For the detection of the author's name in the discourse element, we created 11 JAPE rules. We started from manually annotated sentences extracted from Gender Studies and Computer Sciences documents[8] and analysed the different patterns of grammatical structures produced by ANNIE pipeline.

We looked at the syntactic structure produced by the ANNIE output for each of the different sentences. The following example (see Table 3) describes the entire tag sequence obtained by ANNIE on the following definition of the term "gender".[9]

> On this usage, gender is typically thought to refer to personality traits and behaviour in distinction from the body [20].

For each tag's sequence, we simplified those rules, reduced them and merged some of them, to obtain more generic rules able not only to catch the very specific syntactic pattern, but also to catch the variation of the pattern. We also relaxed the rules and used some "wildcard" tokens (see Table 4).

To increase the precision, we added typical terms that appear in each type of discourse element. For detecting definition, instead of using the tag VBZ (3rd person singular present) which could be too generic, we used a macro that was the inflection of the verb be and have in the singular and plural form. We also used a macro that captured the different variations of the verb "refer" when it appears alongside "to", whether annotated sentences such as those shown in Table 4. These simplifications, reductions and merging were done for the different types of discourse elements. For example, for detecting findings, we selected a collection of verbs that could describe findings, such as "confirm, argue, conclude, demonstrate, explain, …" and used the different inflection of these verbs to create macros that we included in the rules. These macros were used in a wider context (using the part-of-speech tags), allowing detection of the following sentence.

> The fact that male students assessed their own mathematical competence higher than their equal ability female counterparts did explain part of the gender gap in enrolment in high school calculus courses and selection of a 'quantitative' major [4].

We uploaded the domain concept ontology to help to define more precise rules. For example, to detect a definition such as "It follows then that gender is the social organization of sexual difference" [20]; we created a rule that was searching for a concept defined in the domain ontology followed at a short distance by the declension of the verb *be*.

```
(({Lookup.classURI==
"...genderStudies.owl#concept"})
(PUNCT)?
(VERBE_BE))
```

The different ontologies (GenStud, SciObj, SciMeth) were imported using Ontology OWLIM2 and transform in Gazetteers using OntoRoot. These ontologies were used not only for the JAPE rules but also to annotate the concepts in the text.

We implemented a pipeline in Java that first transformed the PDF into raw text using the PDFbox API[10] and regular expressions to clean the raw text. Second, using GATE (see Fig. 5), we annotated each sentence with one or several discourse element types and with the concepts it contains. Third, we transformed each GATE's XML output into an RDF representation of the text (see Fig. 6). The sentences that did not contain one of the four discourse elements (definition, hypothesis, finding or methodology) were annotated with the tag *NonDefinedDE*, allowing annotation of each sentence of the entire document, even those not assigned to discourse elements. For the related work, each sentence that contained a tag *AuthorRefer* detected by the JAPE author rules was defined as a related work, and a new document was created.

The different RDF representations (see Fig. 7) created by the Java application were loaded into an RDF triple store.

We chose Allegrograph[11] because it supports

[8] These sentences are not part of the final annotated corpus and neither part of the "gold standard".

[9] For space reasons, we did not present all the sequences and the definition of the part-of-speech tags can be found at http://gate.ac.uk/sale/tao/splitap7.html.

[10] https://pdfbox.apache.org/.

[11] http://franz.com/agraph/allegrograph/.

Table 4 Definition sentences and JAPE rules (simplified)

Gender	Is	Typically thought to	Refer to
Gender	Has	Become used to	Refer to
Gender	Was	A term used to	Refer to
NN	TO _BE _HAVE (macro)	Token [2,5]	REFER (macro)

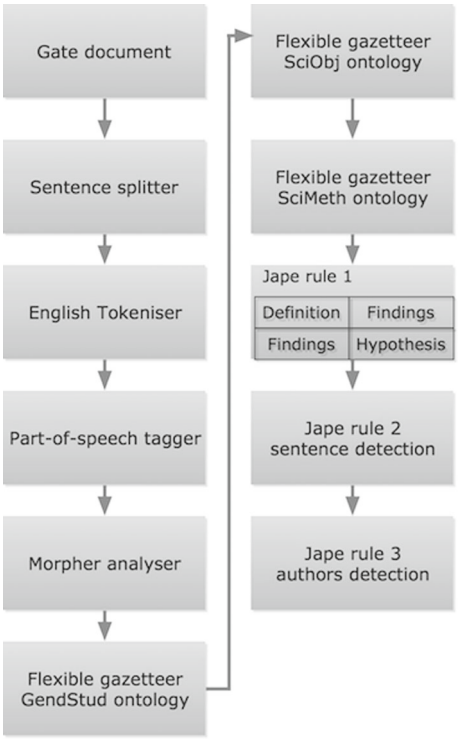

Fig. 5 Annotation GATE pipeline

RDFS++ reasoning in addition to SPARQL query execution.

Table 5 presents the corpus statistics by discourse elements and the number of sentences that contain more than one type of discourse element. Some types overlap more than others; for example, Findings overlap with Hypothesis and Definition; however, in this sample, Methodology does not overlap with any other discourse elements. This confirms the importance of having a multi-type annotation model.

To test the quality of the automatic annotation, we needed to create a new "gold standard" as the different existing corpora [18,26] were in "hard science", and we wanted to evaluate our annotation process and model on a corpus where sentences are not as formalised and are longer than the ones found in "hard science". We manually annotated 555 sentences in gender studies (the sentences manually annotated are extracted from different documents to the one used to analyse the syntactic structure of the sentences and creating the JAPE rules to avoid bias), creating a "gold standard". These sentences were annotated by 2 annotators.

We used Fleiss' kappa to measure the inter-annotator agreement between the two expert annotators (see Table 6).

A reason for the relatively low score could be the number of categories. The more categories involved, the more the annotators may differ in their judgements. Another reason might be the interpretation of the instructions and as

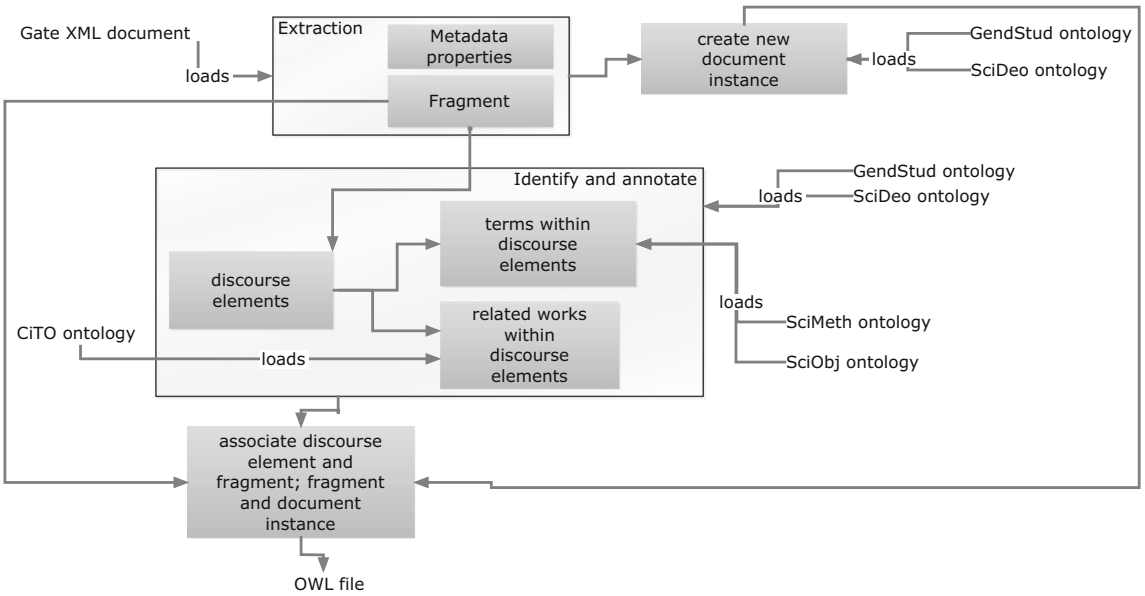

Fig. 6 Annotation algorithm model

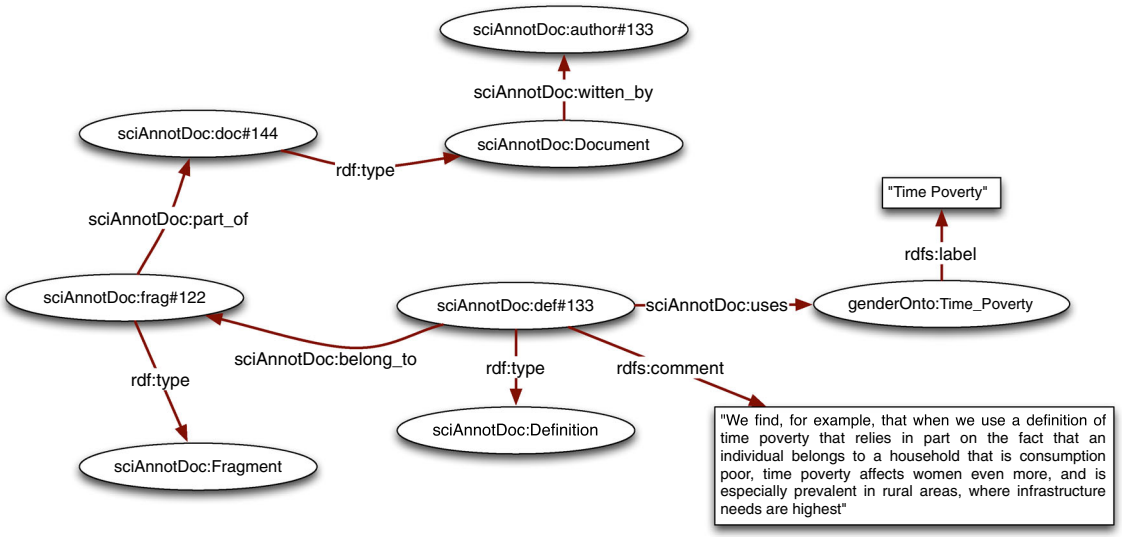

Fig. 7 RDF annotation schema (extract)

Table 5 Annotated corpus statistics by discourse elements

	Find	Hypo	Def	Meth
Findings	17,639	47	30	0
Hypothesis	47	7626	15	0
Definition	30	15	5239	0
Methodology	0	0	0	11,829

a consequence the misinterpretation of the sentences. After discussion with the second annotator, we discovered that she annotated most of the reported findings as something else, general hypotheses. We decided to change all the findings that were annotated as hypotheses back in findings. We performed precision and recall measurements on these sentences (see Table 7).

To increase recall, we added heuristics such as: if a fragment (paragraph) contains unrecognised elements (sentences) and at least three elements with the same type *T* and only this type, then assign type *T* to the unrecognised elements. With these rules, we created 341 findings, 130 methodologies, 29 hypotheses and 6 additional definitions.

The low recall for definition could be explained by the lack of concepts in the GenStud ontology, as this ontology is used to detect definition in case of "is a" (see example above). If the ontology is not complete, it will impact the recall. For findings, an explanation for the low recall could be the relatively short list of manually generated verbs used to describe the findings. Another issue concerning all type of discourse elements which could explain the low recall is the generality of the sentence. For example, in the following sentence: "However, women in the GD condition have higher mean aspirations than men."[4], the detection of a distinctive pattern is almost impossible; however, a human will recognise and annotate that as a finding.

Table 6 Fleiss' kappa results for the two expert annotators in gender studies

Definition	0.697
Findings	0.388
Methodology	0.719
Hypothesis	0.384
Other	0.375

5 User evaluation on complex queries

We conducted user evaluations to check how the annotation system and the SciAnnotDoc model compared to the standard keyword search. We implemented two interactive search interfaces: a classic keyword-based search (with a TF*IDF-based weighting scheme) (see Fig. 8) and a faceted interface (FSAD) (see Fig. 9) based on our model.

5.1 Faceted search interface

Because scientists are not all experts in description logic or other formal query languages (such as SPARQL), we proposed a faceted search interface based on SciAnnotDoc. A major goal of the interface is to hide the complexity of semantic search from scientists and to make it easy to use and effective. The use of a faceted search interface to mask the complexity of SPARQL has the disadvantage of being less expressive than other query languages [10]. However, it is still possible to have sufficient expressiveness to allow end users to find their desired information.

Table 7 Precision and recall

Discourse elements types	No. of sentences	Precision	Recall	F1.0s
Findings	168	0.82	0.39	0.53
Hypotheses	104	0.62	0.29	0.39
Definitions	111	0.80	0.32	0.46
Methodologies	172	0.83	0.46	0.59

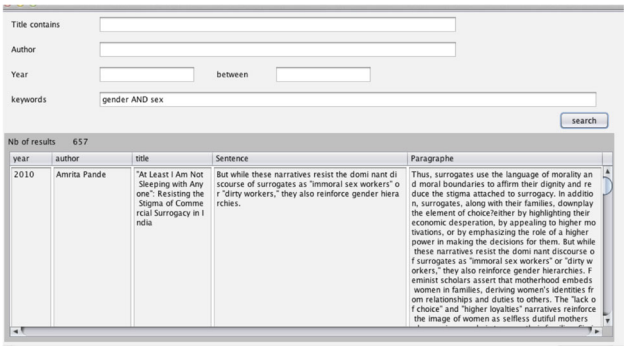

Fig. 8 Keyword-based search interface

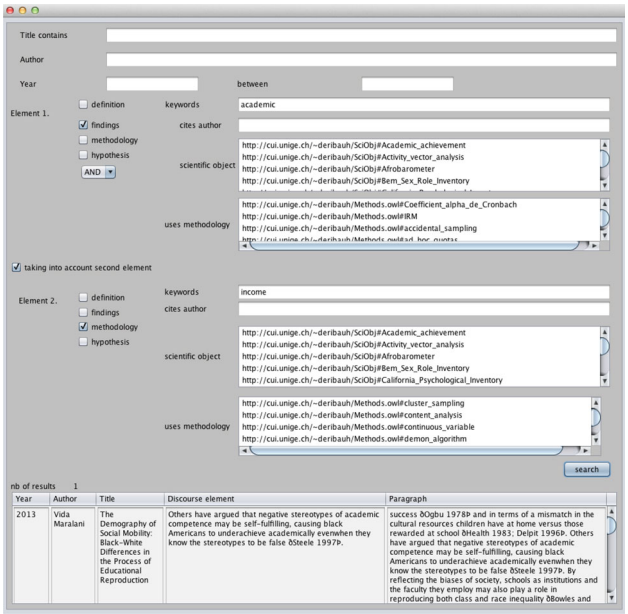

Fig. 9 Faceted search interface (FSAD)

We represent the different dimensions of the SciAnnotDoc model as facets.

The *metadata facet* allows users to query the metadata of the SciAnnotDoc model.

The *conceptual facet* allows users to specify what concepts are contained in the discourse elements using either ontologies concepts or free keywords.

The *discursive facet* allows users to specify in which discourse element they want to apply the other search criteria. Users can search for a sentence that contains one or several types of discourse elements. For example, if users want to search only for a definition, they tick only one checkbox for the first discourse element. If they want to search for a sentence that belongs to two discourse elements, such as a definition that is also a finding, they tick two checkboxes of the first discourse element. Another possibility is that users want to search for two discourse elements disjointedly in the same text, such as a finding describing *the difference of gender in salary* based on a hypothesis describing *women having less academic degree*. Users can then specify which types of discourse elements they desire to query for each sentence associated with keywords. For the above example, they could tick findings associated with keywords "gender AND salary" and tick hypothesis associated with keywords "academic degree".

Thus, the resulting interface is a faceted search interface in the sense that the SciAnnotDoc dimensions structure the information that users can explore. The interface is also an adaptive search interface in the sense that users should be able to add as many discourse elements as they desire to build "complex" queries. In the prototype implemented interface, users are only able to query two disjointed discourse elements. The interface was built using JAVA and SWING. The query model classes were implemented using the OWL API and the Allegrograph API.

FSAD was built in several iterations. During the different cycles of iteration, scientists were asked twice to evaluate the usability interface through heuristic evaluations [21].

5.2 User evaluation

The corpus used for the user evaluation was the annotated corpus (1410 documents) in gender studies. All documents were in English. Both systems index and query at the sentence level instead of the usual document level. The maximum answers in a set was fixed in both interface at 3000. The tests were conducted with 8 scientists (4 in gender studies, and the other 4 were in other fields. 50% were female, and the scientists had an average age of 38 years). Scientists had to perform 3 tasks (see below) with only 1 of the systems. The design of the experiment was based on a Latin square rotation of tasks to control for a possible learning effect of the interface on the participants. The average time for the experiment was 1 hour. The questionnaires were conducted on LimeSurvey.

1. Task 1: Find all the definitions of the term "feminism".
2. Task 2: Show all findings of studies that have addressed the issue of gender inequality in academia.
3. Task 3: Show all findings of studies that have addressed the issue of gender equality in terms of salary.

We gave participants a tutorial on how the system works, but did not give more exact instructions on how to search. The participants determined the end of a task when they thought they had obtained enough information on the given subject. They had to perform the 3 tasks and complete different questionnaires (1 socio-demographic at the beginning of the evaluation, 1 interface evaluation after each task, and a final interface evaluation at the end of the test). The socio-demographic questionnaire contained 11 questions. The questionnaire after each task contained 10 questions:

Q1 Do you think the set of results was relevant to the task? (1 = not useful, 5 = useful)
Q2 Do you think the number of results was too large to be useful? (1 = totally unusable, 5 = usable)
Q3 How many elements correspond to your request? (1 = 0–5%, 2 = 6–15%, 3 = 16–30%, 4 = 31–50%, 5 = 51–75%, 6 = 76–90%, 7 = +90%)
Q4 How many elements were completely irrelevant? (1 = 0–5%, 2 = 6–15%, 3 = 16–30%, 4 = 31–50%, 5 = 51–75%, 6 = 76–90%, 7 = +90%)
Q5 Do you think that these results are obtained faster in this way than when using a common scientific information search system (Google Scholar, etc.)? (1 = not at all faster, 5 = much faster)
Q6 Do you think the number of requests you made was adequate to achieve good results? (1 = very adequate, 5 = not at all suitable)
Q7 Did you obtain satisfactory results for each query you made? (1 = not at all satisfied, 5 = very satisfied)
Q8 Are you satisfied with the overall results provided? (1 = not at all satisfied, 5 = completely satisfied)
Q9 Are you frustrated by the set(s) of results provided? (1 = totally frustrated, 5 = not at all frustrated)
Q10 Did you find this task difficult? (1 = very difficult, 5 = very easy)

The final questionnaire after completion of the 3 tasks contained 10 questions (from questions 6 to 11, the questions were inspired by the USE questionnaire[12]:

Q11 Do you think this system is useful? (1 = useless and 5 = very useful)
Q12 I think the indexing of these documents is not sufficiently precise for my information needs. (1 = totally disagree, 5 = totally agree)
Q13 Do you think this system is better for finding the information you need than your usual system (Google scholar, etc.)? (1 = disagree, 5 = strongly agree)
Q14 I would use this system instead of the search engines that I use every day. (1 = disagree, 5 = totally agree)
Q15 I found the system unnecessarily complex. (1 = disagree, 5 = totally agree)
Q16 I think this system is easy to use. (1 = disagree, 5 = totally agree)
Q17 I think I need support/technical assistance to use this system. (1 = disagree, 5 = totally agree)
Q18 I think that users (scientists) could easily learn to use this system. (1 = disagree, 5 = totally agree)
Q19 I need to learn many new things before I feel comfortable and can use this system easily. (1 = disagree, 5 = totally agree)
Q20 I felt comfortable using this system. (1 = disagree, 5 = totally agree)

A program was written to capture the logs of the two systems. The logs recorded the time stamps indicating when the participants clicked "search", the SPARQL query, which elements were checked (only in FSAD), the different search criteria, and the time stamp indicating when a query was returned by the system. We calculated the length of time during which a participant looked at a set of results based on the interval between the time stamp of a query's end and the next click on the "search" button. Thus, the time spent looking at a set of results was approximate.

5.2.1 Results of the user study

We computed the average responses for the three tasks, and we tested the difference between the participants who had to evaluate the FSAD versus the keyword search, using an analysis of the variance (Anova) tests. As shown in Table 8, the users' answers regarding the FSAD interface was generally more positive than that of the keyword search interface. We can observe that in the questions about the relevance of the set of results and the usefulness of the set of results, participants using the FSAD found the set of results more relevant (Q1) and more usable (Q3) than those using the keywords search, with a significant difference when they were asked if they found the set too large to be useful (Q2), and if the number of elements were completely irrelevant (Q4). Both groups appeared to have mixed feelings about the speed of the systems (Q5). It should be noted that the systems performed poorly at times, resulting in a long time required to obtain a set of results. Thus, the participants may have been confused between the performance in terms of the time

[12] http://hcibib.org/perlman/question.cgi?form=USE.

Table 8 Average response for the three tasks

Question	FSAD μ	Keywords search μ	Sig
Q1	4	3.75	.780
Q2	4.5	2.75	**.045**
Q3	5	4.25	.494
Q4	1.5	3.75	**.031**
Q5	3.5	3.6	.824
Q6	2.16	2.83	.223
Q7	3.83	3.41	.484
Q8	4.16	3.41	.273
Q9	3.16	4.25	.220
Q10	3.5	3.66	.780

Average response for the three tasks (significant P values (<0.05) are in bold)

Table 9 Average for the questions asked after completion of the three tasks

Question	FSAD μ	Keywords search μ
Q11	4.25	4.25
Q12	2.0	3.5
Q13	4.0	3.67
Q14	3.7	3.25
Q15	2.25	2.00
Q16	3.75	3.75
Q17	2.25	2.25
Q18	4.75	4.50
Q19	1.0	2.25
Q20	3.75	3.75

required to obtain the set of results and the time required to achieve the task. More FSAD participants than keyword search interface participants seemed to find that the number of queries was adequate, but the difference was not significant (Q6). The FSAD participants found more results to be satisfactory (by query: Q7 and the overall: Q8) than did the keyword search interface participants, but the difference was not significant. The level of frustration (Q9) seems to be higher for the keywords search interface participants than for the FSAD. Both groups seemed to find the task to be of average difficulty (Q10), and no significant difference was observed.

For the second part of the questionnaire after the completion of the three tasks (see Table 9), no significant differences were found in the 11 questions. However, we observed that answers are generally more positive for the FSAD than for the keywords search interface. Moreover, we observed in Q12 that FSAD participants seems to find the system more precise than the keyword search interface group, but the difference was not significant.

5.2.2 Analysis of the logs

The analysis of the logs (see Table 10) showed for task 1 (the definition of feminism) that FSAD participants queried the system more often than the keyword search participants did. A Student's *t* test was used to calculate the probability of significance. They received fewer answers by query on average, but spent more time on the set of answers and on one answer. The difference in time spent by answer between both groups is significant.

For task 2, we observed that FSAD participants query the system more often than the keyword search participants. In contrast with task 1, the set of results was larger than for keywords search interface. They spent the same amount of time for the full set; however, it was observed that they still spent a little bit more time on average per answer in the FSAD.

In task 3, FSAD participants query the system more often and receive on average fewer answers by query than the keywords search interface participants. In contrast with tasks 1 and 2, they spent less time on average for the set of answers and for an answer; however, one of the participants (it was her first task) spent a considerable amount of time evaluating the set of results that contained 850 answers. As she did not perform any other query and was the only one to spend this amount of time, she could be considered an outlier, even if the samples were larger.

5.2.3 Precision/recall on task 1

We also performed a precision and recall evaluation for the first task. For the FSAD system, when the user chose the facet definition and typed the keyword feminism, the system sent a set of 148 answers. Ninety results were relevant, and the precision was 0.61. For the keywords search system, the set of results was the union of the results obtained with the queries "define AND feminism" and "definition AND feminism" (this combination of terms was the one the most used by participants), and the system sent a set of 29 answers of which 24 were relevant, and thus, the precision was 0.82.

To evaluate the recall, as we did not know the number of definitions contained in the corpus, we simply observed that the ratio between FSAD and the keyword search is 3.77. In other words, the FSAD system was able to find 3.77 times more definitions of the term "gender" than the keywords search. Hence, at the cost of a lower, but still acceptable precision, the FSAD system has a considerably higher recall than the keyword search system. We also observed that the overlap between both systems was 12 sentences.

Table 10 Logs for the three tasks

	Task 1			Task 2			Task 3		
	FSAD	Keywords	Sig	FSAD	Keywords	Sig	FSAD	Keywords	Sig
Av. no. queries	5.7	4.75	0.266	14.25	11.25	0.531	12	8	0.291
Av. no. answer	30.28	731	**0.012**	10.06	16.27	0.131	41.97	224.27	0.223
Av. time spend for the full set	2.26	1.39	0.329	0.7975	0.7575	0.462	1.325	6.512	0.2
Av. time spend for an answer	0.064	0.019	**0.05**	0.008	0.0033	0.164	0.085	0.361	**0.0085**

Average number of queries by task; average number of answers by query; average time spent for the full set of answers; average time spent per answer (Significant P values (< 0.05) are in bold)

6 Discussion

Keyword search interface participants seemed to find the set of results less useful than the FSAD participants did (significant difference). This difference appears to hold true even independently of the number of results in a set. In task 2, FSAD participants received more answers than the keyword search participants did, but they found the number of irrelevant answers less important than the keyword search participants did. This finding may support the fact that the keyword search interface participants found the set of results less useful than the FSAD participants did.

A small difference was observed between the two groups with respect to the relevance of the set of results, with the FSAD participants finding the results slightly more relevant. However, the difference was not significant. There may be several reasons for this lack of significance. First, participants might have built search strategies for searching for scientific documents using a search engine, and those strategies might have yielded "strange" results when applied to the FSAD interface. This finding would correspond to an effect of the novelty of the FSAD interface.

Second, it might be easier for participants to declare that a set of results is not relevant than to find them relevant, especially for non-expert participants. Because the number of participants was small, we did not test the difference between experts and non-experts.

The logs showed that FSAD participants introduced, on average, more queries than the keyword search participants did. In the questionnaire, the FSAD participants, nevertheless, indicated that the number of queries was more adequate (Q6) than the keywords search participants (see Fig. 10). An interpretation of this result could be that when participants have less appropriate answers in the set of results of one query, such "bad" answers could affect the perception of the adequate number of queries. It would be interesting to test this hypothesis with a larger number of queries, and control the quality of the answers to "non appropriate" versus "appropriate", and examine whether the quality alters the perception of the number of queries introduced.

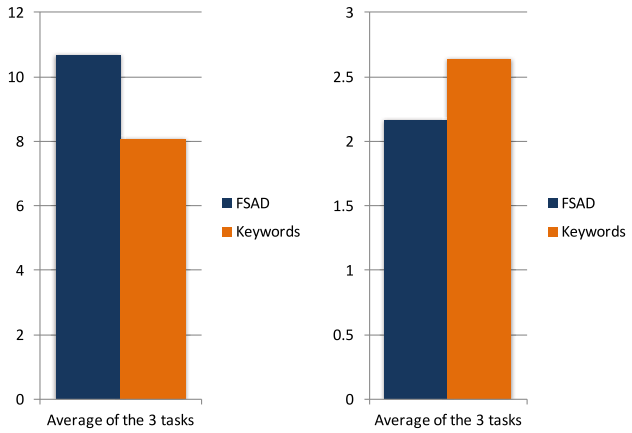

Fig. 10 Average number of queries (**a**) and average feeling of adequacy on the 3 tasks (1 = very adequate, 5 = not at all suitable) (**b**)

Some measure of self-efficacy should be added to the questionnaire for the participants. However, the questionnaire was already long enough. In the future, some of the intervening variables (proposed by Wilson [36]) should be tested, particularly when subjective measures are used. The psychological features of the participants may have an effect on the results as well [14].

Even though the number of the sample is small, we observed a preference among the participants for the FSAD interface and for a search engine that focuses on a sentence rather than an entire document. These findings are encouraging for further refinements to the system.

The user evaluation we conducted shows that despite these inaccuracies and a small sample size, we were able to build a query system that already outperforms keyword searches in many cases, especially in the case where the recall is very important. Google allows users to query a term for the definition with "define" + term. In Google, the first set of answers seems to be extracted from glossaries, dictionaries and Wikipedia for the first ranked answers, and for the next answers the system seems to work by looking at the phase "define"+ term in the document. For scientists, this is not enough, first because the source of the information is not accurate enough and second because of the lack of

answers. For Google Scholar, scientists make the assumption that the sources are more accurate because the IRs are indexing scientific documents. The system queries the index with the pattern "define" AND "feminism", ignoring all the other definitions that use some other sentence construction than ".... define feminism...". And as we have shown above, the number of definitions of the term found with this pattern is insufficient, especially for scientists. Consequently, when the task is to find a definition and the user needs a very high recall, Google or Google Scholar does not perform so well.

One of the difficulties we had to deal with in the evaluation was the lack of a good evaluation corpus, preventing us from calculating the precision and the recall of the system. This problem is very often mentioned in the literature, conferences and workshops. We hope than in future, with the different evaluation campaigns that were created in recent years, this recurrent problem should diminish.

In the case of very precise queries such as tasks 2 or 3, we still have to analyse the precision and recall of our system. We also want to compare the results with some of today's IRs, but we can hypothesise that contrary to the first task, it will be the precision that will be reduced because it is not searching at the sentence level. The reason is that traditional search engines are indexing the text by the terms they find in the metadata (title, abstract, keywords) and sometimes by the terms contained in the entire document, but they do not take into account the context of the term, or even the distance between the terms, or if they take this parameter into their algorithm they do not allow user to add that as a feature in their queries. For example, in task 2, participants typed keywords such as "academic" or "university" and "gender inequality", but the problem is that those terms could appear anywhere in the text, even in the references, and the document that was published in the Oxford University Press, and contains "gender inequality" somewhere in the text, could appear in the top ranked answers.

7 Conclusion and future work

With the growth of publications in scientific domains, scientists need more precise information retrieval systems that are able to search by metadata (as is the case today), but which also allow users to create complex queries such as "retrieve all the findings that women have a tendency to drop their academic career after their first child more than men, using qualitative and quantitative methodologies". This kind of system needs a robust annotation model that takes into account the scientist's needs, and the semantics of the scientific document. In this case, knowing or indexing only the metadata is insufficient, and annotations in the content of the full text such as the discourse element, the references to other documents and the concept are crucial. The SciAnnotDoc model proposed in this paper is built on empirical research that analysed the real needs of scientists and takes into account the semantics and the specificity of the documents. We have also proposed an approach to automatically annotate PDF documents with the SciAnnotDoc model.

The evaluation of the annotations shows not only that the model is realistic because it is amenable to automatic production (many previously proposed annotation models have never been used in practice because they require a manual annotation), but also that the precision is good.

To improve the recall index, more JAPE rules could be created. However, introducing a larger number of rules might also increase the risk of adding noise to the annotation. Another solution could be to test whether some hybrid approaches mixing a rule-based approach and a machine-learning-based approach may improve precision and recall. Another solution is to ask scientists not to classify sentences into categories, but to confirm the type of categories a sentence is already classified into. By using this kind of methodology, we can improve and enlarge a training corpus that we could use to improve the precision and recall of the actual annotation process.

The user evaluation also shows that the user seems to be less frustrated by the FSAD system than the keywords search and seems to find that the level of usefulness in the sets of results is more important in FSAD than in keywords search. Some of the results could be non-significant because of the small sample size.

In the future, we will conduct additional usability testing and collect data to scientifically assess the quality of the system and to determine the influence of the precision and recall of the automated annotation process on the system performance. To improve the design of indexing and ranking models, we will examine whether precision or recall is the most important factor in determining the quality of the results according to the type of task. We will also apply the automatic annotation in different fields of research to estimate the domain specificity of the annotation process.

Acknowledgements We acknowledge Grant No. 159047 from the Swiss National Foundation.

References

1. Bardasi, E., Wodon, Q.: Working long hours and having no choice: time poverty in guinea. Femin. Econ. **16**(3), 45–78 (2010)

2. Biber, D.: Variation Across Speech and Writing. Cambridge University Press, Cambridge (1991)
3. Bishop, A.P.: Document structure and digital libraries: how researchers mobilize information in journal articles. Inf. Process. Manag. **25**, 255–279 (1999)
4. Correll, S.J.: Constraints into preferences: gender, status, and emerging career aspirations. Am. Sociol. Rev. **69**(1), 93–113 (2004)
5. De Liddo, A., Sándor, Á., Shum, S.B.: Contested collective intelligence: rationale, technologies, and a human–machine annotation study. Comput. Support. Coop. Work (CSCW) **21**(4–5), 417–448 (2012)
6. de Ribaupierre, H.: Precise information retrieval in semantic scientific digital libraries. PhD thesis, University of Geneva (2014)
7. de Ribaupierre, H., Falquet, G.: A user-centric model to semantically annotate and retrieve scientific documents. In: Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, ACM, pp. 21–24 (2013)
8. de Ribaupierre, H., Falquet, G.: User-centric design and evaluation of a semantic annotation model for scientific documents. In: Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business, ACM, p. 40 (2014)
9. de Ribaupierre, H., Falquet, G.: An automated annotation process for the scidocannot scientific document model. In: 5th International Workshop on Semantic Digital Archives (2015)
10. Ferré, S., Hermann, A., Ducassé, M.: Combining faceted search and query languages for the semantic web. In: Advanced Information Systems Engineering Workshops, Springer, pp. 554–563 (2011)
11. Groza, T., Handschuh, S., Kim, HL.: SALT: semantically annotated latex. In: Proceedings of 1st Semantic Authoring and Annotation Workshop, Co-located with ISWC (2006)
12. Groza, T., Handschuh, S., Clark, T., Shum, S.B.: Waard ad a short sruvey of discourse representation models. In: Workshop Semantic Web Applications in Scientific Discourse, pp. 1–21 (2009)
13. Harmsze, F.A., Van der Tol, M., Kircz, J.G.: A modular structure for electronic scientific articles. In: Conferentie Informatiewetenschap, pp. 99–20 (1999)
14. Heinström, J.: Fast surfing, broad scanning and deep diving: the influence of personality and study approach on students' information-seeking behavior. J. Doc. **61**(2), 228–247 (2005)
15. Ibekwe-SanJuan, F.: Repérage et annotation d'indices de nouveautés dans les écrits scientifiques. In: Indice, Index, Indexation, pp. 1–11 (2006)
16. Kaplan, D., Tokunaga, T., Teufel, S.: Citation block determination using textual coherence. J. Inf. Process. **24**(3), 540–553 (2016)
17. Kembellec, G.: Technologie et pratiques bibliographiques associées à l'ecriture scientifique en milieu universitaire. arXiv preprint arXiv:1201.4574 (2012)
18. Liakata, M., Soldatova, L.N.: Semantic annotation of papers: interface and enrichment tool (sapient). In: Proceedings of the Workshop on BioNLP, pp. 193–200 (2009)
19. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics **28**(7), 991–1000 (2012)
20. Nicholson, L.: Interpreting gender. Signs J. Women. Cult. Soc. **20**(1), 79–105 (1994)
21. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 249–256 (1990)
22. Novacek, V., Groza, T., Handschuh, S., Decker, S.: CORAAL—dive into publications, bathe in the knowledge. Web Semant. Sci. Serv. Agents World Wide Web **8**(2), 176–181 (2010)
23. Reiplinger, M., Schäfer, U., Wolska, M.: Extracting glossary sentences from scholarly articles: a comparative evaluation of pattern bootstrapping and deep analysis. In: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Association for Computational Linguistics, pp. 55–65 (2012)
24. Renear, A.H., Palmer, C.L.: Strategic reading, ontologies, and the future of scientific publishing. Science **325**, 828–832 (2009)
25. Ribaupierre, H.d., Falquet, G.: New trends for reading scientific documents. In: BooksOnline '11: Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing, ACM Request Permissions. doi:10.1145/2064058.2064064 (2011)
26. Shatkay, H., Pan, F., Rzhetsky, A., Wilbur, W.J.: Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics **24**(18), 2086–2093 (2008)
27. Shotton, D.: CiTO, the citation typing ontology, and its use for annotation of reference lists and visualization of citation networks. Bio-Ontologies 2009 Special Interest Group Meeting at ISMB (2009)
28. Shum, S.B., Motta, E., Domingue, J.: Representing scholarly claims in internet digital libraries: a knowledge modelling approach. In: ECDL'99, Third European Conference on Research and Advanced Technology for Digital Libraries, pp. 853–853 (1999)
29. Sollaci, L.B., Pereira, M.G.: The introduction, methods, results, and discussion (IMRAD) structure: a 50-year survey. J. Med. Libr. Assoc. **92**(3), 364 (2004)
30. Tenopir, C., King, D.W., Edwards, S.: Electronic journals and changes in scholarly article seeking and reading patterns. In: Aslib Proceedings (2009)
31. Teufel, S., Moens, M.: Discourse-level argumentation in scientific articles: human and automatic annotation. In: In: Proceedings of the ACL-1999 Workshop Towards Standards and Tools for Discourse Tagging, Citeseer (1999)
32. Waard, A.d., Tel, G.: The abcde format enabling semantic conference proceedings. In: Proceedings of the First Workshop on Semantic Wikis, European Semantic Web Conference (ESWC 2006) (2006)
33. Wan, S., Paris, C., Muthukrishna, M., Dale, R.: Designing a citation-sensitive research tool: an initial study of browsing-specific information needs. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, Association for Computational Linguistics, pp. 45–53 (2009)
34. Wan, S., Paris, C., Dale, R.: Supporting browsing-specific information needs: introducing the citation-sensitive in-browser summariser. Web Semant Sci Serv Agents World Wide Web **8**(23), 196–202 (2010)
35. Whitmire, E.: Disciplinary differences and undergraduates' information-seeking behavior. J. Am. Soc. Inf. Sci. Technol. **53**(8), 631–638 (2002)
36. Wilson, T.D.: Models in information behaviour research. J. Doc. **55**(3), 249–270 (1999)