Identification of Navigation Lead Candidates Using Citation and Co-Citation Analysis

Robert Moro, Mate Vangel, Maria Bielikova

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
name.surname@stuba.sk

Abstract. Query refinement is an integral part of search, especially for the exploratory search scenarios, which assume that the users start with ill-defined information needs that change over time. In order to support exploratory search and navigation, we have proposed an approach of exploratory navigation in digital libraries using navigation leads. In this paper, we focus specifically on the identification of the navigation lead candidates using keyword extraction. For this purpose, we utilize the citation sentences as well as the co-citations. We hypothesize that they can improve the quality of the extracted keywords in terms of finding new keywords (that would not be otherwise discovered) as well as promoting the important keywords by increasing their relevance. We have quantitatively evaluated our method in the domain of digital libraries using experts' judgement on the relevance of the extracted keywords. Based on our results, we can conclude that using the citations and the co-citations improves the results of extraction of the most relevant terms over the TF-IDF baseline.

Keywords: navigation leads, keyword extraction, domain modeling, citation analysis, co-citations, digital libraries

1 Introduction

There are various needs that motivate users to search vast information spaces, such as the Web or digital libraries. A classical taxonomy of Web search by Broder [2] differentiates three types of needs, namely (*i*) *navigational*, the goal of which is to locate a specific web page whose existence is known to the user, (*ii*) *informational*, the goal of which is to acquire certain information the whereabouts of which are unknown to the user, and (iii) *transactional*, the goal of which is to locate a web page where further transaction will occur, e.g., online shopping. Out of these, the informational need is the most general, but we always assume that users know exactly what kind of information they need (e.g., the title of an article written by an author).

However, the information need of users is often ill-defined at the beginning and it tends to change in the light of new information that they gather during the search. Thus, their search tasks tend to be open-ended and more exploratory in their essence; the term *exploratory search* was coined by Marchionini [10] for this type of searches.

An integral part of exploratory search is *sense-making* [16], i.e., making sense of a problem at hand or a new domain, learning the basic concepts and relationships between them, etc. A typical example of this behavior is *researching a new domain*, a task that researcher novices (e.g., doctoral students) often have to face. Their goal is not to find the specific facts, but to *learn* about the given domain and *investigate* the topics, the existing approaches as well as the gaps in the current state of knowledge.

In order to support exploratory search and navigation, we have proposed a method of exploratory navigation using the *navigation leads*, i.e., the automatically extracted keywords, which help the users to filter the information space [11]. The conceptual overview of the method can be seen in Fig. 1. It is a modification of the classical model of web information retrieval (IR) as defined in [2]. The users do not have to refine their query manually, but we augment the search results with the navigation leads. When the users choose a specific lead visualized in a summary of a search result or underneath it, their query gets modified with the lead so that only documents containing the selected lead are retrieved. The idea is similar to the probabilistic (or blind) relevance feedback [15], but in contrast our approach does not expand the query automatically, but lets the users to decide which terms to use. Also, lacking the relevance judgements, we rely on the topical relevance of the extracted terms [11].

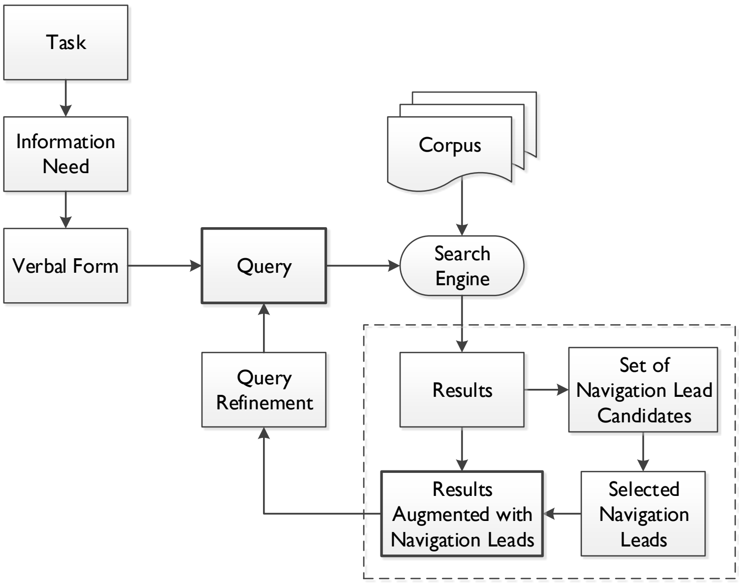

Task
Information Need
Verbal Form
Query
Query Refinement
Corpus
Search Engine
Results
Set of Navigation Lead Candidates
Results Augmented with Navigation Leads
Selected Navigation Leads

Fig. 1. A model of web IR augmented with the navigation leads to support exploratory search.

The process of augmentation of the search results with the navigation leads consists of three main steps (see Fig. 1):

1. *Identification of the navigation lead candidates* – it includes automatic keyword extraction from the documents as well as assessment of their document relevance; it results into a set of the navigation lead candidates.

2. *Selection of the navigation leads* – in this step, the document relevance of the individual keywords is combined with their navigational value, i.e., how relevant the lead candidates are for the whole information (sub)space. The result of this step is a set of the selected navigation leads.
3. *Visualization of the leads with the search results* – the selected navigation leads need to be placed into the search results list, preferably within the summaries (abstracts) of the search results or underneath them. The result of this step is a list of search results augmented with the navigation leads.

While in [11] we have introduced the concept of navigation leads and provided the preliminary results on the step two, i.e., selection of leads, in this paper, we focus mainly on the first step, i.e., the identification of the lead candidates. For this purpose, the document metadata can be utilized alongside the actual contents of the documents. We are interested in the researcher novice scenario in the digital libraries (or more precisely digital library systems that provide access to scientific publications). Therefore, we consider also the metadata that are specific for this domain (see next section). The most prominent of these are citations, i.e., *citing sentences* that provide a unique source of information; they highlight different aspects of the articles (documents) that were deemed important or interesting by other researchers.

We propose a method of keyword extraction using citation analysis that serves as a means of navigation lead candidates identification. Besides the direct citations that have been to some extent examined in the related works, e.g., [1, 4, 9], we consider also the *co-citations* (two articles are co-cited, if there is a third article that cites them both) that to the best of our knowledge have not yet been utilized for this purpose.

In this paper, we examine the following research questions:

1. Does the use of citations and co-citations in the process of keyword extraction for the purpose of navigation lead identification helps to improve the overall quality of the extracted keyword set? Are they capable of finding new words, or promoting the important words by boosting their relevance in comparison with the content-based (TF-IDF) baseline?
2. What are the limitations of using the citation and co-citation analysis with respect to the number of citations of an individual article?

We provide a quantitative evaluation of our method in the domain of digital libraries using experts' judgement on the relevance of the extracted keywords.

2 Identification of Keywords in Digital Libraries

Although our proposed approach of exploratory navigation using navigation leads can be in general applied to an information space in any domain, the knowledge of domain specifics allows us to tailor the method to them and thus, to improve the overall navigation process. Our focus is on the domain of digital libraries, or more specifically (as we have already mentioned) on the digital library systems of journal (or in general research) articles.

2.1 Specifics and Similarities between Digital Libraries and the Web

When designing a method of keyword extraction, we need to consider several specifics of the domain of digital libraries, which make it distinct from the wild Web:

- *Size and structure of information space* – the size of information space of digital libraries is much smaller in comparison with the whole Web, also the rate with which there is a new content is lower. Because the content is in most cases protected by copyrights, we can observe a separation of metadata, which are publicly available and easily processed, from the actual content of the documents.
- *Structure of documents* – in contrast to the wild Web, the documents in digital libraries tend to follow a predefined structure. This structure can differentiate among different publishers and journals, but it is always possible to identify the basic building blocks of the documents, such as title, authors, abstract, etc. There are approaches to automatically extract a table of contents of an article, as well as the actual contents of its sections [7], which can be utilized to reweigh the extracted keywords based on the section of the article in which they occur.
- *Unique set of metadata* – the articles in digital libraries have various metadata associated, which differ from the other sources on the Web in general, such as authors, publishers, where it was published and more interestingly, keywords that are identified by the authors themselves. This all can help to identify the most important aspects of the articles by taking into consideration not only the actual content of an article, but, e.g., also other similar articles from the same authors, etc.

There are also several aspects of this domain that are analogous to those of the Web:

- *User-added tags* – they represent a special type of metadata, because they are added by the users that use them in order to organize the articles for their later retrieval. It is a unique source of metadata, because users tend to use their own vocabulary and can highlight different aspects of the articles than their authors.
- *Links between documents* – the articles are linked by the use of citations. In contrast to the hyperlinks, their mining often requires advanced text mining techniques, because the reference can occur in the text after it has been explained.

The focus of this work is on the latter, i.e., on the citations, or more specifically, on the citation sentences and on their use for extraction of keywords from research articles for the purpose of the navigation lead candidates identification. They provide a unique view of the article content from a point of view of other researchers; citations cover different aspects of an article, but the amount of unique information converges as the number of citations increases [4]. There is an overlap between the topics (and the keywords) that we can extract from the abstract of an article (which has a special place in the domain of digital libraries, because the abstracts are usually freely accessible, unlike the article contents), but the topics of the abstract tend to be more general than those present in the citation sentences [9].

In our work, we examine also the co-citations assuming that there is a stronger relationships between frequently co-cited documents; however, there are other aspects in play, such as proximity of the co-cited articles in the document [8].

2.2 Domain Modeling in Annota

The specifics of digital libraries discussed in the previous section are reflected in the domain model of a bookmarking service Annota (http://annota.fiit.stuba.sk), which we have developed. The users can bookmark and annotate research articles in the digital libraries, such as ACM DL, IEEE Xplore, Springer Links, etc. The metadata of these articles are automatically extracted and processed into our domain model.

The model is two-layered; the domain model at the first layer is overlaid by the user model at the second [6]. It is a graph: the vertices consist of the normalized extracted terms and of the research articles; the weighed edges model the associations between the terms and the articles as well as between the terms themselves. Figure 2 shows a conceptual model of the domain representation used in Annota.

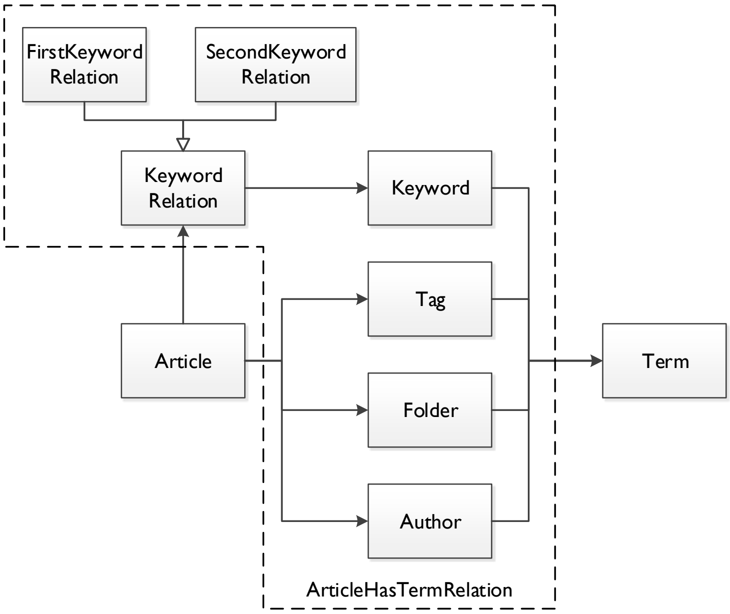


Fig. 2. A conceptual model of the domain representation in Annota. A relation between an article and a term consists of a combination of various relations that can be configured.

The main relation is between an *Article* and a *Term* entity; it is modelled as a combination of partial relations coming from various sources, e.g. the user-added tags, the folder names into which the users organize their articles, author-added keywords, etc. Additionally, we automatically extract keywords from articles (*Keyword* entity) that are used as the navigation lead candidates. The model allows various keyword extraction services to be used and combined into new types of *KeywordRelation*. It is also possible to configure which partial relations (and with which weights) should be used in a combination of *ArticleHasTermRelation*. This makes the model flexible and enables us to test various settings of the model as well as to quantitatively compare various keyword extraction services by using Annota as an A/B testing platform.

3 Method of Keyword Extraction Using (Co-)Citation Analysis

We have proposed a method of keyword extraction using citation and co-citation analysis that we employ for the purpose of identification of a set of potential navigation lead candidates as an extension of the existing domain model. A conceptual model of this extension can be seen in Fig. 3.

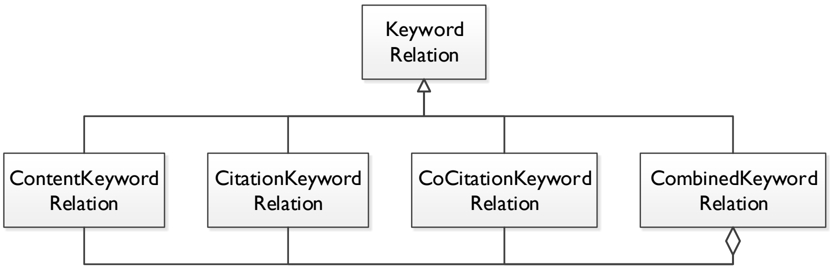


Fig. 3. An extension of the *KeywordRelation* that combines keywords extracted from a document content, its citations, and the co-cited documents.

The method combines keywords extracted from three sources: (*i*) the content of a document itself, (*ii*) the citations of a document, and (*iii*) the co-cited documents.

In order to extract the keywords from a document content, we first preprocess the document—we tokenize the text and lemmatize it (transform the terms into their dictionary form)—and then assess the relevance of the terms using a TF-IDF metric.

During the citation analysis, we identify all the *citation contexts* from the articles that cite a given document (that are present in our domain model). We define a citation context as 100 words before and 100 words after the occurrence of a reference to a given article in a citing document. This value is based on the results in [14]. We use ParsCit [3] in order to extract the citation contexts from the citing articles. The extracted citation contexts are preprocessed the same as the document content; TF-IDF is used to assess the relevance of the terms as well.

Lastly, during the co-citation analysis, we identify all the documents which are cited alongside the document that is being analyzed. We assume that the fact that two documents (articles) are co-cited increases their chance of being similar; therefore, we can use the keywords extracted from the co-cited documents and extend with them a set of the keywords extracted from a document content and its citations. We compute two measures for each co-cited article:

- *Co-citation weight (CW)* – it represents a frequency with which two articles are co-cited. If, e.g., the articles A and B are both cited by the articles C, D, and E, then the value of CW of B w.r.t to A (and vice versa) equals to 3. It is a global (or aggregated) measure of the co-citation relevance.
- *Co-citation proximity index (CPI)* – it is based on [5]; if the articles A and B are referenced in the same sentence of the article C, than CPI equals to 1, if in the same paragraph, it equals to ½, if in the same section, it equals to ¼ and if they occur in the same article, CPI equals to ⅛, which is the smallest possible value for any co-

citation. In other words, the closer the references to the articles occur in the text, the stronger the relationship we assume to be. As it characterizes the individual co-cited articles, it is a local measure of the co-citation relevance.

We have defined a set of rules for deciding, which co-citations to consider for the keyword extraction, to maximize the chance of two co-cited articles being similar:

1. In case of a maximal co-citation weight CW_{max} for a given document being larger than or equal to *N*, we ignore all co-citations with CW lower than *N*.
2. If CW_{max} lies within interval <2, *N*-1>, we use only co-cited documents, the weight of which equals to CW_{max}.
3. If CW_{max} equals to 1, we consider only co-citations that are the closest to a given document for each citing document separately based on their CPI. In other words, if there were, e.g., five citing documents, we would consider for each document only co-citations with the maximal CPI.

The specific values of a threshold value N can differ based on the domain and dataset used. We have empirically chosen *N* to be 5 based on the standard dataset that we used during the evaluation (see next section). It should be fine-tuned if a distribution of citations and co-citations differs significantly from the one in our dataset.

As the last step, we combine the keywords extracted from the content of the document with those extracted from the citation contexts as well as from the selected subset of co-cited articles (applying the same treatment during the text processing). We use a linear combination of weights which we normalize using the min-max normalization prior to the combination itself:

$$w = \alpha w_D + \beta w_C + \gamma w_{CC} \tag{1}$$

where w_D is a weight of a keyword extracted from the document content, w_C is a weight of a keyword extracted from the citation contexts and w_{CC} is a weight of a keyword extracted from the co-cited documents. The weight is for all three sources determined using a TF-IDF metric. The coefficients α, β, and γ are real numbers from the interval <0, 1>; we used a value of 1 in all our experiments.

The result is a set of extracted keywords, which serve as navigation lead candidates, but can be used also for other purposes, as they are a part of the domain model.

4 Evaluation

We have evaluated our proposed method of keyword extraction using the citation and the co-citation analysis on a standard dataset of articles in ACL Anthology Network [13] that we imported into Annota. Overall, the dataset consists of 18,290 articles with 84,237 citations. We have conducted a quantitative evaluation using the experts’ judgements; therefore, we have limited a number of articles from the original dataset to those which have already been known by the experts, i.e., which they had bookmarked in their Annota personal libraries. This way, we could use a subset of dataset consisting of 250 articles.

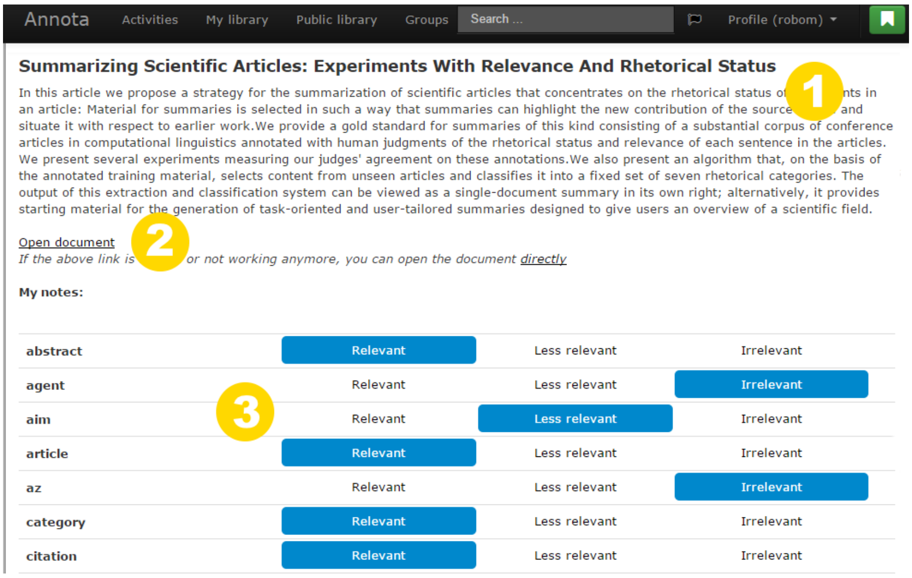
The goal of the experiment was to assess the relevance of the extracted keywords based on the experts' judgements. We have extracted three keyword sets for each document:

- *M1*: Keywords extracted from a content of a document using TF-IDF; this served us as a baseline.
- *M2*: Keywords from M1 enriched with the keywords extracted from the citation contexts of a given article.
- *M3*: Keywords from M2 enriched with the keywords extracted from the co-cited articles; this represents our proposed method discussed in the previous section.

We have hypothesized that adding the keywords extracted from the citation contexts and from the co-cited articles will improve the overall precision *P*, i.e., formally:

$$P(M1) < P(M2) \leq P(M3) \tag{2}$$

We have prepared an evaluation interface in Annota for the experts to be able to easily assess the relevance of the presented keywords (see Fig. 4). The experts could have chosen one of the possible assessments of a keyword relevance: *relevant*, *less relevant* (meaning *somewhat relevant*), or *irrelevant*.



Summarizing Scientific Articles: Experiments With Relevance And Rhetorical Status

In this article we propose a strategy for the summarization of scientific articles that concentrates on the rhetorical status o nts in an article: Material for summaries is selected in such a way that summaries can highlight the new contribution of the source and situate it with respect to earlier work.We provide a gold standard for summaries of this kind consisting of a substantial corpus of conference articles in computational linguistics annotated with human judgments of the rhetorical status and relevance of each sentence in the articles. We present several experiments measuring our judges' agreement on these annotations.We also present an algorithm that, on the basis of the annotated training material, selects content from unseen articles and classifies it into a fixed set of seven rhetorical categories. The output of this extraction and classification system can be viewed as a single-document summary in its own right; alternatively, it provides starting material for the generation of task-oriented and user-tailored summaries designed to give users an overview of a scientific field.

Open document

If the above link is or not working anymore, you can open the document directly

My notes:

abstract	[✓ Relevant]	Less relevant	Irrelevant
agent	Relevant	Less relevant	[✓ Irrelevant]
aim	Relevant	[✓ Less relevant]	Irrelevant
article	[✓ Relevant]	Less relevant	Irrelevant
az	Relevant	Less relevant	[✓ Irrelevant]
category	[✓ Relevant]	Less relevant	Irrelevant
citation	[✓ Relevant]	Less relevant	Irrelevant

Fig. 4. An evaluation interface in Annota. The experts could have read the title and the abstract of an article (1), or navigate to its fulltext (2). They assessed the relevance of the presented keywords by choosing one of the possible values – *relevant*, *less relevant*, or *irrelevant* (3).

Under each article, we have presented top ten scoring keywords extracted by each compared variant (M1, M2, M3), i.e., together up to 30 keywords for each document merged into a single list. All the keywords were presented only once in the list even if there was an overlap between the sets of extracted keywords. The keywords in the list were sorted alphabetically so that the experiment participants could not have found out which keywords were extracted by which method (and with which relevance).

There were 8 domain experts who participated in the experiment. Together, they assessed 844 unique extracted keywords from 45 different articles; 7 articles were assessed by more than one expert. We have evaluated each variant (M1, M2, M3) when considering only relevant or also less relevant keywords using a standard information retrieval metric P@N (precision at N); results are shown in Table 1. The metric computes a ratio of relevant keywords among the top N scoring ones. We can see that the variant M2 and M3 outperformed the TF-IDF baseline (M1) in all P@N measures for N = 1, 2, 5, and 10, thus confirming our hypothesis. As to the comparison of M2 and M3 variant, considering just citations provided in general a slightly better precision, except of P@5, where co-citations significantly improved the results (there was an improvement also for P@10, although marginal). A closer analysis of the extracted keywords revealed that using M3 variant in some cases promoted names of the authors or acronyms of the method names which were deemed irrelevant by the judges, as they were too specific and not really describing the content of the articles.

Table 1. The results of precision for each variant when considering only the keywords assessed by the experts as *relevant* (*R*), or when considering also those assessed as *less relevant* (*R+L*).

	P@1		**P@2**		**P@5**		**P@10**	
Variant	**R**	**R+L**	**R**	**R+L**	**R**	**R+L**	**R**	**R+L**
M1	78.85	88.46	69.23	79.81	60.08	73.26	51.35	68.73
M2	**88.46**	**96.15**	**81.73**	**95.19**	59.46	74.90	**53.49**	69.77
M3	84.62	94.23	79.81	93.27	**63.85**	**80.77**	52.13	**70.35**

We have analyzed also the limitations of our proposed method concerning the second research question. For this purpose, we have compared values of the P@10 measure (when considering the relevant as well as the less relevant keywords) for each variant (M1, M2, M3) with respect to the number of citations of an article (see Fig. 5). Because the number of the evaluated keywords in the selected citation groups was not the same, we have reported also the confidence intervals.

We can see that though the overall scores of the P@10 measure are almost identical for all the variants (see Table 1), there are differences based on the number of citations. The TF-IDF baseline (M1) slightly outperforms the other two (M2, M3) if the number of citations is lower than 20, although the measured differences are not significant. The situation changes with the increasing number of citations in favor of variants M2 and M3 with M3 (employing co-citations) being better. Again, the differences are in the most cases not significant (with the exception of articles with at least 20 and less than 100 citations), but the trend is clear. This is in agreement with the previous finding in [4] that 20 citations can in general cover all the important aspects of an article. Novelty lies in the comparison with the co-citations that seem even better suited for the keyword extraction in this case.

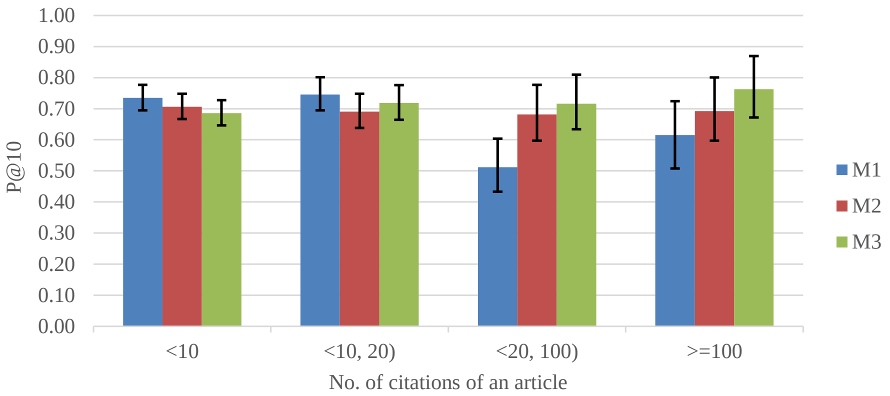


Fig. 5. A comparison of P@10 for each variant w.r.t. the number of citations of an article. The error bars around the mean values represent the confidence intervals.

Lastly, we have found out that out of 844 extracted keywords that have been assessed by the experts, 119 were identified by the M3 variant, meaning that they did not occur at all among the keywords extracted by M1 and M2 or they were deemed irrelevant be them. Out of these 119 keywords, 54.6% were assessed as relevant or less relevant by the experts. This suggests that considering the co-citations is capable of finding new important words, although there is still some noise which could be reduced by making the rules for deciding, which co-cited articles to consider, more strict.

5 Related Work

As we have already established throughout this article, citations reflect value, impact, and importance of research works, which makes them interesting in areas such as scientometrics. They also represent a judgement of other researchers on the actual content of a research work, which is a reason, why they are researched also in the field of natural language processing (e.g., for keyword extraction and summarization).

It has been found out that as much as 20 citations is enough to cover all the important aspects of an article [4]. The topics extracted from the citations differ significantly from those that can be extracted from the abstracts [9] and using at least one citation sentence for keyword extraction renders better results in comparison with the keywords extracted solely from the document content, while the optimal citation context is 100 words before and after a reference in a citing article [14].

The keywords (or keyphrases) extracted from the citation contexts are also capable of producing better summaries of researcher articles as shown in [12] on 25 manually annotated articles from ACL Anthology Network.

There is still much work to be done on the categorization and characterization of citations and their role within an article; Bertin and Atanassova [1] analyzed verbs used in the citations in different sections of articles and found out that there are significant differences which can be attributed to their different roles. This remains a challenge, because the most of the works do not differentiate between these roles when using citations for keyword extraction or other natural language processing related tasks. Also, the exact role and potential contribution of the co-citations for these tasks remains an open problem, which we have tried to tackle also in this work.

6 Discussion and Conclusions

In this work, we have presented a general model of Web search that we have augmented for exploratory search and navigation in an information space using navigation leads. In order to select navigation leads, we first identify a set of potential lead candidates using automatic keyword extraction.

For this purpose, we have proposed a method of keyword extraction in digital libraries domain employing citation and co-citation analysis and tried to answer two research questions, namely whether the citations and the co-citations improve keyword extraction and what are their limitations. Our main contributions are as follows:

1. We have extended a domain model of a digital library system with keywords extracted from the co-cited articles and proposed a set of rules for deciding which co-citations to consider, and which not.
2. We have evaluated usefulness of citation and co-citation analysis on a standard dataset and examined their limitations. Based on the results of our quantitative experiment, we can conclude that using citations and co-citations significantly outperforms the TF-IDF baseline; in addition, the co-citations are capable of finding new keywords that would not have been otherwise extracted. As to the limitations of our proposed method, its precision depends on the number of citations of an article with the ideal number being above 20.

Although co-citation analysis is capable of finding new keywords, there is still some noise in the form of irrelevant keywords that should be addressed in the future (e.g. by filtering out the names of authors or method acronyms by setting a higher threshold for minimal number of occurrences of a term in the documents when using TF-IDF). Open question remains, how to automatically adapt the rules of co-citation selection so that they would take into consideration citation specifics of different domains.

In addition, our method depends on a number of citations of an article, i.e., it performs well when there is enough information in the form of citations and co-citations and is outperformed by the TF-IDF baseline, if there is not. This could be reduced by modifying the method to automatically adapt the weights based on the number of citations, thus promoting content keywords for articles with only a few citations and gradually increasing the weight of the citation and the co-citation keywords with the increasing number of citations. Other promising direction is to analyze the citation intention and its role within an article.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 009STU-4/2014, the Scientific

Grant Agency of the Slovak Republic, grant No. VG 1/0646/15, and by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

1. Bertin, M., Atanassova, I.: A Study of Lexical Distribution in Citation Contexts through the IMRaD Standarda. In: Proc. of the 1st Workshop on Bibliometric-enhanced Inf. Retrieval co-located with 36th European Conf. on Inf. Retrieval (ECIR 2014). pp. 5–12. CEUR-WS (2014).
2. Broder, A.: A Taxonomy of Web Search. ACM SIGIR Forum. 36, 3–10 (2002).
3. Councill, I.G., Giles, C.L., Kan, M.: ParsCit: An Open-Source CRF Reference String Parsing Package. In: LREC ’08: Proc. of the 6th Int. Conf. on Language Resources and Evaluation. pp. 661–667. ELRA (2008).
4. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., Radev, D.R.: Blind Men and Elephants: What Do Citation Summaries Tell us about a Research Article? J. Am. Soc. Inf. Sci. Technol. 59, 51–62 (2008).
5. Gipp, B., Beel, J.: Citation Proximity Analysis (CPA) – A New Approach for Identifying Related Work Based on Co-Citation Analysis. In: ISSI ’09: Proc. of the 12th Int. Conf. on Scientometrics and Informetrics. pp. 571–575. ISSI (2009).
6. Holub, M., Moro, R., Sevcech, J., Liptak, M., Bielikova, M.: Annota: Towards Enriching Scientific Publications with Semantics and User Annotations. D-Lib Mag. 20, (2014).
7. Klampfl, S., Kern, R.: An Unsupervised Machine Learning Approach to Body Text and Table of Contents Extraction from Digital Scientific Articles. In: Research and Advanced Technology for Digital Libraries, LNCS 8092. pp. 144–155. Springer, Berlin Heidelberg (2013).
8. Liu, S., Chen, C.: The Effects of Co-Citation Proximity on Co-Citation analysis. In: ISSI ’11: Proc. of the 13th Int. Conf. of the Int. Society for Scientometrics and Informetrics. pp. 474–484 (2011).
9. Liu, S., Chen, C.: The Differences between Latent Topics in Abstracts and Citation Contexts of Citing Papers. J. Am. Soc. Inf. Sci. Technol. 64, 627–639 (2013).
10. Marchionini, G.: Exploratory Search: From Finding to Understanding. Commun. ACM. 49, 41–46 (2006).
11. Moro, R., Bielikova, M.: Navigation Leads Selection Considering Navigational Value of Keywords. In: WWW ’15 Companion: Proc. of the 24th Int. Conf. on World Wide Web Companion. pp. 79–80. IW3C2, Geneva (2015).
12. Qazvinian, V., Radev, D.R., Özgür, A.: Citation Summarization through Keyphrase Extraction. In: COLING ’10: Proc. of the 23rd Int. Conf. on Computational Linguistics. pp. 895–903. Association for Computational Linguistics (2010).
13. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL Anthology Network Corpus. Lang. Resour. Eval. 47, 919–944 (2013).
14. Ritchie, A., Robertson, S., Teufel, S.: Comparing Citation Contexts for Information Retrieval. In: CIKM ’08: Proc. of the 17th ACM Conf. on Inf. and Knowledge Mining. pp. 213–222. ACM Press, New York (2008).
15. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval, 3(4), pp.333–389 (2009).
16. White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool (2009).