Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2014 Proceedings

ENHANCING LITERATURE REVIEW METHODS - TOWARDS MORE EFFICIENT LITERATURE RESEARCH WITH LATENT SEMANTIC INDEXING

André Koukal
Leibniz Universität Hannover, Information Systems Institute, Hannover, Germany, koukal@iwi.uni-hannover.de

Christoph Gleue
Leibniz Universität Hannover, Information Systems Institute, Hannover, Germany, gleue@iwi.uni-hannover.de

Michael Breitner
Leibniz Universität Hannover, Information Systems Institute, Hannover, Germany, breitner@iwi.uni-hannover.de

André Koukal, Christoph Gleue, and Michael Breitner, 2014, "ENHANCING LITERATURE REVIEW METHODS - TOWARDS MORE EFFICIENT LITERATURE RESEARCH WITH LATENT SEMANTIC INDEXING", Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, ISBN 978-0-9915567-0-0
http://aisel.aisnet.org/ecis2014/proceedings/track19/1

ENHANCING LITERATURE REVIEW METHODS - TOWARDS MORE EFFICIENT LITERATURE RESEARCH WITH LATENT SEMANTIC INDEXING

Prototype

Koukal, André, Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany, koukal@iwi.uni-hannover.de

Gleue, Christoph, Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany, gleue@iwi.uni-hannover.de

Breitner, Michael H., Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany, breitner@iwi.uni-hannover.de

Abstract

Nowadays, the facilitated access to increasing amounts of information and scientific resources means that more and more effort is required to conduct comprehensive literature reviews. Literature search, as a fundamental, complex, and time-consuming step in every literature research process, is part of many established scientific methods. However, it is still predominantly supported by search techniques based on conventional term-matching methods. We address the lack of semantic approaches in this context by proposing an enhancement of established literature review methods. For this purpose, we followed design science research (DSR) principles in order to develop artifacts and implement a prototype of our Tool for Semantic Indexing and Similarity Queries (TSISQ) based on the core concepts of latent semantic indexing (LSI). Its applicability is demonstrated and evaluated in a case study. Results indicate that the presented approach can help save valuable time in finding basic literature in a desired research field or increasing the comprehensiveness of a review by efficiently identifying sources that otherwise would not have been taken into account. The target audience for our findings includes researchers who need to efficiently gain an overview of a specific research field, deepen their knowledge or refine the theoretical foundations of their research.

Keywords: Literature review, literature research, latent semantic indexing (LSI), information retrieval, design science research (DSR).

1 Introduction

A complete literature research process sets the basis for every research project and represents an "essential first step and foundation when undertaking a research project" (Baker, 2000). Before attempting to contribute to any research field, it is crucial to be aware of what is already known in the respective scientific discipline's body of knowledge (Hart, 1998; Levy and Ellis, 2006). Webster and Watson (2002) state that in order to strengthen IS as a field of study, effective literature review methods may provide great value and that well-founded and rigorously conducted literature reviews have a higher chance of getting published. However, the constant increase in the number of scientific publications, facilitated access to scientific resources through new technologies (Park and Lee, 2011) and a resulting complex information environment (Bawden and Robinson, 2009) imply that a manually-conducted literature review becomes an increasingly time-consuming task.

Despite their usefulness, keyword-based approaches have their shortcomings (Homayouni et al., 2004; LaBrie and St. Louis, 2003). Ambiguity, synonymy, polysemy, the inappropriate use of plurals or stopwords like "and" and, ultimately, the indexers' inconsistency when applying subject terms can distort the query results. Hence, keyword searches are likely to cause false-positive or false-negative errors, i.e. potential matches are missed or mismatches are considered (Hofmann, 1999; Yandell and Majoros, 2002). However, most existing search engines still rely on term-matching methods only (Cui et al., 2003). We observed that search mechanisms of most of the repositories for research papers, e.g. AISeL, IEEE, JStor, ScienceDirect, Wiley, are also keyword-based. Thus, overcoming the aforementioned deficiencies is an important challenge in IS research and research in general.

Promising assessments from different authors indicate that latent semantic indexing (LSI) might provide a solution to the aforementioned set of problems and is likely to outperform established lexical matching similarity methods. The appropriateness and applicability of LSI to a wide variety of tasks has been proven already (Deerwester et al., 1990; Gordon and Dumais, 1998; Kontostathis and Pottenger, 2002). We argue that the use of a semantic approach is likely to increase efficiency and, thus, save valuable time in identifying the relevant literature in a designated research field, while potentially avoiding the recent challenge of the proliferation of terms describing similar concepts in IS research (Lebek et al., 2013). Accordingly, the objective of this paper is to introduce the prototype of our "Tool for Semantic Indexing and Similarity Queries" (TSISQ) and discuss this alternative, LSI-based approach for identifying relevant publications. The paper addresses the following research question:

> *RQ: How can a LSI-based approach be adopted and implemented to increase the efficiency of scientific literature research processes?*

The remainder of this paper is structured as follows: After this introduction, the research background is addressed. Next, the theoretical concepts of LSI are presented, followed by an illustration of the TSISQ prototype for the enhancement of any scientific literature research process. Additionally, a demonstration and evaluation of the tools' performance on a large index of complete research papers is performed. This is followed by a critical discussion, theoretical and practical recommendations as well as limitations. Finally, a short conclusion and implications for further research are drawn.

2 Research background

2.1 Theoretical background and related work

Literature reviews are the most basic, yet very important concept to set a theoretical basis. Their quality and usefulness greatly depends on the literature research process (vom Brocke et al., 2009). In the IS community, various frameworks for conducting a quality literature review exist (e.g. Levy and Ellis, 2006; Okoli and Schabram, 2010; Webster and Watson, 2002). Although the respective authors propose different sets of guidelines, it appears to be common sense that it is of particular importance to

get a broad understanding of the pursued research topic. Thus, the identification of relevant literature is an important subtask in every literature review (Wolfswinkel et al., 2013). Taking into consideration the established guidelines, our aim is *not* to introduce an entire new method for literature reviews but to facilitate certain steps of the well-established ones by introducing a prototype that supports the literature research process. In the following, a brief overview of approaches is provided that address the underlying challenge of identifying semantic similarities between texts.

Query expansion (QE) is an information retrieval technique that aims to advance retrieval effectiveness by extending the provided search terms by synonyms or related terms. It addresses some fundamental deficiencies of keyword queries, such as word mismatch and synonymy (Cui et al., 2003; Liu et al., 2011; Mitra et al., 1998; Qiu and Frei, 1992). However, it is not suitable for the underlying problem of this study for the following reasons: first, the aim of our retrieval activity is not to increase the *quantity*, but the *quality* of the query results. While the problem of synonymy may be diluted, the problem of polysemy (ambiguity of a term) remains unsolved (Liu et al., 2011). Many authors state that QE techniques mostly do not increase query effectiveness (Vorhees, 1994; Xu and Croft, 1996).

Semantic similarity is "[…] a concept by which a metric is given to groups of terms or documents based on the similitude of their meanings" (Furlan et al., 2013). Thus, one of the key concepts in the understanding of natural languages is the field of natural language processing (NLP) with three fundamental aspects: information extraction, semantics, and information retrieval, with the latter term referring to document-based and query-based retrieval (Yandell and Majoros, 2002). LSI, also referred to as latent semantic analysis (LSA), belongs to the field of NLP techniques. Put simply, LSI maps meaning into a semantic space (Kintsch, 2010). Due to its generality, it is a valuable analysis technique for many different problems in practice involving textual data, such as search and retrieval (Dumais, 1992 and 1994), classification (Zelikovitz and Hirsh, 2001), and filtering (Zha and Simon, 1998). Hence, LSI has a wide range of possible applications (e.g. Deerwester et al., 1990; Foltz and Dumais, 1992; Hofmann, 2001; Landauer and Dumais, 1997; Wolfe et al., 1998) and has proven to be effective in advancing average retrieval accuracy (Ding, 1999).

In the last three decades, there have been many publications about LSI and its mode of operation, evaluating LSI performance, theoretical approaches towards understanding LSI in detail and studies about optimizing the algorithm or parts of the LSI process, e.g. Brand (2006), Deerwester et al. (1990), Dumais (1992), Hofmann (1999), Řehůřek and Sojka (2010). However, only few publications exist on practical applications. LSI is used in the context of e-mail spam filtering (Gansterer et al., 2008; Gee, 2003), prediction of psychological phenomena (Wolfe and Goldman, 2003), text mining (Lee et al., 2010), automatic text summarization (Bhandari et al., 2008; Gong and Liu, 2001; Steinberger and Ježek, 2004; Yeh et al., 2005), and classification of contents (Shen et al., 2004; Kuechler, 2007).

To the best of our knowledge, the research gap we seek to address, which is to provide a foundation and tool for comparing a query formulated in natural language and a large body of complete IS research papers using LSI, has not been reported in academic literature yet. The works that appears to be closest to our study were published by Sidorova et al. (2008) and by Homayouni et al. (2004). However, the authors did not present tools, but used LSI-based approaches to analyze abstracts of IS research articles and in the field of bioinformatics. Since abstracts contain only aggregated content, we also want to enable an analysis of much larger text-corpora than the ones examined by them.

One challenge is that most of the common algorithms compute very large matrices directly in-memory. As this memory-intensive application demands a lot of computing power, it is desirable to keep the technical requirements as low as possible while not having to cut the tools' performance drastically. To address this issue, the framework that served as one of the main foundations for TSISQ, called gensim, uses an algorithm for a memory-efficient incremental process proposed by Brand (2006) (Řehůřek and Sojka, 2010). Řehůřek and Sojka (2010) state that, to their best knowledge, their implementation of Brand's algorithm is the only publicly available implementation of LSI that is independent of the index size, which allows an execution of TSISQ on an average, up-to-date computer.

2.2 Research design

Our research was conducted using design science research (DSR) principles in order to address relevance and enhance rigor of the research process and results. The design-orientated research process was recommended by Offermann et al. (2009) and, in particular, Peffers et al. (2008). Additionally, we used key recommendations provided by Hevner et al. (2004, 2007) and March and Smith (1995). The actual research design is classified as problem-centered approach (see Figure 1).

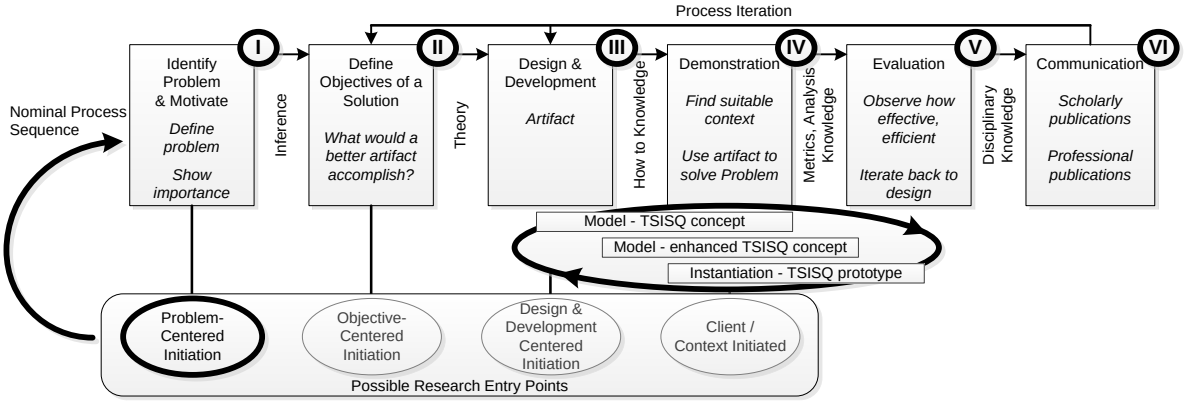

Figure 1. Research design according to the DSR methodology process (Peffers et al., 2008)

The lack of automated support in combination with a consideration of semantic concepts for text retrieval makes the literature search process slow and time consuming. The need for a more efficient approach to find relevant literature triggered the development of the TSISQ prototype. We initiated the research process by identifying the above-mentioned problem (I). To ensure methodological rigor, we conducted a comprehensive literature review, also with our tool, within the fields of methods for conducting literature reviews in the IS, information retrieval, and NLP domain. According to our research question, the design, demonstration and evaluation of artifacts that can provide an enhancement of the literature research process was the main objective (II). After refining the problem domain and defining specific requirements, the scientific input was used to design the first artifact (III): a basic model for TSISQ. It was limited to the central aspects of semantic indexing and similarity queries. For a further development, we used an iterative approach to create and refine artifacts cyclically according to guideline six, "design as a search process", by Hevner (2004). Thus, the basic model was extended with extra parameters, a layer for automated preparation of the content database and a web frontend, resulting in an enhanced model for TSISQ. March and Smith (1995) provide a classification into constructs, models, methods, and instantiations as the result of design-oriented research. In addition to the constructed formal models, an instantiation was implemented: a prototype of TSISQ. The DSR process cycles were then completed by more extensive tests of the artifacts and a demonstration (IV) and evaluation (V) of the capabilities of the prototype to enable documentation of research results (VI).

3 Enhancing the literature research process with TSISQ

Three core stages can be derived from the different guidelines for conducting a systematic literature review in the IS field. These principles should be addressed in every literature research process, which we understand as an important sub-step of a complete literature review. The first stage is the definition of the search space, e.g. the selection of a specific scientific database. The second stage is the search process in which papers that possibly fit the author's needs are identified. The third stage is the screening of the identified papers in order to check the content for relevant aspects. The TSISQ prototype allows an enhancement of the second stage by providing a search method that addresses the lack of not taking semantic concepts into consideration when performing keyword-based searches. Besides that, it may reduce efforts of the third stage due to less irrelevant articles to filter out of search results.

3.1 Underlying theoretical concepts and applied methods

To enable a computer-aided processing of contents, one core concept is the conversion of documents into its representation in the vector space model (VSM) (Salton et al., 1975). This concept represents the initial step of the processing of any document in TSISQ. Each document is defined as a t-dimensional vector in Euclidean space, where t corresponds to the amount of different terms of a document. This is combined with a weighting of each term in order to quantify its importance and relevance. A simple weighting would be the frequency of each term in a document, but we follow the most common approach that determines the term weights (Yandell and Majoros, 2002) by applying the term frequency-inverse document frequency (TFIDF) concept (Salton and McGill, 1986). It enhances the performance of retrieval systems (Maas et al., 2011) and consequently of our TSISQ prototype by discounting the influence of more common non-stopwords and promoting of occurrence of rare terms (Sidorova et al., 2008). The result of the conversion into VSM and the application of the TFIDF concept is a term-by-document matrix whose columns contain the weighting of terms for each of the considered documents. As this conversion of documents to fixed-length lists of numbers does not result in a greater reduction of the dimension of content nor in a greater consideration of the statistical structure of a document or a corpus, the problem of the VSM of not being able to deal with synonymy (e.g. "required" and "substantial") and polysemy (e.g. read a "book" and "book" a journey) persist.

To reduce the dimension and deal with the other shortcomings, we applied the LSI method. LSI is an extension of the VSM and uses co-occurrences of terms in order to take advantage of an implicit higher-order structure in the association of terms with documents ("semantic structure") (Zhang et al., 2011). For this purpose, TSISQ initially decomposes the previously created term-by-document matrix into three other matrices by a process called singular value decomposition (SVD) (Forsythe, 1977). As the standard procedure of SVD quickly exceeds memory limits, we make use of an incremental SVD processing algorithm by Brand (2006). Next, the three matrices are reduced in their dimensions to provide the best rank-k approximation of the original term-by-document matrix (Kontostathis, 2007).

The columns of the third reduced matrix are used for comparisons and similarity queries. Each column represents a vector that characterizes the aggregated semantic concept of the original content. In order to compare a query with documents, TSISQ initially converts a query input into its representation in the VSM and subsequently transforms it into the same space as the document vectors. The comparison of documents and queries is performed with the help of the cosine measure, which is perhaps the most frequently applied measure for comparison of document similarities (Korenius et al., 2004). Instead of determining the angle between query and document vectors, the cosine of the angle is calculated. TSISQ returns the absolute value of the cosine that expresses the similarity between query and documents within an interval of [0, 1]. The higher this value, the higher is the similarity.

3.2 Implementation and architecture of the TSISQ prototype

The TSISQ prototype is implemented in the Python programming language to allow cross-platform use. However, the system is only tested on linux systems, except of the web frontend. The system architecture and data flow are explained subsequently as presented in Figure 2.

Scientific articles in PDF file format are used as a foundation. These files are converted into plain text files with Xpdf. For the application of the presented methods of VSM, TFIDF, and SVD the software framework gensim is used in combination with simserver, a higher level control layer. Gensim is a NLP software framework which is based on the idea of document streaming (Řehůřek and Sojka, 2010). It requires the open source NumPy and SciPy libraries. NumPy provides n-dimensional array manipulation and SciPy provides routines for numerical integration and optimization. The advantages of gensim are fast processing of large datasets and memory independence because the term-by-document matrix does not have to be stored in memory. In addition, it enables the direct application of the SVD concept on a term-by-document matrix with term frequency weightings or with a previous

application of the TFIDF weighting scheme. The latter procedure is used in TSISQ. The indexing process is computationally expensive and requires some time for larger collections, e.g. the indexing process of 10000 documents took approximately half an hour on a system with an Intel Core i7-2640M CPU with 2.80 GHz, 8GB Ram. As a result of the indexing process, an index corpus file for further processing is created. TSISQ stores information about indexes, contents, file sizes of PDF and text files, and the query history in a SQLite database.

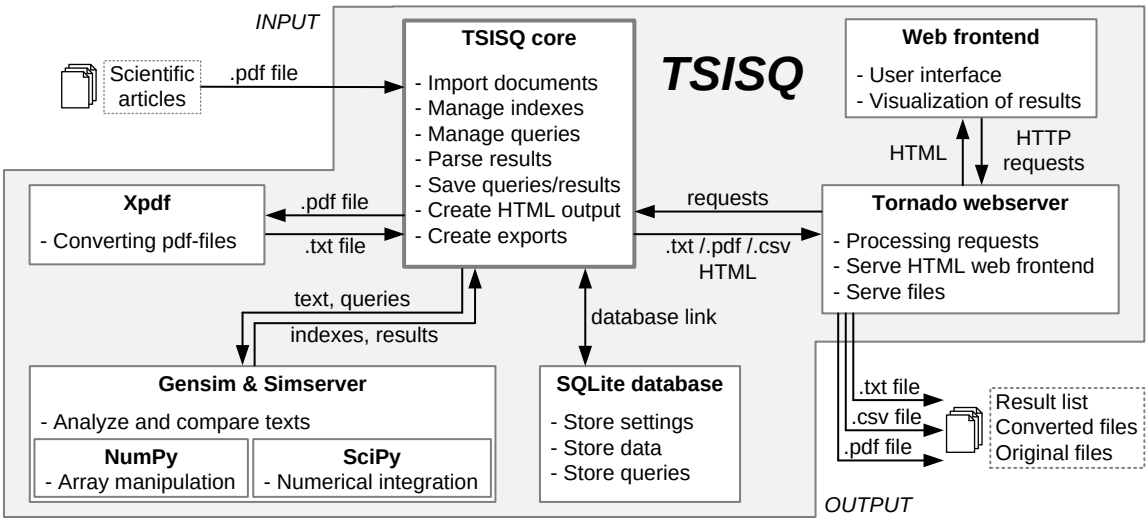

Figure 2. System architecture of the TSISQ prototype

For the delivery of user request and responses from TSISQ, the Python based Tornado web framework is used. It passes the user requests to the TSISQ core and presents the HTML web frontend to the user. The graphical representation of the web frontend is based on the Twitter Bootstrap framework. The web frontend allows to perform queries (Figure A.1 within the appendix), easily compare different query results with the help of a query history (Figure A.2), and visualize processed results (Figure A.3). Beyond that, the entire management of indexes, including the expansion of existing ones, can be performed with the help of the web frontend (Figure A.4). More information about the integrated tools and frameworks can be found on the respective homepages.

It is possible to use and present the TSISQ prototype with all features on a laptop with any common presentation system. The prototype with source code will be available online from June 2014 at:

http://www.iwi.uni-hannover.de/TSISQ (open access)

4 Demonstration and evaluation

In order to evaluate and show applicability of the constructed artifacts in the DSR process, we performed a case study. As it makes sense to start a literature review with the leading journals, because any major contribution is likely to be found in them (Webster and Watson, 2002), our data base to search in was composed of eight journals and four conferences of important IS research, see Table 1.

Journal	**Total**	**EJIS**	**ISR**	**JAIS**	**JIS**	**JIT**	**JMIS**	**JSIS**	**MISQ**
Articles	1660	269	263	189	140	157	248	125	269
Share in %	16.14%	2.62%	2.56%	1.84%	1.36%	1.53%	2.41%	1.22%	2.62%
Conference	**Total**	**AMCIS**	**ECIS**	**HICSS**	**ICIS**				**Total**
Articles	8626	2806	1342	3039	1439				10286
Share in %	83.86%	27.28%	13.05%	29.55%	13.99%				100%

Table 1. Database of journal and conference articles from 2007 to 2013

The design of this example is inspired by a hypothetical use case in which a relatively short text with aggregated, focused contents is used for topic-related literature research. Since abstracts of articles represent such texts, they might be an ideal input to perform a query and to identify relevant, related articles. However, also full articles can be used as query input. The selected research field for this purpose is enterprise content management (ECM), as it is a comparatively well-defined domain with a manageable amount of established literature incorporated into our index. As query input an abstract of an ECM article by Rickenberg et al. (2012) was utilized. The resulting output was reviewed by ECM domain experts, who ranked the top 40 results without knowledge about their individual ranking on a three-point scale: *irrelevant*, *relevant* or *highly relevant*.

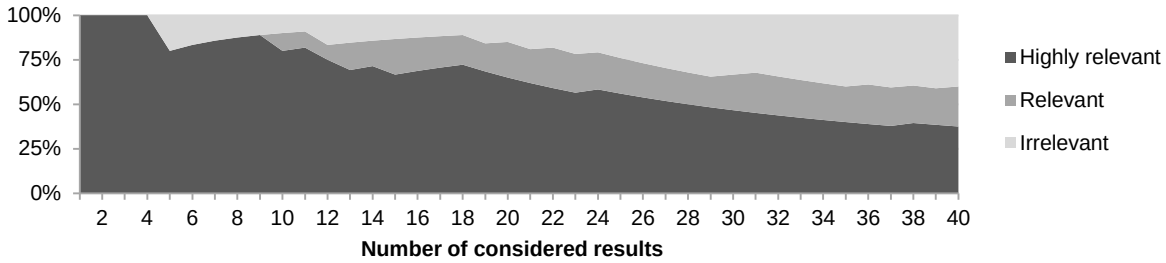

Figure 3. Identified results of the case study

According to Figure 3, the percentage of highly relevant articles within the top ten results is 80 percent, showing a decreasing trend the more results are taken into consideration. The share of articles classified as relevant remains approximately constant around 20 percent beginning from 15 considered results. The combined share of highly relevant and relevant articles shows that in the range from 1 to 25 results, the percentage of these articles remains constantly above 75 percent.

5 Discussion, limitations and recommendations

Based on established approaches of the document retrieval domain, we created a prototype of our "Tool for Semantic Indexing and Similarity Queries" to address the tasks of performing semantic analysis of texts and subsequently searching for similar content. Consequently, the instantiation represented by our implemented prototype has a strong practical focus. From a technical point of view, the whole process of an LSI based literature search tool like TSISQ is complex. While the effort of an implementation into an existing infrastructure is low, it requires considerable computing power to create index corpora. However, this procedure is only necessary when new documents are added to the data base and can be performed automatically by scheduled background tasks. From a researchers' perspective, preparing and conducting any query with TSISQ is not more time consuming than the use of conventional techniques. The possibility to use natural language as input without the need to further specify search terms or meet formal requirements concerning a search engines' syntax can even reduce time and effort for the user.

In general, in IS research, artifacts are not evaluated thoroughly in many cases (Arnott and Pervan, 2012). However, this step seems to be essential to ensure the quality of research. The results of our case study indicate that each of the first 25 results of this literature search is at least relevant to the field of ECM with a probability of 75 percent (hit rate). For the first ten results, the respective probability is even higher at 90 percent. Though these figures make no claim to be a general rule, it can be stated that the earlier an article appears in the result list, the more likely it is that it contains relevant content. Due to the plausibility and at least comparable quality of results, we assume that using our approach is in any case at least as efficient as conventional methods. The probability of receiving results of higher quality in the same or even less time may lead to an increase in efficiency. We are convinced that the articles identified by this procedure serve in any case as a solid foundation for further manual processing, e.g. forward and backward search as proposed by Webster and Watson (2002).

We identified some limitations with regard to our research artifacts. As in many other research projects, the data base is a critical spot for the examination of the results from the case study. Our data base consists of approximately 10000 research articles of a six year period for demonstration and evaluation. This is a comparably small amount. However, a wide range of top IS articles in diverse fields of research is covered. If a comprehensive literature review is to be conducted, the considered period may be insufficient. Nevertheless, to demonstrate and evaluate the feasibility of our prototype for conducting the proposed literature research process, the specified time frame is a good basis for further development. Besides this and as outlined earlier, synonymy and polysemy of terms are still common problems in information retrieval. Since the false-positive error rate is an indicator for the problem of synonymy and the false-negative error rate for polysemy, their reduction is one of the main issues to be addressed by an LSI-based approach. The case study, however, only allows us to draw conclusions about the false-positive error rate: the large size of the data base makes it nearly impossible to manually analyze which articles are missing in the result list. In order to further assess the quality of results returned by the TSISQ prototype, the false-negative error measure needs to be quantified. Additionally, no statement about the efficiency compared to alternative methods or established search engines e.g. AISeL can be derived from the case study. Consequently, it will be necessary to set up a test case that aims on benchmarking the TSISQ prototype against other approaches.

Upon examination of the outcomes of the case study and our own literature research with TSISQ, we can make recommendations for the use of the prototype to support the scientific literature research process. We propose an iterative process cycle by starting the literature search with a complete article related to the desired topic as query input in order to get an initial, broader selection of semantically similar research papers. Then, the results can be manually screened to identify the most relevant articles to a more specific topic. The next step includes the creation of a collection of concise key-words, phrases, or sections found in the identified selection of literature. Then, the resulting collection can be utilized as query input. This process can be repeated until no new relevant literature is found.

6 Conclusion and further research

The literature research process is an essential component of a literature review. The complex and highly important subtask of literature search is time-consuming and requires a lot of effort. The more comprehensively a literature search is conducted, the more likely it is that existing research gaps can be identified and research questions can be formulated and addressed precisely. This applies to methods in the field of IS research as well as in every other scientific discipline.

In this paper, we seek to provide an extension of established research methods by utilizing a theoretically well-founded technique, namely LSI. For this purpose, a prototype of TSISQ, our Tool for Semantic Indexing and Similarity Queries, was implemented. It enables researchers to efficiently gain an overview of a specific research field, deepen their knowledge and furthermore, to refine the theoretical foundations of their research. We demonstrated and evaluated the applicability of the tool in a case study. It can be concluded that using our approach can help save valuable time in finding the basic literature in a desired research field. Furthermore, it can help increase the comprehensiveness of a review by identifying sources that would otherwise not have been taken into account.

Following the identified limitations, further research steps are required with regard to our approach. The performance of the tool should be evaluated in a more extensive way: first, the data base used for the index should be extended to cover more conferences and journals in a longer period of time and second, query results should be compared directly to results of keyword based searches in established databases. We assume that the more focused a query is formulated, the more relevant the results become. Hence, future work should also address the establishment of clear guidelines concerning the composition of query inputs. Beyond that, and according to our recommendations, embedding the approach presented in this paper into an iterative-process cycle can be a promising expansion.

References

Arnott, D. and Pervan, G. (2012). Design Science in Decision Support Systems Research: An Assessment using the Hevner, March, Park, and Ram Guidelines. Journal of the AIS, 13 (11), 923-949.

Baker, M.J. (2000). Writing a Literature Review. Marketing Review, 1 (2), 219-247.

Bawden, D. and Robinson, L. (2009). The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies. Journal of Information Science, 35 (2), 180-191.

Bhandari, H., Shimbo, M., Ito, T. and Matsumoto, Y. (2008). Generic text summarization using probabilistic latent semantic indexing. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP), 133-140.

Brand, M. (2006). Fast low-rank modifications of the thin singular value decomposition. Linear Algebra and its Applications, 415 (1), 20-30.

Cui, H., Wen, J.-R., Nie, J-Y. and Ma, W.-Y. (2003). Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering, 15 (4), 829-839.

Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41 (6), 391-407.

Ding, C.H. (1999). A similarity-based probability model for latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 58-65.

Dumais, S.T. (1992). LSI meets TREC: A status report. In Proceedings of the 1st Text REtrieval Conference (TREC-1), 137-152.

Dumais, S.T. (1994). Latent semantic indexing (LSI) and TREC-2. In Proceedings of the 2nd Text REtrieval Conference (TREC-2), 105-116.

Foltz, P.W. and Dumais, S.T. 1992. An analysis of information filtering methods. Communications of the ACM, 35 (12), 51-60.

Forsythe, G.E., Malcolm, M.A. and Moler, C.B. (1977). Computer methods for mathematical computations. Prentice-Hall, Englewood Cliffs, NJ.

Furlan, B., Batanović, V., and Nikolić, B. (2013). Semantic Similarity of Short Texts in Languages with a Deficient Natural Language Processing Support. Decision Support Systems, 55 (3), 710-719.

Gansterer, W.N., Janecek, A.G.K. and Neumayer,R. (2008). Spam filtering based on latent semantic indexing. In Survey of Text Mining II, M. W. Berry and M. Castellanos (eds.), Springer Verlag, London,165-183.

Gee, K.R. (2003). Using latent semantic indexing to filter spam. In Proceedings of the 2003 ACM symposium on Applied computing, 460-464.

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 19-25.

Gordon, M.D. and Dumais, S.T. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science, 49 (8), 674-685.

Hart, C. (1998). Doing a literature review: Releasing the social science research imagination. Sage.

Homayouni, R., Heinrich, K., Wei, L. and Berry, M.W. (2004). Gene Clustering by Latent Semantic Indexing of MEDLINE Abstracts. Bioinformatics, 21 (1), 104-115.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28 (1), 75-105.

Hevner, A.R. (2007). A Three Cycle View of Design Science Research. Scandinavian Journal of Information Systems, 19 (2), 87-92.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, 289-296.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42 (1-2), 177-196.

Kintsch, W. (2010). Modeling Semantic Memory. Mobile Ad-hoc NETworkS (MANETS), S. Jhean-Larose and G. Denhière (eds.).
Korenius, T., Laurikkala, J. and Juhola, M. (2007). On principal component analysis, cosine and Euclidean measures in information retrieval, Information Sciences, 177 (22), 4893-4905.
Kontostathis, A. and Pottenger, W.M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. Information Processing & Management, 42 (1), 56-73.
Kontostathis, A. (2007). Essential dimensions of latent semantic indexing (LSI), In Proceedings of the 40th Annual Hawaii International Conference on System Sciences, 73-73.
Kuechler, W.L. (2007). Business Applications of Unstructured Text. Communications of the ACM 50 (10), 86-93.
LaBrie, R. and St. Louis, R. (2003). Information Retrieval from Knowledge Management Systems: Using Knowledge Hierarchies to Overcome Keyword Limitations. In Proceedings of the 9th Americas Conference on Information Systems, 2552-2563.
Landauer, T.K. and Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104 (2), 211-240.
Lebek, B., Uffen, J. and Breitner, M.H. (2013). Employees' Information Security Awareness and Behavior: A Literature Review. In Proceedings of the 46th Hawaii International Conference on System Sciences, 2978-2987.
Levy, Y. and Ellis, T.J. (2006). Towards a Framework of Literature Review Process in Support of Information Systems Research. In Proceedings of the 2006 Informing Science and IT Education Joint Conference, 171-181.
Liu, Z., Sivaramakrishnan, N. and Chen, Y. (2011). Query Expansion Base on Clustered Results. In Proceedings of the VLDB Endowment, 4 (6), 350-361.
Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 142-150.
March, S.T. and Smith, G.S. (1995). Design and Natural Science Research on Information Technology. Decision Support Systems, 15 (4), 251-266.
Mitra, M., Singhal, A. and Buckley, C. (1998). Improving Automatic Query Expansion. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 206-214.
Okoli, C. and Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. Sprouts: Working Papers on Information Systems, 10 (26).
Offermann, P., Levina, O., Schönherr, M. and Bub, U. (2009). Outline of a Design Science Research Process. In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technologies, Philadelphia.
Park, J. and Lee, J.-N. (2011). The Impact of Information Overload on Decision Quality in the Web 2.0 Environment: A Cognitive-Emotional Dichotomy Perspective. In Proceedings of the 2011 International Conference on Information Resources Management (CONF-IRM 2011), paper 22.
Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management IS, 24 (3), 45-77.
Qiu, Y. and Frei, H.P. 1993. Concept based query expansion. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 160-169.
Řehůřek, R. and Sojka P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, 46-50.
Rickenberg, T.A., Neumann, M., Hohler, B. and Breitner, M.H. (2012). Towards a process-oriented approach to assessing, classifying and visualizing Enterprise Content Management with Document Maps. In Proceedings of the 20th European Conference on Information Systems, Barcelona, Spain.

Salton, G. and McGill, M.J. (1986). Introduction to modern information retrieval. McGraw-Hill, New York, NY.

Salton, G., Wong, A. and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, 18 (11), 613-620.

Shen, D., Chen, Z., Yang, Q., Zeng, H.J., Zhang, B., Lu, Y. and Ma, W.Y. (2004). Web-page classification through summarization. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 242-249.

Sidorova, A., Evangelopoulos, N., Valacich, J.S., and Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. Mis Quarterly, 32 (3), 467-482.

Steinberger, J., & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. In Proceedings of ISIM'04, 93-100.

vom Brocke, J.M., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R. and Cleven, A. (2009). Reconstructing the giant: On the importance of rigor in documenting the literature search process. In Proceedings of the 17th European Conference on Information Systems, Italy, Verona.

Vorhees, E. (1994). Query Expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, 61-69.

Webster, J. and Watson, R.T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26 (2), xiii-xxiii.

Wolfe, M.B.W. and Goldman, S.R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. Behavior Research Methods, Instruments, & Computers, 35 (1), 22-31.

Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W. and Landauer, T.K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. Discourse Processes, 25 (2), 309-336.

Wolfswinkel, J.F., Furtmueller, E. and Wilderom, C.P. (2013). Using grounded theory as a method for rigorously reviewing literature. European Journal of Information Systems, 22 (1), 45-55.

Xu, J. and Croft, W.B. (1996). Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th ACM SIGIR International Conference on Research and Development in Information Retrieval, 4-11.

Yandell, M.D. and Majoros, W.H. (2002). Genomics and Natural Language Processing. Nature Reviews Genetics, 3 (8), 601-610.

Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information Processing & Management , 41 (1), 75-95.

Zelikovitz, S. and Hirsh, H. (2001). Using LSI for Text Classification in the Presence of Background Text. In Proceedings of the 10th ACM International Conference on Information and Knowledge Management, 113-118.

Zha, H. and Simon, H. (1998). A subspace-based model for Latent Semantic Indexing in information retrieval. In Proceedings of the Thirteenth Symposium on the Interface, 315-320.

Zhang, W., Taketoshi, Y. and Xijin, T. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38 (3), 2758-2765.

Appendix

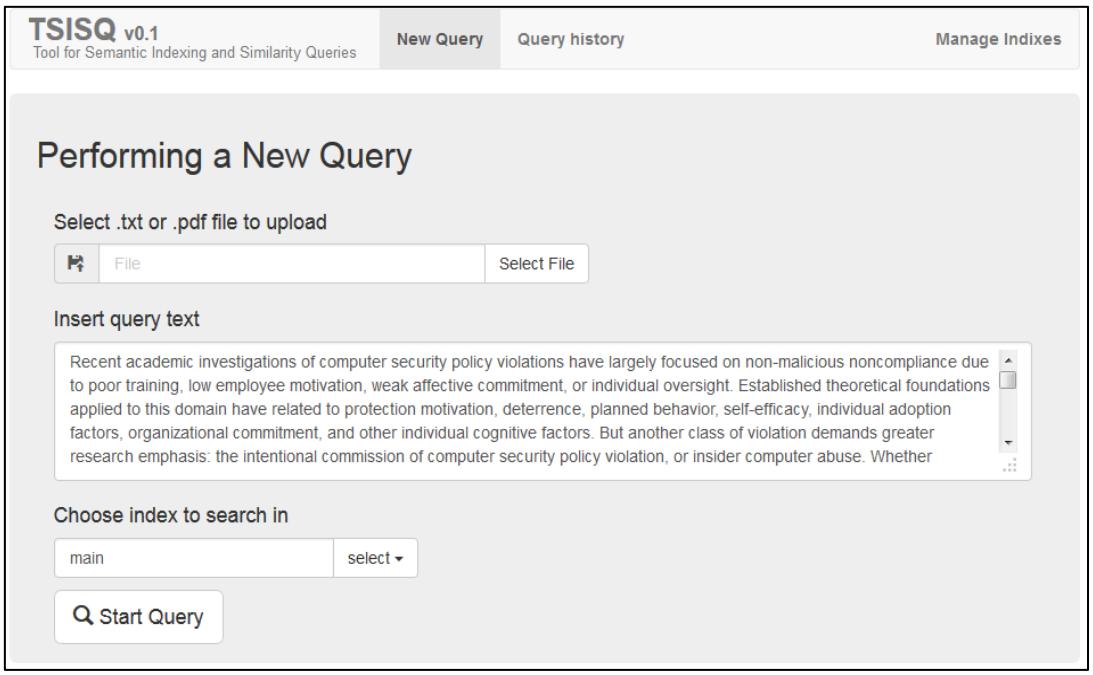

Figure A.1. Web interface to perform a new query

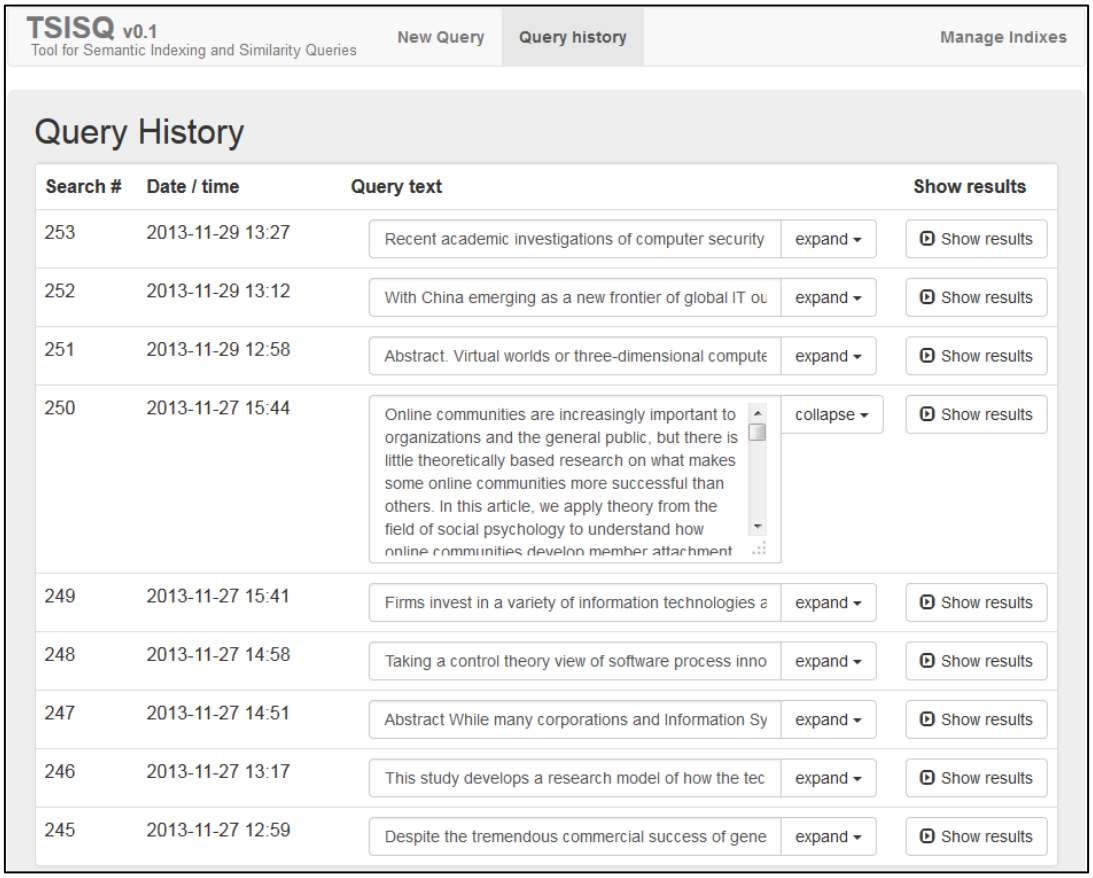

Figure A.2. Web interface to see and select recent queries

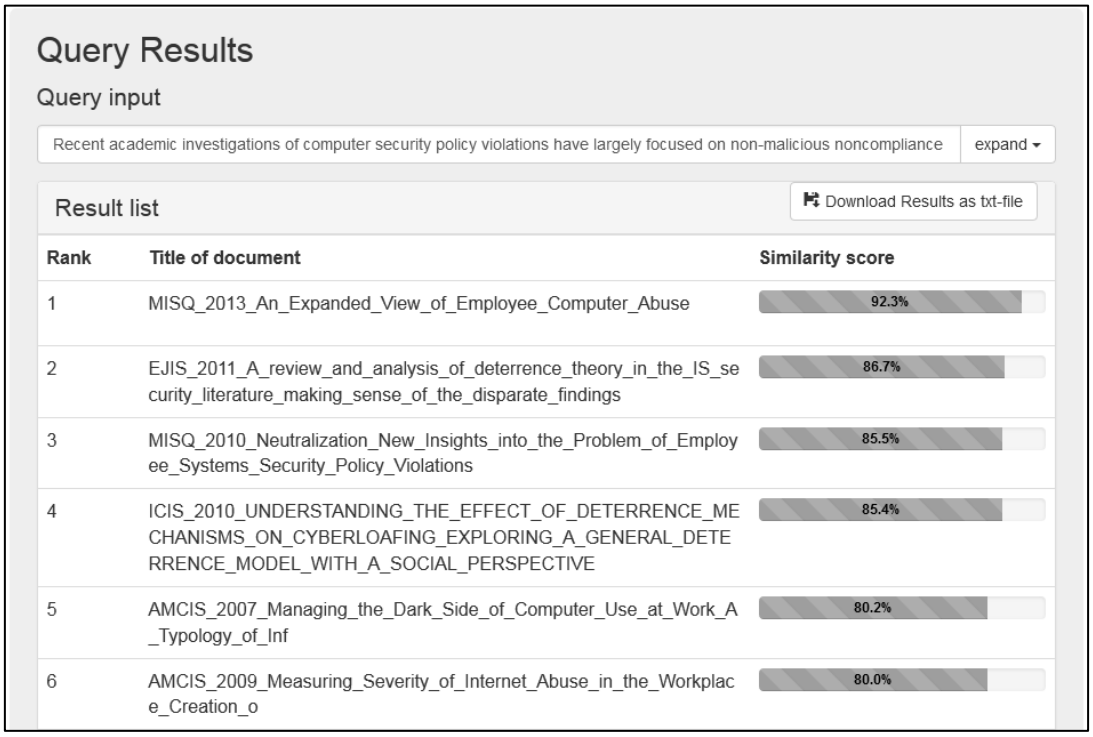

Figure A.3. Visual representation of query results

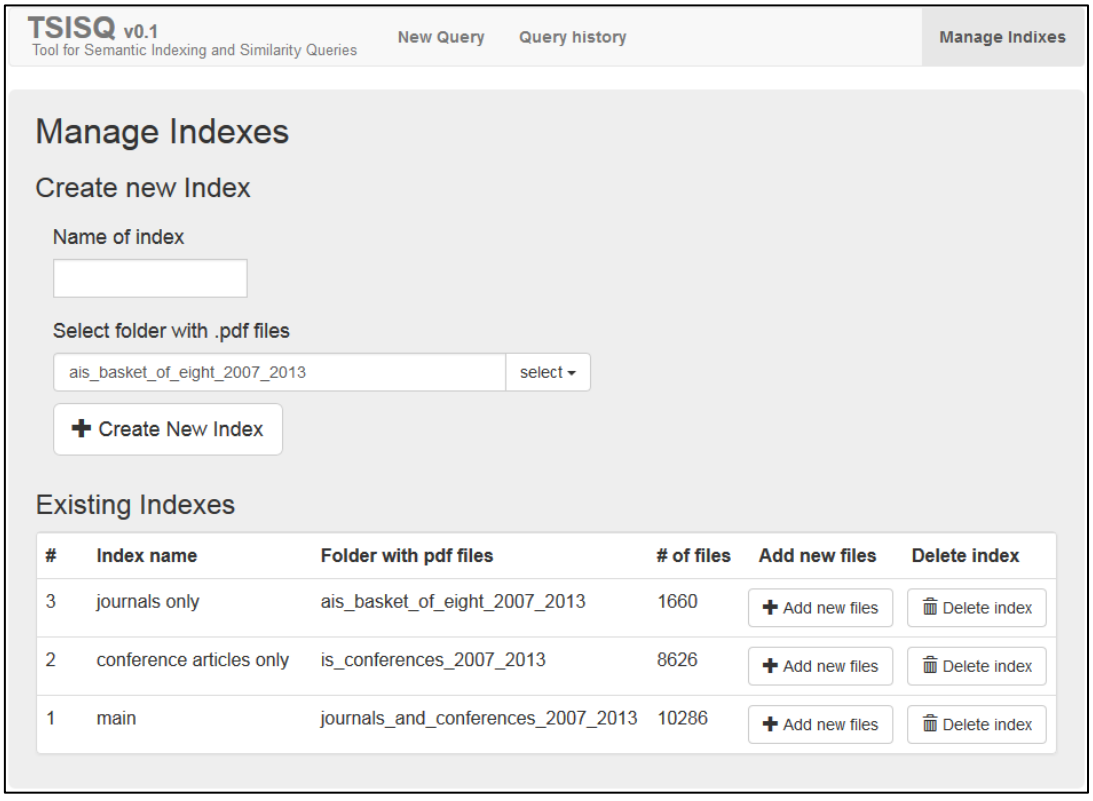

Figure A.4. Web interface to manage indexes