Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing

Completed Research Paper

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

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

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

Abstract

Literature search, as a fundamental and time-consuming step in a literature research process, is part of many established scientific research methods. The facilitated access to scientific resources requires an increasing effort to conduct comprehensive literature reviews. We address the lack of semantic approaches in this context by proposing and evaluating our Tool for Semantic Indexing and Similarity Queries (TSISQ) for the enhancement of established literature review methods. Its applicability is evaluated in different environments and search cases covering realistic applications. Results indicate that TSISQ can increase efficiency by saving valuable time in finding relevant literature in a desired research field, improve the quality of search results, and enhance the comprehensiveness of a review by identifying sources that otherwise would not have been considered. The target audience includes all researchers who need to efficiently gain an overview of a specific research field and refine the theoretical foundations of their research.

Keywords: Literature review, literature search, latent semantic indexing (LSI), information retrieval, similarity search

Introduction

Literature research is a complex and highly important task (Wolfswinkel et al. 2013). As it sets the basis for every research project, the literature research process represents an "essential first step and foundation when undertaking a research project" (Baker 2000), independent from the research domain and the research method that is intended to be followed. 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) observed that specifically in the relatively young field of IS research, there is a lack of a proper theoretical foundation for quality literature reviews. They also state that, in order to strengthen IS as a field of study, effective literature review methods may provide great value to that discipline and furthermore, that well-founded and rigorously conducted literature reviews have a higher chance of getting published. Moreover, taking into account the constant increase in the number of scientific publications worldwide, as well as facilitated access to broad scientific resources triggered by new technologies (Mabe and Amin 2001; Park and Lee 2011) and the resulting complex information environment (Bawden and Robinson 2009; Manwani et al. 2001), an extensive literature review, conducted manually, is a more and more time-consuming task.

Despite their usefulness as compared to a completely manual analysis of a large scientific database, keyword-based approaches have their shortcomings (Blair and Maron 1985; Homayouni et al. 2004; LaBrie and St. Louis 2003) and thus, "[...] are far from ideal" (Dumais et al. 1988). Ambiguity, synonymy, polysemy, the inappropriate use of "stop-words" like "and", "is", "it" or slashes, plurals and parentheses and, ultimately, the indexers' inconsistency when applying subject terms can distort the search results. Hence, keyword searches are likely to cause false-positive or false-negative errors, i.e. potential matches may be missed or mismatches incorporated into the search results (Blair and Maron 1985; Dumais et al. 1988; Hofmann 1999; LaBrie and St. Louis 2003; Salton and McGill 1986; Yandell and Majoros 2002).

In spite of the above-mentioned limitations of this retrieval approach, most existing search engines still rely on term-matching methods only (Cui et al. 2003). Accordingly, we observed that search mechanisms of most of the established repositories for research papers and journal articles relevant to the information systems academic community today, e.g. AISeL, IEEE, JStor, ScienceDirect, Wiley, also seem to be keyword-based. Subsequently, overcoming the aforementioned deficiencies of keyword-based searches is an important challenge in IS research and research in general. To address this set of problems, we aim to evaluate an alternative, more sophisticated approach to finding similarities between texts that fulfills the following requirements: (1) Reliable identification of scientific papers that belong to a specific field of research (containing keywords from the initial query). (2) Identification of semantically similar publications from the same (or even different) field of research (containing synonyms or related terms to the keywords given in the initial query). (3) Overcoming the synonymy/polysemy problem and thus, avoiding false-positive and false-negative errors to a certain extent. (4) Support of informal formulation of search queries, i.e. from keywords, complete sentences, abstracts and entire research papers.

Promising assessments from a number of authors (Corley et al. 2005; Kontostathis and Pottenger 2006; Kuechler 2007; Řehůřek and Sojka 2010; Zhang et al. 2011) 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 learning tasks has been proven already (Deerwester et al. 1990; Dumais et al. 1988; Gordon and Dumais 1998; Kontostathis and Pottenger 2002; Zelikovitz and Hirsh 2001) but it has not yet been applied to our area of interest. We argue that following a semantic approach is likely to increase efficiency and thus, helps to save valuable time in identifying the most important 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 and discuss an alternative approach to the individual researchers' literature research process using LSI and evaluating its applicability to the detection of similarities between texts in a large database of scientific publications.

The core of our research is the "Tool for Semantic Indexing and Similarity Queries" (TSISQ). It is designed to use unstructured texts, e.g. either complete scientific research papers or any kind of natural language, as query input and is able to identify semantically similar texts in a large index of scientific publications. We posit that TSISQ represents a useful addition to several steps of well-established literature review methods (Okoli and Schabram 2010; Levy and Ellis 2006; Webster and Watson 2002) as it can help overcome the limitations of "classic" keyword searches mentioned above while enhancing and facilitating the research process. Furthermore, TSISQ is designed to be used to support every method that implicitly demands an extensive literature review, such as design science research according to Hevner with his claim for research rigor (Hevner et al. 2004) or the rigor cycle (Hevner 2007). The prototype of our tool will be used to evaluate the capabilities of LSI in this context. Hence, we seek to answer the following research question:

RQ: "How can a LSI-based approach increase the efficiency of scientific literature research processes?"

The remainder of this paper is structured as follows: After highlighting the motivation and relevance of the topic and pointing out the contributions and research objective, the theoretical background is presented together with an overview of related work. Next, the research design of this paper is explained in detail. Then, the theoretical concepts of LSI and its mode of operation are described, followed by an illustration of the architecture and implementation of TSISQ. Next, the tool's performance is evaluated within a case study with three search cases and an experimental test case with an index of a controllable size. This is followed by a critical discussion, limitations, as well as theoretical and practical recommendations. Finally, a short conclusion is given and implications for further research are drawn.

Research Background

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 well-established methods for properly conducting a quality literature review exist. Although the respective authors propose different sets of guidelines, it appears to be common sense that it is of particular importance to get (and thus provide) a broad understanding of the pursued research topic. Accordingly, the identification of relevant related literature is an important subtask in every literature review (Wolfswinkel et al. 2013). Amongst others, Webster and Watson (2002) claim that, in order to write an ideal article, relevant prior literature in IS and related areas has to be reviewed, which means an examination of past research is required. Levy and Ellis (2006) propose a systematic data-processing approach consisting of a three-stage framework. The first stage already includes the gathering and screening of "inputs", i.e. the identification and analysis of quality literature in the respective field in order to ensure the validity and reliability of the study and its results. Okoli and Schabram (2010) point out that a literature review has to be systematic in terms of following a specific method and, more importantly in our case, comprehensive in its scope, including all relevant material (see also Fink 2010; Rousseau et al. 2008). The need for comprehensiveness is again underlined in step three, "searching for the literature*,"* in the presented "eight-step-guide to conducting a [scientifically rigorous] systematic literature review."

Taking into consideration the above-mentioned guidelines and proposals, our aim is *not* to introduce an entirely new method for literature reviews, but to facilitate certain steps of the existing, well-established ones by proposing a tool-supported similarity search process. In order to narrow down the scope of this study, it is necessary to provide a brief overview of the considered retrieval approaches to address our underlying challenge of identifying semantic similarities between texts.

Query expansion (QE) is an information retrieval (IR) technique that aims to advance retrieval effectiveness and improves the results of a keyword-based query 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 1993; Xu and Croft 1996). In other words, people who provide information often use different words to describe a concept than the people who search for it. Generally, QE increases the number of relevant results by adding more search terms to the original query (Liu et al. 2011; Mitra et al. 1998; Santos and Riveiro 2011; Xu and Croft 1996). Though this type of retrieval method might help researchers find more relevant documents for a query consisting of very few search terms, it will not perform well if complete research papers are used as query input. We argue that it is not suitable for the underlying problem of this study since the aim of our retrieval activity is not only to increase the quantity, but also and especially the quality of the search results. While the problem of synonymy may be diluted by applying QE to a keyword-based query, the problem of polysemy (ambiguity of a term) remains unsolved (Liu et al. 2011).

Furthermore, 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), which plays an important role in the assessment of document similarity. There are three fundamental aspects to NLP: information extraction, semantics and IR, of which the latter term refers to document-based and query-based retrieval (Yandell and Majoros 2002). LSI, also referred to as latent semantic analysis (LSA), is a related approach which belongs to the field of NLP techniques. LSI is a statistical approach that does not need any explicit knowledge organized by humans to operate. Basically, it uses dimensionality reduction to detect the most prominent dimensions in a document. These dimensions are supposed to correspond to "latent concepts" to represent the meanings of words and documents in the space defined by these concepts (Gabrilovich and Markovitch 2009). Put simply, LSI maps meaning into a semantic space (Kintsch 2010). More detailed information on the theoretical concepts and methods of LSI is provided in the section of the same name. Due to its generality, LSI 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 1999; Landauer and Dumais 1997; Wolfe et al. 1998) and has proven to be effective in advancing average retrieval accuracy (Ding 1999). One application is to match queries to documents in IR applications (Kontostathis 2007), or, in other words, to analyze semantic relationships between a set of documents and included terms. To achieve this, the underlying meaning or concepts behind the words are identified and compared instead of the (key-)words themselves (Farrús and Costa-jussà 2013). Subsequently, LSI offers an entirely different way to use natural language input to find literature related to an initial query (Gordon and Dumais 1998), which leads us to the assumption that LSI may be a helpful approach to support the literature research process.

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; Cao and Ngo 2012; Deerwester et al. 1990; Dumais 1992; Hofmann 1999; Kontostathis and Pottenger 2006; Řehůřek and Sojka 2010). Furthermore, many articles dealing with practical and theoretical applications of LSI in various research domains have been published. Those articles from the recent past are presented in Table 1. In addition to the manual literature research for this paper, our prototype was used for a refinement of the results with the objective of identifying related literature we might have missed. Despite a few publications on LSI that were not taken into consideration in the first iteration, an article from ECIS 2011 with the promising title "A systematic, tool-supported method for conducting literature reviews in information systems" (Bandara et al. 2011) was found. It turns out, however, that the authors neither used LSI nor developed their own tool or IT artifact, but proposed a combination of existing, commercial tools to support literature reviews as described by Levy and Ellis (2006) or Webster and Watson (2002).

Apart from our pilot studies (Koukal et al. 2013, 2014) and to the best of our knowledge, the research gap we seek to address, put precisely, comparing a query formulated in natural language and a large body of published, complete IS research papers using LSI or a related technique, has not been reported in academic literature to date. The work that appears to be closest to our study was published by Sidorova et al. (2008). The authors presented an LSI-based approach to determine the "intellectual core of the information systems discipline" by examining the abstracts of published IS research articles from 1985 through 2006 in three top research journals: MIS Quarterly, Information Systems Research and the Journal of Management Information Systems. Hovorka et al. (2009) refer to the work of Sidorova and her colleagues, analyzing semantic relationships in 24841 abstracts from core business journals to derive possible convergences between IS and other business disciplines. Blake (2010) followed a similar approach, identifying the core topics and themes of data and information quality research. Another technically quite similar attempt was conducted by Homayouni et al. (2004) in the field of bioinformatics. They showed that with LSI, the automatic extraction of functional relationships between genes from abstracts and titles in biomedical literature can be performed with high precision.

However, the text-corpus we seek to analyze is much larger than the one examined by Sidorova et al. (2008) and also larger than the small gene-document collection utilized by Homayouni et al. (2004), due to the fact that it contains full-texts of all publications from the highest ranked IS journals (the "AIS basket of eight") plus the four most important global IS conferences (AMCIS, ECIS, HICSS, ICIS) from 2007 to March 2014. The corpora analyzed by the other aforementioned authors, who followed similar approaches, are also composed of abstracts only. Additionally, none of the authors neither aimed at applying their LSI implementation to identify relevant literature within a literature research process nor evaluated their artifacts in this context.

One of the challenges of our intent is that most of the common algorithms compute very large matrices as a result of a high amount of unique terms and documents 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 Python-based 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 or notebook.

The prior findings and practical applications in NLP and, specifically, the LSI field, as well as likewise the recent technological improvements highlighted above, allow the identified research gap to be addressed.

Table 1. Articles with LSI Applications Identified by the Literature Review

No.	Author	Research topic
1	Gong and Liu (2001)	Automatic text summarization, benchmark with standard methods of IR
2	Gee (2003)	Evaluation of LSI for E-mail spam filtering
3	Wolfe & Goldman (2003)	Prediction of psychological phenomena
4	Homayouni et al. (2004)	Identification of functional relationships between genes from biomedical literature
5	Shen et al. (2004)	Web-page classification through summarization
6	Steinberger and Ježek (2004)	Text summarization, measurement of content similarity between the automatically generated text summary and its original source
7	Yeh et al. (2005)	Automatic text summarization of political articles
8	Kuechler (2007)	Unstructured text acquisition and analysis; Natural language processing, business applications of LSI
9	Bhandari et al. (2008)	Automatic text summarization, topic extraction from documents
10	Gansterer et al. (2008)	E-mail spam filtering
11	Sidorova et al. (2008)	Determination of the "intellectual core of the IS discipline" by examining abstracts of IS research articles
12	Hovorka et al. (2009)	Analysis of semantic relationships in abstracts from core business journals, examining convergences between IS and other business disciplines
13	Blake (2010)	Identification of core topics of data and information quality research
14	Go et al. (2010)	Sentiment analysis, automatic classification of Twitter messages
15	Lee et al. (2010)	Comparison of LSI-related text mining methods
16	Maas et al. (2011)	Automatic sentiment analysis and classification
17	Abate et al. (2013)	Semantic analysis of biomedical literature; integration and extraction of biological information from the web
18	Arijit D. (2013)	SMS based question answering and information retrieval; extraction of meaning from SMS for automatic answering
19	Koukal et al. (2013, 2014)	Enhancement of scientific literature research processes
20	Nugumanova & Bessmertny (2013)	Automatic extraction of word pair collocations from domain texts
21	Shao et al. (2013)	Recovery of traceability information between requirement documents and source codes based on LSI and special features of source codes

Research Design

The proposed enhancement of the scientific literature research process as part of any rigorous research method is only possible if the procedure of this approach itself follows a widely accepted scientific research method. For this purpose, our research was conducted guided by design science research principles in order to address relevance and enhance rigor of our 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). According to Peffers et al. (2008), the actual research design is classified as problem-centered approach and follows their recommended six step guide.

The lack of automated support in combination with a consideration of semantic concepts for text retrieval makes the literature research process slow and time-consuming. The need for a more efficient approach to find relevant literature triggered the development of TSISQ. We initiated our research process by identifying the above-mentioned problem (I). To ensure methodological rigor, we conducted a comprehensive literature review, also by using our tool, within the fields of methods for conducting literature reviews in the IS domain, of information retrieval concepts, and of natural language processing, particularly LSI. Additionally, we conducted a targeted review within the DSR domain. According to the research question, the main objective (II) was to enhance established literature review methods by employing an LSI based approach. With regard to this objective, we focused on the design, demonstration and evaluation of artifacts that can provide an enhancement of the underlying literature research process. Practical and scientific input is used to design and evaluate artifacts in a loop of iterations in the design cycle according to Hevner (2007). In detail, we proceeded as follows: After refining the problem domain and defining specific requirements, the first research artifact, the basic model of TSISQ, was designed (III). It was limited to the central aspects of semantic indexing and similarity queries. This model included only rudimentary parameters and possibilities to control the output. According to a classification of March and Smith (1995) and Hevner (2004) into constructs, models, methods, and instantiations as the result of design-oriented research, the constructed formal models were used to implement an instantiation: a first prototype of TSISQ. This prototype was used for our initial study (see Koukal et al. 2013). For a further development and a detailed elaboration we used an iterative approach to generate 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 a model with an enhanced concept of TSISQ and a corresponding implementation of an extended prototype (see Koukal et al. 2014). The DSR process cycles were then completed by more extensive tests of the artifacts to enable the documentation of research results. According to the classification of research methodologies by Palvia et al. (2006), two case studies in the form of a literature research in the IS domain were conducted in order to demonstrate (IV) the capabilities of the prototype and the underlying model (see Koukal et al. 2013). TSISQ is evaluated (V) in a case study and a laboratory experiment. Within the case study, search results of TSISQ are compared to an established, keyword-based search approach in three different search cases with regard to their quality. The laboratory experiment aims at an evaluation of TSISQ within a controlled environment only consisting of selected articles by identifying not only false-positive search results but also the false-negative error rate. In an iterative step the tool was used to refine our own literature review. Finally, we worked toward publishing our research results (VI).

Enhancing the Literature Research Process with TSISQ

Before trying to enhance the literature research process and thus, established literature review methods, it is vital to understand the relevance and the positioning of a thorough literature review in a target research field. In general, it is of high importance to be aware of what knowledge already exists before initiating any research project (Hart 1999). As illustrated in Figure 1, this not only holds for the conduct of a literature review itself but also for performing a literature review as part of more comprehensive research methods, e.g. within a design science research process according to Hevner (2004).

The several identified guidelines and frameworks for conducting a systematic literature review in the IS field focus on different aspects and thus, highlight an inconsistent number of steps for the processing. However, we derived three core steps from the guidelines. These steps 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. TSISQ 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 (see Figure 1). Besides that, it may reduce efforts of the third stage due to less irrelevant articles to filter out of search results.

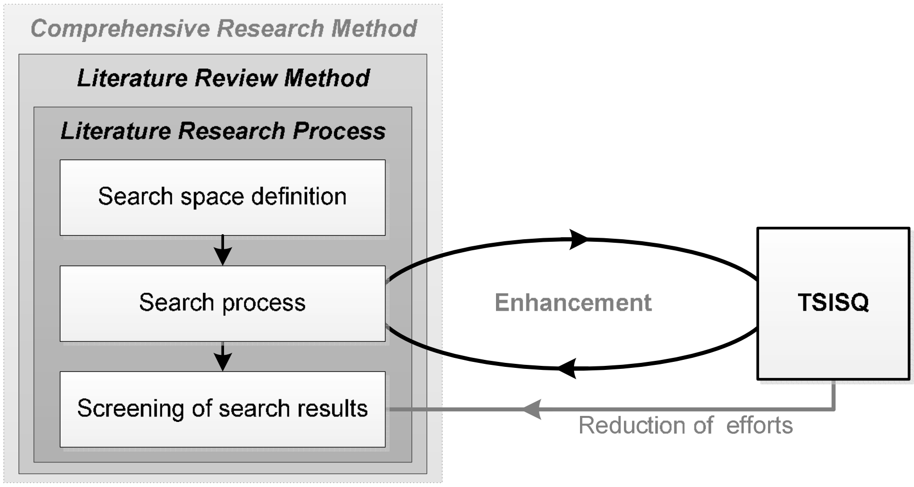

Figure 1. Enhancement of Literature Review Methods with TSISQ

In the following, the theoretical concepts of TSISQ are described in order to ensure an understanding of the underlying model. Additionally, the implementation of TSISQ is described in detail.

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 text document in TSISQ. The content of every considered document D_i in a corpus (of documents) consists of one or more terms T_j. In the VSM, each of these documents (*i*) is defined as a *t*-dimensional vector in Euclidean space, where *t* corresponds to the amount of different terms inside a document. This is combined with a weighting of each term in order to quantify its importance and relevance. Consequently, in TSISQ a document is mathematically described as

$$D_i = (\mathrm{d}_{i1}, \ldots, \mathrm{d}_{ij}, \ldots, \mathrm{d}_{it}) \tag{1}$$

where d_{ij} represents the weight of the *j-th* term. 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) in order to measure the statistical strength of terms in a document. The application of term frequency (TF) and inverse document frequency (IDF) weighting enhances the performance of retrieval and categorization systems (Maas et al. 2011) and consequently of TSISQ by discounting the influence of more common non-stopwords and promoting of occurrence of rare terms. This results in two steps, which are combined by multiplication. Firstly, the normalized frequency of a term in a document is calculated as

$$TF_{ij} = n_{ij} / \sum n_i \tag{2}$$

where n_j is the number of occurrences of a term and Σn_i is, in short, the total amount of all terms in the considered document. Secondly, the rarity of a term within the whole corpus of documents is measured by the inverse document frequency

$$IDF_j = \log(N / df_j) \tag{3}$$

where *N* represents the total amount of documents in the corpus and df_j is the number of documents that contain the term *j*. The multiplication of both measures results in a statistical weight factor for each term of a document

$$TFIDF_{ij} = TF_{ij} \times IDF_j = (n_{ij} / \sum n_i) \times \log(N / df_j) \tag{4}$$

As final result of the conversion of documents into the VSM and the subsequent application of the TFIDF concept we get a term-by-document matrix *X* in which columns contain the weighting of terms for each of the documents in the corpus. This conversion of documents of arbitrary length to fixed-length lists of numbers does not result in a greater reduction of the dimension of content, which may be a problem for processing if the scope of the dataset is very extensive, neither in a greater consideration of the statistical structure of a document or a corpus. Consequently, including the application of the TFIDF weighting concept, the problems 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 of the TFIDF concept, we subsequently apply the LSI method on the converted term-by-document matrix *X*. 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). TSISQ initially decomposes the term-by-document matrix into three other matrices by a process called singular value decomposition (SVD) (Forsythe et al. 1977). The matrices have a special form, where T_O and D_O have orthonormal columns and are matrices of left and right singular vectors and S_O is a diagonal matrix of singular values. Decomposition of matrix *X* into this format is not sufficient, because multiplying the three matrices back together produces the original matrix and there is no gain of information. In the next step, a reduction of the dimension is performed to consider information from higher-order relationships between terms. Figure 2 illustrates truncating of the matrices T_O *x* S_O *x* D_O’ from *m* dimensions to the new matrices *T x S x D’* of *k* dimensions. Multiplying *T x S x D’* produces the best rank-*k* approximation of the original term-by-document matrix *X* (Kontostathis 2007). As the standard procedure of SVD quickly exceeds memory limits, we make use of an incremental SVD processing algorithm by Brand (2006). An illustrative example of the application of VSM, TFIDF and LSI is provided by Sidorova et al. (2008)

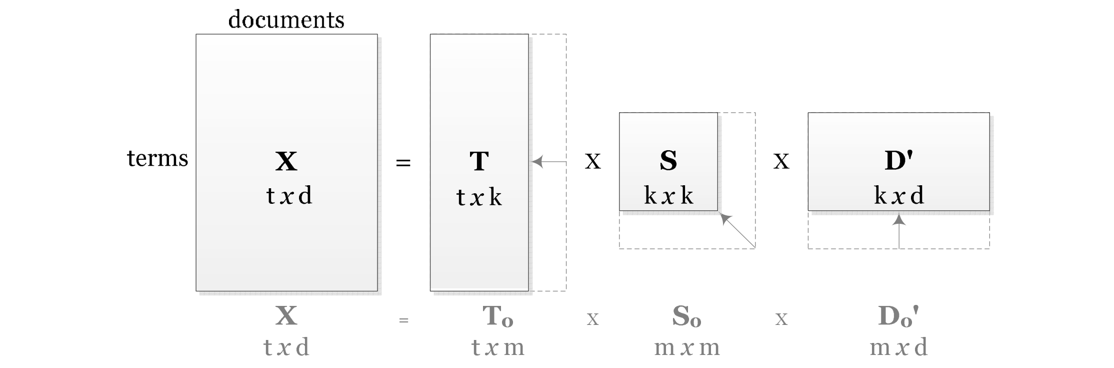

Figure 2. Reduced Singular Value Decomposition of Term-by-Document Matrix

In the next step, the columns of matrix *D’* 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:

$$Q = Q^T T S^{-1} \tag{5}$$

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., 2007). Instead of determining the angle between query and document vectors, the cosine of the angle is calculated. Since the TFIDF values cannot be negative, the angle between two vectors cannot be greater than 90°. Consequently, TSISQ returns cosine values that express the similarity between queries and documents within an interval of [0, 1]. The higher this value is, the higher is the similarity.

Implementation and Architecture of TSISQ

A prototype of TSISQ is implemented in the Python programming language to enable cross-platform use. However, the system has only been used and tested on linux systems, except of the web frontend. The system components, their functions, and the respective data flow are presented in Figure 3 and subsequently described in detail.

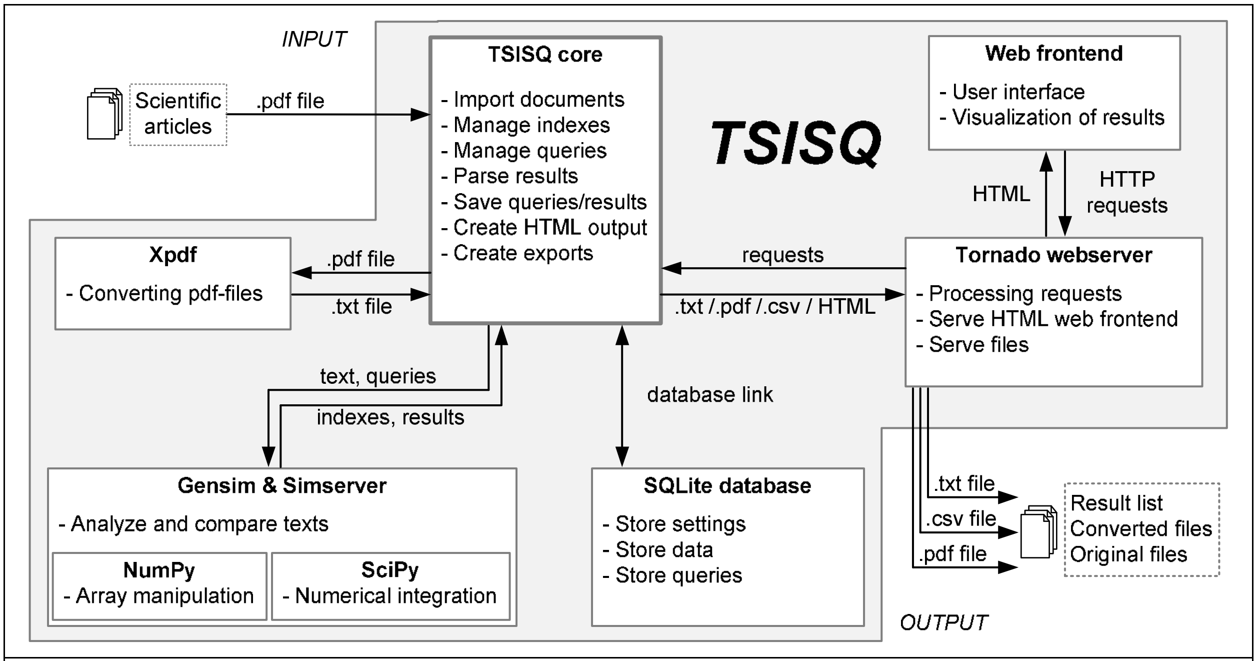

Figure 3. System Architecture of TSISQ

Scientific articles in PDF file format are used as a basic input. 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. According to the gensim documentation, the advantages of the framework 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 lowercases, tokenizes, stems, normalizes and transforms the input texts into Unicode. Further, 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. Although all calculations in TSISQ are memory independent, in memory processing is much faster than the use of a swap file and thus, the amount of RAM determines the size of the matrices which can be calculated with an optimal speed. Given sufficiently large RAM and hard disk, scalability is ensured. The indexing process is computationally and in particular memory expensive. It requires some time for larger collections, e.g. the initial training or the indexing process of 12,300 documents took approximately half an hour each on a system with an Intel Core i7-2640M CPU with 2.80 GHz, 8GB RAM. In contrast, the actual search procedure which uses an already calculated index delivers search results instantly.

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. 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. More information about the integrated tools and frameworks can be found on the respective homepages.

TSISQ and its source code are available for download at: http://www.iwi.uni-hannover.de/TSISQ

Demonstration and Evaluation

In order to evaluate and show the applicability of TSISQ for the literature research process, we performed a comparison of our tool to a keyword-based search engine with the help of domain experts and a laboratory experiment. As it makes sense to start a literature review within the leading journals because any major contribution is likely to be found in them (Webster and Watson 2002), the database we chose as a foundation of our studies was composed according to the following criteria: the eight journals incorporated in our index are the eight leading IS journals in the AIS Senior Scholars' basket of journals, the "AIS basket of eight". Additionally, four of the most important conferences in IS research were added to the index: the International Conference on Information Systems (ICIS), the European Conference on Information Systems (ECIS), the Americas Conference on Information Systems (AMCIS) and the Hawaiian International Conference on System Sciences (HICSS). For more detailed information about the total number of articles and the respective share of each conference or journal, see Table 1.

Table 2. Database of Journal and Conference Articles from 2007 to March 2014

Journal	Total	EJIS	ISJ	ISR	JAIS	JIT	JMIS	JSIS	MISQ
Articles	1685	372	166	323	219	223	280	147	326
Share in %	13.66%	3.02%	1.35%	2.62%	1.78%	1.81%	2.27%	1.19%	2.64%
Conference	**Total**	**AMCIS**	**ECIS**	**HICSS**	**ICIS**				**Total**
Articles	10647	3604	1599	3681	1763				12332
Share in %	86.34%	29.22%	12.97%	29.85%	14.30%				100%

Comparison of TSISQ with a Keyword-based Search Engine

To be able to recommend the use of TSISQ in a practical environment, it is not sufficient to only prove that search results are suitable, but also to verify their quality in comparison to an established scientific search engine. In this case study, AISel was chosen as it is one of the leading databases in IS research. Three different research fields were selected for a direct comparison: (1) Enterprise content management (ECM) as it is a comparatively well-defined research domain within the IS field, (2) IS security and privacy as it is, in contrast, a comparatively vast research field which extends to various domains from IT to e.g. psychology, and (3) structural equation modeling (SEM) to include not only research topics but also a research method which is frequently applied to manifold topics within the IS domain. Subsequently, we prepared a set of keywords for each of the three domains which were used as an input for (a) the search with AISel and (b) the TSISQ keyword search. In addition, we selected suitable natural language inputs for (c) the TSISQ text search. Table 3 shows the different keywords and texts used for the queries with AISel and TSISQ. For the AISel keyword search, the defined keywords were put in quotation marks and an either-or search was conducted for each of the three search cases. In contrast, for the TSISQ keyword search, the keywords were simply used as plain text input.

Table 3. Query Input for Comparison of Search Engines in Different Research Areas

Research Area	AISel and TSISQ keyword search	TSISQ text search
Enterprise content management (ECM)	- Enterprise content management - ECM	Abstract of a literature review by Rickenberg et al. (2012)
IS security and privacy	- IT security - IS security - IT privacy - IS privacy	Description of "IS Security and Privacy" track from ICIS 2014.
Structural equation modeling (SEM)	- Structural equation model - Structural equation modeling - Structural equation modelling - SEM	Abstract of article about evaluation of indicators in SEM by Bollen (2011)

To ensure the comparability of search results from both search engines, the keyword-based search of AISel and TSISQ, only the intersecting set of journals and conferences within the AISel database and our database (AMCIS, ECIS, ICIS, ISJ, JAIS, MISQ) was analyzed and used for further evaluation. The natural language input used for the TSISQ text search was selected as follows: For the ECM domain, the abstract of a literature review was chosen because of its comprehensiveness and the fact that it contains all relevant ECM related terms. In the field of IS security and privacy, the complete description from the 2014 ICIS "IS Security and Privacy" track was used based on the assumption that it contains the most important terms and concepts from the respective field. For the same reason and because it deals not only with an application of SEM but approaches of SEM, the input for the search in the respective domain was an abstract of an article published in one of the leading IS journals, MISQ.

As the relevance of results to the targeted research field is the most important quality criterion for a researcher performing a literature review, it was the main focus of our analysis. To allow a comparison of the results with respect to their quality, a suitable metric which is applicable to all three sets of search results had to be defined. A three-point Likert scale was chosen for the quantification: A ranking of "0" classifies an article as being *irrelevant* to the initial query, while "1" means it is considered *relevant*, and "2" is *highly relevant* to a search topic. To ensure the independence of this classification, it was not performed by the authors themselves. Consequently, domain experts in the respective fields of research were provided with the top results up to a maximum limit of 50 for each search and asked to rank each result on the above mentioned scale. The rankings of the search results were not visible to domain experts to improve objectiveness.

The results of the evaluation for each of the three research fields are presented in Figure 4 (a)-(c). While in the field of ECM (Figure 4a) there is almost no difference in quality of results using TSISQ or AISel, Figure 4b shows a quality improvement, especially in the class of highly relevant literature, in which TSISQ significantly outperforms AISel. Furthermore, it has to be mentioned that the TSISQ output did not contain irrelevant literature at all. In the field of SEM (Figure 4c), the two different TSISQ searches

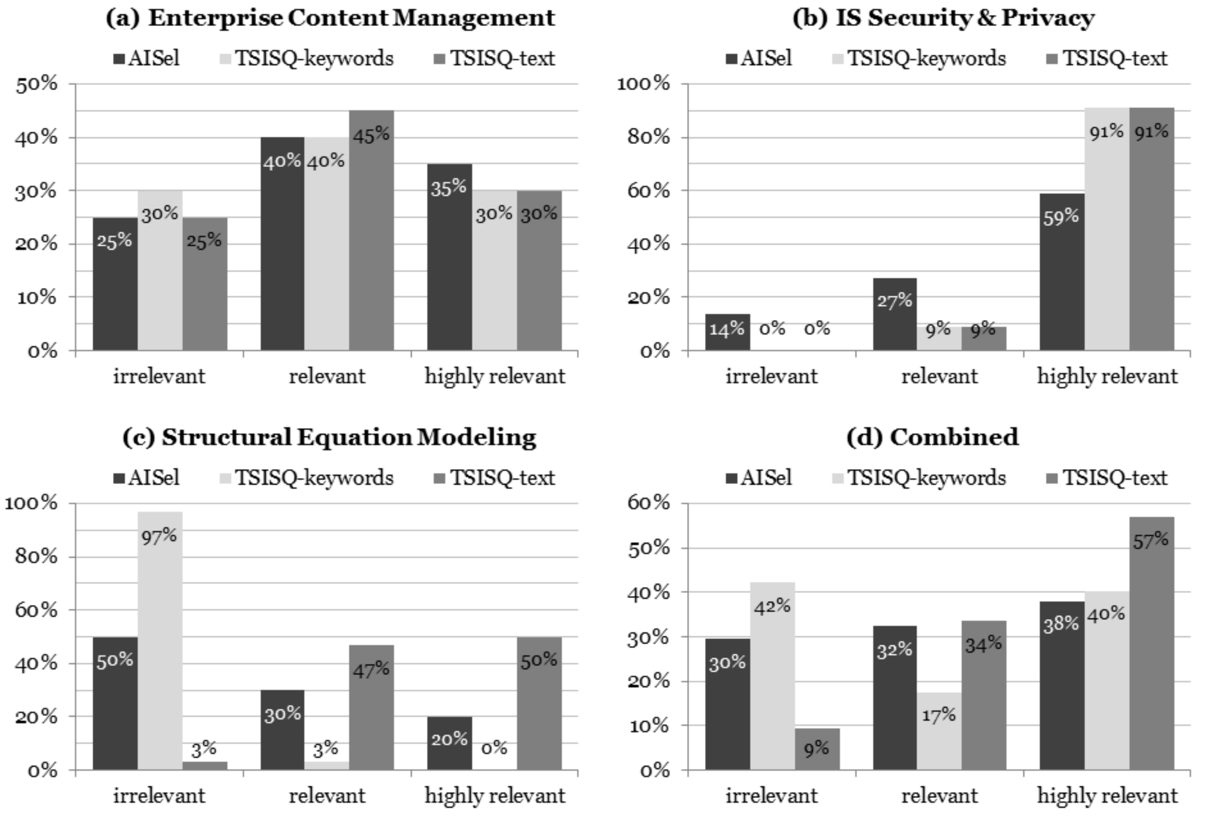

Figure 4a-d. Results of the Evaluation of Search Results

have to be regarded separately: While the keyword search delivered almost exclusively irrelevant articles, the search with natural language input again outperforms AISel regarding relevant and highly relevant literature. Finally, Figure 4(d) aggregates the results of all three search cases. The search results of AISel are distributed among all three classes in approximately equal shares. Though the TSISQ-keyword search delivers a high amount of irrelevant articles, a high number of highly relevant articles is identified as well. In contrast, the use of topic-related, natural language input for the TSISQ-text search delivers only few irrelevant articles, but a significant number of relevant and, especially, highly relevant articles. A comprehensive discussion of the results described above is presented in the section "Discussion, Limitations and Recommendations".

Identification of False-Positive and False-Negative Errors

The intention of this laboratory experiment was to further evaluate TSISQ and underline our theoretical assumptions about the applicability and advantages of LSI for the underlying problem of this study. We highlighted earlier that a LSI-based similarity search can deal with the challenges of synonymy, polysemy and thus, considerably reduce the amount of false-positive and false-negative errors, i.e. finding non-relevant papers or missing relevant ones. While "false-positive hits" are comparatively easy to identify, "false-negative" errors, i.e. thematic mismatches in the upper ranks or outside of our search results, are far more difficult to detect. To achieve this, it is crucial to have a very comprehensive understanding of the contents of all documents compiled in the index, specifically of those that belong to the field of research that is addressed, before analyzing the search results.

As the findings of our pilot study (Koukal et al. 2013) led us to the assumption that the vast majority of the most relevant papers to the query are to be found within the top 25 results, we built a manipulated index composed of only 100 scientific papers chosen from all conferences and journals included in our database. The main purpose of this procedure is to keep the index small enough to receive significant, controllable results in a controlled environment. In order to avoid a selection bias caused by a manual selection of articles by choosing particularly relevant or clearly irrelevant articles regarding a specific topic, we randomly selected 100 articles from our large database as initial step. Subsequently, each article included in this small index was manually screened by a domain expert and classified into two categories. The first category only contained scientific papers that relate to "IS security and privacy" and thus, are likely to be semantically similar. The second category contained all articles that do not belong to this field and deal with various topics from the entire IS domain. To ensure the quality of our experiment's database, we compared the domain expert's categorizations for this experiment with the classifications of articles from domain experts of our case study concerning "IS security and privacy". In the case that an article was found by TSISQ within the case study and the same article is included among the 100 randomly selected articles for this experiment, the classifications as "relevant" or "irrelevant" had to be identical. The classification into two categories resulted in a set of 18 articles relevant to "IS Security and Privacy" and 82 irrelevant articles. As query input, we again chose to use the complete description from the 2014 ICIS "IS Security and Privacy" track.

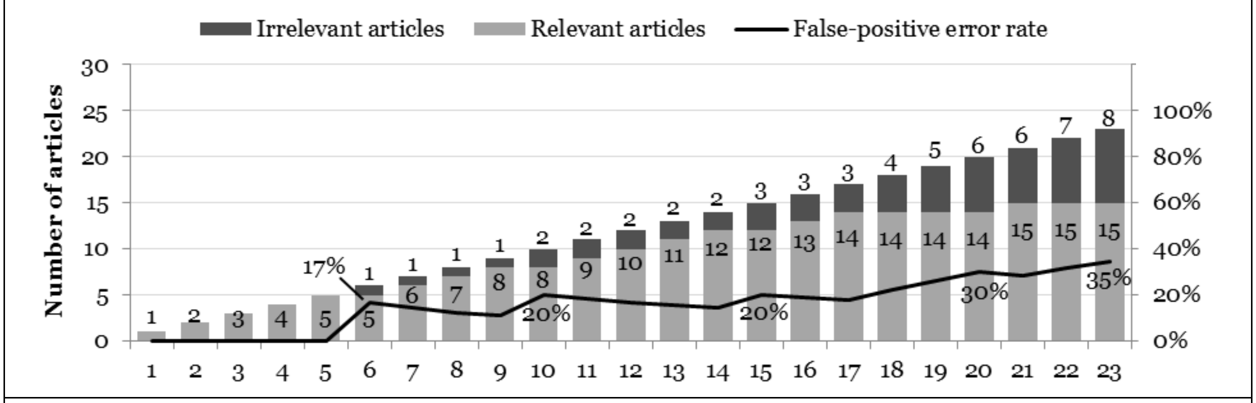

Figure 5. Results of the Laboratory Experiment

In order to avoid getting all 100 articles as search results and thus, having a large number of completely irrelevant articles in the result list, we set a cut-off point at a (very low) cosine value of 0.01. TSISQ delivered an output of 23 results, despite the fact that only 18 articles of 100 belong to the target domain of "IS Security and Privacy". The 23 results contain 15 relevant and 8 irrelevant articles. Thus, 3 of the 18 relevant articles were not identified, which corresponds to a false-negative error rate of 17 percent. In the given environment, the top five search results are most accurate, with a hit rate of 100 percent (see Figure 5). While in the range of the top 17 results the average false-positive error rate is still at an acceptable level at less than 20 percent, this rate increases up to 35 percent when all 23 search results are considered. These results are taken up and discussed in the following section.

Discussion, Limitations and Recommendations

The case study was conducted to gain insight about how well an LSI based approach performs in comparison to an established, keyword-based search engine. The results indicate that the application of TSISQ is very well suited for enhancing scientific literature research processes, especially if natural language text related to the target domain is used as query input. In all of the three search cases (a-c), the TSISQ-text search generates at least the same quality of results regarding the relevance of articles, when relevant and highly relevant articles are considered. In search cases (b) and (c), it significantly outperforms the keyword-based approaches (AISel and TSISQ-keywords). Regarding search case (a) in the ECM domain, the quality of results of all search approaches is almost identical. However, TSISQ-text search performs worse in comparison to the other search cases.

The results have to be analyzed in detail to provide a better understanding of why the individual outcomes show those large differences. As the database used in the three search cases is identical and the setup of the searches is fixed, the remaining and thus, most important and obvious variable factor is the query input. After we had received the evaluation results, we looked retrospectively at all query inputs and identified the following possible reasons: Firstly, the field of ECM can be considered to be narrow compared to SEM or IS security and privacy, for which exist manifold articles containing applications and discussions about their type and conduct in various fields within the IS domain. Secondly, the query input of search case (c) clearly contains more specific content (terms, semantic concepts) regarding the target domain of SEM. Consequently, in search case (c), TSISQ-text delivers considerably better results.

In the light of the fact that SEM is a method and not necessarily a research field itself, and as it is applicable to almost every research domain, the performance of TSISQ-text search in search case (c) is particularly worthy of emphasis. It is to be expected that the research domain analyzed with SEM in an SEM-related article also dominates the semantic concepts extracted by TSISQ. The results of search case (c) do not reflect this aspect as the abstract used as query input for TSISQ-text search describes a meta-research article, dealing with SEM as a method and not its application in a different research domain. Though TSISQ seems to be not feasible for the identification of articles following a certain methodology, our assumption that TSISQ is especially useful for the identification of topic-related texts is confirmed.

Unlike the AISel search engine, TSISQ handles phrases like „structural equation modeling" not as one single search term, but it treats every single word according to the procedure described in the theoretical concepts section. The terms are converted to their representation in the VSM, their weights are determined by the application of the TFIDF concept and then transformed by LSI. Since non alphanumerical characters are removed during this process, it makes no difference for a search with TSISQ if terms are put in quotation marks or not. Though in search case (c) of the TSISQ-keyword search for SEM, the terms "model", "modeling" and "modelling" receive relatively low TFIDF-weights, together they have a slight influence on the vector of a document. These terms can be found together in a high number of documents in the database as the result of a "modeling" process usually is a "model", and even if the term "modeling" is consequently used in the same spelling (which often is not the case), both of the spellings ("modeling" and "modelling") are likely to be found in the references section of articles. As a result, TSISQ identifies articles which contain all of these terms as semantically similar to the query. This implies that, e.g., literature about business process *modeling* or any statistical *model* is a potential but unwanted result. Consequently, as the terms "model" and "modeling" appear in nearly every field of IS research in many different contexts, a keyword search with TSISQ leads to diffuse results.

In contrast, while searching in a database exclusively consisting of IS literature, TSISQ-keyword search will find only relevant articles for the search case of IS security and privacy (b), because (1) the spelling of the terms "security" and "privacy" is unique, (2) the terms "IS" and "IT" are being identified as stop-words and thus, being filtered out as TSISQ does not distinguish between upper and lower case spelling and (3), the aggregation of the terms "security" and "privacy" to a semantic concept is likely to exactly deliver the expected search results. In the case of an expanded database also consisting of not only domain-specific literature, less accurate search results are to be expected. It is mandatory to keep the abovementioned aspects in mind when using solely keywords as input for TSISQ.

The results of the different search cases indicate that the search results of the TSISQ-text search approach will most likely outperform those of a keyword-based search, without the user having to follow specific conventions about the exact formulation of the query input. In AISel and similar keyword-based search engines, if researchers neglect those conventions, search results are likely to be useless. For example, an entire abstract used as query input for AISel will either deliver no results if "AND"-conditions are applied or deliver completely random results if "OR"-conditions are applied. Keeping in mind that our focus is not to replace existing search approaches but to enhance the literature research process by identifying semantically similar literature, we assume that at least one scientific article or text regarding the target domain is known to the researcher who uses TSISQ. Under this assumption, the definition of a *set of suitable keywords* that not only covers an entire research field but the specific target subdomain of it (by including all relevant synonyms or related terms) would require additional time and effort. Given an at least identical or even better quality of the search results (see Figure 4a-d), it can be stated that especially the TSISQ-text search approach, which does not require any transformation of existing texts into a special format for the query, leads to an efficiency increase of the literature research process.

The results of our pilot study (Koukal et al. 2013) illustrate that the chance to achieve good results in the top 25 entries of the output is potentially high. However, if the number of articles dealing with target-related content is unknown due to the large size of the index, no statement about the performance of the search engine can be made. Thus, in the case of a comparatively large amount of adequate articles in the index (e.g. 250), the fact that the search accuracy begins to decrease heavily after 25 results would not be a satisfactory result. In contrast, if the amount of adequate articles in the same index is relatively low, e.g. 20, detecting 16 of them within the first 25 results would be a remarkably good search accuracy.

While the search accuracy and quality of results in comparison to the established, keyword-based search engine of AISel was evaluated in the case study discussed in this paper and in our pilot study, the laboratory experiment allows an additional quantification of the false-positive and false-negative error rate. It also addresses the question as to whether the promising results presented above simply arise from an unexpectedly large number of ECM-, IS security and privacy- or SEM-related publications in the index or from the good performance of TSISQ. In this controlled environment, the average search accuracy is at least 65% within the top 23 results. However, within the first 17 results the average search accuracy is never lower than 80%. Consequently, TSISQ performs slightly worse than in our pilot study (75%) but still underlines the overall remarkable results of the case study. The low amount of relevant articles in the experiment's database results in large variances of the average search accuracy when a relevant article is found at a better rank in the result list. In other words, one additional hit or a slightly different cosine value of an article can lead to very different results, in particular within the upper ranks of the result list. Though the outcomes of the laboratory experiment cannot be generalized due to their sensitivity against slight changes, in combination with the results of the case study and our pilot study, they confirm the suitability of LSI-based approach for supporting the literature research process.

A general aspect worth mentioning is the observation that the amount of text used for the query has a significant influence on the similarity scores according to which the results are being ranked in the search results. The larger the amount of the query input is, the smaller are the differences of similarity scores from one rank in the result list to its neighbors. This can be attributed to the fact that, e.g., a track description or a complete research paper dealing with a certain topic contains a larger and, more importantly, more general set of terms and semantic concepts, as it usually includes one or more less focused sections to depict related work and define the boundaries of research. Semantically, more than just one specific topic is treated, which may lead to more diffuse results. Abstracts, however, are comparatively short and focused on the main research area of the associated article, so they contain a more specific set of terms and semantic concepts. A composition of suitable keywords is the highest level of such an aggregation, which is especially reflected in the composition of the results of search case (b) with the TSISQ-keyword approach. Nevertheless, the input has to be extensive and coherent enough, so semantic concepts can be derived by the LSI algorithm, which is why the TSISQ-keyword approach did not perform very well in search case (c). Concluding, it can be stated that the more focused and consciously a query is formulated according to the conventions presented above, the more specific and relevant the results become.

Due to the large number of publications dealing with LSI in various application contexts and the related algorithms in detail, there is a lot of positive feedback but also some limitations. Lee et al. (2010) and Bhandari et al. (2008) state that LSI is suitable for addressing the issue of synonymy but its performance in solving the problems with polysemy is limited because of the orthogonal characteristic of factors in the term-by-document-matrix. Kintsch (2010) criticizes the simplifying assumption that the vector that is calculated to represent the texts is the centroid of its word vectors. He gives an example showing that the sum of the vectors is identical for the two phrases “the lion killed the deer” and “the deer killed the lion”. The significant difference in meaning of those two sentences would be missed by LSI. However, it is highlighted that for texts of a minimum length of 100 to 150 words (e.g. abstracts), LSI provides a surprisingly good approximation and useful results. This aspect is also illustrated by the results of our case study (see Figure 4), where in contrast to a few keywords, more comprehensive, natural language inputs (abstracts or track descriptions as illustrated in Table 3) lead to comparably good and useful results.

Like in many research projects, the database is a critical spot for the examination of results. Our database consists of about 12300 articles and thus, covers only a small part of the IS literature and even less of the scientific literature in general. However, in this study, 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 of about seven years could also be insufficient. Nevertheless, to evaluate the feasibility of TSISQ for conducting the proposed literature research process, the specified time frame meets our requirements and represents an important step towards a more extensive examination. Though our laboratory experiment is an initial step that aims at providing information not limited to precision or relevance, the application of a statistical approach to explicitly determine the recall or sensitivity by employing an F-measure is missing. More extensive tests including searches in various research domains will have to be conducted to confirm all prior findings. To date, our tool and thus, the whole approach to support the literature review process is only applicable if all textual content is available in English. This arises from the fact that a dictionary of stop-words is needed in the semantic indexing-layer of TSISQ. Since it takes a lot of time and effort to create such NLP resources, those dictionaries still do not exist for many languages (Furlan et al. 2013). However, because most of the top IS literature is still published in English, this aspect does not have any negative influence on the outcomes of our study.

Nevertheless, upon examination of the outcomes of all search cases and the laboratory experiment we can make recommendations for the use of TSISQ to support the scientific literature research process as crucial part of any scientific research method. As LSI is a principle used for the identification of semantic similarities and thus, for a search based on semantic concepts, we recommend to choose domain-specific texts in natural language as query input, not only a few keywords. Based on our finding that the results are thematically more specific when using abstracts as input, while they tend to be more diffuse when using track descriptions or e.g. entire research papers, we propose an iterative process cycle to achieve optimal results when using TSISQ for literature searches: if available, the first query should be conducted using a complete research article, closely related to the desired topic 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 creating a collection of concise text modules, phrases, or sections found in the identified selection of literature with regard to the conventions highlighted in this paper. As final step, the resulting collection can be utilized as query input. This process can be repeated until no new relevant literature is discovered in the academic body of knowledge. This iterative approach was successfully applied to refine the theoretical foundation of our research and to complete our review of existing literature in all relevant research domains and topics.

Conclusion and Further Research

In this paper, a tool-supported approach to enhancing and advancing the scientific literature research process as an essential component of existing literature review methods is presented and evaluated. The complex and highly important subtask of literature search as a fundamental step of every literature research process 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 and questions are precisely addressed. This applies to methods in the field of IS research as well as in every other scientific discipline.

With regard to our research question, we seek to provide an enhancement of established research methods by utilizing a theoretically well-founded, LSI-based approach. For this purpose, 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 evaluate the applicability of TSISQ in a case study consisting of three different search cases and a laboratory experiment. The results are discussed, limitations are identified, and recommendations for its usage are drawn. Against the backdrop of the overall purpose of TSISQ to increase the efficiency of scientific literature research processes, it can be concluded that using our approach (a) can help save valuable time in finding the relevant literature in a desired research field and (b), if time is not a scarce factor, it can help increasing the comprehensiveness of a review by identifying sources that would otherwise not have been taken into account. Consequently, although human cognitive abilities are still indispensable, TSISQ is a useful complement to the established search engines used in the scientific literature research process and can increase its efficiency.

Following the identified limitations, further research steps are required with regard to our approach. The database used for the index should be extended to cover more conferences and journals in a longer period of time. In order to improve the validity of evaluation results, future work should include more extensive and sophisticated tests and assessments of more domain experts. To enable statements about the accuracy of our tests, a statistical analysis that is widely used in the information retrieval domain could be employed. A calculation of the F-measure would not only consider the precision but also the recall of a test. The additional information about recall or sensitivity and thus, the fraction of relevant instances that are retrieved, could provide better insights about the general potential to use TSISQ for literature searches. A deeper analysis concerning the effects on the output, when query inputs for the same target domain are slightly adjusted, similar to a sensitivity analysis, could provide valuable insights as well. Such a controlled adjustment of individual search terms allows a composition of more aggregated and focused queries. Future work should also address the establishment of clear guidelines concerning the composition of query inputs. Measuring the effects on the search results for SEM in search case (c) when e.g. the terms “models”, “modeling”, and “modelling” are left out one at a time could help to provide such guidelines on how to formulate a query. Based on the previously mentioned steps and according to our recommendations, embedding the approach presented in this paper into a structured and well-designed iterative-process cycle can be a promising additional expansion. Finally, the system could be implemented as a web-based solution that serves links to scientific databases and thus, extends the service for researchers while leaving the full articles behind the scientific databases’ paywalls.

References

Abate, F., Ficarra, E., Acquaviva, and A., Macii, E. 2013. “Improving Latent Semantic Analysis of Biomedical Literature Integrating UMLS Metathesaurus and Biomedical Pathways Databases,” in *Biomedical Engineering Systems and Technologies*, Fred, A., Filipe, J., Gamboa, H. (eds.), pp. 173-187.

Arijit D. 2013. “SMS Based FAQ Retrieval Using Latent Semantic Indexing,” in *Multilingual Information Access in South Asian Languages,* pp. 100-103.

Baker, M. J. 2000. “Writing a Literature Review,” *Marketing Review* (1:2), pp. 219-247.

Bandara, W., Miskon, S., and Fielt, E. 2011. “A systematic, tool-supported method for conducting literature reviews in information systems,” in *Proceedings of the 19th European Conference on Information Systems*.

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

Bhandari, H., Shimbo, M., Ito, T., and Matsumoto, Y. 2008. "Generic Text Summarization Using Probabilistic Latent Semantic Indexing," in *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 133-140.

Blair, D. C. and Maron, M. E. 1985. "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM*, (28:3), pp. 289-299.

Blake, R. 2010. "Identifying the core topics and themes of data and information quality research," in *Proceedings of the 16th Americas Conference on Information Systems.*

Bollen, K. A. 2011. "Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models," *MIS Quarterly* (35: 2), pp.359-372.

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

Cao, T. H., and Ngo, V. M. 2012. "Semantic Search by Latent Ontological Features," *New Generation Computing* (30:1), pp. 53-71.

Corley, C. and Mihalcea, R. 2005. "Measuring the semantic similarity of texts," in *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 13-18.

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), pp. 829-839.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990. "Indexing by latent semantic analysis," in *Journal of the American Society for Information Science* (41:6), pp. 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*, pp.58-65.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. 1988. "Using latent semantic analysis to improve access to textual information," in *Proceedings of the 6th SIGCHI conference on Human factors in computing systems*, pp. 281-285.

Dumais, S. T. 1992. "LSI meets TREC: A status report," in *Proceedings of the 1st Text REtrieval Conference*, pp. 137-152.

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

Farrus M. and Costa-jussà, M. R. 2013. "Automatic Evaluation for E-Learning Using Latent Semantic Analysis: A Use Case," in *The International Review of Research in Open and Distance Learning* (14:1), pp. 239-254.

Fink, A. 2010. *Conducting Research Literature Reviews: From the Internet to Paper (3rd edition)*, Thousand Oaks, California: Sage Publications.

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

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

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), pp. 710-719.*

Gabrilovich, E. and Markovitch, S. 2009. "Wikipedia-based Semantic Interpretation for Natural Language Processing," *Journal of Artificial Intelligence Research* (34:2), pp. 443-498.

Gansterer, W. N., Janecek, A. G., and Neumayer, R. 2008. "Spam filtering based on latent semantic indexing," in *Survey of Text Mining II*, pp. 165-183. London: Springer.

Gee, K. R. 2003. "Using latent semantic indexing to filter spam," in *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 460-464.

Go, A., Bhayani, R., and Huang, L. 2010. "Exploiting the unique characteristics of tweets for sentiment analysis," Technical Report, Stanford University.

Gong, Y. and 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,* pp. 19-25.

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

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

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

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

Hofmann, T. 1999. "Probabilistic latent semantic analysis," in *Proceedings of the 15th conference on Uncertainty in artificial intelligence*, pp. 289-296.

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

Hovorka, D. S., Larsen, K. and Monarchi, D. 2009. "Conceptual convergences: Positioning information systems among the business disciplines," in *Proceedings of the 17th European Conference on Information Systems*.

Kintsch, W. 2010. "Modeling Semantic Memory," in *Mobile Ad-hoc NETworkS (MANETS)*, S. Jhean-Larose and G. Denhière (eds.).

Kontostathis, A. 2007. "Essential dimensions of latent semantic indexing (LSI)," in *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pp. 73-73.

Kontostathis, A., and Pottenger, W.M. 2002. *A mathematical view of latent semantic indexing: Tracing term co-occurrences*, Bethlehem, USA: LeHigh Technical Reports.

Kontostathis, A., and Pottenger, W. M. 2006. "A framework for understanding Latent Semantic Indexing (LSI) performance," *Information Processing & Management* (42:1), pp. 56-73.

Korenius, T., Laurikkala, J. and Juhola, M. 2007. "On principal component analysis, cosine and Euclidean measures in information retrieval," *Information Sciences* (177:22), pp. 4893-4905.

Koukal, A., Gleue, C., and Breitner, M. H. 2013. "Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing," in *Lecture Notes in Informatics Proceedings*, pp. 1937-1942.

Koukal, A., Gleue, C., and Breitner, M. H. 2014. "Enhancing Literature Review Methods – Towards More Efficient Literature Research with Latent Semantic Indexing," in *Proceedings of the 22nd European Conference on Information Systems*.

Kuechler, W. L. 2007. "Business Applications of Unstructured Text," *Communications of the ACM* (50:10), pp. 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,* pp. 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), pp. 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*, pp. 2978-2987.

Lee, S., Baker, J., Song, J., and Wetherbe, J. C. 2010. "An Empirical Comparison of Four Text Mining Methods," in *Proceedings of the 43rd Hawaii International Conference on System Sciences*.

Levy, Y. and Ellis, T. J. 2006. "A systems approach to conduct an effective literature review in support of information systems research," Informing Science (9), pp. 181-212.

Liu, Z., Sivaramakrishnan, N. and Chen, Y. 2011. "Query Expansion Base on Clustered Results," in *Proceedings of the VLDB Endowment* (4:6)*,* pp. 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*, pp. 142-150.

Mabe, M. and Amin, M. 2001. "Growth dynamics of scholarly and scientific journals," *Scientometrics* (51:1), pp. 147-162.

Manwani, S., Bech, H. and Dahlhoff, J. 2001. "Managing information overload: Is technology the answer?," in *Proceedings of the 7th Americas Conference on Information Systems*, pp. 34-38.

March, S. T., and Smith, G. S. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 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*, pp. 206-214.

Nugumanova, A., and Bessmertny, I. 2013. “Applying the Latent Semantic Analysis to the Issue of Automatic Extraction of Collocations from the Domain Texts,” in *Knowledge Engineering and the Semantic Web*, pp. 92-101.

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.

Okoli, C. and Schabram, K. 2010. “A Guide to Conducting a Systematic Literature Review of Information Systems Research,” in *Sprouts: Working Papers on Information Systems* (10:26).

Palvia, P., Midha, V. and Pinjani, P. 2006. “Research Models in Information Systems,” in *Communications of the Association for Information Systems* (17:1), pp. 1042-1066.

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,* paper 22.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. 2008. “A design science research methodology for information systems research,” *Journal of Management Information Systems* (24:3), pp. 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*, pp. 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*, pp. 46-50.

Rickenberg, T. A., Neumann, M., Hohler, B., and Breitner, M. H. 2012. „Enterprise Content Management - A Literature Review,” in *Proceedings of the 18th Americas Conference on Information Systems*, Paper 10.

Rousseau, D.M., Manning, J. and Denyer, D. 2008. “Evidence in Management and Organizational Science: Assembling the Field’s Full Weight of Scientific Knowledge Through Syntheses,” *The Academy of Management Annals* (2:1), pp. 475-515.

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

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

Santos, J. C. A. and Ribeiro, M. F. D. S. B. 2011. “Improving search engine Query Expansion techniques with ILP,” in *Proceedings of the 21st International Conference on Inductive Logic Programming (ILP).*

Shao, J., Wu, W. and Geng, P. 2013. “An Improved Approach to the Recovery of Traceability Links between Requirement Documents and Source Codes Based on Latent Semantic Indexing,” in *Computational Science and Its Applications (ICCSA 2013)*, pp. 547-557.

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*, pp. 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), pp. 467-482.

Steinberger, J. and Jezek, K. 2004. “Using latent semantic analysis in text summarization and summary evaluation,” in *Proceedings of ISIM’04*, pp. 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.*

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*, pp. 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), pp. xiii-xxiii.

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), pp. 309-336.

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), pp. 22-31.

Wolfswinkel, J. F., Furtmueller, E. and Wilderom, C. P. M. 2013. "Using grounded theory as a method for rigorously reviewing literature,“ *European Journal of Information Systems* (22:1)*,* pp. 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*, pp. 4-11.

Yandell, M. D. and Majoros, W. H. 2002. "Genomics and Natural Language Processing," in *Nature Reviews Genetics* (3:8), pp. 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), pp. 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*, pp. 113-118.

Zha, H. and Simon, H. 1998. "A subspace-based model for Latent Semantic Indexing in information retrieval," in *Proceedings of the 13th Symposium on the Interface*, pp. 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), pp. 2758-2765.