Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing

André Koukal, Christoph Gleue, Michael H. Breitner

Institut für Wirtschaftsinformatik
Leibniz Universität Hannover
Königsworther Platz 1
30167 Hannover
{koukal, gleue, breitner}@iwi-uni-hannover.de

Abstract: Literature search as a fundamental, complex and time-consuming step in a literature research process is part of many established scientific methods. 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 the literature research process with a prototype of our Tool for Semantic Indexing and Similarity Queries (TSISQ), which is based on latent semantic indexing (LSI). Its applicability is evaluated in two cases. Results indicate that our approach can help to save valuable time discovering relevant literature in a desired research field or to increase the comprehensiveness of a review by identifying sources that otherwise would not have been considered. The target audience for our findings includes researchers who need to efficiently gain an overview of a research field, deepen their knowledge and refine the theoretical foundations of their research.

1 Introduction

The literature research process represents an “essential first step and foundation when undertaking a research project” [VSN09]. 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 and the resulting complex information environment, 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. Ambiguity, synonymy, polysemy, the inappropriate use of “stop-words”, plurals, parentheses and, ultimately, the indexers’ inconsistency when applying subject terms can distort the query results. Latent semantic indexing (LSI) might provide a solution to these problems and is likely to outperform established lexical matching retrieval methods, increase efficiency and save valuable time in identifying literature in a designated research field. Accordingly, the objective of this paper is to introduce an alternative approach to the individual researchers’ literature search process using LSI.
2 Research Background and Research Design

Semantic similarity is “[…] a concept by which a metric is given to groups of terms or documents based on the similitude of their meanings” [FBN02]. LSI is a semantic approach which belongs to the field of natural language processing (NLP) techniques and does not need any explicit knowledge organized by humans to operate [DDF01]. There have been many publications about LSI and its mode of operation, evaluating its performance, theoretical approaches towards understanding LSI in detail and studies about optimizing the algorithm. Although LSI has a wide range of possible applications, only few publications about practical use cases exist. It is being used in the context of e-mail spam filtering, prediction of psychological phenomena, text mining and automatic text summarization, outperforming keyword-based approaches in this field as well. LSI is also used by Google in their category discovery algorithm and in the automated scoring of student essay examinations [Kuo03]. The work that appears to be closest to our study was published by [SEVO08]. However, the text-corpus we seek to analyze is much larger than the one examined by them, as it not just contains abstracts but full-texts of all publications from the highest ranked IS journals plus the four most important global IS conferences from 2007 to now. The prior findings and practical applications in the LSI field allow us to address the research gap, namely, to compare a query formulated in natural language and a large body of published, complete IS research papers using LSI or a related technique, which has not been reported in academic literature to date.

Our research is conducted using design science research (DSR) principles according to [PTR05] in order to address relevance and enhance rigor of the research process and its results. Artifacts like the basic model of TSISQ and the instantiation of our TSISQ prototype are designed, demonstrated and evaluated following the DSR guidelines. The research design is classified as problem centered approach [PTR05].

3 Enhancing the Literature Research Process with TSISQ

In academic literature, several guides and frameworks for conducting a systematic literature review in the IS field exist. In summary, three core principles are identified and should be addressed in every literature research process (see Figure 1).

![Diagram](image)

Figure 1. Enhancement of an Exemplary Literature Research Process with TSISQ

The first stage is the definition of the search space, e.g. the selection of a specific scientific database. The search process represents the second stage, looking to identify papers that possibly suit to the author’s needs. The third stage is the screening of the
identified papers 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.

The TSISQ prototype is implemented in the Python programming language to allow cross-platform use. 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 concept of the tool provides a structuring into three layers. The content database preparation-layer processes all journal articles and conference papers, that is, files in PDF format. The file names are first standardized by removing all special characters and whitespaces to facilitate the automated conversion into plain text in the second step, using the command-line toolset Xpdf. The semantic indexing-layer implements the detection of semantic concepts. The first step, however, is a preprocessing of the content of each document. Stop-words and characters without any semantic significance are removed. The software framework gensim is the used for indexing, applying the methods of the vector space model (VSM), the term-frequency-inversed-document-frequency (TFIDF) weighting scheme and singular value decomposition (SVD) [RZ01]. As a result of the indexing process, an index corpus file for further processing is created. The third layer is responsible for the similarity search query process. This layer requires an input of plain text that is preprocessed in the same way as documents of the index corpus. This is followed by the conversion of content into the VSM and its transformation into the same space as the document vectors. The final step is the comparison of the query and the index corpus in order to create the query output. The latter is represented as a list of titles of the most similar research papers, ranked by their respective similarity score.

4 Evaluation, Discussion and Further Research

In order to demonstrate and evaluate the constructed IT artifacts in the DSR process, two case studies were performed. The underlying data base comprises over 6 years of publications from leading IS conferences and journals (see Table 1). The six year period could be insufficient if a comprehensive literature review is to be conducted. However, it allows a first evaluation of the feasibility of our prototype for conducting the proposed literature research process in a controlled environment.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Total</th>
<th>EJIS</th>
<th>ISR</th>
<th>JAIS</th>
<th>JIS</th>
<th>JIT</th>
<th>JMIS</th>
<th>JSIS</th>
<th>MISQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>1660</td>
<td>269</td>
<td>263</td>
<td>189</td>
<td>140</td>
<td>157</td>
<td>248</td>
<td>125</td>
<td>269</td>
</tr>
<tr>
<td>Share in %</td>
<td>16.1%</td>
<td>2.6%</td>
<td>2.6%</td>
<td>1.8%</td>
<td>1.4%</td>
<td>1.5%</td>
<td>2.4%</td>
<td>1.1%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conference</th>
<th>Total</th>
<th>AMCIS</th>
<th>ECIS</th>
<th>HICSS</th>
<th>ICIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>8626</td>
<td>2806</td>
<td>1342</td>
<td>3039</td>
<td>1439</td>
</tr>
<tr>
<td>Share in %</td>
<td>83.9%</td>
<td>27.3%</td>
<td>13.1%</td>
<td>29.6%</td>
<td>14.0%</td>
</tr>
</tbody>
</table>

Table 1. Database of Journal and Conference Articles from 2007 to Present

The design of the first case study is inspired by a hypothetical use case in which a short text with aggregated contents can be used for topic-related literature research. Since
abstracts represent such texts, they are an ideal query input. The research field we selected for this purpose is ECM, as it is a comparatively well-defined domain with a manageable amount of established literature incorporated into our index. The query input that was utilized is the abstract of an ECM literature review by [RNH07]. The query output was reviewed by the authors and ECM domain experts, who ranked the top 40 results on a three-point scale: A ranking of “1” classifies the article as being irrelevant, while “2” means it is considered relevant, and “3” is highly relevant to the topic.

The results presented in Figure 2 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 (average hit rate). For the first 10 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. Accordingly, the articles thus identified serve as a solid foundation for further manual processing, e.g. forward and backward search [VSN09]. One limitation concerning the evaluation of the results was noted by the domain experts during our tests. It occurred that articles located within the higher ranks of the result list were subconsciously rated better than those located within the lower ranks. This probably leads to a bias in the evaluation results of this first examination. Moreover, the design of this use case has another mentionable drawback. 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. This first 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. Hence, a more controllable environment is needed, which is addressed by the second examination.

For the second case study, a theory-based literature review [LUB04] which analyzed 113 papers in the domain of IT security and privacy was taken as query input. The main idea behind this endeavor is to prove that the TSISQ prototype identifies the relevant literature referred to in the literature review. Due to the limited number of IT-security-specific conferences in our index, only 35 of the 113 identified references are included. The first 100 results of this query were manually screened for affiliation to the topic and checked for its appearance in the literature review.
The first result is a duplicate of the paper used as query input (Figure 3). 95 of the remaining 99 publications were classified by domain experts as being relevant to the desired domain. This corresponds to a false-positive measure of merely 4 percent. Moreover, 28 of the 35 theoretically detectable papers were identified and displayed among the query results, which corresponds to an average search accuracy of 80 percent and accordingly, an average probability of 20% for the occurrence of false-negative errors within the top 100 results. However, the respective articles were scattered across the whole range of the query output without any recognizable pattern. One possible explanation for this behavior is that our index contains articles that were not regarded in the literature review we used as query input. Another reason could be that the search scope of the prototype cannot be as narrow as if papers were chosen using human cognitive abilities. In particular, the literature review by [LUB04] mainly focuses on publications that specifically deal with employees’ information security related awareness and behavior. However, our search results appear to be less fine-grained, as they cover publications of the whole field, e.g. reaching from IT security in the healthcare sector to IT security policies in banking to the influence of human factors. Nevertheless, we believe that this aspect is not a defect but instead, can be helpful to the literature research process, as it increases the probability of identifying relevant, related publications even from fields that primarily had not been taken into consideration.

The next aspect to be discussed is that the first five results do not match with identified articles of the literature review, which leads to a delayed initial increase of the average share of hits curve. The first hit has to be excluded as it is a duplicate of the query input. The next four hits match with the overall topic but are not included in the literature review. This means that either (1), they were missed during the literature research process for the paper or (2), they were irrelevant due to the above-mentioned aspect of less fine-grained results or (3), the respective articles were published after the review was completed. However, after manually screening those four papers it can be stated that the latter aspect applies to only one of them.

While the first case study demonstrates the search accuracy, the main purpose of the second case study is to quantify the false-negative error rate. First, the chance to achieve good results in the top 25 entries of the output is potentially high due to the large size of the index. However, the number of articles dealing with query-related content is unknown. 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 index is relatively low, e.g. 30, detecting 19 of them within the first 25 results (which corresponds to the observed 75 percent accuracy) would be a good hit rate.

In summary, against the backdrop of the overall purpose of our TSISQ prototype, it can be concluded that following our approach can help save valuable time in discovering the basic literature in a desired research field. Additionally, it can assist to increase the comprehensiveness of a review by identifying sources that would otherwise not have been taken into account. Although human cognitive abilities are still indispensible, our prototype is a useful complement to the established information retrieval methods used in the scientific literature research process. Following the identified limitations, further research steps are required with regard to our approach. Firstly, the data base used for the index should be extended to cover more conferences and journals in a longer period of time. Secondly, the tools’ usability and accessibility could be improved, e.g. by implementing a web frontend. Our research showed that the more focused and consciously a query is formulated, the more specific and relevant the results become. Hence, future work should address the establishment of clear guidelines concerning the composition of query inputs. Beyond that, embedding the presented approach into an iterative-process cycle can be a promising additional expansion.

References


