Leibniz Universität Hannover

Boom and Bust Cycles in Scientific Literature

A Toolbased Big-Data Analysis

Bachelorarbeit

zur Erlangung des akademischen Grades „Bachelor of Science (B.Sc.)“ im Studiengang Wirtschaftsingenieur der Fakultät für Elektrotechnik und Informatik, Fakultät für Maschinenbau und der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Khoshandam Ghashang [redacted]

Vorname: Mohammad Amin [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Ort, den: Hannover, 01.09.2015

Contents

1. Introduction

1.1. Motivation

The growing number of research articles evokes to unstructured and complicated literature research and analysis for researchers (Parolo et al. (2015); Mabe and Amin (2001)). In times of the internet and electronic library's it is much easier to search for scientific articles as before, where researchers had to use books and magazines in paper form. As a consequence of this, the researcher publish even more articles and studies so that a structured overview of existing work becomes difficult (Parolo et al. (2015); J. Park and J.-N. Lee (2011)). It is particularly not possible to read any article in detail to select only these which are really relevant in the considered research area. However, a structured literature review is the basis for any research process and therefore it is indispensable (Webster and Watson (2002)). The published articles and studies are the foundation to identify a research gap and to ascertain an overview of the already existing knowledge in the research area (Levy and Ellis (2006)). Especially in the field of IS-research only a few articles are available, which describe the procedure of a structured literature review (Webster and Watson (2002)). A further option to identify relevant and high quality literature are the citations. Different calculations exist to evaluate the present article (Bar-Ilan (2008); Dong et al. (2005)). However, to read and study any article in detail and hence recognize the most relevant literature is still a manual, difficult and lengthy process. Oftentimes researchers are searching for the literature by using keywords which select not every relevant article since synonyms and periphrases are used by the authors (S. T. Dumais, Furnas, et al. (1988); Foltz (1990)). Thus, it is useful and more efficient to use latent semantic indexing to evaluate the present literature (Koukal et al. (2014); S. T. Dumais (1991)).

Once the researcher selected the relevant articles and studies, it is useful to know from which country and of which source the literature is from. By knowing the source and country it can be evaluated if an article is on international high top level or if it is only a national or limited one. Additionally, to assess the identified literature in more detail it is beneficial to ascertain in which years these are published (Koukal et al. (2014)). Research should always be

connected to previous ones, since many research already exist, is published and is available in forms of articles and studies from conferences and journals. Hence, it should be analyzed in which years the literature is published and if this specific research topic is too old for further research or if it is still a hype and much potential is seen from the current researchers. However, it could be also possible, that the literature is elderly but the topic can be continued due to new developments e.g. around information system (IS) topics.

Due to this reasons and the exponential rising number of published research articles, an automated analyzes may increase the process of studying the identified literature accurately. A further option is to examine if authors still search in the same field and only named their research different. Furthermore, diverse authors may use different keywords and synonyms for the same meaning. Hence, a latent semantic indexing supports to identify any relevant article. First, the articles can be selected quickly by an automated literature research within one or more databases. The next step is already to analyze these articles by having information about the source, hypes, trends, authors and published years before reading them to interpret them correctly. Due to the missing analytics options of current automated literature research tools, in this thesis an existing software tool from Koukal et al. (2014) named TSISQ (Tool for Semantic Indexing and Similarities Queries) is enhanced. It has its focus especially on the tasks to find and analyze scientific literature.

1.2. Objectives

The core of this thesis is to further develop and improve the existing literature research tool from Koukal et al. (2014) to filter the identified literature e.g. for years or for sources and to analyze the articles automatically. To demonstrate the options and functions of the new tool, it is used in this thesis for a case example, the literature research and analyze for the topic of cloud computing. Furthermore, the present tool is limited to only 100 articles as result, so that one further aim is, to detect even more relevant literature.

With the case example cloud computing, the relevant literature will be identified within the database AISeL, where articles from diverse conferences and journals

are available. These articles can be assorted e.g. of the similarity score to have first the most relevant and matching ones. The score during the case example is appointed to be at least 50%. The identified literature will be then used for further analyzes to ascertain boom and bust cycles within scientific literature for cloud computing. Using the tool, the research field of cloud computing will be presented and development over the years will be detected. The research tool TSISQ can support to interpret the found articles and gives an overview of the research field quickly. It may be used by any researcher as a basis for further research. Hence, this thesis seeks to answer the following research question:

How can a LSI-based approach be used for an analysis of scientific literature?

1.3. Course of investigation

Figure 1.1 shows the overview of the thesis. First, the research question (RQ) is defined. To derive the RQ a systematic literature review is necessary. This is the second step. After this step, the research tools are analyzed, regarding in which way they can support to answer the research question. In the third step, the past is analyzed in order to answer the RQ. In the fourth step, the discussion of results and a way to answer the research questions is presented. The thesis (sub)-tile is *A **tool-based big-data analysis***: *Tool-based* is defined in step four whereas the *big-data analysis* is processed in step five. The last and sixth step is the discussion, which is in the fifth chapter.

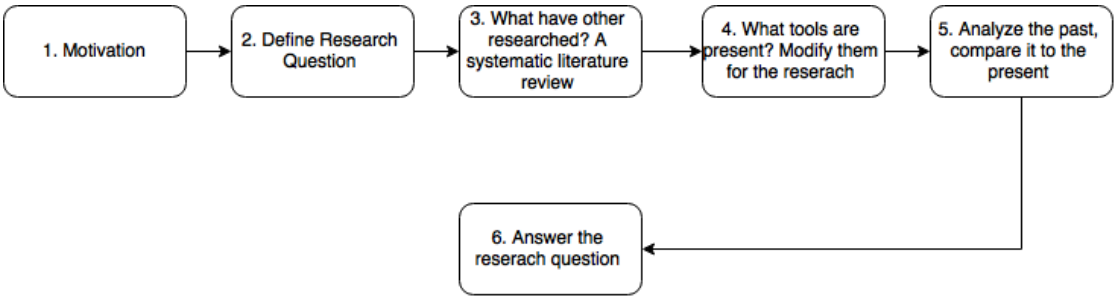

Figure 1.1.: Overview of the thesis

6. Conclusion and Outlook

In this thesis the research objective was to enhance an existing literature research tool. The tool was developed to improve the analysis of the identified literature. Existing limits were identified and resolved. The tool displays now more than 100 results. For that purpose a new module was implemented to gain an in-depth knowledge of the database and the query results. With this method, parts of the AISeL database were analyzed and a case example was preformed on the topic cloud computing. This topic was compared to another scientific search engine, the AISeL search engine. The comparison demonstrated the strength of this enhanced tool. In terms of quantity and quality more and better results could be realized. Additionally, the distribution over the time was evaluated. A trend was indicated, but not enough input data is available to determine this trend for sure. These results were compared also to public search trends. This comparison revealed that the science started later the discussion and analysis of the topic Cloud Computing as the public search interests for this specific topic. Moreover, a boom and bust cycle could be identified (at least in the public search trends). Furthermore, the same data revealed that the topic Big Data rises, when the topic cloud computing declines.

The tool was developed to support researchers in their literature review process. An automated analyzes may increase the process of studying the identified literature accurately. The amount of research articles in the scientific literature is rising and therefore new methods to analyze them are necessary. The semantic similarity is a method, which can handle such rising data and is able to process reasonable queries and results automatically. A structured literature review is the basis for any research process. The developed tool increases and accelerate exact this indispensable process and can be hence utilized of researchers for their research field.