2020

Topical Research Cluster of BLED Community – A Text Mining Approach

Nora Fteimi

Marikka Heikkilä

Jukka Heikkilä

TOPICAL RESEARCH CLUSTER OF BLED COMMUNITY – A TEXT MINING APPROACH

NORA FTEIMI[1], MARIKKA HEIKKILÄ[2] & JUKKA HEIKKILÄ[2]

[1] University of Passau, School of Business; Economics and Information Systems, Passau, Germany, e-mail: nora.fteimi@uni-passau.de
[2] University of Turku, School of Economics, Turku, Finland, e-mail: marikka.heikkila@utu.fi, jups@utu.fi

Abstract The number of research publications is growing exponentially, also in the discipline of Information Systems (IS). Evidently, we need new automated means for carrying out extensive inquiries into bodies of knowledge to understand the thematic foci of publications. The aim of this study is to apply an automated cluster analysis as a method of text mining and identify thematic foci of 654 BLED conference proceedings obtained from Scopus since 2005. Subsequently, we discuss advantages and challenges associated with the automatic analysis of huge volumes of texts. Our results support scientists and practitioners to focus future research efforts on these topics and thus help to establish and investigate the identity of the IS discipline, particularly against the background of the growing diversity of topics. The results help the conference to align future calls accordingly. In the future, a prototype can be implemented based on the results to suggest suitable search results.

Keywords: Bled proceedings, research foci, document clusters, text mining analysis, data analysis.

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-362-3.34
ISBN 978-961-286-362-3

1 Introduction

The need for scientific knowledge for making enlightened societal decisions and for developing goods and services is growing in the society. Simultaneously, the body of scientific literature is growing exponentially. More and more publishers are appearing, new seminars, conferences and publication series introduced – with non-proven scientific record. We are also witnessing a qualitative change in research reporting beyond original, authentic research. Derivative and synthetic reuse of data-sets is increasing, meshing open data from various open sources and reports by automatic analysis and categorization (Buchkremer et al. 2019). Evidently, we need new means for speedy and extensive inquiry into bodies of knowledge, but at the same time, we have to consider the trustworthiness of the sources and reputation of the publishers. In this paper we show a mean to identify the categories of research from a reputable source (i.e., Bled Conference itself), and then discuss its drawbacks and requirements.

The analysis approach we use here is ***text mining*** which is defined as a "*knowledge intensive process in which a user interacts with a document collection […] using a suite of analysis tools*" (Feldmann and Sanger 2007, p.1). Text Mining uses different algorithms and methods from interdisciplinary fields like information retrieval, statistics and natural language processing with the aim of discovering insightful knowledge, new patterns and correlations out of texts. For instance, the analysis outcome can represent semantically related themes, clusters of similar documents, topics with related terms or in the simplest form a frequency count list thus structuring and exploring the text corpus according to certain criteria and objectives.

With its focus on e-related topics, the Bled eConference, which was established in 1988 and has been held annually since then, attracts international interest from researchers and practitioners in Information Systems (IS). The proceedings cover a broad spectrum of established and novel topics that address various facets of social and organizational life, including e-health, e-business, e-government and e-learning. A challenge arising from the increasing amount of textual data is the rapid identification and allocation of publications with similar thematic foci into clusters.

Over the past years, BLED authors applied various methods (e.g., meta-analysis and automated semantic analysis) to investigate the conference topics, its research streams and how they evolved over time (e.g., Clarke 2012; Dreher 2012). With our paper, we will continue contributing to these efforts and present the results of performing a text mining analysis, more precisely an extensive cluster analysis, on 654 BLED conference publications in the time span between 2005 and 2018 (c.f. section 3 for more information about the dataset applied for the analysis). Thus, we define the following research questions for our study:

RQ1: Which topical foci were dominating the proceedings of BLED over the past years and how can these topics be organized according to clusters?

RQ2: What are the requirements and drawbacks of automated analysis of data with methods such as text mining?

The study contributes to research and practice by identifying main research topics, which characterize the research interests of BLED community. We point out prospective future research priorities and deliver the community with an instrument that allows the fast identification of relevant and similar documents according to their topic similarity. The latter can be seen for example as a first step towards implementing an intelligent semantic search engine prototype that suggests similar results and papers to the user and outputs appropriate search results based on the identified topic clusters.

We structure our paper as follows: First, we shed light on selected streams of literature, which contributed to discover the core of knowledge within the BLED community in particular and the IS domain in general (section 2). Subsequently, we provide an overview of the research design applied in this study (section 3). In section 4, we present the main analysis results of the clustering, and then conclude the limitations for using abstract and keyword analysis for recognizing reliable and qualitative different research with implications for further research.

2 Background and Related Work

A first example for the analysis of BLED publications is the study of Clarke (2012), who carried out a long-term analysis of all topics published at the conference between 1995 and 2011. For this purpose, the author examined titles and abstracts of 773 articles in order to elaborate on the topics covered there on the one hand and to examine the impact of the conference in general on the other hand. For this purpose, descriptive information on authorship, citation frequencies and the frequency of article downloads were collected. The thematic analysis was based on the distribution of the articles according to different periods and the examination of the topics dealt with in each period. The author manually grouped these topics subsequently applying a content analysis. This analysis was continued in Clarke and Pucihar (2013) which recognized three phases in the development of research foci of BLED conference: the EDI era (1988 – c. 1995), the period of the Internet and eCommerce (1996 – c. 2004) and the eInteraction era covering web 2.0 and social media (2005 – c. 2011). The authors call for more research going beyond technological interventions and their direct impacts, and to complement economic perspective with personal, community and social perspectives.

With a slightly different focus, Dreher (2012) presented the results of a semantic analysis of the BLED articles between 2001 and 2011. Via a text mining analysis, the author produced a corpus of available full texts and analyzed them automatically using the tool *Rubrico*. The goal of the semantic analysis was to identify embedded concepts and identify terms with related meanings to provide insights into thematic trends, which were present there in latent form. The main finding was that the concept "user" featured strongly in the corpus. A related concept was "people" which was frequently used in health related papers instead of user.

We discuss a third example for the thematic analysis of conference topics within the IS field. Sidorova et al. (2007) thereby use another text mining analysis form, the latent semantic analysis, to examine 1,615 research abstracts published in three top IS journals (*MISQ*, *ISR*, and *JMIS*) between 1985 and 2006 with respect to emerging and declining research topics. Similar to the previously presented studies, the focus was on the temporal differences and dynamics in different time spans. The results show high dynamism of the IS field; new topics replaced the old ones at the top-5

list every five years. The authors used their results to formulate a research agenda, emphasizing the need to pay more attention to the "rigor" factor henceforth.

As a concluding example, we describe the approach of Goyal et al. (2018), who also used latent semantic analysis and methods from the field of natural language processing to identify research topics in IS. With a special focus on four top IS journals (*MISQ, ISR, JAIS* and *JMIS*) and drawing on a previous study of Sidorova et al. (2008), the authors illustrated thematic trends once for the entire corpus and once broken down by the individual journals. A clustering algorithm was finally applied to group the research topics according to eleven corresponding clusters and subsequently assign them manually a label. The top five predominant research themes identified by the authors are knowledge management, technology adoption, e-commerce, recommender systems and security. As an exemplary result of the trend analysis, a high level of activity in the topic of knowledge management was observed for all four journals.

The previous studies indicate how different methods and procedures are used in the IS discipline to address and investigate research themes. Since the focus of our study lies on grouping texts with similar topics rather than on the temporal component, our paper shows how cluster analysis can be used as an effective text mining method to determine clusters of topics based on the documents semantic similarities considered and thus uncover similar topic groups.

3 Research Approach

Our methodological approach for the first research question relies on a text mining procedure in which we have implemented a cluster analysis to classify the documents of our text corpus. We thereby follow the process flow of text mining, which is visualized in Figure 1. The process typically consists of several main phases starting with the project definition and data collection, followed by main text preprocessing steps up to the application and interpretation of the results (c.f., Feldmann and Sanger 2007).

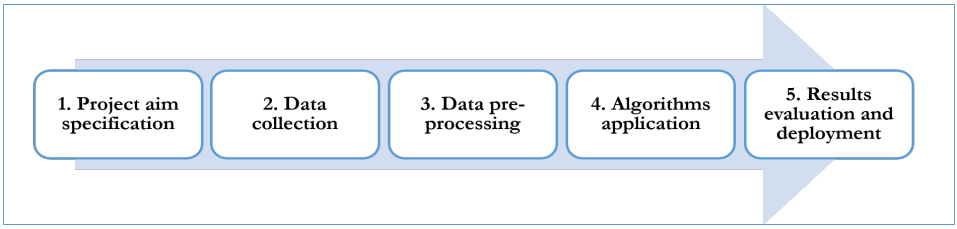

Figure 1: Five phases framework of text mining (c.f., Feldmann and Sanger 2007)

Below we describe the phases of data collection and preprocessing as well as the detailed application of text mining methods by means of the clustering algorithms.

3.1 Data Selection and Corpus Description

We collected our dataset from *Scopus*[1] database, which includes a selection of metadata of 654 articles published in the BLED proceedings (main conference proceedings) since 2005. With reference to Clarke and Pucihar (2013), who describe the new era of interaction from 2005 onwards, we specify this year as starting point for our analysis. Figure 2 shows the yearly distribution of included items and indicates how the number of data items varies per year. However, we point out that these values only reflect the items included in *Scopus* and may therefore differ from the original publication numbers. As our focus lies on analyzing the topical research foci of the conference, we obtained available titles, abstracts and keywords of all 654 data items[2]. As abstracts represent a summary of key findings in the papers, we are confident that our data selection process meets well the analysis purpose. Whereas we used titles and abstracts for cluster analysis, we separately analyzed the keywords to compare both analysis results with each other (c.f., section 4).

[1] https://www.scopus.com
[2] Dataset can be provided on request.

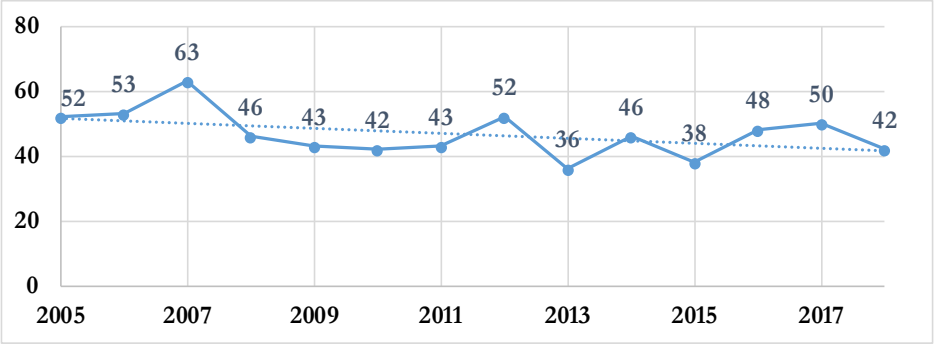

Figure 2: Yearly distribution of dataset (x-axis: years; y-axis: item numbers)

We performed the analysis in the software *R*, an open-source environment for statistical data analysis (Feinerer et al. 2008). For analysis purposes, we built two corpora in R: A first corpus for the running text containing all titles and abstracts (used as input for cluster analysis) and a further corpus containing only all semicolon-separated keywords (used for comparison base with cluster results).

3.2 Preprocessing Steps and Text Mining Analysis

In preparation for the cluster analysis, we undertook some core preprocessing steps of text mining. This included, for example, the harmonization of upper and lower case, the removal of numbers, punctuation marks and stop words, as well as stemming. The latter is crucial to reverse words to their root form and harmonize different word notations.

Subsequently we implemented an agglomerative hierarchical cluster analysis (Zhao et al. 2005) by computing the cosine similarity (Huang 2008) for our dataset, which is a popular measure to compute the document similarity in text mining based on the vector angles (Han et al. 2011). Hierarchical clustering works iteratively, whereby in each iteration step two homogenous documents are merged or clustered based on their similarity. The process ultimately ends with one supercluster. The result of cluster formation in the various iterations is illustrated by a dendrogram representing a cluster tree. The iterative cluster formation was an important reason for choosing this clustering method in this study, as it provides a comprehensive overview of the individual clusters and allows the results to be easily traced and verified stepwise.

Further methods like k-means clustering can be used in the future to compare the results, but their combined use would go beyond the scope of this paper and its topic. We applied *ward's* clustering method (also called minimum variance method) that tends to build compact even-sized clusters (Murtagh and Legendre 2014). This method has wide application in linguistic analysis domains (Szmrecsanyi 2012) and overcomes the drawbacks and computing effort of other clustering methods.

Based on the results of cluster analysis and due to the large number of documents considered in the clustering process, we finally formed groups of superclusters to make the results more comprehensive and to classify them according to key topics. After several iterations in which we varied the number of superclusters, 15 groups of superclusters proved to provide the best and saturated results for the representation of the clustering process. In section 4, we discuss these results.

4 Main Results of Text Mining Analysis

4.1 Clustering Analysis – Supercluster Results and Description

In summary, the application of the clustering algorithm to our corpus led to grouping all 654 datasets based on their thematic similarity. Subsequently, the algorithm formed superclusters in order to summarize the cluster results according to 15 key research topics. Figure 3 shows an excerpt of the overall dendrogram in which four superclusters are illustrated. Each supercluster, represented in the form of colored rectangles, comprises document clusters that are thematically related to the corresponding supercluster. The numbers in the figure represent the respective internal document ID's in the corpus. Each document is labeled using its three top significant terms.

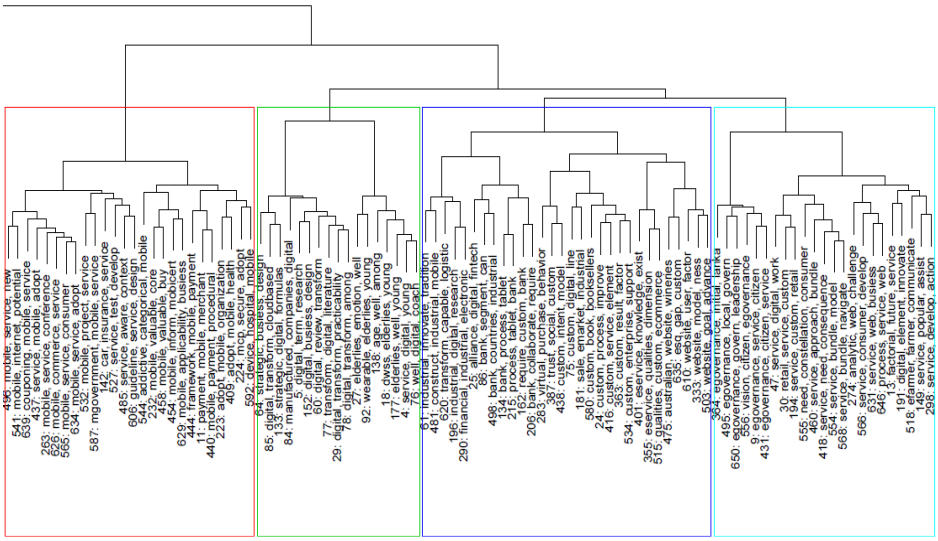

Figure 3: Dendrogram excerpt for 4 out of the 15 superclusters

For instance, the turquoise supercluster on the right represents documents that mainly address the research topic of governance and public administration. In summary, 23 papers were clustered there. This corresponds to 4% of all documents in the corpus. Thus, these documents often deal with the topics of e-governance, leadership processes, (web-based) services e.g., in the agricultural or retailing sector, overall effects on citizens or work procedures in general. The development and use of models for the optimization and implementation of governance structures are also addressed by the related documents.

The blue-bordered rectangle includes all clusters that deal with the research topics of industry and banking. In total, the 30 documents clustered here frequently list terms such as finance, logistics and supply chain, alliances, banking and the purchase or sale of products. Moreover, the virtual orientation of financial services and the use of corresponding technologies as well as the trust of investors and customers into these developments are also covered repeatedly.

Closely related to this topic is the red supercluster (left in Figure 3). Researchers discuss here aspects of cashless payment using mobile technologies, insurance policies and the security of such services in an e-commerce context.

With the green-bordered cluster we were able to identify those documents that are primarily dedicated to the digital transformation and its effects on different generations (e.g., through the use of wearables). The well-being of these person groups is also addressed in this context, e.g., by examining expressed emotions.

In Table 1 we summarize the most prominent results of the cluster analysis. We briefly describe each of the 15 superclusters by outlining the thematic focus of the supercluster and key terms of its related documents. In addition, we summarize how many documents are contained in each supercluster.

Table 1: Research topics and clustering analysis results

ID	Supercluster Topic	Dominant Terms per Supercluster	Paper No.
1	Governance & public administration	E-governance; citizen; service; processes; local; government; retail; consumer; e-farmer	23 (4%)
2	Industry & banking sector	Industry; enterprise; finance; fintech; bank; purchase; logistics; sector; alliance	30 (5%)
3	Digital transformation & generation change	Digital; transform; business; strategic elderlies; young; age	17 (3%)
4	Mobile services	Mobile; service; device; buy; commerce; payment; insurance	26 (4%)
5	Miscellaneous e-business, IT & application areas	Music; sale; workplace; e-business; e-marketplace e-participation; portal; applicability; standardization; knowledge; benefit; trust; evaluate; bitcoin; stock; financial; crowdsource; service oriented architecture; framework; surveillance; data; cloud; reputation; smart	145 (22%)

6	E-Learning & academia	Learn; student; e-learning; course; educate; teach; academia; game-based	18 (3%)
7	Cloud-based collaboration & services	Cloud; adopt; enterprise resource planning; enterprise; change; network; platform; communicate; crowdsource; agile; architectural	61 (9%)
8	Online communities & markets	Communities; virtual; online; market; brand; e-commerce; e-business	24 (4%)
9	E-Commerce adoption & acceptance	Market; e-transformation; e-procurement; economic; adopt; cultural; attitude; game; motivate; incentive; fit; accept	67 (10%)
10	Business-IT alignment & technological trends	RFID; software; inter-organizational IS; e-technology; orchestration; competition; toolmakers; customer relationship management; collaborate; network;	61 (9%)
11	Digital ecosystem	System; information & communication technologies; ecosystem; success; consumer	37 (6%)
12	Big data	Data; analytic; technological; big; platform; open; link; xml	15 (2%)
13	Healthcare	Patient; healthcare; e-health; care; mental; lifestyle; risk	35 (5%)
14	Social media networks & behavior	Social; media; control; network; self-esteem; Facebook; WhatsApp; group; analytic; use	37 (6%)
15	Business models	Business; model; design; business process management; innovation; rule; outsource; interoperability	58 (8%)

4.2 Keyword Analysis

Complementary to the cluster analysis (on titles and abstracts) we separately analyzed the keywords corpus, which BLED researchers used to summarize their papers. Thus, we were able to perform a comparison with the results of the cluster analysis. For interpretation purposes, we consolidated different spelling forms of keywords

(e.g., plural and singular forms, abbreviations) accordingly. Table 2 lists in descending frequency order all terms that occur at least ten times over the entire corpus. In summary, 1,750 unique keywords occur throughout the corpus, of which 51% were mentioned only once. The 15 keywords in Table 2 account for 10% of the overall cumulative term frequency.

Table 2: Keywords in BLED proceedings with a frequency >= 10

Keywords	Term Frequency
Social media, web 2.0	45
Small and medium sized enterprise	28
E-Commerce	25
Business model	24
E-Health	24
Adoption	16
Case study, case survey	15
Cloud computing	15
E-Government	14
Healthcare	13
Trust	12
E-Learning	12
Business model innovation	11
Mobile health	11
Mobile service	10

4.3 Comparison of analyses

Considering Table 2, it is observable that the overall keyword focus lies on topics from the e-business and digital transformation context (e.g., e-health, e-government, mobile service, social media, cloud computing), which is in line with the scope of the conference.

We also find several similarities but also some differences regarding the topics identified in cluster analysis compared to the listing of keywords. For instance, with 45 counts, a predominantly technical keyword "social media/web 2.0" leads the list of keywords. This supports Clarke's and Pucihar's (2013) conclusion that web 2.0 and social media research would dominate the research published in BLED eConference. However, the cluster analysis of the titles and abstracts corpus reveals that the biggest topic cluster (22%) contains miscellaneous research papers on e-business, IT and application areas. The topic "social media", in contrast, represent only one of the smaller superclusters with 6% coverage. Likewise, with 1,750 unique keywords, we find a very heterogeneous set of terms used by the authors to describe their articles.

However, our analysis shows that particular topic clusters – e.g., e-commerce adoption & acceptance (10%), cloud-based collaboration & services (9%), business-IT alignment & technological trends (9%) and business models (8%) – are clearly observable and that BLED authors frequently use certain terms across various papers indicating similar document clusters.

We can also conclude that some results of both analyses are comparable. For instance, similar to the clusters described before, both the superclusters as well as the keyword list deal with hot organizational and technological issues. Electronic service-related topics (e.g., e-governance, e-health, e-business, e-learning) and their coverage in an organizational context (e.g., adoption, trust, success) are dominating in both corpora.

5 Limitations and Future Improvements for Using the Analysis Results

This paper supplemented the previous analysis of Bled eConference papers (Clarke 2012; Clarke and Pucihar 2013) with advanced text analysis methods of abstracts. Even though some similarity with keyword- and abstract-based analysis could be observed in our study, the analysis shows not only discrepancy between the keywords and abstracts, but also lack of data about the research domain, type of data used, nature of studies, and their purposes, all which are of importance when utilizing research results in real world decision-making. As it is now, our analysis provides merely an overview of research subjects.

For publishers' purposes, the above analysis might suffice, but in real life scenarios for decision-making, more information should be provided for the utilization of the results. A journal, conference, or seminar is expected to be scientifically valid and reliable, but this is not always the case (Bowman 2014). Therefore, Masten and Ashcraft (2017) suggest a due diligence of reputable channels for scientific research, but we see the need to go beyond that improvement for the authors, too. As it is evident for the researchers and decision makers to use automated tools for retrieving scientific research, the following measures could be taken:

1) We must provide AI-tools or data scientific -tools full access to texts to find additional information on data, domain, nature of the study, its limitations, etc., because these are too often missing from the abstracts, not to mention keywords.
2) Improving metadata on research.
    a. One way practiced by some journals is to insist for standardized metadata (e.g., standardized abstract revealing purpose, design/methodology/approach, findings, research limitations, theoretical and practical/managerial implications, originality/value and paper type). These are easy to implement in the editorial IS, or complemented by Internet-based questionnaire. Often research projects' *data management plans* can provide this information.
    b. Letting analysis tools to classify the article against the body-of-knowledge automatically, in some cases even without author intervention.
3) Publishers should standardize and gather the information systematically and open it to the public and automatic analyzers.

Regarding our second research question, we conclude that more semantic methods of text mining needs to be incorporated for advanced analytics purposes. The context, in which a certain terminology is used (e.g., the use of negations, sarcasm, slang or the authors' background) and the underlying domain of a text provide important information for analysis, but is difficult to capture using existing automated methods. Manual intervention is therefore often essential. One useful solution that merits further research are domain ontologies, which help to capture the semantic context of similar words, reduce textual and noise and disambiguity (Afolabi et al. 2019).

6 Conclusion

Based on a text mining analysis, we performed an automated cluster analysis on 654 abstracts and titles of BLED proceedings since 2005. Subsequently we compared the results with the keywords provided by the authors of the respective publications. Our aim was 1) to identify research foci of BLED community and 2) to point out drawbacks and requirements of using automated text mining techniques.

Limitations of our work concerns the limited amount of datasets used for the analysis. We deliberately did not include temporal trends, as they were considered in previous work, and focused on the third era of eInteraction (Clarke and Pucihar 2013), in order to determine key research foci in this era. The cluster analysis also leaves some freedom of interpretation, since we have manually labeled the superclusters after performing the automatic cluster analysis and formation of the superclusters. As an unsupervised learning algorithm, this last step of a cluster analysis remains with the analyst. However, since we have carried out the labeling process independently by the authors of the paper, we are confident that it has a reasonable degree of validity.

Concluding, our results show that the community has addressed a broad mix of research topics over the years. In total, we detect 15 superclusters on various topics dealing with different aspects of electronic life. It is also noticeable that many miscellaneous topics (represented by the largest topic cluster no° 5) are addressed.

Methodologically we conclude that the cluster analysis provides first promising results for the classification of semantically similar documents. At the same time, further research is required, for example to develop a comprehensive domain ontology to capture the semantics behind the text. Moreover, the results can be practically implemented in form of a semantic search engine, which supports researchers in finding articles based on the clusters. Such a prototype could, for example, provide users with search suggestions for thematically similar studies by using the thematic clusters and thus generating recommendations and suitable search suggestions based on the similarity of content. The development of a taxonomy or classification scheme would be a further interesting contribution to the harmonization of research topics, as it is already taking place in the IS discipline and its sub-domains (e.g., Knowledge Management). The clusters and assigned terms can

represent the structure of the taxonomy according to which suitable research papers can be delivered. Finally, further analysis methods can also be used based on the cluster analysis, e.g., by means of topic modelling (Blei and Lafferty 2009) to delve deeper into the documents and identify dominant terms.

References

Afolabi, I. T., Sowunmi, O. Y., and Adigun, T. 2019. "Semantic Text Mining Using Domain Ontology," in Roceedings of the World Congress on Engineering and Computer Science, pp. 1–6.

Blei, D. M., and Lafferty, J. D. 2009. "Topic Models," in Text Mining: Classification, Clustering, and Applications, A. Srivastava and M. Sahami (eds.), Chapman and Hall/CRC, pp. 71–89.

Bowman, J. D. 2014. "Predatory Publishing, Questionable Peer Review, and Fraudulent Conferences," American Journal of Pharmaceutical Education (78:10), pp. 1–10.

Buchkremer, R., Demund, A., Ebener, S., Gampfer, F., Jägering, D., Jürgens, A., Klenke, S., Krimpmann, D., Schmank, J., Spiekermann, M., Wahlers, M., and Wiepke, K. 2019. "The Application of Artificial Intelligence Technologies as a Substitute for Reading and to Support and Enhance the Authoring of Scientific Review Articles," IEEE Access (7), pp. 65263–65276.

Clarke, R. 2012. "The First 25 Years of the Bled EConference: Themes and Impacts," in 25th Bled EConference Proceedings - Special Issue, BLED, Slovenia, pp. 12–192.

Clarke, R., and Pucihar, A. 2013. "Electronic Interaction Research 1988 – 2012 through the Lens of the Bled EConference," Electronic Markets (23), pp. 271–283.

Dreher, H. 2012. "Automatic Semantic Trend Analysis of the Bled EConference: 2001-2011," in 25th Bled EConference Proceedings - Special Issue, BLED, Slovenia, pp. 193–208.

Feinerer, I., Hornik, K., and Meyer, D. 2008. "Text Mining Infrastructure in R," Journal of Statistical Software (25:5), pp. 1–54.

Feldmann, R., and Sanger, J. 2007. The Text Mining Handbook, New York: Cambridge University Press.

Goyal, S., Ahuja, M., and Guan, J. 2018. "Information Systems Research Themes: A Seventeen-Year Data-Driven Temporal Analysis," Communications of the Association for Information Systems (43), pp. 404–431.

Han, J., Kamber, M., and Pei, J. 2011. "Getting to Know Your Data," in Data Mining: Concepts and Techniques (3rd ed.), J. Han and M. Kamber (eds.), Morgan Kaufmann, pp. 39–81.

Huang, A. 2008. "Similarity Measures for Text Document Clustering," in Proceedings of the Sixth New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, pp. 9–56.

Masten, Y., and Ashcraft, A. 2017. "Due Diligence in the Open-Access Explosion Era: Choosing a Reputable Journal for Publication," FEMS Microbiology Letters (364:21), pp. 1–7.

Murtagh, F., and Legendre, P. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," Journal of Classification (31:3), pp. 274–295.

Sidorova, A., Evangelopoulos, N., and Ramakrishnan, T. 2007. "Diversity in IS Research: An Exploratory Study Using Latent Semantics," in ICIS 2007 Proceedings.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T. 2008. "Uncovering the Intellectual Core of the Information Systems Discipline," MIS Quartely (32:1), pp. 467–482.

Szmrecsanyi, B. 2012. Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry, Cambridge: Cambridge University Press.

Zhao, Y., Karypis, G., and Fayyad, U. 2005. "Hierarchical Clustering Algorithms for Document Datasets," Data Mining and Knowledge Discovery (10:2), pp. 141–168.