TEXT MINING TOURISM LITERATURE

AJDA PRETNAR & TOMAŽ CURK

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia.
E-mail: ajda.pretnar@fri.uni-lj.si, tomaz.curk@fri.uni-lj.si

Abstract Literature reviews are essential for understanding a specific domain as they map the main topics of current re-search. Our aim was to provide a framework for retrieving articles from online databases and analyzing them in a single script. We provide the analytical pipeline as open-source (https://github.com/tourism4-0/BibMine). The main research focus was on analyzing 318 abstracts from scientific papers on tourism and innovation, which we report in Zach et al. (2019). We used LDA topic modeling to uncover ten main topics, which we analyzed using pyLDAvis visualization. We used saliency and relevance scores to determine the main words that de-scribe a topic. The uncovered topics range from climate change and land use to smart destinations, travel expe-riences, and ICT. We performed similar analyses for the term "stakeholders," where we also observed the main verbs related to the query. Since verbs best define an activity, we used them to determine how stakeholders are involved in tourism development. Finally, we analyzed papers with the keyword "technology," where energy efficiency, VR, web technology, and augmented tourist experiences were the main topics.

Keywords:
text mining, literature review, meta-analysis, topic modeling, tourism

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-515-3.9
ISBN 978-961-286-515-3

1 Introduction

Computational literature reviews have become increasingly popular in recent years. As the amount of scientific literature grows, automatically summarizing and mapping relevant documents can save the researcher a lot of time (O'Mara-Eves et al. 2015). Korhonen *et al.* (Korhonen 2012) showcase how important assisted information extraction tools are in biomedicine. Talafidaryani (Talafidaryani 2020) uses topic modeling to expand the understanding of information systems research and to propose relevant tangent topics. Computational literature reviews can also assist in observing temporal changes in topics and trends (Karami et al. 2020). Kumar *et al.* (Kumar, Kar, and Ilavarasan 2021) use text mining to map the literature on text mining approaches in service management in a sort of meta-review. They use various techniques to better understand how text mining has been used in their field and what are the most frequent applications. Overall, text mining has been mostly used in tourism for the analysis of customer reviews (Claster et al. 2013; Godnov, Redek, et al. 2016; Tamrin and Septianasari 2021), while computational literature reviews remain sparse.

The main limitation of current literature reviews is retrieving structured data. When analyzing text data computationally, one needs suitable raw data, such as plain text. Retrieving data from PDF format is only a partial solution. These files often contain a certain amount of OCR noise and redundant sections of the paper, such as headers, footnotes, figures, and references.

To overcome this, we designed a system called BibMine[1] to retrieve structured XML files from the Elsevier web service, which enables extracting only the body of text, abstracts, and keywords. Such structured data can be fed directly to text mining algorithms without having to do extensive data preprocessing. In such a way, researchers can use BibMine without having to program scripts for data retrieval themselves, and feed structured tabular data directly to their favorite data analysis software.

In continuation, we briefly describe BibMine and how it works, and then present several use cases of text mining tourism literature.

[1] https://github.com/tourism4-0/BibMine

2 Building BibMine

The main advantage to using BibMine is the retrieval of structured scientific papers, offered by Elsevier. The system enables retrieving data from ScienceDirect, which offers access to full-content papers, and from Scopus, which offers access to abstracts alone, see Figure 1.

```
<prism:aggregationType>Journal</prism:aggregationType>
<prism:issn>0025326X</prism:issn>
<prism:volume>23</prism:volume>
<prism:startingPage>403</prism:startingPage>
<prism:endingPage>410</prism:endingPage>
<prism:pageRange>403-410</prism:pageRange>
<dc:format>text/xml</dc:format>
<prism:coverDate>1991-12-31</prism:coverDate>
<prism:coverDisplayDate>1991</prism:coverDisplayDate>
<prism:copyright>Copyright © 1991 Published by Elsevier Ltd.</prism:copyright>
<prism:publisher>Published by Elsevier Ltd.</prism:publisher>
<prism:issueName>Environmental Management and Appropriate Use of Enclosed Coastal Seas</prism:issueName>
<dc:creator>Apicella, M.</dc:creator>
<dc:creator>Benassai, E.</dc:creator>
<dc:creator>Di Natale, M.</dc:creator>
<dc:creator>Panelli, E.</dc:creator>
<dc:description>Abstract  The functional and architectural beauty of the harbours built on the Mediterranean
coast by ancient seafaring people (e.g. Fenix, Greeks and Romans) is of unestimable value in innovative planning
```

Figure 1: XML format enables extracting various metadata from the document.

The user provides the query and the system returns a structured tabular file. As mapping scientific papers to corresponding topics is the primary goal of literature review, one can run a basic analysis and topic discovery after retrieving the data.

Ananiadou et al. (Ananiadou et al. 2009) identify three stages of literature review: searching, screening, and synthesizing. BibMine assists mostly with the first two. First, it offers easy access to ScienceDirect and Scopus databases, which already contain a large number of papers, most of which are well-structured. Second, it helps narrow the search by using targeted queries, which search only in titles, keywords, and abstracts, thus avoiding a deluge of distantly related papers.

3 Case Studies

We used BibMine for several literature reviews. In the framework of the Tourism 4.0 project, we were mainly interested in innovation, technology and tourism impacts. Specifically, we looked at the following topics:

- innovation in tourism (query: tourism+AND+innovation)
- technology (query: tourism+AND+technology)
- innovation and technology (query: tourism+AND+innovation+AND+technology)
- tourism impacts (query: technology+OR+("sustainable tourism")+OR+("tourism impacts"))
- travel flows (query: ("travel flow?")+OR+("tourism flow?"))
- stakeholders (query: tourism+AND+system+OR+stakeholder+OR+behavior+OR+behaviour+ OR+factor+ OR+driver)

With every data set collected from the BibMine service, we did the initial statistical analysis of the data set, such as counting word frequencies, finding word contexts (collocations), or mapping keyword occurrences in time. Word frequencies give an insight into the most frequently used words in the subset, which already serves as a preliminary topic mapping. Before counting word frequencies, we extensively preprocessed the data. We removed phrases relating to each individual publisher (*i.e.*, @Copyright), removed numerals, removed stopwords, and lemmatized the tokens with WordNet lemmatizer. Finally, we also included the 100 most frequent bigrams in the analysis. To compute word frequencies and word contexts, we used the NLTK package.

For the stakeholders corpus, we extracted verbs from the sentences containing the word "tourist" and observed their polarity to determine what agency is related to them, see Figure 2. That is, we applied the sentiment analysis lexicon from NLTK to each sentence, then aggregated the score based on the verbs appearing in the sentence. Unsurprisingly, a frequent verb related to tourists is "enjoy," which always appears in a positive context. Conversely, the verb "prefer" is usually neutral, which likely means tourist preferences are described as the basis for a targeted action.

For the technology corpus, we mapped the terms of interest by their frequency in time, see Figure 3. Terms of interest were selected manually, but they were automatically extracted from the text and mapped to the year the paper was published. Occurrences were summed by year and plotted in a heat map, which shows the proliferation of certain technologies, such as virtual reality (VR), augmented reality, and artificial intelligence (AI), in recent years.

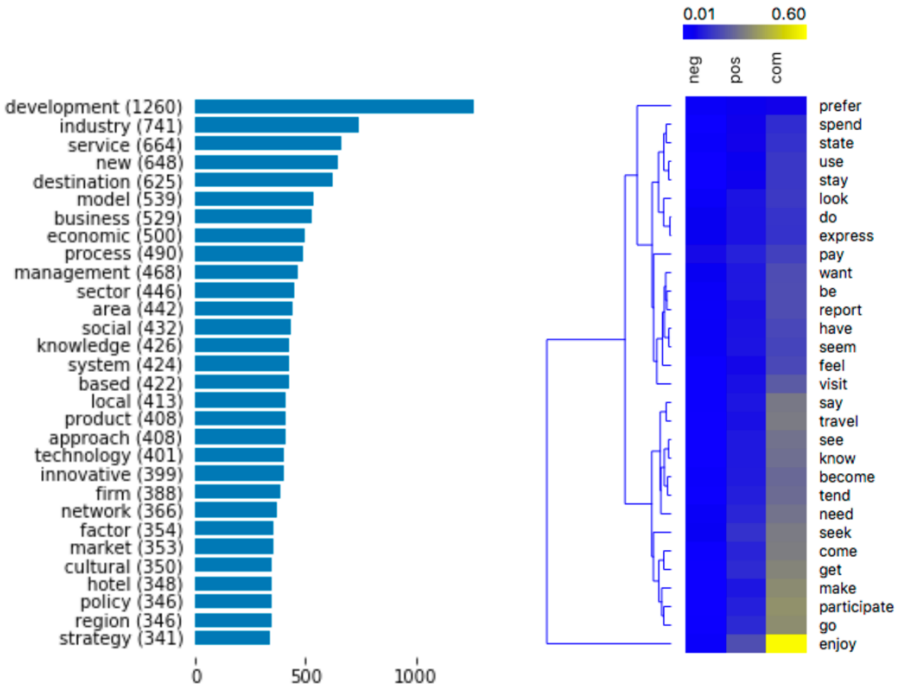

Figure 2: Word frequencies from the innovation in tourism corpus and sentiment analysis of stakeholder actions.

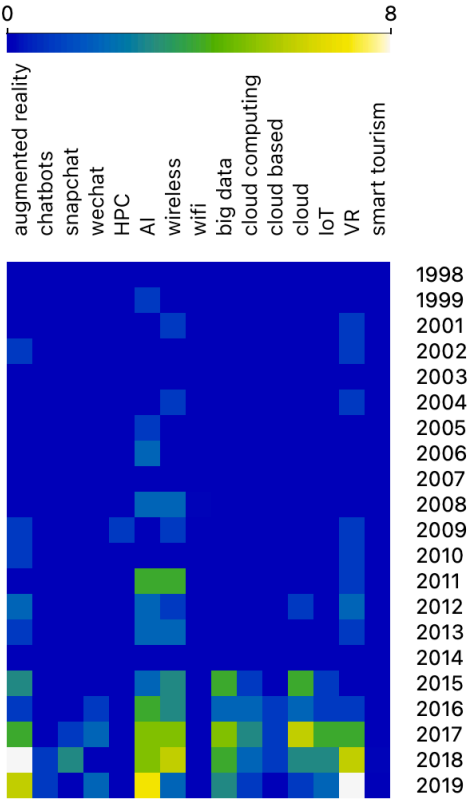

Figure 3: Occurrence of selected technologies in papers over time.

Finally, we used LDA topic modeling (Blei, Ng, and Jordan 2003) to extract topics from text. Log perplexity was used to estimate model quality - lower perplexity means a better model. Next, we used two techniques to determine the optimal number of topics; GridSearch (Bergstra et al. 2011) and Mallet (McCallum 2002). The two measures were used as validation for the number of topics. Finally, we plotted the topics with pyLDAvis interactive visualization (Sievert and Shirley 2014), which enables exploring the topics and their corresponding keywords, see Figure 4.

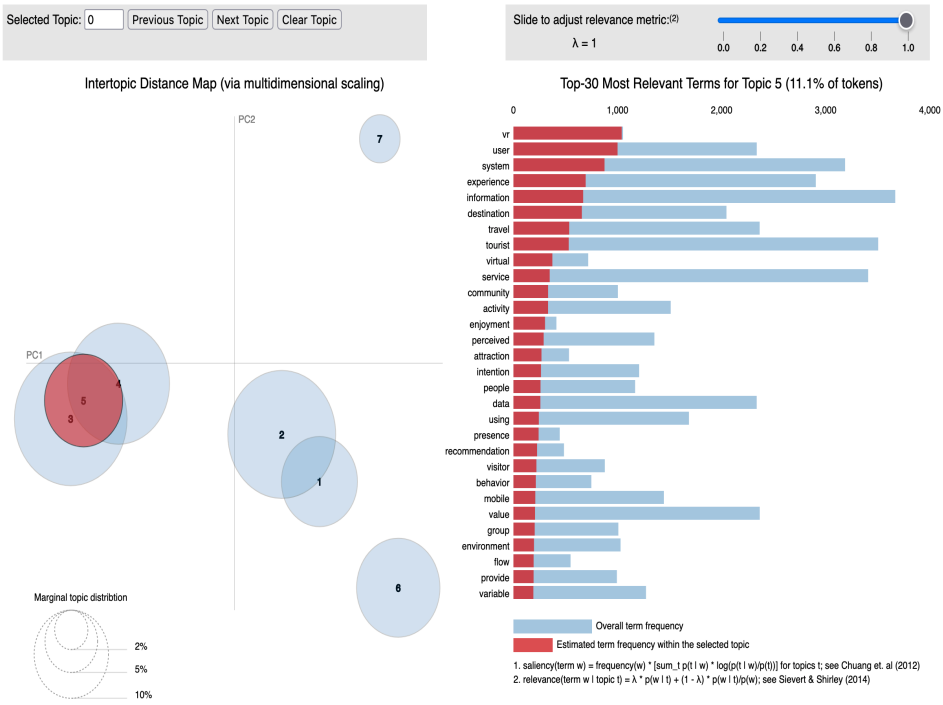

Figure 4: pyLDAvis visualization for seven topics from the technology corpus. Topic number 5 is selected, which includes references to VR, user (systems), and experiences.

4 Conclusion

Streamlining API queries and structuring the data is of great benefit to researchers who wish to focus on data exploration and analysis rather than on coding themselves. The increasing availability of structured scientific data enables rapid topic overviews, which can be substantiated with visualization. We used several approaches to analyze the collected data, from basic word frequencies to topic optimization and interactive maps. This provided the researchers a venue to explore several aspects of technology, innovation, and stakeholder attitudes in the tourism literature, but the approach could be applied to any scientific field.

Note

This is a Published Scientific Conference Contribution.

Acknowledgement

The research presented in this article has been conducted within the research and development project TRL3-6 Tourism 4.0 - Enriched tourist experience (OP20.03536), which is co-financed by the Ministry of Education, Science and Sport of the Republic of Slovenia and the European Regional Development Fund of the European Union.

References

Ananiadou, Sophia, Brian Rea, Naoaki Okazaki, Rob Procter, and James Thomas. 2009. "Supporting Systematic Reviews Using Text Mining." Social Science Computer Review 27 (4): 509–23.

Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. "Algorithms for Hyper-Parameter Optimization." Advances in Neural Information Processing Systems 24.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." The Journal of Machine Learning Research 3: 993–1022.

Claster, William, Phillip Pardo, Malcolm Cooper, and Kayhan Tajeddini. 2013. "Tourism, Travel and Tweets: Algorithmic Text Analysis Methodologies in Tourism." Middle East Journal of Management 1 (1): 81–99.

Godnov, Uroš, Tjaša Redek, et al. 2016. "Application of Text Mining in Tourism: Case of Croatia." Annals of Tourism Research 58: 162–66.

Karami, Amir, Morgan Lundy, Frank Webb, and Yogesh K. Dwivedi. 2020. "Twitter and Research: A Systematic Literature Review Through Text Mining." IEEE Access 8: 67698–717. https://doi.org/10.1109/ACCESS.2020.2983656.

Korhonen, Diarmuid AND Silins, Anna AND Ó Séaghdha. 2012. "Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research." PLOS ONE 7 (4): 1–16. https://doi.org/10.1371/journal.pone.0033427.

Kumar, Sunil, Arpan Kumar Kar, and P Vigneswara Ilavarasan. 2021. "Applications of Text Mining in Services Management: A Systematic Literature Review." International Journal of Information Management Data Insights 1 (1): 100008.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit."

O'Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. "Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches." Systematic Reviews 4 (1): 1–22.

Sievert, Carson, and Kenneth Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63–70.

Talafidaryani, Mojtaba. 2020. "A Text Mining-Based Review of the Literature on Dynamic Capabilities Perspective in Information Systems Research." Management Research Review.

Tamrin, M, and L Septianasari. 2021. "Mapping Problem Using Text Mining to Boost Tourism Industry: Is It Possible?" In IOP Conference Series: Earth and Environmental Science, 778:012009. 1. IOP Publishing.