Visual analysis of search results in Scopus database focused on sustainable tourism

Ondřej Klapka[1], Antonín Slabý[1]
e-mail: ondrej.klapka@uhk.cz, antonin.slaby@uhk.cz

[1] *Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic*

Klapka, O., & Slabý, A. (2021). Visual analysis of search results in Scopus database focused on sustainable tourism. *Czech Journal of Tourism*, 9(1), 41-53. DOI: 10.2478/cjot-2020-0003.

Abstract

The enormous growth of research and development is accompanied by growing number of scientific publications in recent decades. These publications are collected and processed by a number of digital libraries. Though digital libraries provide basic search tools, more advanced methods such as visualization and visual analysis can be implemented by using special software only. This article presents the possibilities of visual analyzing content of digital libraries using the CiteViz tool developed in Klapka (2013) and shows the implementation using of the Scopus database. Results of testing the implemented solution in selected areas of sustainable tourism and demonstration of the possibilities of the implemented solution are presented at the end of the article.

Keywords

Visualization, data mining, scopus, CiteViz, sustainable tourism

JEL classification: L83 **Accepted: 25 November 2021**

Introduction

More than 80% of information entering the human brain is of visual nature Jensen (2008). Scientific studies show that open human eyes require two-thirds of the electrical activity of the brain – a full 2 billion of the 3 billion firings per second – as show the findings of neuroanatomist R. S. Fixot in the paper published in 1957 (Fixot, 1957). These studies claim the sight to be the most significant and mostly used human sense for mining any kind of information.

These abilities of the human visual system can be well used in a process called visualization. Visualization tries to maximize the flow of information that human brain is able to receive. The growing importance of visualization itself is associated with increasing importance of information and the need to process it through suitable presentation as well. Visualization is often able to show or present very large and complex data structures in order to make them easily comprehensible to people (Keim et al., 2006).

Visualization forms compact and intuitive way to represent big and complex data of various nature Techopedia (2014). A very often used visualization method is based on networks expressed using graph theory methods as it is quite intuitive and can represent complex relationships between data. This structure is quite easily machine-processable at the same time and enables to apply the theory of discrete mathematics, which deals with this data structure and its processing. It also makes it possible to calculate different scales and statistics concerning analyzed network and present them to the user (Newman, 2010). This feature helps even more in the more complex process of visualization.

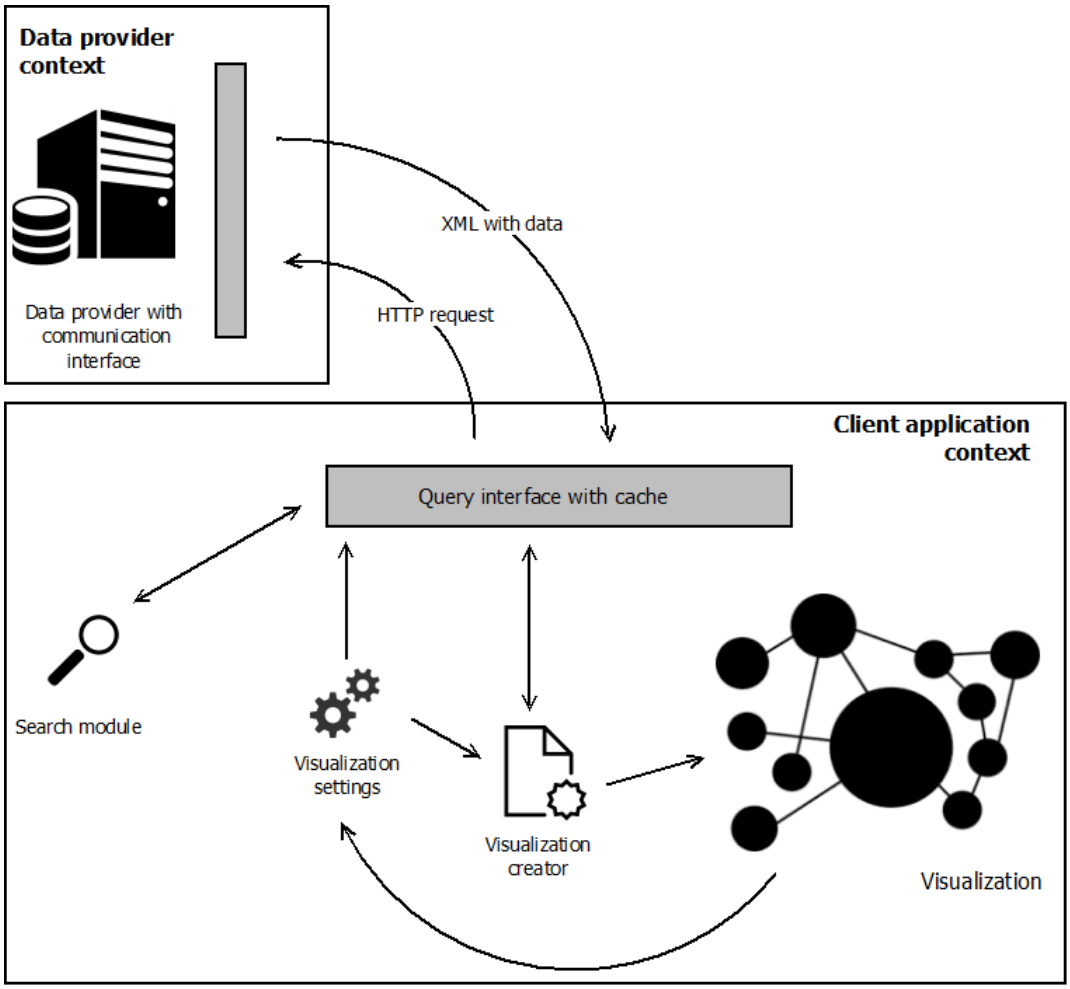
Visualization can be understood as the process of displaying data or information in the form of various objects such as graphical charts, figures, bars and graphs and others. Data visualization is typically applied to data extracted from the underlying information system. This data have various forms that include text, numbers, statistics etc. It is also necessary to process and transform data into the suitable form for visualization so that the visualization tool is able to interact subsequently with users and respond quickly to their requests (Bénédicte, 2015). Processing the data, as well as data mining from the information system, is not always a simple operation. The reason may be in a large amount of data being processed or varied approaches to data representation which the information system architecture is not prepared for.

Theoretical part

Citeviz Tool

CiteViz visualization tool invented by authors of the article and presented in Klapka (2013), allows to visualize the relationships between scientific publications. This tool has the form of the client applet that has the interface implemented to retrieve data from a remote server using web services. The solution proposed in Klapka (2013) has its own REST-based server request interface. The server uses its own, manually managed database that provides data to the client's visualization application.

Figure 1 Overall diagram of the visualization tool CiteViz with basic components (Source: author)



CiteViz can visualize database records in various ways and provide selected graphic representations of data. This visualization tool provides easy navigation and enables significant help to better understand of the relationships between scientific publications. The overall diagram of the visualization tool is captured in the Figure 1.

Better user applicability of the CiteViz tool is achieved by communication based on appropriate link to some of the large scientific databases or digital libraries, that collect scientific publications

The core functionality of the tool concerns elaboration and subsequent visualization of the relationships between publications. Digital library linked to the CiteViz should be able to provide in addition to the common attributes the following information about the relationship between records in a structured form:

- **Citation relationships between publications:** It enables to obtain and display publication relationship from the database in both directions (i.e., to obtain for a selected publication a list of publications that cite it and the list of publications to which it refers to).
- **Author's relationships of the publication – authorship:** It enables to provide structured information about the authors, (i.e., it enables to obtain a list of publications for selected author).

- **Citation relationships between/among authors:** It makes it possible to obtain relationships in both directions, (i.e., the list of both citing and referenced authors for the selected author). Compound relationships can also be obtained by combining the citation relationships between publications and the relations of the type of publication – author.
- **Relationships of co-authorship:** They concern the authors' participation in one publication. These relationships can also be obtained by multiple queries for the relations of the type of publication – author.

The core requirement placed on the library is therefore to obtain citation of the two types: relationships between publications and the relationships publication – author in a structured form. The other two types of mentioned relationships can be derived from them. Some individual databases were selected and their ability to provide data in a structured form are described in the following part of the text.

Digital libraries analyze

Every field of human activity produces, requires and uses a great deal of information, and produce and use enormous and ever-increasing number of publications as well as other types of results.

These publications serve as a way of communicating within the scientific community and are very important for further research and development in the area. The processing of information on these publications is tight with the existence of a number of scientific databases (digital libraries) that store information on publications as well as their full texts (Dunne et al., 2018). The number of publications in these libraries is estimated up to several tens of millions (Clarivate Analytics, 2018). These libraries enable search based on several criteria, but the possibilities of presenting relationships between publications are limited (Dunne et al., 2018). For these reasons, it is suitable to use visual analysis options to facilitate the search and data mining in large libraries. In the next part of contribution are presented basic information on and characteristics of selected digital libraries.

Web of Science

Web of Science (part of the Web of Knowledge) at present administered by Clarivate Analytics is a web-based library that has the longest tradition and is well-known for its Science Citation Indexes. This library includes both the citation of scientific articles and regularly updated bibliographic data of articles from more than 10,000 leading world scientific and professional journals covering all fields of science, with more than 70 years of retrospection. Data are collected and available since the year 1945. The citation database covers in its seven parts natural sciences, social sciences and humanities (Web of Knowledge, 2013).

The database offers its own API, to access information to its about 59 million records. Access the API requires a prepaid extra product from Web of Science Core Collection (Clarivate Analytics, 2018).

Google Scholar

Google Scholar provides a simple way to broad search for scholarly literature. It enables search across many disciplines and various resources that include articles, theses, books, abstracts from academic publishers, professional societies, online repositories of universities and other scientific institutions and other web sites. Google Scholar is aimed to find relevant work across the world of scholarly research. Google Scholar ambition is to rank documents the way researchers do, to provide the full text of each document and information about where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature (Google, 2018). Although the Google company is known by providing large number of the Internet technology interfaces, it does not offer any interface available for structured data retrieving from the Google Scholar service (Google Groups, 2015).

Scopus

The basic definition of Scopus is published on their official webpage. It is assumed to be EU competitor to older Web of Science. Scopus in the first paragraph claims (Scopus, 2018): "Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings. Delivering a comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences, and arts and humanities, Scopus features smart tools to track, analyze and visualize research."

In order to access information in a structured form, Scopus has its own API, that is freely available as the part of institutional subscription. Through this API is possible to access all information stored in the Scopus library. It is possible to get all the metadata about publications, including citation relationships. The database also introduces the subject of the author type and provides information about the author's relationships associated with the publications Elsevier (2018). The basis of the entire communication is the API key which is generated and subsequently used for each individual data request (Elsevier, 2018). The communication process is established as a web service via HTTP protocol. API also includes a quite extensive documentation of individual requests.

The SAO/NASA Astrophysics Data System

The SAO/NASA Astrophysics Data System (ADS) is a Digital Library portal for researchers in Astronomy and Physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant. The ADS maintain three bibliographic databases containing more than 12.8 million records covering publications in Astronomy and Astrophysics, Physics, and the arXiv e-prints (SAO/NASA ADS, 2018). Abstracts and full-texts of major astronomy and physics publications are indexed and searchable through special ADS "Bumblebee" interface as well as the traditional "Classic" search forms (SAO/NASA ADS, 2018). ADS tracks citations and usage of its records to provide advanced discovery and evaluation capabilities, is integrated into its databases and provide access to rich external resources of various kinds. ADS currently have links to over 12.3 million of items (SAO/NASA ADS, 2018).

This digital library also offers the API enabling to access data in a structured form. The library allows users to obtain metadata on publications, as well as the citation relationships between publications. Unfortunately, the database does not provide detailed information about the authors who are represented only as one of the text attributes of the publication in the form of author's and institution's name. The principle of communicating with the database is the same as in the previous case of Scopus, i.e., using the authentication token gained through the HTTP protocol.

CiteSeerX

CiteSeerX specializes in publications in the field of computer science and informatics. The library has been available since 1997 and currently contains more than 750,000 publications (The Pennsylvania State University, 2007). It also contains the automated indexing engine that extracts and stores citation and publication statistics. It also includes functionality for extraction and differentiation of author's publications (The Pennsylvania State University, 2007). The library is freely accessible and also includes the API for retrieving information in a structured form. The API is accessible only on the basis of the agreement with the operator (The Pennsylvania State University, 2007).

Citeviz with Scopus API

Based on previous research of the possibilities of obtaining information from digital libraries and due to the scope of the article the Scopus database was selected for connection and analysis. Scopus provides several types of queries over the library through standard institutional subscription (Scopus, 2018). Scopus API allows to retrieve information about the authors and publications in a structured form, but it does not allow to obtain directly all the required types of relationships among these records. It is but possible to obtain complex information by combining several queries. The original man operated database solution enables to get detailed information about multiple records within one request. For Scopus API, obtaining detailed information is limited to one record per request, which requires a large number of requests to retrieve records from the database

The problem of sufficient speed of the entire communication during the implementation process of connection CiteViz tool with the Scopus database raised as a result of Scopus API restrictions. The low speed of communication with the Scopus database is caused by the fact, that many calls are needed to get one complete record (including the relationships to other records), whereas one call lasts about 1 second. From the user's point of view the serial call of the API using the CiteViz tool is almost unusable due to the very long response time for retrieving data from the database.

The CiteViz tool enables a sequence loading of records, whereas visualization can be processed during this data retrieval. The data is sequentially replenished and visualized as it is retrieved from database or library. This feature was utilized for this purpose and further expanded. When the search request contains more records (filter by the attribute ID) there are created multiple threads, among which are all the requests equally distributed. The number of necessary threads is then calculated based on the number of requests up to the largest possible limit of threads. Thus, loading occurs in parallel in several threads, which makes the process of data loading significantly faster.

The previous optimization partially solved the problem of speed, but it may still take longer to get some responses to queries. Although the CiteViz tool has an in-memory cache, this cache is deleted after the session when the tool is finished. In order to build easily on the previous work of the user without long waiting to download the information already retrieved, the current in-memory cache has been extended and disk cache was added. In-memory cache then works in LRU (Last Recently Used) mode, and the longest unused records are put into a file on the user's computer drive. After user exits the tool, all the in-memory cache is stored in the disk cache. Consequently, once the tool has been restarted, previously downloaded records are loaded from user's computer drive without the need to download them from the Scopus database.

Methodology

The tool was used in the tourism application area to acquire a relatively comprehensive panoramic view of the main research directions, the most important research activities, finding substantial and prestigious results. Other objectives were the discovery of research cooperation at national and international level and the improvement and efficiency of the research work of research groups at the University of Hradec Králové. For these reasons, the appropriate Scopus database, which is the EU's citation index and the main competitor of WoS, was used.

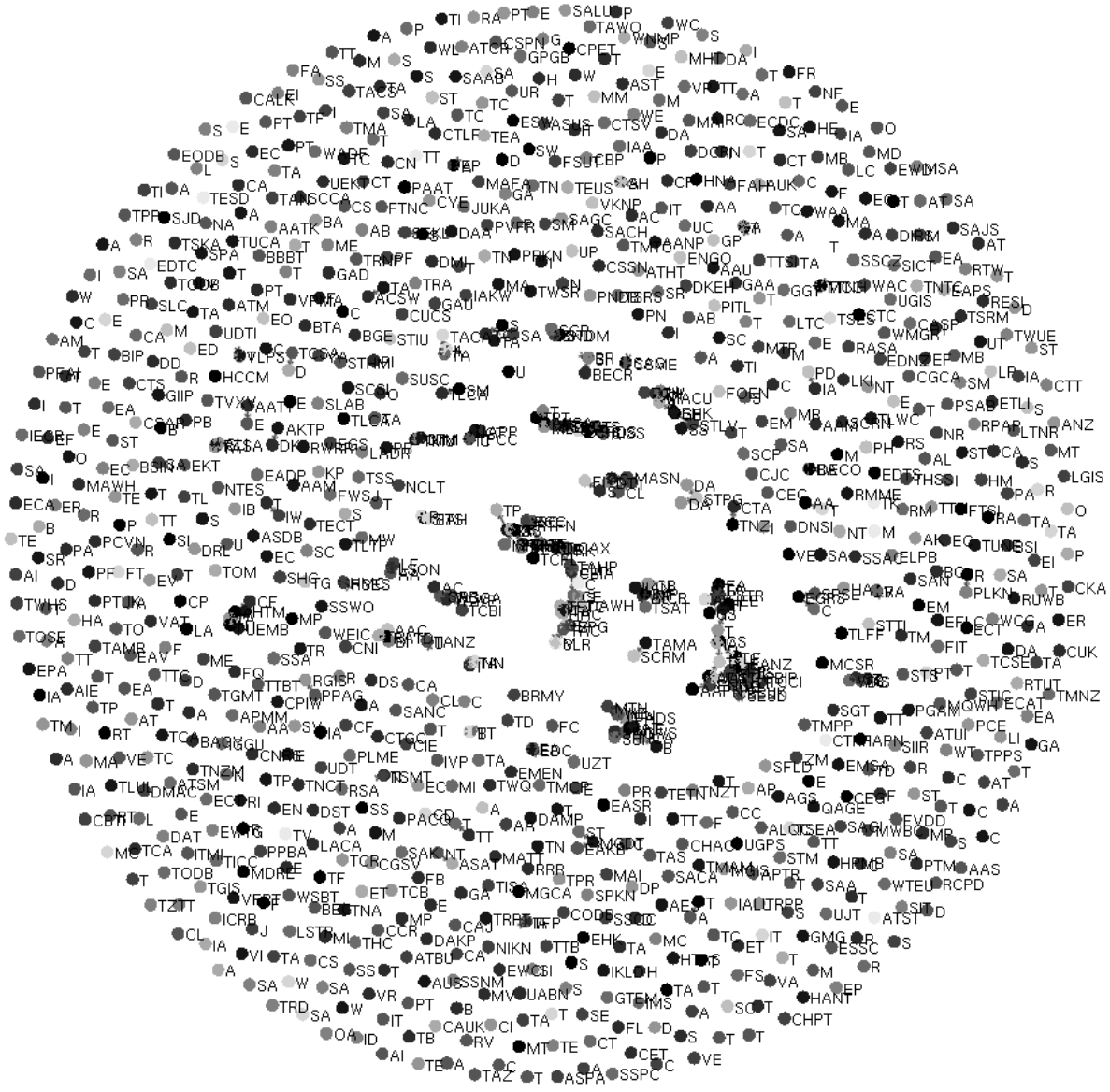
The analyzed dataset was obtained through an iterative process of refining the search query. The keyword "sustainable tourism" was chosen as the starting point of the process. The results were restricted to the interval 1988–2017. First year in the interval –1988 was chosen as the year following the publication of the 1987 Brundtland Report (World Commission on Environment and Development, 1987), which in fact laid the foundation and defined the importance of the phrase sustainable development (McLennan et al., 2015; Jiaying, Sanjay, 2009). The total of 8137 results were returned from the database in the first phase, with 93%, i.e., 7591 search results in English. Subsequently the dataset was limited to English publications only due to the insignificance of other languages. The search results exclude publications that match the key words but do not fall into expected fields of study required, i.e., the fields of Social Sciences, Business, Management and Accounting, Environmental Sciences. These phrases are at the same time the most frequently represented as expected. The selection was then again restricted so that to include only publications in Journals, Books and Conferencing Proceedings. The total number of analyzed publications in the resulting dataset was 5650. Finally, a lot of testing was executed over this dataset to check the advanced analytics of CiteViz.

Results

Overall citation coherence

The task of this scenario was to measure the overall citation linking of Scope's publications in the field of tourism. The total of 1000 most popular publications (measured not only by internal citations in Scopus, but also by external citations) were selected from the basic dataset. The resulting citation network is shown in Fig. 2.

Figure 2 Structure of citations among one thousand of the most cited publications in the analyzed dataset. The publication year is represented by a color transition interval from white (oldest) to black (the latest) (Source: author)

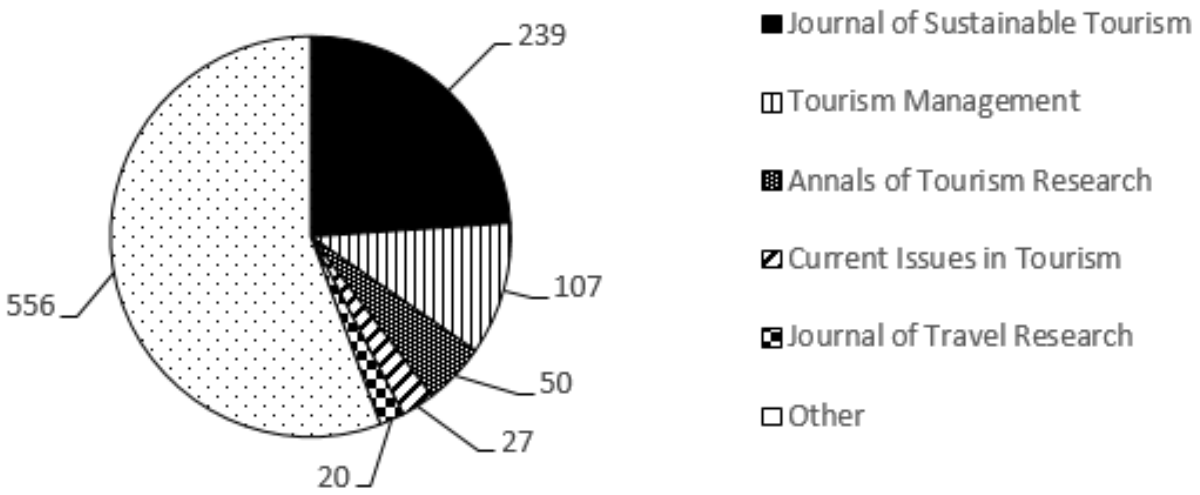


The mutual citation between records in the database is relatively small, as can be seen from the figure. There are only 9 components larger than 5 in the graph. The two largest components have 33 and 30 nodes. Based on the highlighted year of publishing in the citation network, it has also been shown that publications having citations are more likely published after year 2000.

Examining of mutual citations surprising fact was discovered. Most of the authors of the New Zealand publication are cited among themselves, although the percentage of these authors is only 4.5%. Publications from United Kingdom (18.7%), United States (13.9%) and Australia (13.7%) and the source type, Journals (96%) belong to the mostly represented in the network. Representation of individual sources (journals) is shown in Fig. 3. The most citation between

publications has the Journal of Sustainable Tourism, which can be explained by the fact that almost a quarter of the publications were published in this journal.

Figure 3 Representation of the most important sources of the most cited publications on sustainable tourism (Source: author)

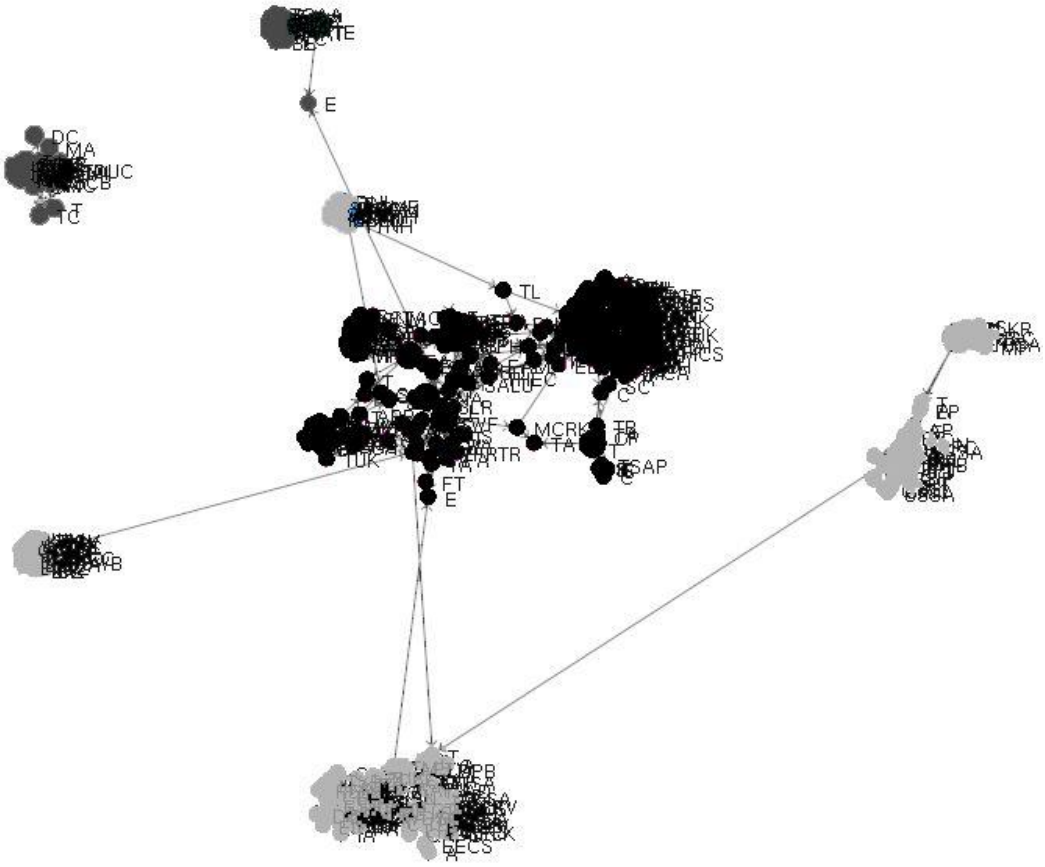


Subject area influence

The impact of sustainable tourism on the various disciplines was monitored in this scenario. The total of 15 most-cited publications and their citations were selected from the original dataset. These citations were not restricted to the original dataset. Subsequently, the representation of individual fields was analyzed in the created citation network, totaling 380 nodes and 395 edges. The analysis of this scenario did not lead to surprising findings and the vast majority of the fields into which the publications in the citation network are classified correspond to areas that can be classified as "sustainable tourism".

The fields, from which the selected publications are drawn, we are also monitored within this scenario. In this case, a much greater interconnection of individual publications was detected compared to the previous case, as shown in Fig. 4, where the citation network was subjected to a Girvan-Newman algorithm cluster analysis (Girvan, Newman, 2001).

Figure 4 The overall consistency between the publications referred to by the 15 most-cited publications (Source: author)



A surprising finding was the relatively frequent occurrence of Environmental Chemistry (3.44%) and Earth and Planetary Sciences 2.54%. The following table shows the most represented fields in the citation network.

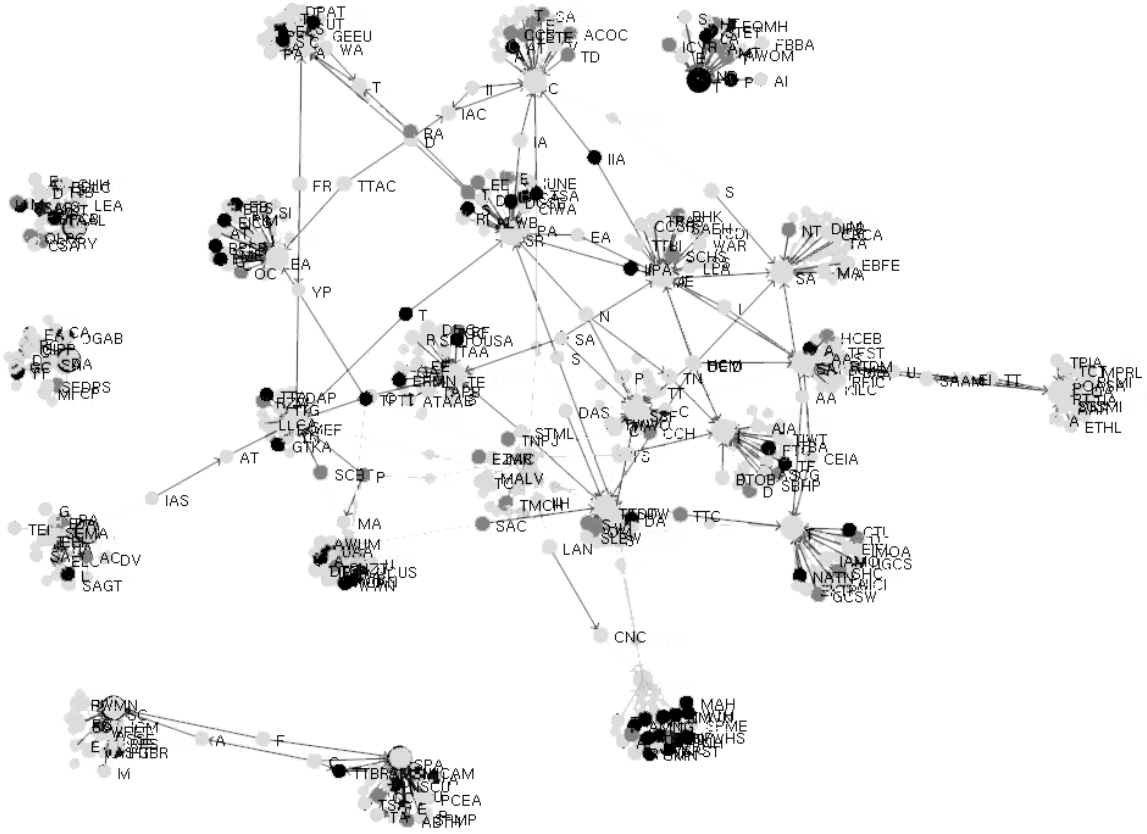
Table 1 Representation of individual fields in the references of the most cited publications (Source: author)

Subject	Count	Percentage
Tourism, Leisure and Hospitality Management	87	13,0%
Geography, Planning and Development	67	10,0%
Development	43	6,4%
Ecology	43	6,4%
Environmental Science	41	6,1%
Management, Monitoring, Policy and Law	34	5,1%
Transportation	32	4,8%
Strategy and Management	27	4,0%
Ecology, Evolution, Behavior and Systematics	25	3,7%
Environmental Chemistry	23	3,4%
Earth and Planetary Sciences	17	2,5%

International influence

The total of 25 most popular publications were collected and their citations were retrieved from Scopus. Consequently, cross-country linkages were monitored in this scenario. It can be inferred from the citation network shown in Figure 5 that, for example, within the UK territory (black) there is a relatively closed community of scientists who cite each other, but do not use citations of publications from other countries. On the other hand, China (dark gray) has the 4th highest representation in the citation network, but neither publication is in the original data set of 25 most popular publications.

Figure 5 A sample of brightness-coded countries in the citation network, with hidden records that are insignificant in terms of analysis (Source: author)



Conclusion

In this paper, the possibilities of retrieving data from digital libraries for the CiteViz visualization tool developed within Klapka (2013) have been explored. In consideration of all the possibilities of individual libraries and possible scope of the article it was chosen to implement connection with the Scopus library.

During the implementation of a connection to the Scopus, the speed of the data retrieving has proven to be the biggest problem. This is caused by the fact that several Scopus API calls have to made to get one complete record including all citation relationships.

Consequently, it was necessary to optimize the data retrieval and caching system in CiteViz. Optimization process resulted in using several parallel threads to retrieve the data, which shortens the communication of the visualization tool with the underlying databases. The data caching system has been expanded by the disk cache that enabled not download repeatedly the already downloaded data even in the case, when the work with the tool is interrupted and subsequently the tool restarted, and the work continues.

There is also way to optimize the speed of data acquisition from the underlying database is to implement the so-called "Shared Cache" when records retrieved by one of the users are uploaded to the shared CiteViz server and made available to other users who work with the tool lately. The data obtained by one user tool could therefore be used in other user's sessions.

The resulting solution was tested on selected scenarios in the field of "sustainable tourism". The tool has revealed a number of otherwise concealed patterns mentioned in this test, such as the very small inter-linkage between the various publications and the very frequent citation among the New Zealand authors. It has also been revealed that the most cited publications often draw on Environmental Chemistry and Earth and Planetary Sciences, which do not fall under the concept of sustainable tourism.

Acknowledgment

The financial support of the Specific Research Project "Information and knowledge management and cognitive science in tourism" of FIM UHK is gratefully acknowledged.

References

Bénédicte, L. G. (2015). How can data (and graph) mining techniques support research in information systems. In *9th International Conference on Research Challenges in Information Science, Athens, Greece, 13 – 15 May 2015*. DOI: 10.1109/RCIS.2015.7128857.

Clarivate Analytics (2018). It's time to get the facts. Retrieved from http://images.info.science.thomsonreuters.biz/Web/ThomsonReutersScience/%7bd6b7faae-3cc2-4186-8985-a6ecc8cce1ee%7d_Crv_WoS_Upsell_Factbook_A4_FA_LR_edits.pdf.

Clarivate Analytics (2018). Web of Science Web Services (APIs). Retrieved from http://wokinfo.com/products_tools/products/related/webservices/. Accessed 6 Feb 2018.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B. (2018). *Rapid Understanding of Scientic Paper Collections: Integrating Statistics, Text Analytics, and Visualization Department of Computer Science*.

Maryland, USA: University of Maryland. Retrieved from http://www.cs.umd.edu/~ben/papers/Dunne2011Rapid.pdf. Accessed 20 Jan 2018.

Elsevier (2018). Elsevier Developer Portal. Retrieved from https://dev.elsevier.com/. Accessed 14 Jan 2018.

Fixot, R. S. (1957) *American Journal of Ophthalmology*.

Girvan M., Newman J. (2001). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.122653799.

Google (2018). *About Google Scholar*. Retrieved from https://scholar.google.com/intl/en/scholar/about.html. Accessed 7 Feb 2018.

Google Groups (2015). *Is there a possibility to provide API access for Google Scholar*. Retrieved from https://productforums.google.com/forum/#!msg/websearch/CZFYeaZp_7c/CPRHHVfXAQAJ. Accessed 7 Feb 2018.

Jensen, E. (2008). *Brain-based learning: The new paradigm of teaching*. Thousand Oaks: Corwin Press.

Jiaying, L., Sanjay, K. N. (2009). Sustainable tourism research: an analysis of papers published in the Journal of Sustainable Tourism. *Journal of Sustainable Tourism*, 17(1), 5–16.

Keim, D. A., Mansmann F., Schneidewind, J., Ziegler, H. (2006). Challenges in Visual Data Analysis. In *10th International Conference on Information Visualization, London, England, UK, 5 – 7 July 2006*. DOI: 10.1109/IV.2006.31.

Klapka, O. (2013). *Visualization Analysis, Master Thesis, Faculty of Informatics and Management*. Hradec Králové: University of Hradec Králové.

McLennan, Ch.-L. J., Moyle, D. B., Weiler, B., Ruhanen, L. (2015). Trends and patterns in sustainable tourism research: a 25-year bibliometric analysis. *Journal of Sustainable Tourism*, 23(4), 517–535. DOI: 10.1080/09669582.2014.978790.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press.

SAO/NASA ADS (2018). *ADS Home Page*. Retrieved from http://adswww.harvard.edu/. Accessed 6 Feb 2018

Scopus (2018). *About Scopus*. Retrieved from https://www.elsevier.com/solutions/scopus. Accessed 6 Feb 2018

Techopedia (2014). *Data Visualization: Data That Feeds Our Senses*. Retrieved from https://www.techopedia.com/2/29217/trends/big-data/when-data-visualization-works-and-when-to-avoid-it. Accessed 15 Jan 2018.

The Pennsylvania State University (2007). *About CiteSeerX*. Retrieved from http://csxstatic.ist.psu.edu/about. Accessed 7 Feb 2018.

Web of Knowledge (2013). *Web of Science Quick Reference Guide*. Retrieved from http://wokinfo.com/media/pdf/qrc/webofscience_qrc_en.pdf. Accessed 6 Feb 2018.

World Commission on Environment and Development (1987). *Our Common Future*. Oxford, UK: Oxford University Press.