An Automatic Method for the Analysis of Scientific Literature Collection

Adrian-Mihai CUSTURĂ, Lidia BĂJENARU, Florin POP

The National Institute for Research & Development in Informatics, 8-10 Mareşal Averescu Avenue, Sector 1, Bucharest, 011455, Romania

mihai.custura@ici.ro, lidia.bajenaru@ici.ro, florin.pop@ici.ro (**Corresponding author*)

Abstract: When it comes to scientific and technological developments of any kind, the first step in producing them consists of research. This is the only possibility to understand a problem and develop a solution. As the number and percentage of publications available in a digital format keep increasing, research is bound to become easier and more organized. Grouping publications on similar topics into one literature collection and performing an analysis in order to gain knowledge about a subject can be a difficult and time-consuming process. In this paper, an innovative approach that automatically provides insights, results and statistics on a given set of scientific publications is proposed.

Keywords: Scientific literature analysis, Water management, SurVis, Data projection, Visualization techniques.

1. Introduction

Given that research has already proven its benefits, it is a continuously developing domain, as the number of scientific papers published each year keeps increasing, which results in more discoveries and more scientific and technological advancements. But with the available depth of knowledge and the ever-growing collection of scientific publications, being able to store and organize related publications becomes a necessity, as well as a challenge. Building and managing a literature collection can provide insights into the development and current situation of a specific domain or into the work of a researcher or a research group and requires going through the following steps:

1. Selecting the collection topic;
2. Choosing the data format;
3. Gathering publications;
4. Extracting the most relevant information;
5. "Presenting" (Neagu, 2016) the selected information in a user-friendly, easily understandable manner;
6. Projecting the data on as many levels as possible and elaborating statistics.

This paper focuses on providing an alternative to the existing solutions in scientific literature management and analysis, whose main purpose is to perform a thorough analysis of the publications and to offer relevant results and statistics. In order for such an analysis to be conducted, certain objectives and requirements need to be met:

1. Aggregation of the information regarding the publications;
2. Data reduction and projection;
3. Information retrieval;
4. Multiple "visualization" (Neagu, 2016; Niazi & Hussain, 2011; Renfro, 2017) techniques;
5. Display of "statistics" (Neagu, 2016);
6. High customization capabilities.

Our proposal, an automatic analysis tool, is an extension of the "SurVis" (Beck, Koch & Weiskopf, 2016) online browser for scientific literature collection analysis. It combines 2 types of literature collections, as well as 2 kinds of visualization techniques: a collection of 100 publications from the last 40 years on water management topics (with domain visualization – focusing on water management) and a collection of 50 publications whose authors/collaborators are part of the Computer Science Department at The Faculty of Automatic Control and Computer Science of Politehnica University of Bucharest, Romania (with research group visualization – which brings together the work of researchers from the same community).

The paper is structured as follows. Section 2 will describe the existing solutions for performing scientific literature surveys, focusing on frameworks and projects used for reference management. Section 3 provides an overview of the architecture of the proposed solution, specifying its components and the concepts used in its development. Section 4 offers details regarding the technologies used in the implementation. Section 5 is focused on functionalities and results

https://doi.org/10.24846/v27i4y201806

and provides a comparison with the existing solutions. Section 6 draws a conclusion and summarizes the future work.

2. Existing solutions for literature survey

2.1 Zotero

"Zotero" (Trinoskey, Brahmi & Gall, 2009) is an open-source reference management tool used to save and manage bibliographic data about research materials found on the Internet. It is a browser extension that can automatically identify if a webpage is a bibliographic item and it allows for the full reference information about such items to be saved to the Zotero library. It can also save a copy of the webpage or the full text PDF and it allows the user to manually add entries (along with notes, tags, attachments and metadata) to the library via the Zotero framework. Furthermore, the framework provides the possibility of creating a bibliography from a group of selected publications and viewing it in different formats, printing or saving it in a word processor. Zotero comes with a very useful feature, namely the integration with word processors, which enables it to be used as an extension of such a processor, in order to add in-text citations. Upon finishing the paper, Zotero inserts all items that were referenced from the Zotero library.

2.2 Mendeley

"Mendeley" (Zaugg et al., 2011) is a reference management application, available as a desktop, web and mobile app, used to gather and manage research publications and data. One of the fundamental differences when comparing it with "Zotero" (Trinoskey, Brahmi & Gall, 2009) resides in the fact that it allows for online sharing and collaborating, through a social network for researchers. It works with PDF files, providing the automatic extraction of bibliographic data, smart filtering and tagging. It offers full-text search across papers and has an integrated PDF viewer with sticky notes and text highlighting. While it supports fewer file formats than Zotero, Mendeley innovates through the social networking features, such as newsfeeds, profile pages, comments, the public groups used to share reading lists and the private groups for analysing papers in a collaborative manner. The users receive statistics regarding their reading at paper, author and publication level.

2.3 SurVis

"SurVis" (Beck, Koch & Weiskopf, 2016) is a flexible and extensible online browser used to develop and view statistics regarding a scientific literature collection. It is highly customizable, since the user controls the input (the list of publications and the information about them), as well as the output, by modifying the source code (since the use of SurVis requires an installation of the entire project, all modifications to the source code affect only the current installation). Its main advantages are the user-friendly interface and the statistics it displays, such as publication years, keywords, authors, series and clusters of publications. The most important feature is represented by the selectors (up to 6, which may be years, keywords, authors, series, clusters or even publications) that allow for the publications to be selected and ordered by their relevance towards the respective selectors.

3. Architecture of the proposed solution

3.1 Main Architecture and Components

The architecture of the proposed automatic analysis tool is shown in Figure 1, while each component is detailed in the following subsections.

3.1.1 Input

The input consists of 2 "BibTeX" (Patashnik, 1988) files, corresponding to the 2 visualization techniques exemplified in this paper: domain and community. The entries in the referecesWater.bib file were collected from the 100 most relevant publications on Google Scholar by issuing a „water management" query, while the community literature collection was gathered by collecting the top 5 most cited papers belonging to each author from the community. The connection between the input component and the main component consists of 2 steps: parsing the input and checking if any modifications occurred in the input files.

3.1.2 Application Logic

The entire application logic is written in the JavaScript programming language. The architecture of this component is highly modular,

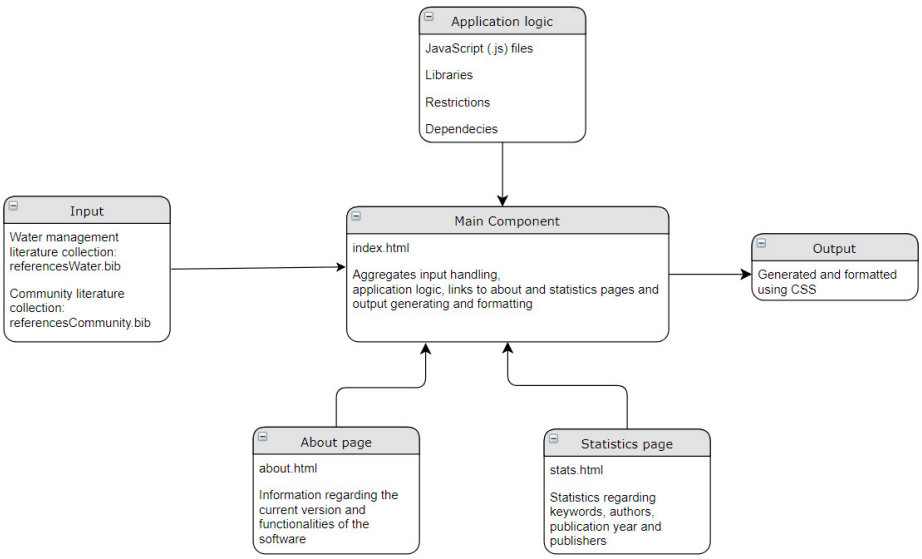

Figure 1. Main architecture and components

separating tasks between subcomponents and allowing for easy interactions between them.

3.1.3 About Page

This window contains information regarding the current version of the online browser. It specifies that it is an extension of "SurVis" (Beck, Koch & Weiskopf, 2016) and details the modifications that were implemented.

3.1.4 Statistics Page

This window displays the most relevant information regarding the literature collection used in the survey, in the form of a key-value table. The data was extracted from the output of the page (e.g. keyword and author frequency), as well as from the "BibTeX" (Patashnik, 1988) files (e.g. publisher with most papers).

3.1.5 Main Component

The index.html brings together the functionalities of all the other components. It sets the project properties and requirements and then starts the application logic, through which the input is handled and the output is produced.

3.1.6 Output

The output behaviour is controlled by the application logic, namely the JavaScript files. The properties of each object on display are defined by the Cascading Style Sheets component of the solution and the objects are rendered via HTML sections.

3.2 Data Structure and Aggregation

3.2.1 Data Structure

As mentioned in subsection 2.2.1, the input data is stored in the "BibTeX" (Patashnik, 1988) format. This format was chosen because it provides a relational key-value data model: each tag within an entry is followed by the „=" sign, which assigns a value to the tag.

3.2.2 Data Aggregation

The online browser aggregates the data sets for each publication in the input files and presents them in a simplified and organized manner. The publication list section provides visualization of the most relevant information regarding each publication, such as title, abstract, authors, journal/ conference and keywords, as well as links to the website it originates from or to search engines. The statistics section displays data such as number of publications per year, keyword frequency and number of publications each author contributed to.

3.3 Data Reduction and Search Capabilities

In order for a thorough analysis to be performed, the data needs to be projected using one or more criteria (selectors), thus obtaining the most relevant publications in regard to the selected criteria:

a) year selector;

b) keyword selector;

c) author selector;

d) series selector;

e) cluster selector: group of similar publications

f) publication selector: computes similarities between the selected publication and the others and sorts the results descendingly.

3.4 Visualization Techniques

The concept of "visualization" (Neagu, 2016; Niazi & Hussain, 2011; Renfro, 2017) is influenced by the user requirements, which shape the publication collection provided as input. There are several visualization techniques that can be applied in this context:

a) individual visualization: a researcher may gather all of his/her publications in order to obtain statistics regarding research domains and subjects or collaborators;

b) research group visualization: members of a research group can collect all of their papers in one place, thus being able to share more details about their discoveries and results, as well as gaining an insight into the general results and direction the group is focusing its work on;

c) domain visualization: gathering papers from the same domain results in a detailed view of that domain – the main branches of the domain, the components of each branch, existing or potential problems, existing and proposed solutions, factors influencing each problem and solution, the most important researchers;

d) institution visualization: this type of visualization can be used in order to centralize information about all the research domains the institution focuses on, in order to gain a clear perspective on the progress and results achieved in every domain;

e) subject visualization: this technique is similar to the domain visualization technique, the only difference being that it has a higher level of detail, focusing on a single branch, instead of an entire domain (e.g. agricultural water management is a branch/subject of the water management domain);

f) excellence level visualization: a user may put together a collection of publications that he/she found inspiring or interesting; the results consist of identifying the user's areas of interest and favourite researchers, journals and/or conferences.

3.5 Analytics Functions

Performing a complete "analysis" (Neagu, 2016) of the literature collection provided as input is the primary objective and feature of the proposed solution. Its accomplishment requires extracting and displaying as much information as possible while preserving an easily readable and interpretable output. In order to enable the user of the online browser to gain as much knowledge as possible about the publications, the following analysis results are displayed:

1. timeline – a chart presenting the number of publications per year;

2. keywords – a list of all keywords used in the publications, sorted by category and number of appearances;

3. authors – complete list of researchers whose work is part of the literature collection;

4. clusters – the possibility of grouping publications based on similarities between them using one or more criteria;

5. publication list – each entry displays the most relevant information regarding the publication it corresponds to;

6. statistics – a table containing the most relevant statistics according to the user.

4. Technologies and Research Methodology

4.1 Technologies

4.1.1 HTML (Hypertext Markup Language)

HTML is the standard mark-up language used for developing webpages and web applications. The workflow is as follows: the web browser receives an HTML document from a web server or from local storage and interprets the content of the document in order to render it into a webpage. The contents of a webpage are called blocks and are defined using tags. The most common way to create webpages with HTML is by embedding programs that define the behaviour of the web page when certain events occur (e.g. JavaScript

as the programming language) and by including CSS in order to define the layout of the objects.

4.1.2 JavaScript

JavaScript is a high-level interpreted programming language. A high proportion of websites use it, as it enables and handles interaction between the user and the webpage. JavaScript was initially designed for the client side of web applications, but its high availability and versatility, as well as its low restrictions, led to it being embedded in server-side and database software, word processors, PDF software, as well as in mobile and desktop applications.

4.1.3 CSS (Cascading Style Sheets)

CSS is a stylesheet language designed for formatting the layout of contents in a document written in a mark-up language. The most important functionality of this language is that it enables the separation of content and presentation, which makes it possible for multiple pages to use the same format defined in a single .css file. Furthermore, using a separate file to define the content layout properties helps reduce complexity and avoid unnecessary duplicate code. The separation of content from its format allows CSS to display the same webpage using different styles for different rendering methods. The properties that CSS can define include layout, colours, margins, borders and fonts.

4.2 Research Methodology

Figure 2 specifies the research methodology, similar to the one used by Ochoa et al. in 2018.

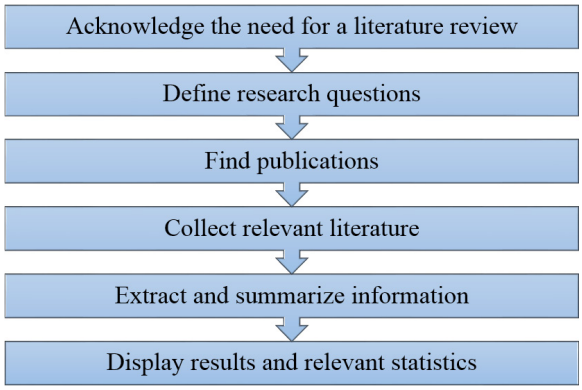

Figure 2. Research methodology

When performing a scientific literature analysis, the first step is to identify a topic or a situation that requires a literature review. Once the focus of the survey is set, the reviewer needs to define the research questions. The questions are then used as queries in order to find publications, but before a paper is to be added to the literature collection, the reviewer needs to read the abstract (or even the full text, if possible) in order to ensure that each publication is relevant to the purpose of the research. Once the publication list is completed, the information it contains needs to be parsed and summarized. Thus, only the most important pieces of information get to be included in the results.

5. Result Evaluation and Interpretation

5.1 Functionalities Proof

5.1.1 Timeline

The timeline displays all the publication years corresponding to the papers in the literature collection; in the context of this paper, the years range from 1978 to 2018. For each year, the number of publications is displayed, the entire timeline being formatted as a bar chart. Any year, including those with no associated papers, can be used as a selector: the publication list is sorted such that the papers published in the selected year are displayed at the top of the list; if more years are selected, publications corresponding to one of the years used as selectors are moved to the top of the list. Figure 3 shows the timeline for the literature collection used by the proposed solution with a year utilized as selector.

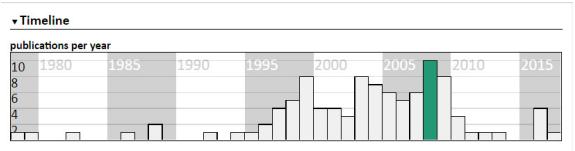

Figure 3. Timeline section and year selector

5.1.2 Keyword List

The keywords section of the webpage lists every keyword associated with any publication in the input files. The keywords are split into categories, with those belonging to the same category being sorted descendingly according to their total number of appearances in the literature collection. Keywords can also be used as selectors, which results in publications containing the selected keyword appearing at the top of the list. Multiple keyword selectors may be used, and, unlike the year selectors, a publication may agree with more than one keyword selector. Figure 4 shows the keywords section (keywords with at least 3 appearances).

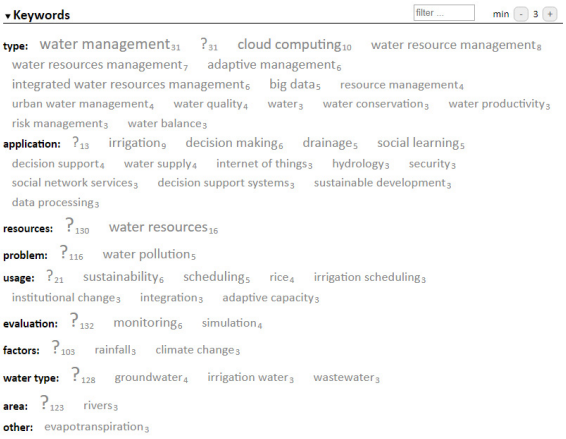

Figure 4. Keywords section

5.1.3 Author List

The authors section of the webpage mentions every author that wrote or collaborated in at least one of the publications forming the literature collection. Much like the keywords section, it lists the authors in descending order in relation to the number of publications and offers the possibility of marking one or more authors as selectors. Figure 5 represents the output of this section (for authors with at least 2 publications).

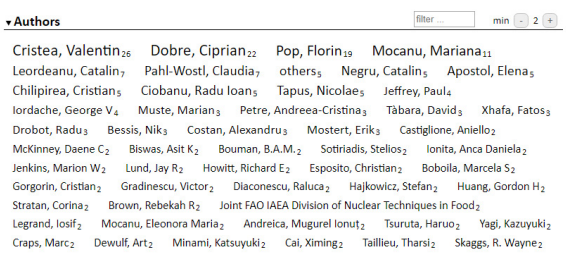

Figure 5. Authors section

5.1.4 Series

The series section is used to display the series the papers used in the literature collections were published in. The elements of the list are displayed descendingly according to the number of publications. Figure 6 displays the series section.

Figure 6. Series section

5.1.5 Clusters

The clusters section is the last part of the control group of components. It is the most complex of these components, since its purpose is to provide the user with the possibility of grouping publications by computing similarities between them in relation to one or more of the following criteria: keywords, authors, author community. This feature is particularly useful in order to discover related publications, which may result in a better understanding of the literature collection and its underlying details. Multiple classifications can be created with different numbers of generated clusters.

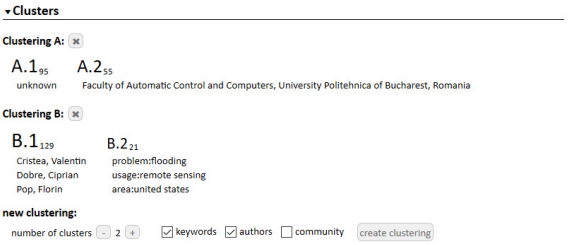

Figure 7. Clusters section

Figure 7 displays the clusters section, along with an example of generated clusters.

5.1.6 Selectors

When a selector is chosen, the following events occur:

- the selected tag appears in the selectors section and is associated with a colour (Figure 8);

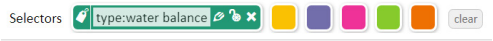

Figure 8. Selector appearance

- publications are sorted descendingly according to the selector agreement level (in the case of multiple selectors, the sorting criteria is represented by the average of the selector agreement levels);

- other tags (e.g. years, keywords, authors, series, clusters) are marked with a vertical bar, in case the selector agreement level is different than 0 (the selector agreement level is not equal to 0 in case there are any publications associated with both tags).

Figure 9 displays an example of multiple selectors from different tag categories being used at the same time. Using multiple selectors, especially related ones, can provide a more in-depth analysis of the literature collection. Furthermore, the selectors section also provides search capabilities, which leads to the search query being used as a selector. Finally, the selectors have 2 more functionalities:

1. invert: by applying the „invert" function on a selector, the selector agreement level of each publication is replaced with its 1 complement (e.g. 1 becomes 0, 0.43 becomes 0.57) – the

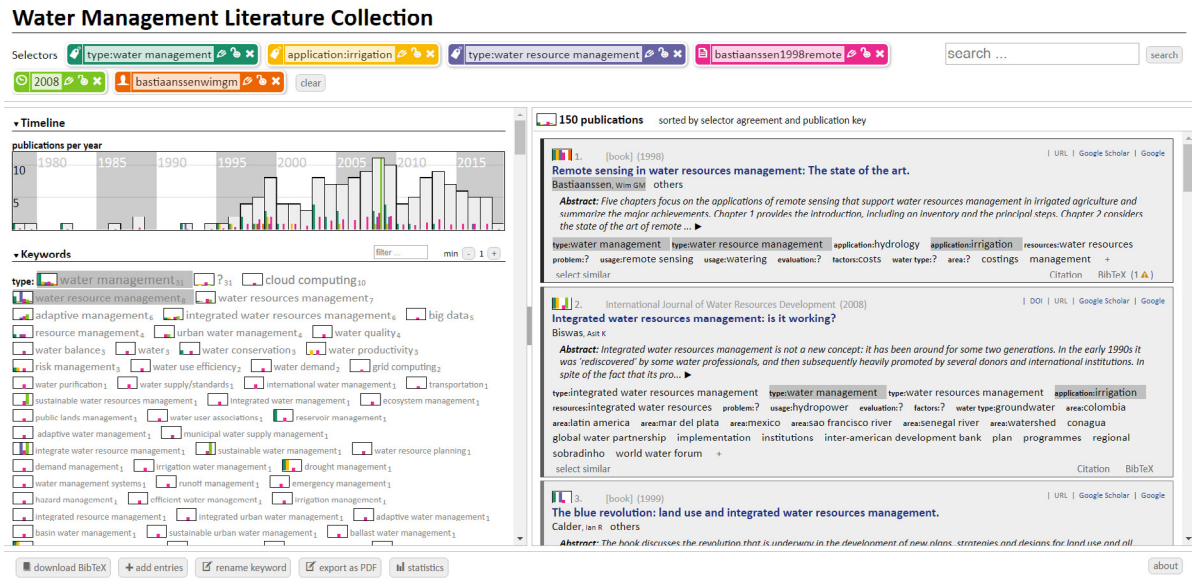

Figure 9. Applying selectors

least relevant publications have the highest levels, which is useful for queries such as „publications that do not contain keyword X";

2. lock: applying a „lock" on a selector means that the result list will display only the publications that have an agreement level of 1 in relation to that selector; if no publication meets this requirement, all publications are displayed.

5.1.7 Results

The results section of the webpage displays the list of publications in the literature collection. If no selector is in place or different papers share the same selector agreement level, the results are sorted alphabetically in a numbered list, according to the citation keys.

5.2 Looking Glass – Presenting the Overall Systems

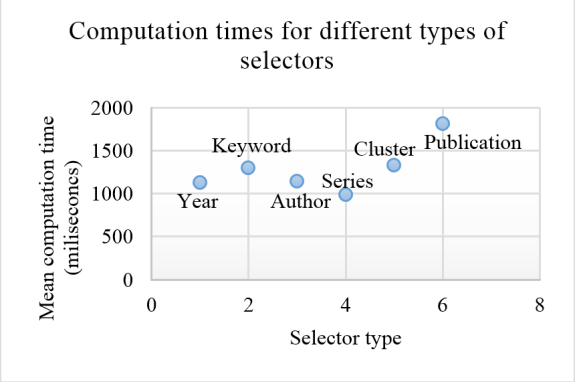

Figure 10. Mean computation times for different types of selectors

Since the most important feature of the proposed solution is represented by the literature analysis, the performance of the tool is given by the time it takes to perform different types of analyses. Figure 10 illustrates a chart displaying the mean times for computing publication similarities and adjusting the interface and results for all available types of selectors.

The lowest times are the ones obtained using a series selector, since most publications in the literature collections were not published in any series. The highest times are the ones measured using publication selectors, since computing similarities between publications involves complete comparisons between the selected publication and the others. Similar times were measured for year and author selectors, since both are easy to find amongst the information about publications and, by contrast with series, are both present in every "BibTeX" (Patashnik, 1988) entry. An interesting result is that similar times were computed using keyword and cluster selectors. This can be explained by the fact that on the one hand every publication has multiple keywords, hence the search for a particular one is more time-consuming than, for example, the search for the publication year, while on the other hand the inclusion of a paper inside a cluster is also more difficult than the use of year and author selectors, as clusters also require computing similarities between papers.

In comparison with the existing solutions, the proposed one has its own advantages and

disadvantages. The most important differences in relation with each alternative are enlisted below:

1. "Zotero" (Trinoskey, Brahmi & Gall, 2009): from the literature collection analysis point of view, the proposed solution is of more interest, as it provides insights into a particular topic, while Zotero only offers the possibility of creating a bibliography from a group of publications, without performing an analysis; on the other hand, Zotero is a much more complete framework, with better integration and more functionalities, such as automatic publication discovery and information gathering, as well as supporting multiple file formats;
2. "Mendeley" (Zaugg et al., 2011) : although it provides statistics about papers, authors and publications, this solution does not perform complex analyses of literature collections; instead, it offers cross-platform integration and automatic extraction of metadata from PDF files, as well as a researcher social network that enables sharing and collaboration between users; both solutions work only with the "BibTeX" (Patashnik, 1988) format;
3. "SurVis" (Beck, Koch & Weiskopf, 2016): since the proposed solution is an adapted and extended form of SurVis, the two are very similar; the proposed solution is an adaptation on water management and community publications and, more importantly, an extension of SurVis from a theoretical viewpoint, as it introduces 6 different types of visualization techniques mapping to 6 distinct use cases; in terms of publication analysis, it added a classification of the keywords, author community as a clustering criteria and the extraction of statistics containing the most important results of the analysis.

Compared with the other solutions, the proposed one lacks in terms of number and variety of functionalities, as well as in terms of user interaction and experience. But we argue that its most important functionality, namely the analysis of the literature collection, offers more relevant and comprehensive results, as well as thorough awareness regarding the research publications it reviews.

6. Conclusion and future work

In this paper, an automatic analysis was conducted, an innovative approach that automatically provides insights, results and statistics on a given set of scientific publications. Reviewers benefit from high availability and customisation capabilities, detailed analysis of the literature collection, easily understandable and valuable results, as well as viewing similarities and correlations between publications.

The main future developments include: complete implementation of the experimental features; full text analysis of the papers; automatic literature discovery and relevance assessment through a web crawler that uses "data mining" (Ioniță & Ioniță, 2016) techniques; improved analysis by using "natural language processing" (Ionescu, Demian & Czibula, 2017) techniques; user-oriented platform.

REFERENCES

1. Beck, F., Koch, S. & Weiskopf, D. (2016). Visual analysis and dissemination of scientific literature collections with SurVis, *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 180-189.
2. Ionescu, V. S., Demian, H. & Czibula, I. G. (2017). Natural Language Processing and Machine Learning Methods for Software Development Effort Estimation, *Studies in Informatics and Control*, *26*(2), 219-228.
3. Ioniță, I. & Ioniță, L. (2016). Applying Data Mining Techiques in Healthcare, *Studies in Informatics and Control*, *25*(3), 385-394.
4. Neagu, G. (2016). Book Review: Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, *Studies in Informatics and Control*, *25*(1), 131-135.
5. Niazi, M. & Hussain, A. (2011). Agent-based computing from multi-agent systems to agent-based models: a visual survey, *Scientometrics*, *89*(2), 479-499.
6. Ochoa, L., Gonzalez-Rojasa, O., Pereira, J. A., Castro, H. & Saake, G. (2018). A systematic literature review on the semi-automatic configuration of extended product lines, *Journal of Systems and Software*, *144*, 511-532.
7. Patashnik, O. (1988). BIBTEXing, *Documentation for general BIBTEX users*.
8. Renfro, C. (2017). The Use of Visual Tools in the Academic Research Process: A Literature Review, *The Journal of Academic Librarianship*, *43*(2), 95-99.
9. Trinoskey, J., Brahmi F. A. & Gall, C. (2009). Zotero: A Product Review, *Journal of Electronic Resources in Medical Libraries*, *6*(3), 224-229.
10. Zaugg, H., West, R. E., Tateishi, I. & Randall, D. L. (2011). Mendeley: Creating Communities of Scholarly Inquiry Through Research Collaboration, *TechTrends*, *55*(1), 32-36.