Word Cloud Analytics of the Computer Science Research Publications’ Titles over the Past Half Century

L. Abazi-Bexheti*, A. Kadriu * and M. Apostolova*

* South East European University/Faculty of Contemporary Sciences and Technologies, Tetovo, Macedonia
l.abazi@seeu.edu.mk, a.kadriu@seeu.edu.mk, m.apostolova@seeu.edu.mk

Abstract – The field of computer science is becoming increasingly important as almost every aspect of our lives is more and more dependent on computers and the Internet. As a result, computer hardware and software technology are rapidly increasing in power and sophistication. Consequently, the data and the research topics and publications in this field are also evolving rapidly.

In this study we used the data form the Computer Science Bibliography DBLP database for the research publications in the field of Computer Science the past 50 years. Further on, we analyze the articles (journal publications) and proceedings as categories with larger numbers of entries and that provide more significant outcomes. The overall dataset is split in decades and for better demonstration of the information and the evolving topics, text visualization tool was applied.

The results generated by using word clouds provide a much faster and easier way to detect the topics and other changing details about the publications in each decade of the past half century.

Keywords - CS, word clouds, DBLP, Computer Science Research Topics

I. INTRODUCTION

The field of Computer Science is one of the most changing and evolving research fields. This change is empowered by even more changing technology supported with exponential growth of hardware capabilities and significant growth in software productivity as well. Consequently, almost every aspect of our lives is progressively influenced and enriched using technology.

On the other side, the data generated by systems, sensors, and other computers is increasing exponentially and the number of research papers published in Computer Science conferences and journals has increased constantly every year [3].

Usually these research papers are in the form of publications in conference proceedings or journal articles and they provide a starting point for many young scholars and researchers. In this regard, there are many publications that deal with the content analysis of a certain number of papers in a given field in a certain period of time and identify gaps or further paths for research [10][11][12]. However, due to the large number of data and publications it is difficult and time consuming to do an overall study of the evolution of the key topics in computer science research. The idea of this study was to track and show how the landscape of Computer Science research topics changed over time by analyzing 'the titles of research publications’ in the past half century.

Based on our previous research where the Computer Science Bibliography DBLP data was used for quantitative structured literature review of research on e-Learning from its appearance till nowadays [6], in this study we broadened our research on the overall research publications in Computer Science [8].

In addition to this, for better demonstration of the information and the evolving topics, word cloud visualization tool was employed.

We considered that this representation will give to the reader a much faster and easier way to detect the topics and other changing details about the publications. By visualization of information we can see the connections and patterns that matter. Additionally, the information can be designed so that it makes more sense and gives the possibility to concentrate only on the important information. Our sense of sight is the fastest of the five senses and the eye is exquisitely sensitive to patterns in variations in color, shape and pattern. Combining “the language of the eye” and the “language of the mind” provide a powerful combination to gather insights from data [4].

The data for this study was gained through the DBLP. This service provides open bibliographic information on major computer science journals and proceedings. DBLP indexes more than 4 million articles and contains titles of articles, their authors, years of publication etc.

Word clouds are interesting to use as a visual aid with possibility to underscore the main topics. A word cloud is built from the words used in a text or subject, in which the size of each word indicates its frequency or importance. Lately, they are very popular, and the users find them very appealing. They provide a visualization form for text that is recognized for its aesthetic, social, and analytical values. On the other hand, there are several researchers that argue on the inadequacy and limitations of the word clouds. They mainly argue that there is a lack of natural reading order in how words are laid out; the font size is not appropriate for communicating quantitative information; the difference in word size might be due to word length rather than value and similar [2][9].
Taking into consideration all the ‘pros’ and ‘cons’ for the word cloud, for a dataset with over 4 million entries, we found word count as a very suitable visualization tool that provides a very interesting fast and informative venture through the key research topics in Computer Science. The readers will easily notice the larger, bold words and understand the main topics in certain periods.

II. METHODOLOGY

The DBLP dataset is downloadable and represents a very interesting resource for investigating the trends in terms of research fields, most productive authors as well as evolution analysis of various aspects of research publications [1][7].

In our analysis, we have created three datasets from the DBLP database, in Json format as shown in Figure 1, Figure 2 and Figure 3.

These datasets contain data for three categories based on types of publications [8]:

- articles (journal publications) - counting 1,952,444 papers.
- conference proceedings - counting 2,361,456 papers.
- books - counting 15,617 books.

Since the focus of our analysis is identifying research trends in the field of Computer Science, we have continued our analysis only on the first two categories: articles (journal publications) and proceedings, which have a much larger number of entries and hence would provide a more significant outcome.

Further on, for each category separately, a word cloud visualization method is used. From the dataset records, for each paper, only the title and the year of publication is extracted. In this way a list of paper titles for each decade is created. From the words used in the titles further on word cloud is created where each word is associated with a frequency value. Also, visual representation of the information is created. We used the representation in which words are positioned in an unordered layout, to optimize the use of space, and frequency values are mapped to font size. Also, different colors are assigned to the words in order to create more visual clouds which turn to be more acceptable by the users.

III. ANALYSIS

A. Analysis of the journal articles dataset

In this part is presented the study of the dataset of articles (journal publications). As mentioned above, this dataset contains 1952444 records distributed from 1950 to 2019. For each decade the word cloud technique is applied and below are presented the results for the past seven decades.

As it can be seen in Figure 4, in the initial stage of journal articles in computer science dominates the term analog, computers and computational methods. As we move toward the next decade 1960-1969, we can notice a significant increase in the number of publications. The number is doubled and the word ‘programming’ and ‘programming languages’ gain focus (Figure 5).
Also, the word ‘contribution’ appears in the title of the publications more and more, which indicates the application and the contribution of computer science studies in various fields.

‘programming’ topic as the focus of the research but at the same time additional topics such as ‘systems’ and ‘design’ are gaining weight as well.

In the next decade, 1970-1979, the number of research papers was four times larger compared to the previous decade and almost 10 times more compared with the initial decade (Figure 6). These numbers speak a lot of the rapid and exponential growth of computer science research and for evolving topics of research. In some decades there is a concrete distinction of the main words (topics) while in some there is a variety of topics with similar frequency but still contribution and applications remain the focus of the research. Another interesting aspect of the word cloud presentation is that besides the key main topics, one could identify some new significant trends that are appearing in the research such as the word ‘digital’ or ‘systems’.

As we continue the journey through the next decade (1980-1989), the number of the records reaches 897 (Figure 7). In this decade, we see the continuation of the ‘programming’ topic as the focus of the research but at the same time additional topics such as ‘systems’ and ‘design’ are gaining weight as well.

In the next decade (1990-1999), with 3709 number of records we see the topic ‘systems’ as the main topic but also topics such as: ‘software’, ‘engineering’, ‘networks’ and ‘design’ are gaining importance in the research journals (Figure 8). Another interesting aspect that can be noticed is that the publications in German language are taking a notable place in the field of computer science research.

Further on, in the first decade of this century (2000-2009), the number of records is 3999 (Figure 9). The research topics are more concrete and narrowed: ‘java’, ‘linux’, ‘windows’, ‘oracle’ and similar. Also, topics such as ‘web’, ‘data’ and ‘learning’ are gaining attention of the researchers. Once again, through the word clouds we see the initiation of the research topics that are
becoming the main topics of research in the next years to come.

Figure 9. Word cloud of titles of journal articles through the time period 2000-2009 (3999 records)

In the last decade (2010-2019) are included the records form the DBLP till the date that it was downloaded and further on analyzed (Figure 10). We see that the topic ‘systems’ is the most frequent word on the research articles titles, showing the growth of interdisciplinary dimension of computer science and its applications. Also, we see that topics such as ‘web’, ‘models’, ‘design’ continuing to be important but at the same time new topics such as ‘intelligence’, ‘social’, ‘security’ and many others are also entering the stage of research in computer science articles.

Figure 10. Word cloud of titles of journal articles through the time period 2010-2019 (3752 records)

B. Analysis of the conference proceedings dataset

Another aspect of our research was to analyze the part with conference proceedings, which is a much larger dataset and can also lead to interesting results.

As it can be seen in the following word clouds the conference proceedings publications provide a step ahead in terms of research topics compared with journal articles. This is mainly due to the fact that the review and publication time of conference publication is usually much shorter compared to the journal publications.

The first decade dataset for conference publications starts in 1960-1969 and as it can be seen it has main ‘topics’ of the research such as ‘systems’ appearing much earlier than in journal articles.

Figure 11. Word cloud of titles of conference proceedings through the time period 1960-1969

In the next decade (1970-79), in addition to the topic ‘system’, the topics such as ‘database’, ‘data model’, ‘design’ and ‘simulation’ are gaining weight while other topics are fading down (Figure 12). For instance, ‘language’, ‘automatic’ and other topics are lighter compared to the previous decade.

Figure 12. Word cloud of titles of conference proceedings through the time period 1970-1979

In the 1980-1989 decade, ‘algorithm design’ is a new tendency as well as the ‘usability’ is becoming a new keyword in the field (Figure 13) and continues as such in forthcoming decades (Figure 14, Figure 15). From the last two figures the perception is that in the latter two decades the keyword ‘system’ starts to lose weight. Usability continues to be an important part of the research, together with ‘based’ and ‘model’, which indicate that nowadays CS research is focusing more on applicability in interdisciplinary fields, with solutions
based on concrete models. Also, we can notice a variety of topics, with a heterogeneous frequency of appearance which once again show that the computer science research is entering in many aspects and dimensions. In the last decade fields 'neural networks' and 'evaluation' increase their relevance.

Figure 13. Word cloud of titles of conference proceedings through the time period 1980-1989

Figure 14. Word cloud of titles of conference proceedings through the time period 1990-1999

Figure 15. Word cloud of titles of conference proceedings through the time period 2000-2009

IV. CONCLUSION

Taking into consideration the rapid growth of the number of publications in one hand and the rapid change of the computer science field on the other hand, the aim of this study was to provide a journey through the changing topics in the research publications in computer science.

Through the word cloud visualization, the study brings to the user an easy and fast way to get an insight of the key topics in the field for a certain decade in thousands of papers.

The results summarized above lead to several conclusions as presented above for each decade separately.

An appealing outcome for instance was that although the journals in terms of publications are more relevant, in terms of identifying the trends the conference proceedings seem to offer a step ahead. This is a very important result and guide especially for young researchers who are in the stage of identifying their research objectives and current research gap.

With the provided analysis it can be also concluded that computer science research is becoming very wide and diverse, giving solutions to different problems in many other disciplines.

REFERENCES


