PREGLEDNI RAD – REVIEW SCIENTIFIC ARTICLE 10.7251/ZRPIM2101015N

STRUCTURAL AND STATISTICAL ANALYSIS OF LARGE DATASETS OF TERMS AND RELATED ARTICLES: EXAMPLES FROM WIKIPEDIA

Zoran Nikolić

University of Belgrade, Faculty of Physics, Studentski trg 12, 11 158 Belgrade, Serbia,
nizoran@ff.bg.ac.rs

ABSTRACT

Among the most famous collections of publicly available data on the Internet is Wikipedia, which contains millions of articles in many languages covering a wide variety of topics. Complete dumps of all texts from the Wikipedia database in XML format are updated monthly. In this paper, the contents that exist on Wikipedia in the official languages of the former Yugoslavia are analysed and the knowledge base is integrated. Although there are over 10 million articles in this data collection, the number of described terms and topics is significantly smaller, because many articles only redirect to other articles, some are user conversations and some are article templates. A detailed classification of articles, terms, and topics was performed and their mutual connections were obtained (for that, an auxiliary dataset of the English version of Wikipedia was used). Detailed statistical, structural, and cluster analyses were performed on the generated graph of interrelationships of articles, terms, and topics. Using force-directed algorithms for redistribution of graphs, the final result was a comprehensive mapping and visualization of the knowledge base map.

Keywords: XML, big data, structural graph analysis, graph clustering, graph layout.

INTRODUCTION

Wikipedia was founded in 2001. Within the Wikimedia Foundation, as a non-profit open-source project, it was designed as an online encyclopedia of a general nature that can be edited by registered users in the Wikipedia community. Administration, preferences, and user rights are described in hierarchical detail. Within the user community of Wikipedia, there is completely regulated communication, it is possible to leave a comment on articles written by some other users, and with the permission of the administrator indirectly correct them. By 2021, Wikipedia has a total of over 300 editions (more than 200 languages). Wikipedia has sister projects within the Wikimedia Foundation, which supplement the content of the online encyclopedia. Table 1 shows the sister projects with a description of the scope of activities. The exchange of data between projects is complete, so the data are largely intertwined on different platforms. An Interwiki connectivity platform has been developed.

Table 1. Wikimedia Foundation projects.

Project	Description	Project	Description
Commons	Free media repository	Wikiquote	Collection of quotations
MediaWiki	Wiki software development	Wikisource	Free-content library
Meta-Wiki	Wikimedia project coordination	Wikispecies	Directory of species
Wikibooks	Free textbooks and manuals	Wikiversity	Free learning tools
Wikidata	Free knowledge base	Wikivoyage	Free travel guide
Wikinews	Free-content news	Wiktionary	Dictionary and thesaurus

Wikipedia has powerful software and administrative support and a large number of users in the community who edit articles. Such an organization of editing an encyclopedia enables the construction of a huge knowledge base, practically down to the smallest details, but the growing number of users who edit editions of Wikipedia diminishes the importance of expert opinions on various topics. Some important content on Wikipedia has been enriched with controversial interpretations. Scientific content usually does not have a strong connection outside the author's field of interest, because it often happens that highly professional article editors do not connect enough topics from different fields of science. In addition, the focus on dominant detailed concepts with a strong hierarchy is shifting from encyclopedic general education topics to administrative, political, entertainment, and sports topics. This development of Wikipedia in recent years exists in practically all editions.

Wikipedia is a very large set of data with daily changes and backups of the Wikipedia database that is updated monthly. Wikipedia is characterized by uneven growth in different areas, where there are articles that are updated almost daily, while some articles have unchanged content for many years. The density of interrelated articles in different thematic frameworks is very inhomogeneous. In addition, it is not possible to search different editions of Wikipedia at the same time with a single query. The connection between the articles on Wikipedia is not completely unambiguous, nor is it clear. For articles listed in the selected article, it is possible to generate a list of links, while it is impossible to generate a list of articles that cite the article without downloading the complete content of the publication or downloading the table of links between articles. All the above provides significant motives for a detailed statistical and structural analysis of collections of terms and related articles that exist on Wikipedia.

A very interesting area for research is the simultaneous searches of terms in several editions of Wikipedia and the tabular and graphical presentation of the mutual relations of articles. For that purpose, analyzes were performed in editions of Wikipedia in the official languages of the former Yugoslavia, as well as analyzes of the connection according to the articles with the most massive edition of Wikipedia - in English.

MATERIAL AND METHODS

The Wikipedia backup database is updated monthly at https://dumps.victorimedia.org/backup-index.html. It contains huge collections of datasets from all Wikipedia editions. It includes an XML printout of the current version in each language, a history of all changes, as well as auxiliary data repositories on the interconnection of items. The 1 May 2021 editions of Wikipedia, which were used in this analysis, are available in individual files ranging from 863 MB (Bosnian edition) to 13.2 GB (Serbian edition). The English edition of the XML backup takes about 180GB (60 parts). Figure 1 presents a fragment of an XML file.

```
<page>
    <title>Trebinjsko polje</title>
    <ns>0</ns>
    <id>422150</id>
    <revision>
      <id>3217372</id>
      <parentid>3217371</parentid>
      <timestamp>2020-08-13T13:24:46Z</timestamp>
      <contributor>
        <username>Zavičajac</username>
        <id>76629</id>
      </contributor>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="768" xml:space="preserve">'''Trebinjsko polje''' je [[kraško polje]], na jugoistoku
[[Hercegovina|Hercegovine]], [[Bosna i Hercegovina]].

Površina Trebinjskog polja je 18 km&lt;sup&gt;2&lt;/sup&gt;. Dugo je 6,5, a široko između 1 i 3,8 kilometara.
[[Nadmorska visina]] dna polja je između 268 i 275 metara. Kroz polje protiče rijeka [[Trebišnjica]].&lt;ref&gt;Dajana
Vukojević: [http://www.doiserbia.nb.rs/img/doi/0350-7599/2011/0350-75991103095V.pdf Geomorphological specific features
of Trebinje as tourist attraction], Zbornik radova Geografskog instituta Jovan Cvijić, 61(3), 95.-107. str.&lt;/ref&gt;

Na obodu polja nalazi se grad [[Trebinje]] na 275 metara nadmorske visine.

== Reference ==

{{reference}}

{{Kraška polja}}

[[Kategorija:Kraška polja u Bosni i Hercegovini]]
[[Kategorija:Trebinje]]</text>
      <sha1>plb0xmd5hnz1bk0zy5uu2vp9q3chw4r</sha1>
    </revision>
  </page>
```

Figure 1. Part of the XML dump that exists for a single term on Wikipedia.

To process XML files, it is necessary to use a developer library such as TinyXML or, as in this case, to develop efficient parser code. The tabular data obtained after parsing contains important records for further processing, such as title, ID, timestamp, text, and similar tags, as ancillary tags about the position of the beginning of the record and the number of bytes occupied by the XML table file. Also, for the development of any search and analytical application, it is necessary to develop such an indexing system that will enable fast access to records and further analysis.

Efficient development of software solutions for further manipulation of such mass databases is possible in those programming languages that support pointers, enable tokenized data reading, and manipulation of HASH tables. In this case, the C ++ language (Cserép, & Krupp, 2015), was used with the support of STL libraries for elementary data manipulation, as well as the GSL library for statistical analysis and the Boost (Polukhin, 2013) and Igraph (Gábor, & Nepusz, 2019) library for structural analysis in the generated charts.

For the analysis in this paper, a web-based web solution for the unified search of 6 editions of Wikipedia was developed with the possibility of accessing the original Wikipedia pages, displaying XML dumps with the selected article, reviewing links in the selected article, and list of articles citing the selected article. Such a solution enables centralized storage of processed data, XML dumps, and indexes, and thus fast search from the server. It allows access to a large number of computers that do not have to perform well and involve multiple users in data search. Maintaining such a solution is not a problem and it is possible to update the data on the website because they exist inside the XML dump. To implement such a solution used from the Boost library package (ASIO and Beast) to develop a portable server solution.

For the needs of the WEB solution, a portable server was developed that does not use Apache, WAMP, LAMP, IIS, or similar technologies. The server was developed as a CHAT solution for the desired port communication via the HTTP protocol. The application takes up almost no resources (processing power is less than 2% in sleep mode). It is located at the WEB address of the Faculty of Physics on a computer that is not intended for a server (ThinkPad X220). In addition, several search applications have been developed that run the server through the Fork process. There is no HTML file on the server (no homepage) - the website is 100% dynamic.

A dynamic WEB solution has been developed, which is an aggregate in searching all editions of Wikipedia in the official languages of the former Yugoslavia. Entering the first letters of a term (the ability to enter a string of characters that includes multiple words) allows you to search 6 editions of Wikipedia. As a result, you will get consolidated lists of articles on Wikipedia. The list contains a link to the original article, as well as links to the XML dump of the selected article. The server addresses are: HTTP: //147.91.68.147:1000 and HTTP: //147.91.68.144:1000.

RESULTS AND DISCUSSION

This paper analyzes 7 editions of Wikipedia. 6 editions are in the official languages of the former Yugoslavia, and the English edition of Wikipedia was also analyzed. All relations between articles were analyzed. As mentioned earlier, for 6 editions, a Wikipedia search aggregator was created in the form of a WEB solution. Table 2 shows data on the number of articles, the number of actual articles, the number of users, and the number of active users in the communities of 7 editions of Wikipedia. Real articles are those that do not include user conversations (most articles are just that). Real articles are also not templates, they are not redirected to real articles, nor alternations of articles in the form of duplicates, etc.

The size distribution of real articles in the English edition of Wikipedia is shown in Figure 2. It is noticed that the largest number of articles is less than 5 kB in size, which are common articles in any edition, which describe the less important term. Interestingly, there are pronounced peaks in the sizes of articles at 7 kB, 13 kB, and 45 kB, which can be explained by the existence of classification articles in the fields, which are most often realized using predefined templates.

Table 2. Basic data of several editions of Wikipedia.

Wiki	English	Bosnian	Croatian	Macedonian	Serbo-Croatian	Slovenian	Serbian
No of Articles	53257422	355853	482806	488846	4638394	419577	3906579
Real articles	6289918	85811	208581	114052	454945	172164	645873
%real art.	12%	24%	43%	23%	10%	41%	17%
No of Users	41455629	132702	254715	94261	160513	199875	284324
Active users	139984	212	646	240	220	512	989
%act. users	0.34%	0.16%	0.25%	0.25%	0.14%	0.26%	0.35%

The timestamp represents the time the article was last modified by Wikipedia. If they are aggregated by years of the time of the last changes, graphic representations are obtained in Figures 3 a (all articles) and 3 b (real articles), which represent the time distributions of the last changes on articles in 6 editions of Wikipedia. It can be noticed that the Serbo-Croatian edition of Wikipedia had the most significant changes in the period 2013-2016. After that period, certain national editions have complete domination and practically the Serbo-Croatian edition ceases to be current. Also, the Serbian edition of Wikipedia has a significantly reduced number of changes during 2020, which can be explained by the significant reduction of entertainment and sports contents during the year of the Covid-19 pandemic, because a significant number of articles in those areas were often most edited during the year.

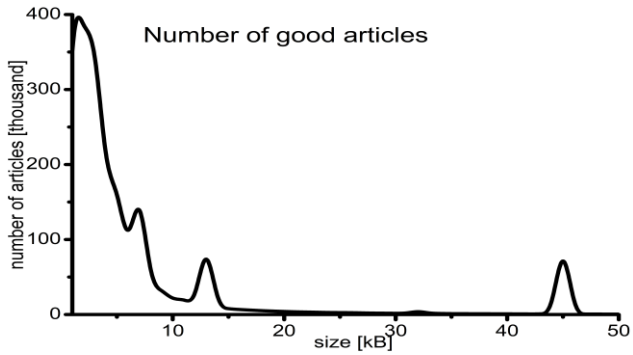
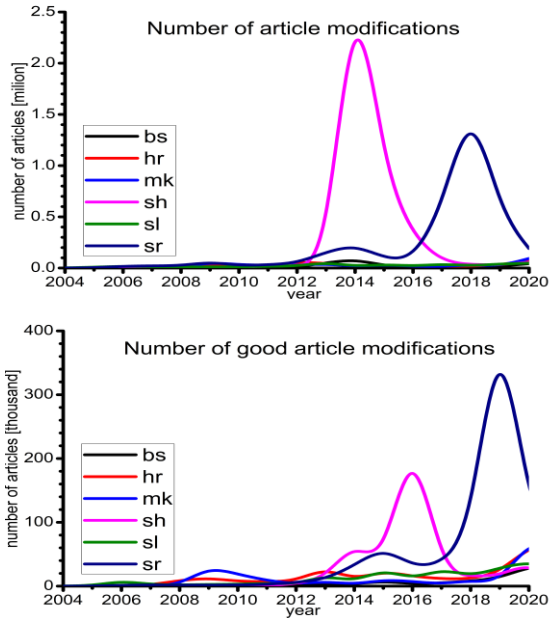


Figure 2. Distribution of the size of actual articles in the English edition.

This paper analyzes the data on the interconnection of identical terms from Wikipedia in the languages of the former Yugoslavia and the links to the English edition of Wikipedia. Multiplicity is also defined as a measure of the absolute deviation from the mean of the link to and the link to the English edition. Data on the number of words on Wikipedia in each language that meet the condition of external significance were obtained. A combined number of terms from 6 editions of Wikipedia based on the English edition was also obtained.



Figures 3a. & 3b. Last changes by years: 3a. - all articles, 3b. - real articles.

Table 3 shows the number of related actual articles from the 6 editions of Wikipedia, according to the English edition. Data were obtained that the total number of articles in the languages of the former Yugoslavia relating to the English edition is around 650,000. The total number of articles in 6 languages is 1,800,000, but there are only 750,000 different articles.

Table 3. Relation of articles from 6 editions of Wikipedia with the English edition.

Wiki	English	Bosnian	Croatian	Macedonian	Serbo-Croatian	Slovenian	Serbian
To	628513	131738	178371	161127	402201	177979	559371
From	655186	132638	180190	160974	485030	185857	560814

The interconnectedness of articles in the same edition of Wikipedia was additionally analyzed. It was mentioned earlier that it is possible to determine the list of articles cited by the selected article, but the problem is to find all the articles that cite the selected article. To determine the list of citations for an article, you need to create a collection of all articles in one edition and find all the links from the selected article to others.

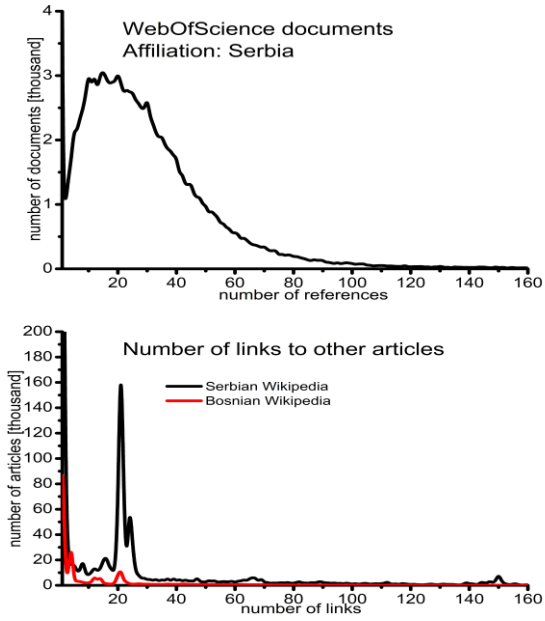
For the number of citations of articles (number of other articles that have a link to the selected one), it is possible to apply a complete analogy with citations of scientific publications. Articles are interconnected based on some meaningful, conceptual, thematic, or logical connections. An article is important and visible if it is highlighted by another article. The measure of the number of links is an indirect measure of the visibility of the article. Sometimes this happens only because there are links to articles only from classification articles. Classification articles or hierarchical articles are similar to review scientific publications. They refer to other articles and usually have a pronounced citation rate on their own.

In Table 4 compares the total number of links with citations in different editions of Wikipedia. It can be noticed that the average number of citations is practically at the same level in all editions, except in the case of the Bosnian edition, which can be explained by the incompleteness of the complete collection of articles, where currently there are only classification articles with a more significant number of citations than others.

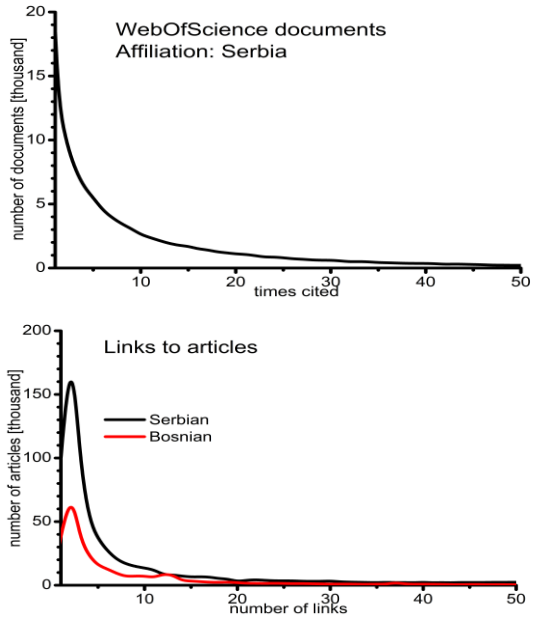
Table 4. Overview of the total linking of articles in the same edition of Wikipedia.

Wiki	English	Bosnian	Croatian	Macedonian	Serbo-Croatian	Slovenian	Serbian
No of Articles	53252742	355853	482806	488846	4638394	419577	3906579
Real articles	6289918	85811	208581	114052	454945	172164	645873
No of links [million]	480	15	14.9	10.2	38.5	14.4	50
Links per real article	76.3	174.4	71.4	89.5	84.6	83.5	77.5

By applying the mentioned analogy with citations of scientific publications, analysis of the distribution of the number of links from the mentioned article (number of references in the scientific publication) and the number of links to the mentioned article (number of citations of the selected publication) was obtained (Garfield, 1972; Price, 1965). These graphs are shown in Figures 4 a and 4 b (number of references, ie number of links in the selected article) and 5 a and 5 b (number of citations, ie number of links to the selected article). It is noticed that the distributions of the number of references and links have no similarities, because there are some predefined characteristics in the editions of Wikipedia (templates for classification type of articles), while in scientific publications the distribution is approximately log-normal. However, in the domain of distributing the number of citations and the number of links to the selected article, there are complete analogies, except at the beginning that in the case of Wikipedia articles, due to the existence of a template influence, a log-normal profile appears (links are required to some extent) instead of the distribution for citations where the profile is exponential (by importance) (Leydesdorff, Wagner, & Bornmann, 2018; Bornmann, Haunschild, & Hug, 2017).



Figures 4a. & 4b. Number of references, ie number of links in the selected article.

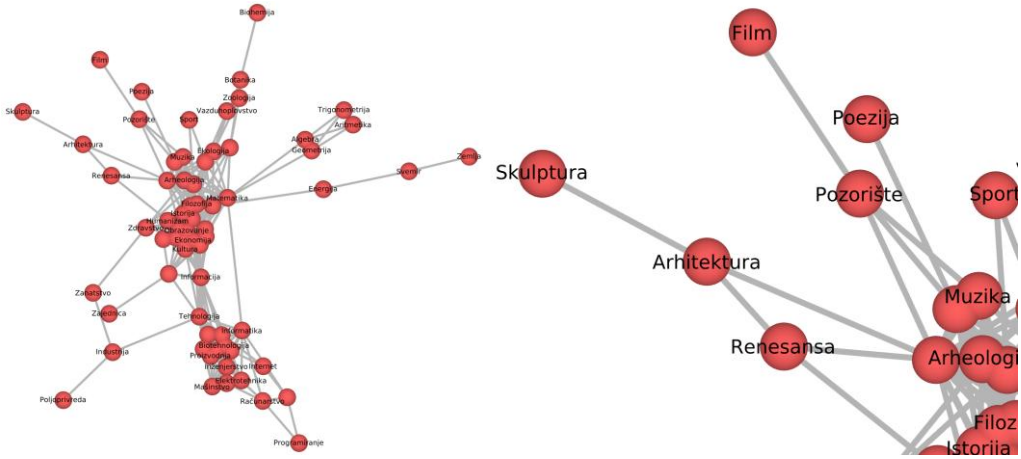


Figures 5a. & 5b. Number of citations, ie number of links to the selected article.

Graph analyzes of related articles and concepts are also part of the results in this paper (Torsten Zesch, & Gurevych, 2007). Graphs are directed and are represented by articles as nodes and links as links in the chart. The easiest way to present nodes and links as a text record of a DOT file. In the DOT file, the format of the directional link record is in the form Term1 -> Term 2 in pseudo-textual format, which provides the ability to easily generate a file from developed code. It is efficient to use Tulip Data Visualization Software or a similar graphical environment to generate a graphical visual result (Auber et al., 2012).

Directional force-based algorithms are used to represent graph nodes using physical analogies. In such a system, the nodes are connected by springs or are represented by charges or dipoles. The system tends to go into a state with minimal energy. Nodes with high connection density tend to be positioned in the middle. The procedures are iterative. The most well-known algorithms are Kamada-Kawai (Kamada, & Kawai, 1989) and Fruchterman-Reingold (Fruchterman, & Reingold, 1991). The application of the mentioned algorithms enables detailed structural analysis and visualization of complex graphs.

Structural analysis of Wikipedia editions was performed in three ways. The Wikipedia classification with a list of basic topics was used first. All links to each main topic and each main topic were taken into the analysis. Then, unions and cross-sections of common connections of each basic topic with each were made, and by applying the similarity criteria for the sizes of sets $A \cap B$: $A \cup B$ with an emphasized lower limit, the connection graph was obtained for basic topics. For each edition of Wikipedia, there are similar structures for linking basic topics. For example, Figure 6a shows a graph of the relationship between the main topics of the Serbian edition of Wikipedia, and Figure 6b shows the zoomed part.



Figures 6a. & 6b. Graph of the connection of basic topics in the Serbian edition of Wikipedia and one zoomed part.

In the second case, the terms that refer to the largest number of articles in the Wikipedia edition with a predetermined limit or a predetermined number of the most important terms are selected, and then a graph of the most important terms of each Wikipedia edition obtained based on the list of terms. There is variability in different editions because there are unequal distributions on basic topics in different editions of Wikipedia. Figure 7a shows a graph of the relationship between the most important terms of the Serbian edition of Wikipedia, and Figure 7 b shows the zoomed part. It should be noted that the zoomed part of the image clearly shows the pronounced connection of concepts in biochemistry.

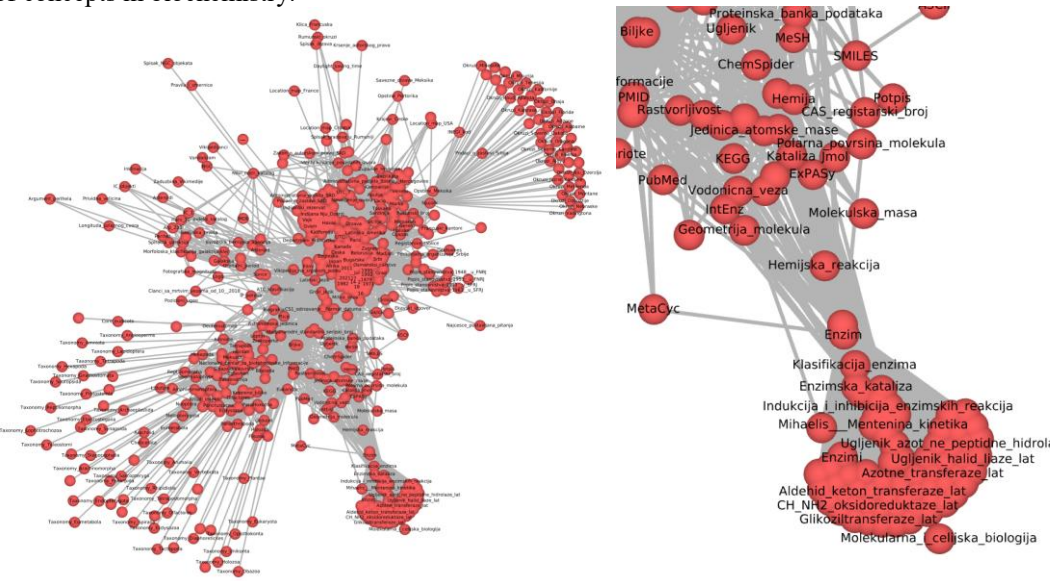


Figure 7a. & 7b. Graph of the connection between the most important terms of the Serbian edition of Wikipedia and one zoomed part.

A solution has also been developed that generates a list of all expressions that refer to it or are related to it for a given term. The lower limit of the number of generated terms or the sum of the number of links can be set in advance in the analysis. Using the same similarity criterion, a graph of the mutual relations of all articles concerning a given term is generated. Figure 8 a shows a

graph of the connection of terms related to the term "Trebinje" of the Serbian edition of Wikipedia, and Figure 8 b shows the zoomed part. It should be noted that the zoomed part of the image clearly shows the pronounced connection of geographical toponyms in the vicinity of the town of Trebinje.

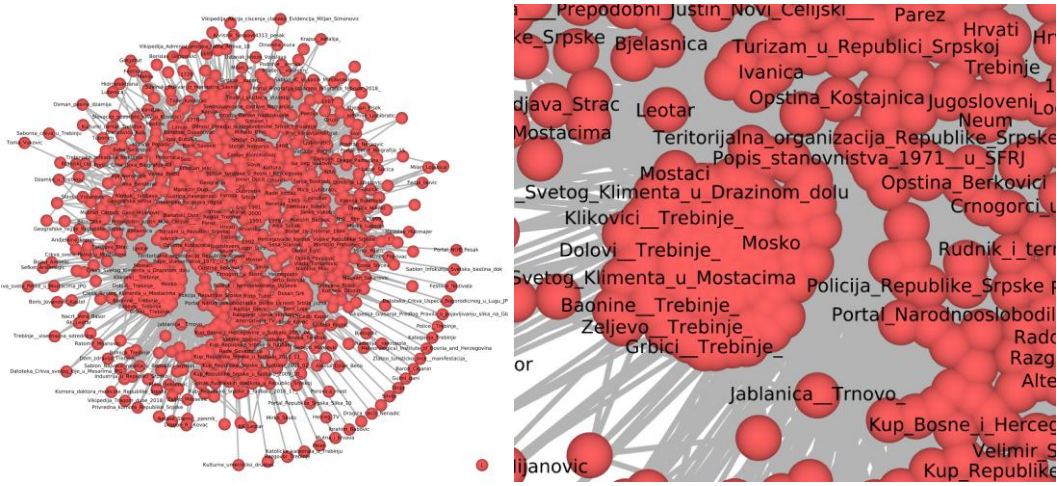


Figure 8a. & 8b. Graph of the connection of terms that refer to the term "Trebinje" of the Serbian edition of Wikipedia and one zoomed part.

It should be noted here that in all three cases of the presented extractions of concepts and the applied similar criteria of similarity, conceptual connections are obtained very clearly, which are completely scientifically classified, clear and logical. All the results shown in the graphs and the accompanying analyzes are valid for any edition of Wikipedia, because the interrelationships of the terms are based on a meaningful and logical basis, and are therefore invariant. In this way, it is clearly shown that it is possible to determine complex structural relations and determine precise classification criteria by simple set operations (Palomares, Ahres, Kangaspunta, & Ré 2016).

CONCLUSION

Repositories of Wikipedia editions in 6 languages of the former Yugoslavia were downloaded. A search tool has been developed for several editions of Wikipedia. Solutions for the analysis of mutual relations of articles have been developed. Links to articles from different editions of Wikipedia were analyzed. The conceptual links of articles in each Wikipedia were analyzed. The developed solutions enable a comparative analysis of the representation of articles in different editions of Wikipedia. In this way, it is possible to indirectly determine which articles are missing in each edition of Wikipedia. The developed solution enables better tracking of redirects to a given article and creates a better working environment for potential Wikipedia editors. Developing a solution for uniquely searching massive datasets can open up new ideas in combining large amounts of heterogeneous data in a WEB environment. By combining filtered data and their current analysis in developed WEB solutions, it can contribute to a better knowledge of current events in the field of very heterogeneous data collections, as well as in the field of administration in modern society.

LITERATURE

Auber, D., Archambault, D., Bourqui, R., Lambert, A., Mathiaut, M., Mary, P., ... & Melançon, G. (2012). *The tulip 3 framework: A scalable software library for information visualization applications based on relational data* (Doctoral dissertation, INRIA).

Bornmann, L., Haunschild, R., & Hug, S. E. (2018). Visualizing the context of citations referencing papers published by Eugene Garfield: a new type of keyword co-occurrence analysis. *Scientometrics*, *114*(2), 427-437.

Csárdi, G., & Nepusz, T. (2010). Igraph Reference manual. Retrieved May 20, 2021, from http://igraph. sourceforge. net/documentation. html.

Cserép, M., & Krupp, D. (2015, January). Component visualization methods for large legacy software in C/C++. *Annales Mathematicae et Informaticae*. Retrieved May 20, 2021, from https://ami.uni-eszterhazy.hu/.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129-1164.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471-479.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, *31*(1), 7-15.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity—A tribute to Eugene Garfield. *Scientometrics*, *114*(2), 567-592.

Palomares, T., Ahres, Y., Kangaspunta, J., & Ré, C. (2016, April). Wikipedia knowledge graph with DeepDive. In *Tenth International AAAI Conference on Web and Social Media*.

Polukhin, A. (2017). *Boost C++ Application Development Cookbook*. Packt Publishing Ltd.

Price, D. D. S. (2011). *Networks of scientific papers* (pp. 149-154). Princeton University Press.

Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing* (pp. 1-8).