7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Bibliometric of Semantic Enrichment

Mohammad Javad Shayegan*
Department of Computer Engineering,
University of Science and Culture,
Tehran, Iran
shayeagn@usc.ac.ir

Mohammad Mehdi Mohammad
Department of Computer Engineering,
University of Science and Culture,
Tehran, Iran
mehdimohammad73@gmail.com

***Abstract*— Much research has been done on the Semantic Web. Semantic enrichment is one branch of this field that has gotten much attention recently. This study aims to use bibliometrics to look at the current state of research in this field. Bibliometrics is the study of scientific sources' bibliographic data, which can be used to assess the current state of a field. Various bibliometric analyses were performed after extracting the metadata required for bibliometry from the Scopus database. According to the study's findings, more articles in this field have been published since 2018, and they have recently received special attention. Ontology/semantics/semantic web/semantic enrichment are also becoming more important keywords. Furthermore, based on the countries that published the articles, the United States, Germany, the United Kingdom, France, Italy, and Brazil published the most in this field. The article goes on to provide more analysis and illustrations, which will be helpful to researchers in this field.**

Keywords*— *Bibliometrics, Semantic Enrichment, Semantic Web.

I. INTRODUCTION

To better understand texts, semantic enrichment refers to the use of machine learning, artificial intelligence, and language processing. In general, semantic enrichment aids operational teams in removing the noise and problems associated with events, facilitating their access and completion, and speeding up the acquisition of required data [1]. As a result, the analysis of articles and research related to the subject (semantic enrichment) will assist writers and researchers in this field by providing an overview of various research in this field and demonstrating the existing connections.

Bibliometrics is the study of bibliographic data from scientific sources to determine the current state of a scientific field. Bibliometrics [2] is another form of measurement tool that assesses the quantitative interconnectedness of written communication. We may discuss and evaluate the relevance of books, the countries that publish articles, and their relationship to one another using bibliometrics, which is focused on important terms and their relationship to one another, journals, and publishing sources. It is also possible to explain and demonstrate the relations in the topic under review by using the citation analysis tool, which is one of the bibliometric methods.

Articles in the field of semantic enrichment are analyzed in this study using the bibliometric method. The results are displayed using a set of metadata. The Scopus database is the focal point of this study.

II. METHODOLOGY

This section explains the proposed method for conducting bibliometric analyses on the topic at hand. The proposed method's first step is to extract the initial database from the Scopus database and save it in CSV format. The database used in this study is a collection of 783 articles about "semantic enrichment" that were extracted from the Scopus database on October 3, 2021. The VOSViewer software is then used to analyze the data and visualize the results. VOSviewer is a bibliometric and scientometric software that summarizes data and draws maps from research-related data. This software allows you to create maps based on citations, bibliographic connections, joint citations, or co-author relationships. VOSviewer's other great feature is that it's free and doesn't require any special installation, which has attracted the attention of researchers. It also lets you import information from citation databases like Web of Science, Scopus, and Pop Fashion. The research findings are discussed in the following section.

III. FINDINGS

Let's take a look at how many articles have been written in the area of "semantic enrichment" before we get into the metadata of articles. The distribution of papers written in different years is depicted in Fig.1.

According to the Scopus citation database, the frequency of documents published on this topic between 1991 and 2021 is shown in Fig.1. The number of publications published in this area has increased over time, as seen in the graph above, from 2012 to 2019. In addition, the graph shows that since about 2018, the number of papers published in this area has increased dramatically. This result demonstrates that this subject has gotten much publicity in the last few years.

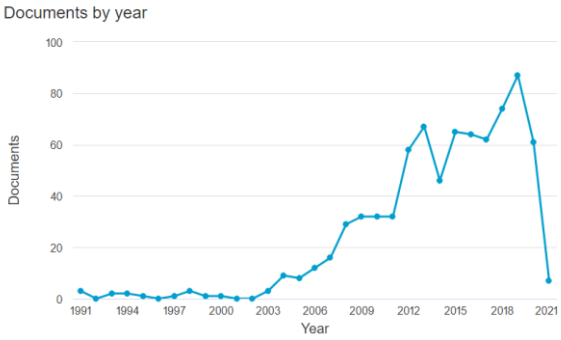

Fig. 1. Dispersion rate of articles published in different years

A. *Impressive articles*

Table I shows the results of analyzing statistics from the Scopus scientific database and using the VOSviewer software to identify ten useful articles in the field of semantic enrichment based on the number of citations.

B. *Keyword analysis in the articles*

The research is summarized in this section based on the papers' most commonly used keywords. The papers are examined in terms of keywords and how they relate to one another, as shown in Fig.2. Fig.2 shows how similar keywords are grouped together and represented in larger dimensions. The terms ontology/semantics/semantic web/semantic enrichment are more prevalent in this area, as can be seen. The use of ontology in semantic enrichment has been considered by scholars, as shown by the presence of the term ontology.

The essential keywords in the articles were examined in the previous figure, and their relationship to one another was analyzed. In addition to the previous analysis, as shown in Fig.3, keyword analysis can be done based on time. In this diagram, purple represents keywords from earlier years, and yellow (a brighter color) represents keywords from more recent years. Building information model/architectural short texts/3d modeling/has been used more in recent articles, as shown in the graph.

C. *Analysis based on keywords determined by the author*

The papers will be checked and evaluated based on the keywords determined by the authors of the articles after they have been analyzed based on their keywords. The most important keywords from the authors' perspective are displayed in Fig.4, along with their relationship to one another.

The terms semantic web, semantic enrichment, linked data, semantic annotation, and semantics are more important than other words, as seen in Fig.4. This indicates that annotation and related data are more important to researchers in this area.

TABLE I. TEN PROLIFIC AND INFLUENTIAL AUTHORS IN THE FIELD OF SEMANTIC ENRICHMENT

Id	*Author*	*Title*	*All Citations*	*Year*
1	Parent C and et al. [3]	Semantic trajectories modeling and analysis	302	2013
2	Abel F and et al. [4]	Analyzing user modeling on Twitter for personalized news recommendations	270	2011
3	Kim J and et al. [5]	Corpus annotation for mining biomedical events from literature	211	2008
4	Abel F and et al. (2011)[6]	Semantic enrichment of Twitter posts for user profile construction on the social web	168	2011
5	Gerner M and et al.[7]	A species name identification system for biomedical literature	155	2010
6	Shotton D and et al. [8]	Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article	110	2009
7	Kapanipathi P and et al. [9]	User interests identification on Twitter using a hierarchical knowledge base	90	2014
8	Abel F and et al. (2012) [10]	Semantics + Filtering + Search = Twitcident exploring information in social web streams	88	2012
9	Schulz A and et al. [11]	Real-time detection of small scale incidents in microblogs	80	2013
10	Belsky M and et al. [12]	Semantic Enrichment Engine for Building Information Modeling	71	2016

Given that keywords were also used in this analysis, these words can be measured over time, as shown in Fig.5. Building information modeling/ load indexing has become more popular in recent years, as can be seen.

D. *Analysis based on the importance of the articles*

The importance of articles is one of the subjects studied in bibliometrics. The number of citations in articles can be used to estimate their importance.

Articles with larger dimensions are referred to more frequently and are considered more important, as shown in Fig.6. Furthermore, articles that are more closely related are grouped together. Because the titles of the articles are long, this illustration is based on the names of the authors of the articles, and the articles Bruno [13], Schlabach [14], Abel [4], and parent c [3] are bolder, as shown in the figure.

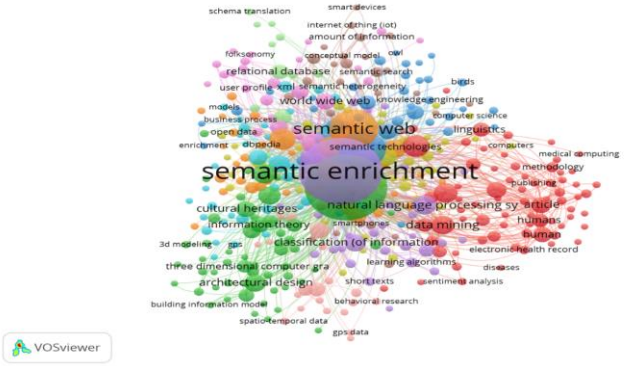

Fig. 2. Keyword-based analysis

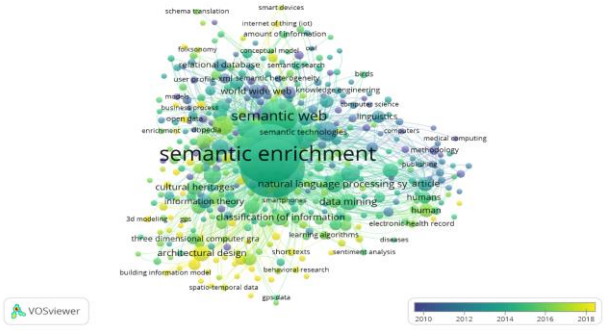

Fig. 3. Keyword-based analysis in terms of year

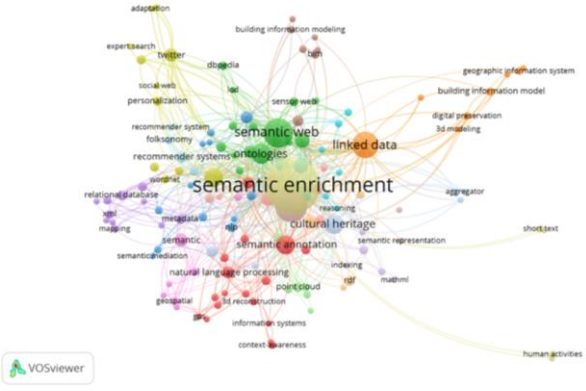

Fig. 4. Analysis based on author-defined keywords

E. Analysis based on publisher

The publisher-based analysis is another type of analysis that provides instrumental and citable data. The connection between journals and publishers is extracted, as shown in Fig.7.

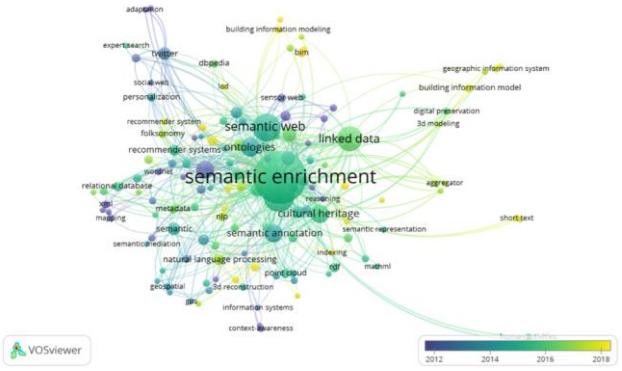

Fig. 5. Analysis based on author-defined keywords in different times

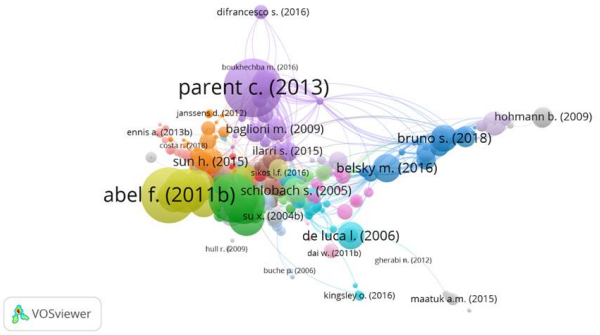

Fig. 6. Analysis based on the importance of articles

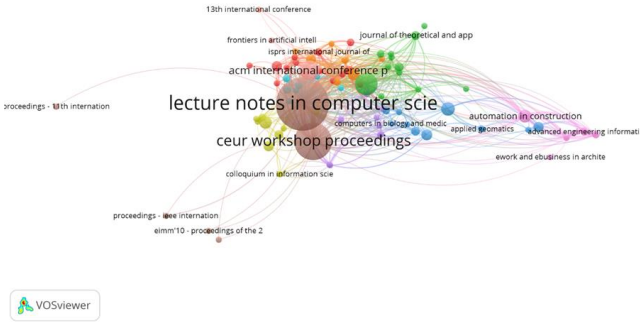

Fig. 7. Analysis based on the publisher

The following publishers published the most articles in the field of semantic enrichment, as shown in Fig.7:

CEUR workshop proceeding/Lecture Notes in Computer Science/ACM international conference/communication in computer and information science

F. *Analysis based on the relationship between countries that publish articles:*

Another form of analysis that may aid in article analysis focuses on the articles' countries of origin. The country in which at least three or more papers were published is shown in Fig.8 using the VOSviewer program, along with the analysis performed on the checked articles and based on the software settings. As can be seen, out of the 56 countries that are present as article publishers, 56 are shown as production, all of which follow the rules and are connected to one another.

The United States, Germany, the United Kingdom, France, Italy, and Brazil have all published more articles in this field, as shown in Figure 8.

G. *Word cloud analysis*

The semantic enrichment field's word cloud is shown in Fig.9. The words information-twitter-semantic-event-classification have a remarkable amount of boldness to them.

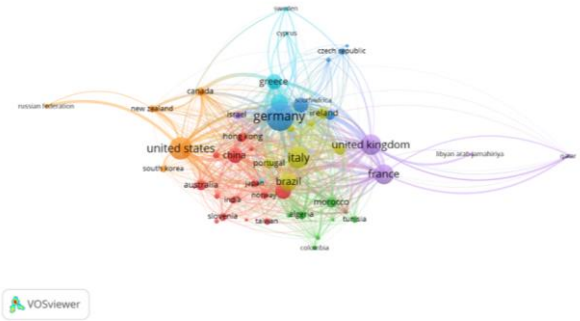

Fig. 8. Analysis based on the donor countries of articles

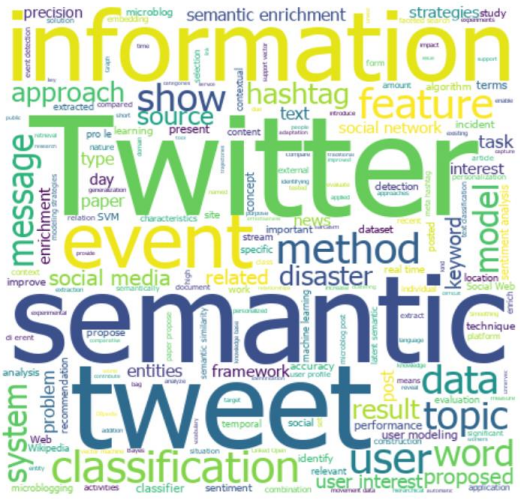

Fig. 9. Word cloud of the semantic enrichment

IV. Conclusion

A bibliometric analysis of semantic enrichment was carried out in this study. The initial database, which contained 783 articles, was extracted from the Scopus database to perform bibliometric analysis in the first stage. Then, various analyses were performed on a variety of metadata of articles, including keywords, author-specified keywords, article publishers, article relationships, and article submission countries.

Initially, the research was focused on keywords, with the terms ontology/semantics/semantic web/semantic enrichment becoming more significant based on the findings of important keywords and their relationship with each other, as well as the study of keywords based on time. The research was then conducted using the author's keywords, which revealed that the terms semantic web/semantic enrichment/linked data/semantic annotation/semantics were the most relevant. Following this scenario, the study was focused on the relationship between the articles, which were portrayed. Based on the number of articles published in each of the outlets, it was determined that CEUR workshop proceedings/Lecture Notes in Computer Science/ACM international conference/communication in computer and information science are the most relevant. Furthermore, the most recent study was focused on the countries that published the papers, with the United States, Germany, the United Kingdom, France, Italy, and Brazil being the most relevant in terms of the number of articles published in these countries, according to the findings.

References

[1] S. Romero, and K. Becker. "A framework for event classification in tweets based on hybrid semantic enrichment." Expert Systems with Applications, 2019, pp. 522-538.

[2] D. Nicholas, and M. Ritchie. "Literature and bibliometrics." C. Bingley, 1978.

[3] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, and Z. Yan. "Semantic trajectories modeling and analysis." ACM Computing Surveys (CSUR), 2013, 45(4), pp. 1-32.

[4] F. Abel, Q. Gao, G. J. Houben, and K. Tao. "Analyzing user modeling on Twitter for personalized news recommendations." In international conference on user modeling, adaptation, and personalization, Springer, Berlin, Heidelberg, July 2011, pp. 1-12.

[5] J. D. Kim, T. Ohta, and J. I. Tsujii. "Corpus annotation for mining biomedical events from literature." BMC bioinformatics, 2008, 9(1), pp.1-25.

[6] F. Abel, Q. Gao, G. J. Houben, and K. Tao. "Semantic enrichment of Twitter posts for user profile construction on the social web." In Extended semantic web conference. Springer, Berlin, Heidelberg, May 2011, pp. 375-389.

[7] M. Gerner, G. Nenadic, and C.M. Bergman. "LINNAEUS: a species name identification system for biomedical literature." BMC bioinformatics, 2010, 11(1), pp.1-17.

[8] D. Shotton, K. Portwin, G. Klyne, and A. Miles. "Adventures in semantic publishing: exemplar semantic enhancements of a research article." PLoS Comput Biol, 2009, 5(4), e1000361.

[9] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on Twitter using a hierarchical knowledge base. In European Semantic Web Conference. Springer, Cham, May 2014, pp. 99-113.

[10] F. Abel, C. Hauff, G. J. Houben, R. Stronkman, and K. Tao. "Semantics+ filtering+ search= twitcident. exploring information in social web streams." In Proceedings of the 23rd ACM conference on Hypertext and social media, 2012, pp. 285-294.

[11] A. Schulz, P. Ristoski, and H. Paulheim. "I see a car crash: Real-time detection of small scale incidents in microblogs. In Extended semantic web conference. Springer, Berlin, Heidelberg, May 2013, pp. 22-33.

[12] M. Belsky, R. Sacks, and I. Brilakis. "Semantic Enrichment for building information modeling." Computer-Aided Civil and Infrastructure Engineering, 2016, 31(4), pp.261-274.

[13] S. Bruno, M. De Fino, and F. Fatiguso. "Historic Building Information Modelling: performance assessment for diagnosis-aided information modeling and management." Automation in Construction, 2018, pp. 256-276.

[14] S. Schlobach. "Debugging and semantic clarification by pinpointing." In European Semantic Web Conference. Springer, Berlin, Heidelberg, 2005, pp. 226-240.