Dutta, Biswanath and Sinha, P. K. A bibliometric analysis of automatic and semi-automatic ontology construction processes, *Annals of Lib. and Inf. Studies.* (Accepted for publication)

A bibliometric analysis of automatic and semi-automatic ontology construction processes

Biswanath Dutta[a] and Prashant Kumar Sinha[b]

Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, RVCE post, Bangalore-560059, Email: bisu@drtc.isibang.ac.in and pk@drtc.isibang.ac.in

Received: 15 November 2017; revised: 13 June 2018; accepted: 26 June 2018

Through a bibliometric analysis, the paper reveals the current state and the global research trend in the areas of automatic ontology construction process (AOCP) and semi-automatic ontology construction process (SOCP) during the period of 2000-2016. Scopus, GoogleScholar and Scitepress digital library were used to extract the data for analysis. The study revealed that the majority of the works were published in conference proceedings. China was found to be the most contributing country in this area followed by USA, France, and Spain. The University of Karlsruhe contributed the maximum publications in both AOCP and SOCP whereas Peking University contributed largely to AOCP and Jozef Stefan Institute contributed largely only to SOCP. The majority of the researchers were from computer science background but a significant number of researchers were also from other disciplines including engineering and allied operations, library and information science, management and auxiliary services, making this research area truly interdisciplinary.

Introduction

Ontology is a major area of research which has become a multidisciplinary pursuit[1,2]. There are three ways of designing ontologies[3] namely manual ontology construction process (MOCP), automatic ontology construction process (AOCP) and semi-automatic ontology construction process (SOCP). In MOCP, all the tasks of designing an ontology such as, identifying the source of key terms, selection of the terms, discovery of the classes and hierarchies, property selection, modelling, formalization, etc., are done manually by an ontology designer[4,5]. In AOCP all the tasks involved in designing an ontology starting from extracting the domain terminologies, identifying the classes and properties, discovering the class hierarchies, etc., are done automatically with the help of software. In SOCP, the major tasks of an ontology construction are done with the help of the software, although the ontology designers stay in the loop, for instance, to define the ontology extraction pattern from the text corpora, to evaluate the output and overall, to oversee the entire process.

Bibliometric analysis is used to measure the impacts of research with the help of quantitative indicators[6]. It generally results in critical information which gives an

idea of the quality and quantity of the research. The current study focuses on the bibliometric analysis of ontology construction process research and to the best of our knowledge, there exists no such studies in the literature on this subject. There are a few studies available in the literature but in the related subjects, for instance, digital libraries[7,8] ontology[9,10,11] in general, and semantic web[12].

The current work is a bibliometric analysis of the two types of ontology construction processes i.e., AOCP and SOCP, and not the MOCP. MOCP was not considered as the design of MOCP is expensive, especially in terms of time, infrastructural support, human labour etc.,[13] and importantly, literature is scant on MOCP in the recent years. This study assesses the research output for the period of 2000–2016 and during 1994-1998, the methodologies[14,15] for ontology construction were largely manual and from 2000 onwards, the emphasis on constructing the ontologies has using automatic and semi-automatic methods.

Objectives of the study

- To explore the research growth trend, authorship pattern, collaborative nature of research on AOCP and SOCP;
- The find the most active and productive researchers, countries, organizations and also the types of organizations working in the area; and
- To study the multidisciplinary contributions in research and the key literature in the area.

Methodology

Data were collected from Scopus (https://www.elsevier.com/) and GoogleScholar (http://scholar.google.com/). Besides these two databases, data was also downloaded from Scitepress digital library (http://www.scitepress.org/). To begin with, search terms and their combinations including AOCP, SOCP, ontology, taxonomy, vocabulary, development, creation, and building were used. We further enriched this list by identifying few more key terms from the initial set of identified literature, for instance, data mining, relational database to ontology construction, fuzzy logic, formal concept analysis, and approaches to ontology construction. Also, we extended the literature search by going through the related work sections and references of the downloaded papers. Only research articles published in journals and conference proceedings have been considered. Reviews, editorials, newsletters, etc., were excluded.

The final set consisted of 324 articles published on AOCP and SOCP during the period 2000-2016, out of which, 169 articles were on AOCP and 155 were on SOCP. The study was conducted at three different levels: (i) based on the publications on AOCP (169 articles), (ii) based on the publications on SOCP (155 articles) (iii) based on both type of procedures (where AOCP and SOCP

publications were combined that totalled 324 articles). Microsoft Excel was used for data analysis.

Analysis

Document type

Table 1 depicts that out of total 324 papers, 127 papers (39.20%) were published in journals and 197 papers (60.80%) in conference proceedings. When considered separately too, for AOCP and SOCP, the distribution of papers in journals and conference proceedings almost remained the same with conference proceedings being the preferred medium over journals. The reason for this seems to be that majority of the proposed approaches for AOCP and SOCP are still at the experimental level as the techniques and tools contributing towards the AOCP or SOCP are under development and immature[16] and therefore researchers prefer to present and publish their papers in conference proceedings rather than journals.

Table 1—Document-wise distribution of ontology construction research output

	AOCP		SOCP		Total	
	No.	%	No.	%	No.	%
Journal	66	39.05%	61	39.35%	127	39.20%
Conference proceedings	103	60.95%	94	60.65%	197	60.80%
Total	169	100%	155	100%	324	100%

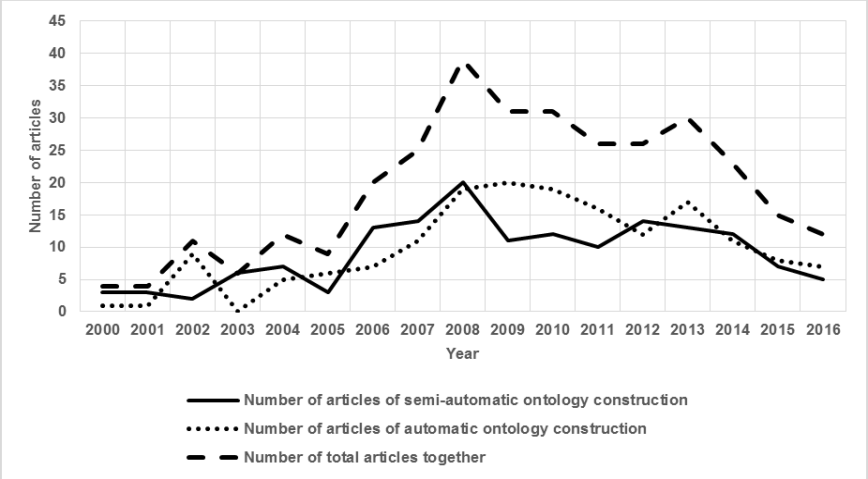
Temporal trend of publications

The temporal change in publications gives an idea about the trend of a specific topic. Fig. 1 provides an overview of the trend of research on AOCP and SOCP, and when taken together for the study period 2000–2016. As can be seen from the figure, the research output between 2000 and 2005, when taken together was relatively low. The most productive time period was between 2006 and 2008 where there was steep increase in the number of publications. The year 2008 was the most productive year when 39 articles (12.03%) were produced. Between 2009 and 2013, there was a slight decrease in the number of publications but still, a number articles were produced (35.16%). During 2014 to 2016, there has been a decrease in the number of publications. It seems that after the relative wide-spread enthusiasm during the initial years where researchers from different fields worked in this area, the area had fewer researchers working on the two topics owing perhaps to the speialized nature of the area.

Measuring the research output separately for AOCP and SOCP reveals a slightly different trend. In the 17 years of research on AOCP, for the period between 2000 and 2004, the growth rate was inconsistent. But between 2005 and

2010, there was a consistent increase in the number of publications. After 2010, the research productivity slowed down. In the case of SOCP, only during the 2005-2008, there was a steady increase in the number of publications.

Fig. 1 Publication trend on AOCP and SOCP.



Authorship pattern

The authorship pattern is an important bibliometric measure to determine the contemporary communication patterns, productivity, and collaboration among the researchers. We counted the publications based on the number of authors separately for AOCP and SOCP and for both together (Table 2). The majority of the papers were multi-authored papers suggesting a high degree of collaboration in the subject and there were very few singl-authored papers. For instance, when we consider the publications together on both types of ontology construction processes, 89.42% of the total publications are authored by multiple authors (i.e., two or more than two authored papers), and only 9.58% are single-authored papers. Of the 89.42% multi-authored papers, 28.40% papers are authored by three authors. Following this, the publications authored by two (26.85%) and four (23.15%) authors constituted the maximum number of multiple authored papers.

Table 2--Authorship pattern in AOCP and SOCP research

Sl. no.	Type of authorship	AOCP papers	SOCP papers	Total	Authorship pattern(%)
1	Single authored	18	13	31	9.58%
2	Two authored	49	38	87	26.85%
3	Three authored	42	50	92	28.40%
4	Four authored	41	34	75	23.15%
5	Five authored	13	14	27	8.33%
6	Six authored	6	3	9	2.78%
7	Seven authored	0	1	1	0.31%
8	Eight authored	0	1	1	0.31%
9	Nine authored	0	0	0	0
10	Ten authored	0	0	0	0
11	Eleven authored	0	1	1	0.31%
	Total	169	155	324	100.00%

Top contributing authors

To assess the productivity of an author, we counted the author's total publications. For multiple authored papers, we gave equal weight to each author and counted the contribution as one for each of the authors.

There were in all 508 authors (of which 440 were unique authors) who contributed 169 papers on AOCP research, and of them, 51 authors published two or more than two papers. Table 3 presents the top six authors who published three

(not all are mentioned in the list) or more than three papers on AOCP. As can be seen, Yao Liu and Zhifang Sui top the list with five publications each. Besides, there were 38 authors who contributed two papers.

Table 3--Six top contributing authors on AOCP research

Sl. no.	Name(University/Organization)	No. of contributions
1	Yao Liu (Institute of Scientific and Technical Information of China)	5
2	Zhifang Sui (Peking University)	5
3	Than Tho Quan(Nanyang Technological University)	3
4	Sui Cheung Hui(Nanyang Technological University)	3
5	A.C.M. Fong(Nanyang Technological University)	3
6	Yongwei-Hu(Peking University)	3

Similarly, there were in total 489 authors (of them 437 were the unique authors) who contributed 155 papers on SOCP, and of them, 36 authors published two or more than two papers. Table 4 presents the eight top authors who published three or more than three on SOCP. Dunja Mladenić tops the list with seven publications. Following him, Blaž Fortuna and Marko Grobelnik are the two most productive authors with six publications each. The rest of the authors contributed three papers each. There were 28 authors who contributed two papers.

Table 4--Eight top contributing authors on SOCP research

Sl. no.	Name(University/Organization)	No. of contributions
1	Dunja Mladenić (Jozef Stefan Institute	7
2	Blaž Fortuna (Jozef Stefan Institute)	6
3	Marko Grobelnik (Jozef Stefan Institute)	6

4	Alexander Maedche (University of Karlsruhe)	3
5	Steffen Staab (University of Karlsruhe)	3
6	Eva Bolmqvist (Jonkoping University	3
7	Fuji Ren (The University of Tokushima)	3
8	Rodrigo Martínez-Béjar (University of Murcia)	3

Top contributing organizations

The purpose was to measure the contributions of the organizations working in the areas of AOCP and SOCP. The organizations were identified by the author's affiliations. In our dataset, there were multiple authored papers where either all the authors were from the same organization or from different organizations. In case, all the authors of a paper were from the same organization, we counted the contribution of that organization as one. If the authors were from different organizations, we gave equal weight to each organization involved and took count contribution as one for each of them.

Figure 2 shows the organizations who contributed minimum three publications on AOCP. Of them, Peking University (PU, China) tops the list with seven publications. Following this, Harbin Institute of Technology (HIT, China), Institute of Scientific and Technical Information of China (ISTIC, China), and Nanyang Technological University (NTU, Singapore) produced five publications each. Besides them, there are another 185 organizations (not shown in the figure) who contributed minimum one publication. In the case of SOCP (Fig. 3), Jozef Stefan Institute (JSI, Slovenia) tops the list with 11 publications. University of Karlsruhe (UOK, Germany) follows with six publications. There are another 185 organizations who contributed minimum one publication to SOCP.

Fig. 2 Top contributing organizations for AOCP works

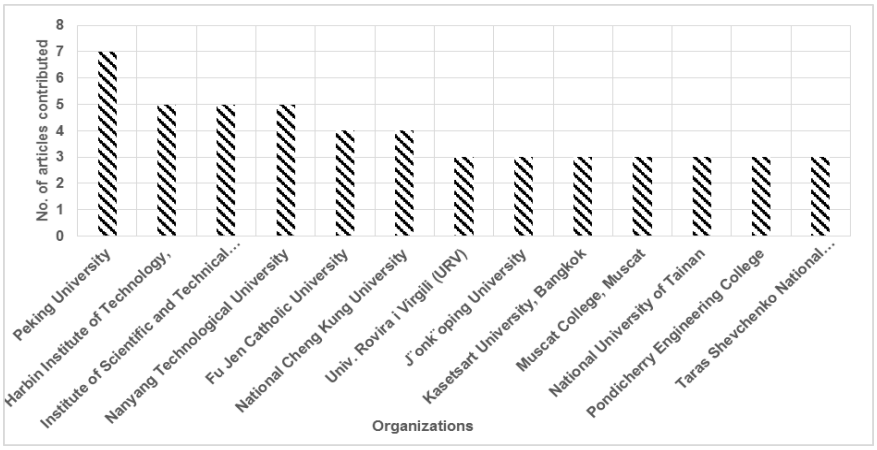


Fig. 3--Top contributing organizations for SOCP works

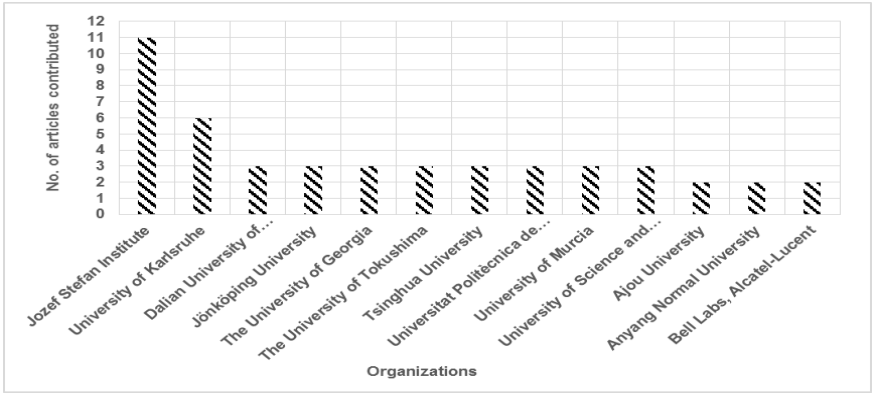


Table 5 presents a list of top 15 organizations (in total 26) that contributed to both the types of ontology construction process. UoK (Germany) tops the list with seven publications- six publications on SOCP and one publication on AOCP. Following this, both Jönköping University (JU,Sweden) and HIT(China) contributed six publications to both types of processes. The majority of the organizations (in total 14 not depicted in the table) contributed two publications one on each of the types of processes. Note that Table 5 does not list the organizations who contributed only one type of the processes. For instance, although JSI led the works on SOCP, but had no publications on AOCP, and hence was not included in the table.

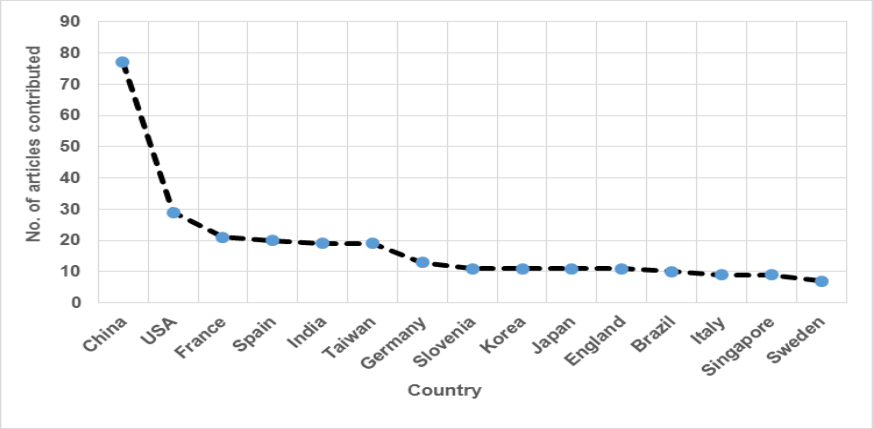
Table 5--Top contributing organizations considering SOCP and AOCP together

Sl. no.	Organization	No. of contributions in SOCP	No. of contributions in AOCP	Total
1	University of Karlsruhe	6	1	7
2	Jönköping University	3	3	6
3	Harbin Institute of Technology	1	5	6
4	Pondicherry Engineering College	1	3	4
5	National University of Singapore	2	2	4
6	Tsinghua University	3	1	4
7	University of Science and Technology Beijing	3	1	4
8	Chinese Academy of Sciences	1	2	3
9	Keio University	2	1	3
10	Wuhan University	2	1	3
11	Shanghai University	2	1	3
12	Anna university	1	2	3
13	Renmin University of China	1	1	2
14	China Agricultural University	1	1	2
15	Université Tunis El Manar	1	1	2

Distribution of papers by country

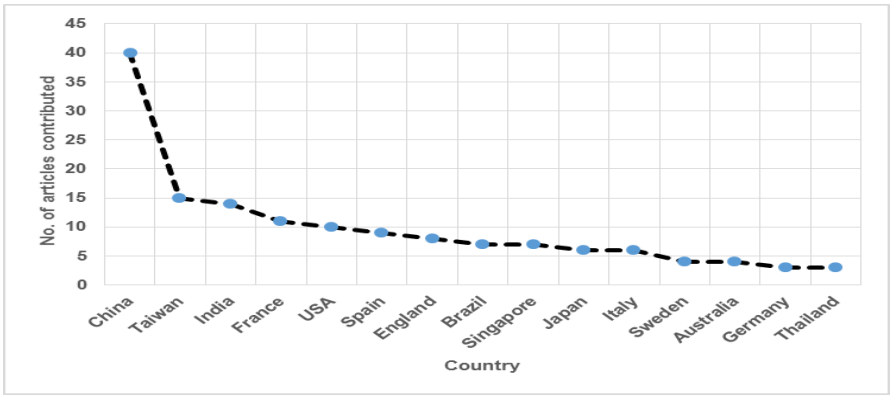
In all, 53 countries contributed to both types of ontology construction processes. The top 15 countries are shown in Fig. 4. China topped the list by contributing 77 papers (21.21%). The second and third most productive countries were USA with 30 papers (7.98%) and France with 21 papers (5.68%). The countries like France, Spain, India, and Taiwan contributed 20 or more papers. The other countries like Germany, Slovenia, Korea, etc., formed the long tail of the graph contributing in both types of ontology construction process.

Fig. 4--Most productive countries on ontology construction processes



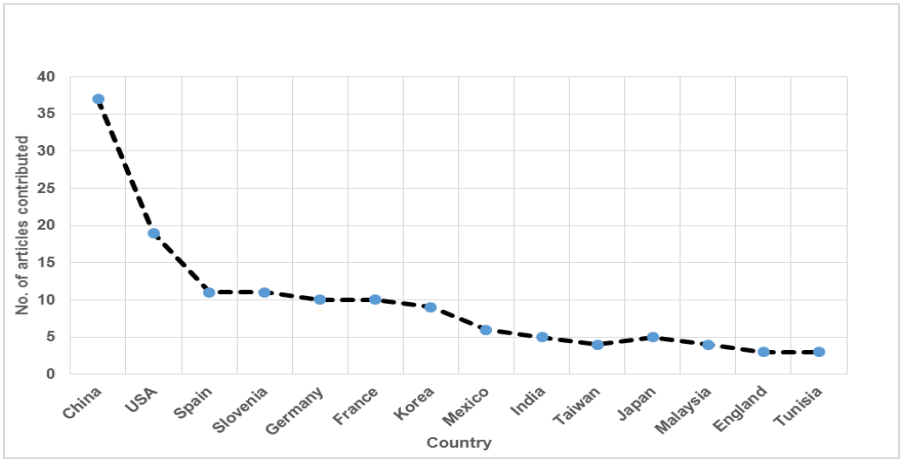
When AOCP and SOCP were considered separately, China still led the list contributing 40 papers (21.85%) on AOCP and 37 papers (20.67%) on SOCP. For AOCP, Taiwan [15 papers (8.19%)], and India [14 papers(7.65%)] emerged as the major countries and France (6.01%), USA (5.46%), Spain (4.91%) and England (4.37%), etc., formed the long tail of the graph (Fig. 5).

Fig 5.--Most productive countries on AOCP



For SOCP, USA with 19 papers (10.61%) emerged as the second highest productive countries following China (20.67%). The other productive countries included Slovenia (6.14%), Spain (6.14%), France (5.58%), Germany (5.58%), and Korea (5.02%) (Fig. 6).

Fig. 6--Most productive countries on SOCP



From the above, we can see that besides China, there are countries like USA, France, Spain, and Japan that have contributed to both the kinds of ontology construction mechanisms. Countries like India, Taiwan, England, Brazil, Singapore, and Italy contributed more on AOCP than on SOCP. Countries like Slovenia, Germany, and Mexico produced more works on SOCP rather than on AOCP.

Distribution of papers by subjects

To examine the multidisciplinary nature of the subject, we looked at the author affiliation to identify the discipline to which the author belonged to. Since we came across different subject affiliations, it was difficult to confine them to a certain number of subjects and therefore we used Dewey Decimal Classification System[17] to group them into the major subject categories.

Our study revealed that authors from the different subject background are working on AOCP and SOCP. As shown in Tables 6 and 7, computer science researchers contributed the maximum to both AOCP [154 publications (62.09%)] and SOCP [120 publications (49.38%)]. The second highest contribution came from the library and information science (LIS) researchers with 24 publications (9.68%) in AOCP and 32 publications (13.17%) in SOCP. Researchers in engineering and allied operations, management and auxiliary services, medicine and health, chemical engineering and related technologies, economics, language, agriculture, etc., also took part in both the types of ontology construction research.

Table 6--Subject-wise contribution to AOCP

Sl. no.	Subject	No. of research

		papers
1	Computer science	154
2	Library and Information science	24
3	Engineering and allied operations	23
4	Unknown*	19
5	Management and auxiliary services	9
6	Medicine and health	5
7	Chemical engineering and related technology	3
8	Economics	3
9	Language	2
10	Agriculture and related technologies	1
11	Earth sciences	1
12	Science	1
13	Social science	1
14	Manufacturing	1
15	Transportation	1

*****For authors, affiliated departments not found in the articles labelled as Unknown.

Table 7--Subject-wise contribution to SOCP

Sl. no.	Subject	No. of research papers
1	Computer science	120
2	Library and Information science	32
3	Unknown*	27
4	Engineering and allied operations	23
5	Management and auxiliary services	10
6	Medicine and health	9
7	Biological Sciences	5
8	Chemical engineering and related technology	5
9	Economics	4
10	Earth sciences	2
11	Language	2
12	Mathematics	2
13	Agriculture and related technologies	1
14	Physics	1

Highly cited papers

The highly cited AOCP and SOCP papers as per Google Scholar and Sopus are given in Tables 8 and 9. In the case of AOCP, the paper "Yago: A large ontology from Wikipedia and wordnet" received the higher number 675 Google Scholar citations.

Table 8—Highly cited AOCP papers

Sl. no.	Paper	No. of Google Scholar citations	No. of Scopus citations
1	Suchanek F M, Kasneci G, and Weikum G, Yago: A large ontology from wikipedia and wordnet, *Web Semantics: Science, Services and Agents on the World Wide Web*, 6 (3) (2008) 203-217.	675	307
2	Cimiano P, Hotho A and Staab S, Learning concept hierarchies from text corpora using formal concept analysis, *Journal of Artificial Intelligence Research (JAIR)*, 24 (1) (2005) 305-339.	559	311
3	Tho Q T, Hui S C, Fong A C M and Cao T H, Automatic fuzzy ontology generation for semantic web, *IEEE Transactions on Knowledge and Data Engineering*, *18* (6) (2006) 842-856.	392	260
4	Shamsfard M, and Barforoush A A, Learning ontologies from natural language texts, *International Journal of Human-Computer Studies*, 60 (1) (2004) 17-63.	247	121
5	Khan L and Luo F, Ontology construction for information selection. In *Proceedings of 14th IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, 4-6 Nov. 2002 , p. 122-127.	205	86
6	Velardi P, Navigli R, Cuchiarelli A and Neri, R, Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. *Ontology Learning from Text: Methods, evaluation and applications*, (2005) 92-119.	190	Not indexed

Table 9 presents the top six articles on SOCP. The paper "A method for semi-automatic ontology acquisition from a corporate intranet" received the highest number of 284 Google Scholar citations. It is interesting to see that the first three ranked articles, based on GoogleScholar citation, were not found in Scopus because they were conference proceedings articles. Also, the other articles, as per Scopus, received very few citations.

Table 9--Highly cited SOCP papers

Sl. no.	Title	No. of GoogleScholar Citations	No. of Scopus Citations
1	Kietz J U, Maedche, A and Volz, R, A method for semi-automatic ontology acquisition from a corporate intranet. In *EKAW Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000*.	284	Not indexed
2	Maedche A and Staab S, Semi-automatic engineering of ontologies from text. In *Proceedings of the 12th international conference on software engineering and knowledge engineering* ,Chicago, IL, USA, 6-8 July 2000, p. 231-239.	238	Not indexed
3	Bisson, G., Nédellec, C., & Canamero, D. (2000, August). Designing Clustering Methods for Ontology Building-The Mo'K Workbench. *ECAI'2000 Workshop on Ontology Learning, Proceedings of the First Workshop on Ontology Learning OL'2000*, Berlin, Germany, August 25, 2000	193	Not indexed
4	Fortuna B, Mladenič D and Grobelnik, M, Semi-automatic construction of topic ontologies. In *Semantics, Web and Mining* , Springer, Berlin, Heidelberg, 2006, p. 121-131.	126	29
5	Fortuna B, Grobelnik M, and Mladenic D, OntoGen: semi-automatic ontology editor. In *Symposium on Human Interface and the Management of Information* . Springer, Berlin, Heidelberg, July 2007, p. 309-318.	118	39
6	Maedche A, and Staab S. (2001, May). Learning ontologies for the semantic web. In *Proceedings of the Second International Conference on Semantic Web-Volume 40* (pp. 51-60). CEUR-WS. org.	113	Not indexed

Conclusion

The present study revealed that during the initial years there was a gentle growth in the number of publications in both AOCP and SOCP reducing the time, human labour and infrastructural cost of the process of ontology construction. However, with passage of time, research publications tapered. These can be attributed to the unavailability of mature tools and technologies (especially the learning techniques) required to carry forward the research and also the lack of infrastructure, funding, and expertise which are the essential component of this research. Ontology research, as we know, is an interdisciplinary area of research and requires expertise in data and knowledge representation, natural language processing, information extraction, and so forth. Naturally, we found a high degree of research collaboration between the researchers from various disciplines namely computer science, library and information science, philosophy, mathematics, linguistics and so forth. In future, as a continuation of this study, we plan to analyse and study the collaborative network of the research in detail.

References

1. Dutta B, Examining the interrelatedness between ontologies and Linked Data, *Library Hi Tech*, 35 (2) (2017) 312-331.
2. Gruber T R, A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5 (2) (1993) 199-220.
3. Qu C, Liu F, Yu H, Yuan R and Wang A, User oriented semi-automatic method of constructing domain ontology, In *Int. Symposium on Intelligence Computation and Applications*, Springer, Singapore, 2015 p. 553-561.
4. Dutta B, Chatterjee U, and Madalli D P, YAMO: yet another methodology for large-scale faceted ontology construction, *Journal of Knowledge Management*, 19 (1) (2015) 6–24.
5. Dutta B, Giunchiglia F and Maltese V, A Facet-based methodology for Geo-spatial modeling, In C Claramunt, S Levashkin, and M Bertolotto (Eds.) *International Conference on GeoSpatial Semantics*, LNCS, 6631, (Springer-Verlag, Berlin), 2011 p.133–150
6. Campbell D, Picard-Aitken M, Côté G, Caruso J, Valentim R, Edmonds S, Williams G T, Macaluso B, Robitaille J P, Bastien N and Laframboise M C, Bibliometrics as a performance measurement tool for research evaluation: the case of research funded by the National Cancer Institute of Canada, *American Journal of Evaluation*, 31 (1) (2010) 66–83.
7. Singh G, Mittal R, and Ahmad M, A bibliometric study of literature on digital libraries, *The Electronic Library*, 25 (3) (2007) 342-348.
8. Chandrashekara M , Harinarayana N S, Mulla K R and Ramachandra S, Bibliometric study of literature on digital libraries. *In National Seminar on Webometrics, Informetrics and Scientometrics : Measuring Scientific and Technological Progress of India*, Karnataka University, Dharwad , 21-22 December 2009.

9. Syamili C and Rekha R V, Trend of research on Ontology: a study based on ProQuest and DART, In *Proceedings of International Conference on Knowledge Modelling and Knowledge Management*, Documentation Research and Training Centre Bangalore, India , December 2013, p. 161-171.
10. Lima J F, Amaral, C M G and Molinaro L F R, Ontology: An analysis of the literature. In *International Conference on ENTERprise Information Systems*, Berlin, Heidelberg, October 2010, p. 426-435.
11. Zhu Q, Kong X, Hong S, Li J and He Z, Global ontology research progress: a bibliometric analysis, *Aslib Journal of Information Management*, 67 (1) (2015) 27-54 .
12. Kshitig A and Gupta, B M, Semantic web: A quantitative analysis of world publications output (2001-2010), *DESIDOC Journal of Library & Information Technology*, 31 (4) (2011) 253-261.
13. Dutta B, Toulet A, Emonet V, and Jonquet C, New Generation Metadata vocabulary for Ontology Description and Publication. In E. Garoufallou et al. (Eds.) *Communications in Computer and Information Science (CCIS)*, 755, (Springer publication, Tallinn), 2017 p.173-185.
14. Fernández-López M, Overview of methodologies for building ontologies. *Procceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods*, Stockholm, Sweden 1999
15. Boudabous M M, Belguith L H and Sadat F, Exploiting the Arabic Wikipedia for semi-automatic construction of a lexical ontology. *International Journal of Metadata, Semantics and Ontologies*, 8 (3) (2013) 245-253.
16. Gherasim T, Harzallah M, Berio G and Kuntz P, Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application, *Advances in Knowledge Discovery and Management*, (Springer publication, Berlin) 2013 p. 177-201.
17. Dewey M, Dewey Decimal Classification and Relative Index, in Joan S. Mitchell (Ed.), 22nd ed., (OCLC Forest Press, USA) 2003.