This is a preprint version of a published paper. For citing purposes please use:

Ivanjko, Tomislav; Špiranec, Sonja. Bibliometric Analysis of the Field of Folksonomy Research // Proceedings of the 14th International Symposium on Information Science (ISI 2015) (Schriften zur Informationswissenschaft ; Bd. 66) / Pehar, Franjo ; Schloegl, Christian ; Wolff, Christian (ur.). Gluckstadt : Verlag Werner Hulsbusch, 2015. 370-379

Bibliometric analysis of the field of folksonomy research

Tomislav Ivanjko, Sonja Špiranec
University of Zagreb
Faculty of Humanities and Social Sciences
Department of Information and Communication Sciences
Ivana Lučića 3
10000 Zagreb
Croatia
tivanjko@ffzg.hr; sspiran@ffzg.hr

Abstract

The area of researching folksonomies is still in development, so theoretical perspective and research methods are still being defined. This study conducts a webometric and bibliometric analysis of the folksonomy research in the Library and Information science (LIS) field by collecting data from Web of Science (WOS), SCOPUS and Google Scholar in July 2014. It utilizes a total of 346 papers with 2660 citations from WOS and 1581 papers with 8848 citations from SCOPUS. In addition, Google Scholar database search was also included for providing a wider coverage of works published in conference proceedings, books and to include a wider journal base. Based on these results, research identifies most influential papers and authors across all three databases.

Keywords: folksonomies, social tagging, collaborative tagging, bibliometric analysis, webometrics.

Introduction

With the rise of Web 2.0, a new wave of user participation in creating and describing online resources instigated a new approach in knowledge representation – folksonomy. Folksonomy relies on the process of collaborative tagging, where many users add metadata in the form of keywords to shared content (Golder and Hubermann, 2006; Mathes, 2004). The totality of these user-generated keywords (tags), gathered around any different platform or resource creates a folksonomy (Peters, 2009). Within this framework, different approaches are possible, where only one of the elements can be analyzed (for example, analyzing the linguistic characteristics of a chosen tag corpus) or, more often, the relationship between two elements is investigated (such as the relationship between tags and resources, identifying possible differences in tagging different types of resources). Since the coining of the term the new research topic emerged in the field of Information Science dealing with the structure, use and application of folksonomies in the field of knowledge organization and representation and information retrieval. This paper aims to explore the body of literature currently present on the topic of folksonomies inside the field of Information Science by using webometric tools and methods to identify key concepts and bibliometric methods to identify key authors and papers in the field.

Identifying key concepts

Since the coining of the term folksonomy (Vander Wal, 2004) different competing terms emerged to describe the field of research. Peters (2009) provided an exhaustive literature review regarding the terminology use and listed the most prominent ones "ethnoclassification", "communal categorization", "democratic indexing", "mob indexing", "social classification system", "social indexing", "user-generated metadata", "collaborative tagging", "social tagging" and "folksonomy". Following the methodology from our previous work (Lasić-Lazić, Špiranec and Ivanjko, 2014) where the focus was on the content analysis of the field, a webometric analysis of the competing terms was conducted using the tool Webometric Analyst 2.0 (http://lexiurl.wlv.ac.uk). Following the method from Thelwall (2013) a cross-domain web impact assessment via web mentions was conducted in July 2014 including the most mentioned terms. Web impact assessment (WIA) is the evaluation of the “web impact” of documents or ideas by counting how often they are mentioned online. The underpinning

idea is that, other factors being equal, documents or ideas having more impact are likely to be mentioned online. The tool returns a number of different metrics, the most reliable being the number of domains due to the possibility that text or links are copied across multiple pages within a web site (Thelwall, 2013). The results are presented in Table 1.

Table 1. Cross-domain web mentions of competing terms

TERM	**RESULT**
folksonomy	575
user-generated metadata	473
social tagging	453
collaborative tagging	298
social classification system	240
social indexing	188
ethnoclassification	166
democratic indexing	51
mob indexing	50
communal categorization	40

As we can see from the analysis the most widely used term is “folksonomy” with 575 cross-domain mentions, followed closely by “user-generated metadata” and “social tagging”. By examining the results in detail it became obvious that the terms “social classification system” and “ethnoclassification” yielded such high results not because they relate to a concept found in the literature but its origin derives from sociology where they denote a completely unrelated notions so it was clear they should be

excluded from any literature search as it would generate a lot of false results not related to our field of interest.

Identifying key authors and papers

The results of the webometric analysis gave us a starting point for constructing a Boolean query (folksonom* OR "social indexing" OR "social tagging" OR "user-generated metadata" OR "collaborative tagging") in order to include all the relevant concepts when searching the databases. Three different sources included in the search were Web of Science (http://wokinfo.com/products_tools/multidisciplinary/webofscience/), SCOPUS (http://www.info.sciverse.com/scopus) and Google Scholar (http://scholar.google.com). In addition to searching the standard bibliographic databases in the field, Google Scholar was also included in order to provide a better insight into publications outside high impact journals, such as works published in conference proceedings, books and to include a wider journal base as suggested by Harzing (2008). Some studies have shown that although Google Scholar ranking algorithm weighs heavily on articles' citation counts (Beel and Gipp, 2007), top ranked articles are not necessarily those with the highest citation count so the search of Google Scholar database was conducted using software Publish or Perish 4 (Harzing, 2007) to identify relevant papers.

Table 2. Summarized data on sources included in the analysis

DATABASE	NO. OF PAPERS	NO. OF CITATIONS	h-INDEX
WOS	346	2660	21
SCOPUS	1581	8848	41
GOOGLE SCHOLAR	1000+	31234	80

As we can see, fewest papers on the topic are published through WOS, with the lowest h-index. As for Google Scholar, the software Publish or Perish 4 is

limited to processing the first 1000 results so the total number of articles could not be calculated but instead first 1000 results were analyzed. These results show that there is a notable interest in the field of research with an already respectable number of published articles in high impact journals. To get some insight into the most influential papers and authors in the field 20 most cited articles from WOS and SCOPUS were compared and a total of 7 articles were found cited both in top 20 WOS and SCOPUS. If we take Google Scholar into account, then only 3 papers are present in top 20 for all 3 databases (*Usage patterns of collaborative tagging systems* (2158); *Information retrieval in folksonomies: Search and ranking* (725); *Ontologies are us: A unified model of social networks and semantics* (619)). It is clear from the results that the paper published by Golder and Huberman (2006) (*Usage patterns of collaborative tagging systems*) is by far the most cited paper in the field, having attracted most citations across all three databases. Also it should be noted that the paper from Mika (2008) (*Ontologies are us*) is present in both WOS and SCOPUS in two slightly different versions (a conference paper from 2005 was rewritten as a journal article in 2007) but both papers share the same basic concepts and ideas, so from the intellectual point of view they should be regarded as one article. If we take that into account then the citation number for that paper raises significantly making it the second most cited article across analyzed databases. Although there is a fairly large amount of papers published, it is obvious from the results that the field is very heterogenic, with only several papers being present as top cited in all the databases. Since both WOS and SCOPUS provide access based on subscription fees and Google Scholar is free to access, researchers in the field trying to get insight into the topic could start with very different papers based on their institution financial power with only three articles being present in the top 20 most cited articles across all three databases. When we look at the categories from which the journals with the most citations stem, there are two main fields that are interested in the topic of folksonomies: Computers Science and Library and Information Science. Articles written from a Computer Science perspective are concerned mostly with using folksonomies in exploring the ways in which user tags can improve the effectiveness of different systems and information retrieval (for example, extracting meaningful data for creating partial ontologies as a basis for the Semantic Web). On the other hand, Library and Information Science field is more interested in researching user motivations for tagging (to enable better communication with its patrons) and the potential of user tags in enhancing

resource description (to complement standard KOS methods). A more detailed content analysis of the approaches in the field can be found in the work published by Lasić-Lazić, Špiranec and Ivanjko (2014).

Co-citation analysis

One of the basic methods of bibliometrics is counting co-citations, a method for identifying influential authors and displays their interrelationships from the citation record (White and McCain, 2009). In order to provide that kind of insight in the field of folksonomy research, a co-citation analysis of the papers from both WOS and SCOPUS was carried out. From the SCOPUS database a total of 1581 articles with 8848 citations were analyzed. The analysis was carried out using the software Bibexcel (http://www8.umu.se/inforsk/Bibexcel/) a bibliometric toolbox for most types of bibliometric analysis (Persson, Danell and Wiborg Schneider, 2009). Bibexcel was used for processing the data, while Pajek (http://pajek.imfm.si/doku.php) was used for visualization of the data as used in Batagelj and Mrvar (2003). Figure 1 shows the co-citation graph from SCOPUS records where the size of vertices indicates the number of citations while the thickness of lines indicates the number of co-citations between authors. To reduce the complexity of the visualization, figure shows only authors that have more than 20 co-citations.

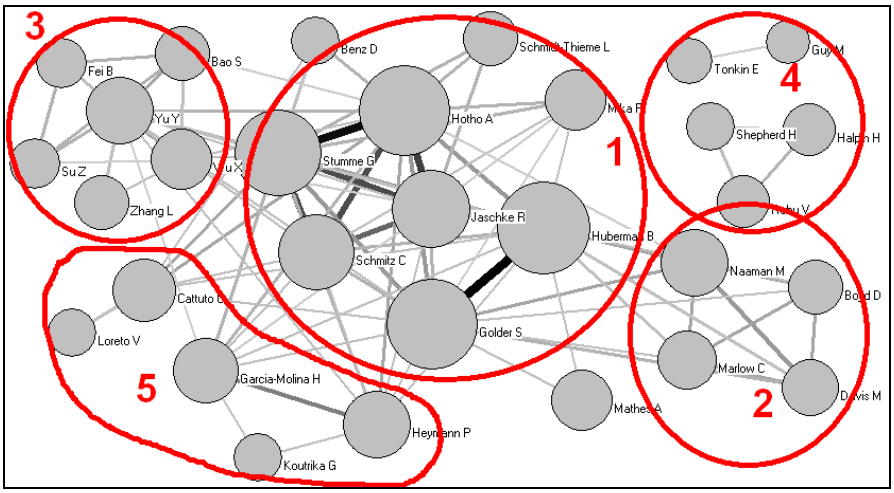

Figure 1: Co-citation graph based on 8848 citations from SCOPUS database

As we can see from the graph, there are roughly 5 main clusters of authors that are interconnected with high number of co-citations. Again, in the centre of the graph (1) there are authors of the two most cited papers (*Golder and Hubermann; Hotho, Jäschke, Schmitz, and Stumme*) that have the strongest co-citation links. Then there are three clusters of authors that are on the outskirts of that centre cluster (2, 3, and 5) that have strong mutual connections and are also strongly connected to the central cluster. And finally there are authors that have a high number of citations but are not that strongly co-referenced by other authors (4). This analysis revealed some new influential authors and papers in the field such as Marlow, Naaman, Boyd and Davis (*HT06, tagging paper, taxonomy, Flickr, academic article, to read*) and Xu et al. (*Exploring folksonomy for personalized search*) but it also confirmed previous results, identifying the authors previously mentioned. The final co-citation analysis was conducted on 2660 citations extracted from the WOS database. This time, data was analyzed not only to identify co-citation clusters but also included publication year and shortened journal names for the cited articles so that a time and origin component is added to the analysis enabling better overview of field development. Figure 2 shows the co-citation graph from WOS records where the size of vertices indicates the number of citations while the thickness of lines indicates the number of co-citations between authors. To reduce the complexity of the visualization, figure shows only authors that have more than 10 co-citations.

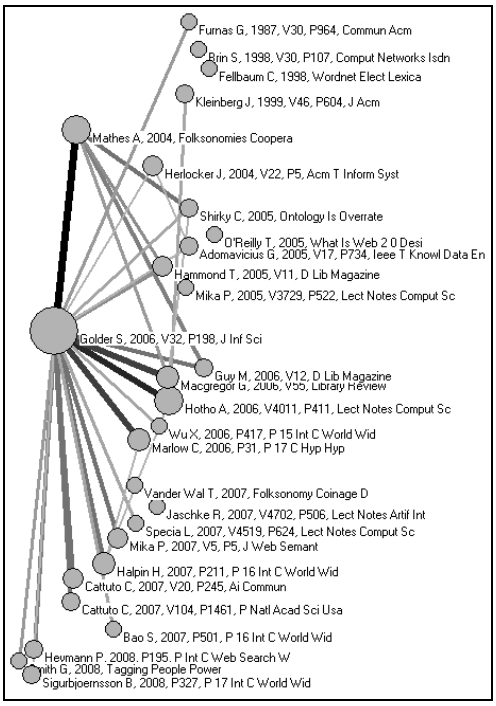

Figure 2: Co-citation graph based on 2660 citations from WOS sorted by publication year

As we can see from the graph, the starting point for field development is the 2004 article by Mathes (*Folksonomies-cooperative classification and communication through shared metadata*) and the central article is again the 2006 paper by Golder and Hubermann (*Usage patterns of collaborative tagging systems)*. Such visualization that includes a time component is a great start for a possible reading list for new researchers in the field where the development of the topic is clearly outlined with closely 30 key papers in the field. We can see that a large amount of most cited papers is from 2006-2007 where the field of research was defined and when the scientific debate was at its peak. Again, here we can see that the journal that published most cited articles are from the field of Computer Science and Library and Information Science.

Conclusion

This paper aimed to provide insight into the field of research on the topic of folksonomies by combining webometric and bibliometric tools in analysis of the data found in the most prominent databases in the field of Information Science. Since the field is fairly new with terminology and methods still being discussed, first analysis used the webometric method of counting cross-domain web mentions of competing terms. The results have shown that the most commonly used terms when describing the field of research are folksonomy, user-generated metadata, social tagging and collaborative tagging with the term folksonomy being the single most used term in use.

Based on these insights Web of Science, SCOPUS and Google Scholar were queried with a Boolean query including all the commonly used terms to ensure all the relevant papers were reached. These queries resulted with 346 papers with 2660 citations from WOS, 1581 papers with 8848 citations from SCOPUS and 1000 papers with 31234 citations from Google Scholar. Such numbers clearly showed that the field is already well developed with a respectable number of papers published on the topic.

Next, the top 20 most cited articles from each database were compared in order to identify key paper and authors. It was shown that only 7 papers are present in both WOS and SCOPUS top 20, and when taking into account Google Scholar that number falls down to only three articles. This has shown that the field is very heterogenic, with only several papers being present as top cited in all the databases. When we examined the journals where the most cited papers were published, two main subfields of Information Science that are interested in the topic of folksonomies arose: Computer Science and Library and Information Science. Computer Science perspective was concerned more with using folksonomies and tags to improve the effectiveness of different systems and information retrieval, especially in the domain of Semantic Web, while Library and Information Science papers were more interested in researching user motivations for tagging and the potential of user tags in enhancing resource description.

Finally, a co-citation analysis was conducted on the citation data from both SCOPUS and WOS databases. The data from SCOPUS has given insight into the most influential authors in the field and their mutual connections, while

the data from WOS included a time component that enabled the tracking of the field development. The best identification of key papers and authors is achieved in Figure 2 which gives a chronological reading list for all new researchers in the field trying to explore the heterogenic field of folksonomy research.

This analysis confirmed that the field of folksonomy research is a relevant topic inside the Information Science field, with a respectable number of papers published in the most prominent databases for the field. It identified key authors and papers, as well as provided a chronological list of key papers and their mutual connections by conducting a co-citation analysis. Further analysis should include a topic analysis of the papers published on the topic in recent years thus providing insight into the current state of research and possible future directions.

References

Batagelj, V., Mrvar, A. (2003). Pajek - analysis and visualization of large networks. In: Graph Drawing Software, Jünger, M., Mutzel, P., (eds.) Springer, Berlin. 77-103.

Beel, J. and Gipp, B. (2007). *Google Scholar's ranking algorithm: the impact of articles' age: an empirical study*. Proceedings of the 6th International Conference on Information Technology: New Generations. 160-164.

Golder, S.A. and Huberman, B.A. (2006). *Usage patterns of collaborative tagging systems*. Journal of Information Science. 32, 2, 198-208.

Harzing, A. W. (2008). *Google Scholar: a new data source for citation analysis*. Retrieved January 22, 2015 from: http://www.harzing.com/pop_gs.htm.

Harzing, A.W. (2007). *Publish or Perish*. Retreived January 22, 2015 from: http://www.harzing.com/pop.html.

Lasić-Lazić, J.; Špiranec, S. and Ivanjko, T. (2014) *Tag-Resource-User: a review of approaches in studying folksonomies*. QQML Journal. 3, 683-692.

Mathes, A. (2004). *Folksonomies - cooperative classification and communication through shared metadata*. Computer Mediated Communication, 47, 10.

Persson, O. D., Danell, R. and Wiborg Schneider, J. (2009). *How to use Bibexcel for various types of bibliometric analysis*. In: Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday, Åström, F. et al. (eds.), International Society for Scientometrics and Informetrics. 9-24.

Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0*. Berlin: De Gruyter.

Thelwall, M. (2013). *Webometrics and Social Web research methods*. Retrieved January 22, 2015 from: http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf

Vander Wal, T (2007). Folksonomy coinage and definition. Retreived January 22, 2015 from http://vanderwal.net/folksonomy.html.

White, H. D., and McCain, K. W. (2009). *Visualizing a discipline: An author co-citation analysis of information science, 1972-1995*. Journal of the American Society for Information Science. 49, 4, 327-355.