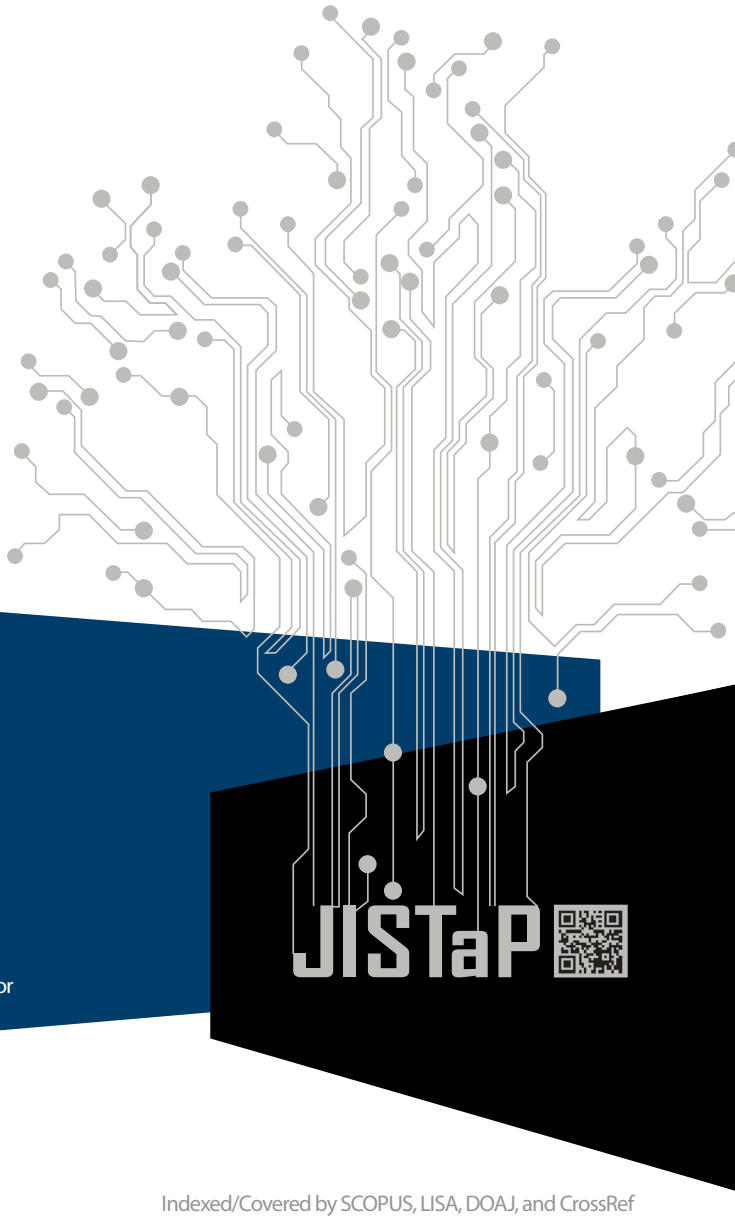
eISSN : 2287-4577 pISSN : 2287-9099
http://www.jistap.org
Vol. 6 No. 4 December 30, 2018

Journal of Information Science Theory and Practice



KISTI
www.kisti.re.kr

Indexed/Covered by SCOPUS, LISA, DOAJ, and CrossRef

General Information

Aims and Scope

The *Journal of Information Science Theory and Practice (JISTaP)* is an international journal that aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics. JISTaP will be published quarterly, issued on the 30th of March, June, September, and December. JISTaP is indexed in the Scopus, Korea Science Citation Index (KSCI) and KoreaScience by the Korea Institute of Science and Technology Information (KISTI) as well as CrossRef. The full text of this journal is available on the website at http://www.jistap.org

Indexed/Covered by

Scopus LISA DOAJ DIRECTORY OF OPEN ACCESS JOURNALS Crossref

Publisher

Korea Institute of Science and Technology Information
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6102
(F) +82-2-3299-6067
E-mail: jistap@kisti.re.kr
URL: http://www.jistap.org

Managing Editor: Suhyeon Yoo, Eungi Kim

Copy Editor: Ken Eckert

Design & Printing Company: SEUNGLIM D&C

4F, 15, Mareunnae-ro, Jung-gu, Seoul, Republic of Korea
(T) +82-2-2271-2581~2
(F) +82-2-2268-2927
E-mail: sdnc@sdnc.co.kr

Editorial Board

Table of Contents

JISTaP

Vol. 6 No. 4 December 30, 2018
Journal of Information Science Theory and Practice • http://www.jistap.org

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 06-16, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.1

Citations to arXiv Preprints by Indexed Journals and Their Impact on Research Evaluation

Antonia Ferrer-Sapena*
Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: anfersa@upv.es

Fernanda Peset
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: mpesetm@upv.es

Rafael Aleixandre-Benavent
Instituto de Gestión de la Innovación y del Conocimiento-Ingenio (CSIC-Universitat Politècnica de València), UISYS, Universitat de València, Valencia, Spain
E-mail: Rafael.Aleixandre@uv.es

Enrique A. Sánchez-Pérez*
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain
E-mail: easancpe@mat.upv.es

ABSTRACT
This article shows an approach to the study of two fundamental aspects of the prepublication of scientific manuscripts in specialized repositories (arXiv). The first refers to the size of the interaction of "standard papers" in journals appearing in the Web of Science (WoS)—now Clarivate Analytics—and "non-standard papers" (manuscripts appearing in arXiv). Specifically, we analyze the citations found in the WoS to articles in arXiv. The second aspect is how publication in arXiv affects the citation count of authors. The question is whether or not prepublishing in arXiv benefits authors from the point of view of increasing their citations, or rather produces a dispersion, which would diminish the relevance of their publications in evaluation processes. Data have been collected from arXiv, the websites of the journals, Google Scholar, and WoS following a specific ad hoc procedure. The number of citations in journal articles published in WoS to preprints in arXiv is not large. We show that citation counts from regular papers and preprints using different sources (arXiv, the journal's website, WoS) give completely different results. This suggests a rather scattered picture of citations that could distort the citation count of a given article against the author's interest. However, the number of WoS references to arXiv preprints is small, minimizing this potential negative effect.

Keywords: preprint, research evaluation, arXiv, impact, bibliometric analysis, citation

Open Access

Accepted date: September 12, 2018
Received date: May 03, 2018

***Corresponding Authors:**
Antonia Ferrer-Sapena
Professor
Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain
E-mail: anfersa@upv.es

Enrique A. Sánchez-Pérez
Professor
Instituto Universitario deMatemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain
E-mail: easancpe@mat.upv.es

1. INTRODUCTION

The prior publication of scientific manuscripts in electronic preprint repositories has proved, since the beginning of this practice, to be a useful way of increasing the visibility and accessibility of research work. Since the last years of the past century a big amount of papers have shown that, as a direct consequence of prepublication, citations of previously posted articles increase. In this paper we are interested in showing a particular aspect of the citation of research documents deposited in arXiv whose consequences would not be so positive. Specifically, we wanted to analyze what is the total number of arXiv documents that get citations in "standard journals"—journals appearing in Web of Science (WoS), now Clarivate Analytics—and what is the citation dynamics of those documents. Since these kinds of citations are, in a way, beyond the reach of typical counting tools, we try to provide quantitative information on how many citations could be missed. Our interest is to measure to what extent these missed citations may harm the interests of authors undergoing a bibliometric evaluation. In this sense, it is well known that some national research evaluation agencies—such as the Polish or the Spanish—use citation counts for research evaluation.

Therefore, we analyzed citations to preprints that appeared in WoS and were previously posted in arXiv. The reason we chose arXiv is that it has become a prototype of a universal e-preprints repository for physics, mathematics, and computer science. Consequently, we restrict our attention to these disciplines. As sources of citations for further analysis, we mainly use WoS and Google Scholar. To complete the chart, other sources have also been used, such as citations provided by the websites of individual journals.

First, let us explain some basic facts about the context of our work and the previous research that has been done on the subject. Since arXiv was an early initiative in the field of e-preprint repositories, some authors have carried out several analyses of the motivation of researchers to use it over the last twenty years. The general opinion is that the main reason for uploading a manuscript to arXiv is the same as the one that caused classical preprint circulation (hard copies). The outline of the main practical motivations of the authors with regard to preprint publication as presented in the paper by Pinfield (2005) should be mentioned here. Although this paper is not recent, an inspection of the authors' reasons for preprint publication in the current literature suggests that they have not changed at all. The first motive explained there for registering a manuscript in arXiv is that this is a way of setting priorities when presenting a new idea or research result. Preprints provide a way to register them without having to wait for standard journal publication. The second objective of e-preprint publication is rapid dissemination. Preprint circulation is clearly faster than formal peer-reviewed publication. The third reason is that the circulation of a preprint is a way of improving the finished article by considering the comments of colleagues for the drafting of the final version. More works about motivation for preprint publication supporting these ideas can be found in Ardichvili, Page, and Wentling (2003), Kim (2011), and Zha, Li, and Yan (2013) (see also the references therein). It should also be mentioned that the new social media and other technological tools of the digital era have changed the role of preprint publication in the scholarly communication process. They have produced a clear diversification of "knowledge objects" (e-preprints, datasets, open access, on-going manuscripts, short letters…). Although there is no discussion here of how this might affect the prepublication of manuscripts, it is clear that the role of traditional documents in the dissemination of science, and therefore of preprints themselves, will change. The reader can find more information in the paper by Haustein (2016) and the references therein. In this regard, other interesting contributions have also been made from sociology. The prepublication of e-prints and, in general, open access initiatives have greatly changed the classic world of scientific publishing in the way that Bohlin (2004) had already noticed: New internet technologies had changed the needs and interest of potential users of scientific publishing, producing a transformation in academic communication. The study of how these changes might also modify the evaluation of the research would also be interesting, and would help to show a general picture of the problem we are facing. However, this issue is outside the scope of this document, where we provide only some bibliometric information and general explanations.

However, there are other reasons for using the "electronic version" of this classic practice, which is represented by repositories such as arXiv. In some cases, and also depending on the scientific field, articles are deposited in arXiv in the author's version after their acceptance and even after their publication in a standard journal. In fact, this practice is proposed and accepted by major publishers such as Springer. In the "Self-archiving Policy" section of the website, the following sentence appears in the Copyright Transfer Statement: "Authors may also deposit this version (the author's version) of the article in any repository, provided it is only made publicly available 12 months after official publication or later." This should be understood in

the context of the open access movement, to facilitate the dissemination of research results outside the business of scientific publishers (Klein Broadwell, Farb, & Grappone, 2016). Although the reasons why authors use the arXiv repository in this way is also an interesting topic of study, we will not analyze it in this article, since a priori it does not seem to interfere too much with article citations.

From the point of view of bibliometric parameters, the advantages of prepublication have been explained in terms of the following facts, which are widely accepted. The reader can find the following classification in the paper by Kurtz et al. (2005): Open Access Postulate: Free access papers can be read more easily, and so get cited more frequently. Early Access Postulate: posted preprints are available sooner and thus gain primacy, increasing citations. Self-selection Bias Postulate: The authors select their most important (and so more citable) papers to post them. This explanation serves to justify the empirical fact that preprint publication indeed increases the total number of citations. Some early studies have already noted this (see for example Fig. 4 in Henneken et al. [2006] and the references therein); however, other works warn that this is not always the case (Kurtz & Henneken, 2007). A 2010 report lists 27 studies in which this positive conclusion is found, compared to four studies in which the conclusion is the opposite (Swan, 2010).

Considering all these issues, we study how the publication of a manuscript in arXiv has effects in terms of benefits for the authors, from the point of view of increasing the number of citations. Of course, prepublication ensures a better opportunity in the diffusion of the work, but it is not easy to know to what extent this practice can actually improve some of the bibliometric parameters of authors, such as the number of citations of their papers or the publication of their articles in journals with higher citation rates. It is already well understood that prepublication affects the citation dynamics of a given paper, and should be taken into account in any comprehensive citation analysis (Neuhaus & Daniel, 2008). In particular, some specific statistical studies have been carried out on arXiv. The main current references are the exhaustive papers by Larivière et al. (2014) and Li, Thelwall, and Kousha (2015), but also the earlier works by Kurtz et al. (2005), Henneken et al. (2006), and Kurts et al. (2007). The statistical studies presented there—mainly the first one by Larivière—give a clear idea of the relationship between arXiv and the main databases of scientific articles. This type of analysis is not reproduced here: Our aim is to provide more specific information on the aspects of this relationship explained above and to discuss them together with some empirical opinions often expressed by researchers in the fields of physics and mathematics.

Let us finish this section by explaining the main conclusions presented in the existing literature on the subject. Some studies confirm that documents deposited in arXiv receive more citations and are cited before (see p. 2053 in the paper by Moed, 2007). According to this reference, the main advantage of using arXiv from the point of view of bibliometric parameters is that citations occur earlier. The author explains that, although the number of citations does not seem to increase due to the use of arXiv, the scientific community begins to process the information earlier, so the citations appear earlier. This obviously means an improvement in the promotion of the document. However, there are other studies on particular contexts in which this effect is not detected (Davis & Fromerth, 2007), although they are in the minority. There is also evidence that the quality of papers previously published in arXiv is generally above average (Moed, 2007; Davis & Fromerth, 2007); measuring quality is always delicate, so these results must be considered in the appropriate context. More studies on arXiv and the dissemination of the manuscripts deposited in it can be found in the papers by Haque and Ginsparg (2009, 2010), Manuel (2001), and Youngen (1998). In general, it must be said that all of them demonstrate some aspects of the advantages of prepublication that we have explained above: impact, parallel form of distribution, independence from the delays produced by the standard publication process, etc. A different methodology has been used in the present document. We have considered only the total set of preprints that appear in arXiv and that have been cited at least once in a regular WoS journal.

Thus, the sample of papers is not the same as the one that has been analyzed in other works. The results of the use, citations, and journal publication of the articles in the selection will be explained and some conclusions will be presented. Mainly, the dynamics of the citations will be explained, considering preprints as if they were regular papers in standard journals, as well as the statistical data on the areas to which these papers belong. A recent paper that studies the dynamics of publication/citation in arXiv in comparison to other sources and that is related to our methodology in a sense is the one found in the paper by Bar-Ilan (2014). This work is dedicated to the area of astrophysics. It analyzed the work of one hundred European astrophysicists indexed by Scopus, including the number of manuscripts deposited in arXiv and the number or brands of Mendeley readers. Although arXiv is widely used in astrophysics, it shows that more documents appear in Scopus than in arXiv; it also shows that the number of

marks in Mendeley is significantly lower than the number of citations in Scopus. In this case, the comparison between the data sources was made based on the names of the authors and the titles of the publications, thus being more related to our methodology.

In order to facilitate easy understanding of the arguments in this paper, we recall that the term "standard publication" of an arXiv manuscript will be used when it is published in a journal appearing in WoS. The term "standard citation" from an arXiv document will also be used if the journal in which the citation appears belongs to the WoS Core Collection. In general, the word "standard"—or "regular"—will be used for citations, journals, and articles that are measured and covered by journals in the WoS Core Collection. We have adapted the terminology found in the paper by Kling, Spector, and McKim (2002).

Specifically, our bibliometric analysis is guided by the following general questions.

Q1. "arXiv to standard" publication dynamics: How many documents in arXiv are cited in WoS? Which are the scientific fields in which research preprints posted in arXiv—with at least one citation in WoS—are most cited in standard journals? What is the proportion of papers that meet this requirement and are finally published in standard journals? What about the delay in publication?

Q2. "non-standard citation" of arXiv manuscripts as non-standard documents: How can citation of documents in arXiv with at least one citation in WoS be measured outside the WoS context?

2. MATERIAL AND METHODS

2.1. Data Collection

Our study followed the steps explained below. The data collection procedure started by setting the end date: December 2015. We have collected all article citations in arXiv that have appeared in WoS up to this date. Using the option Cited Reference Search in WoS, a search was made of the word "arXiv" in the field Cited Work. This provided the total amount of papers that, coming from the repository—and therefore accessible by the scientific community without peer review—enter the world of standard publications by appearing in a list of references of a published paper. It must be said that we searched these references one-by-one, attending to the specific properties of each of them in order to decide whether or not they were acceptable for the sample. The reason is that the way researchers cite preprints in arXiv is not homogeneous, and there are no fixed rules for doing this. This implies that the process of identification of a paper is in general difficult, if not impossible. This is the case if the final published version of the article does not have the same title, in which case it is difficult to realize that this article actually coincides with an earlier preprint. Although arXiv allows you to upload updated versions, this is not always done.

For instance, references with the following structure "MAYOR M, 2008, ARXIV," or "BEIRAO, ARXIV ASTROPHYSICS" were difficult to find. To detect the first one in arXiv, the name "MAYOR" has been introduced in the field "Authors," limiting also the date of storage. The result obtained in which "Mayor" appeared as the first author—also with the initial of the name "M"—was considered as the document referred to. If it appeared as the author, and there are no more preprints, it was also considered as such. In the event that there were two or more preprints with these characteristics, the paper was classified as "untraceable." The easiest references to find were those that appeared as follows: "Compere G, 2007, ARXIV07083153HEPTH."

After setting the correct reference, Google Scholar was used to determine whether the article was already published in a regular journal. To check this, the DOI number was used if it was in arXiv; otherwise, the title was used for this purpose. This gave us a set of 561 preprints as a working sample. Some of them were later withdrawn for other reasons—for example, some were classified as biology papers—and so the final sample was set at 554 manuscripts. We will present only the most relevant data to support our arguments.

2.2. Citation Analysis

Once the total set of relevant preprints was identified, several analyses were conducted.

a. The first was to calculate the proportion of manuscripts deposited in arXiv that appeared in references of articles published in journals that are listed in WoS. This analysis was carried out after grouping some of the different scientific fields determined by arXiv, in order to have a relevant number of papers in each group. The proportion of articles cited in this way that were eventually published in WoS journals was also calculated.
b. The difference between the year of standard publication and the year in which the preprint was deposited in arXiv was also calculated.

c. Citations of articles published in a standard journal were also counted: number of citations recorded in arXiv, number of citations recorded on the website of the regular journal that published the manuscript, and the difference between these amounts. In case the journal did not provide the number of citations for the articles, Google Scholar was used.
d. Finally, we also counted the number of citations of articles that did not appear in any regular journal: number of citations registered in arXiv of the preprints, number of citations registered in Google Scholar, and the difference between these amounts.

3. RESULTS

3.1. Global Impact of Standard Citations to arXiv Documents

A total of 554 documents were considered from our search, after clearing references to documents that were impossible to fix due to deficiencies in the citation. The set is small, compared to the total amount of documents that can be found in arXiv. Taking into account that the number of documents in arXiv in the date of completion of the research was about 1,150,000 (see https://arxiv.org/stats/monthly_submissions), the overall impact of the citations that we are studying is not relevant. However, the number of documents is large enough to analyze some of the properties of these citations.

3.2. Publication Ratio and Publication Delay

As we have explained, the subject classification provided by arXiv was followed, unifying some fields by subject proximity if necessary for getting statistically meaningful results. The way the areas are grouped is the following. Areas with a big number of preprints are considered separately (astrophysics, computer science, and condensed matter).

Table 1. Deposited manuscripts and finally published papers

Scientific areas	Deposited	Published	Ratio (%)
Astrophysics	208	154	74
Computer science	50	36	72
Condensed matter	101	76	75.2
Mathematics, statistics, nonlinear sciences	56	36	64.3
Physics	139	72	51.8
Total	554	374	67.5

The rest of the areas were grouped in a standard way under the names "mathematics, statistics, nonlinear sciences." and "physics." The amount of deposited papers depends strongly on the area, and also the publication rates. Table 1 shows the total number of deposited and published papers, respectively, for some scientific fields that are particularly relevant for our study. The complete tables with all the disciplines can be seen in the attached datasets.

It can be seen that the result depends greatly on the subjects. However, our result coincides broadly with the ones obtained in Larivière et al. (2014). It is shown that about 64% of all arXiv preprints are published in a WoS-indexed journal. In our case the rate is 67.2%. There is a small deviation, probably due to the bias produced by our selection criteria. Indeed, since the set of manuscripts which have citations from standard journals has been chosen, this already means that they are in a sense more relevant than non-cited ones. The results reinforce the idea suggested by the value of the total rate computed in Larivière et al. (2014). It could be interpreted in terms of the coherence of the authors' publication policies: The more citations in standard journals, the greater the likelihood that the paper will be published in a standard way. This could mean that the manuscript is considered a standard scientific document both by the authors and by the rest of the researchers of the scientific field. Publication in arXiv would be just a first step in the standard publication process, not an alternative form of dissemination of information. The value of the ratio itself suggests this conclusion: At least two of each three papers published in arXiv—that is, most authors—understand arXiv as the first step in the publication process, and not as a final publication medium. However, this subtracts some of the potential standard citations, contrary to the interests of authors who need to pass an evaluation process. In return, a rapid and early dissemination of the work would help the authors to gain prestige in the field. Each researcher must find the right balance between these two factors.

The results for specific arXiv specialties follow a similar rule, and are compatible with those published earlier; see Fig. 1 in Larivière et al. (2014). Again, our results show a higher standard publication rate, due to the relevance argument explained above with respect to the results presented in Table 1. Note also that the results are given for the grouped specialties, which does not allow for a direct comparison with previously published material. However, there are some interesting differences in two opposite directions that should be noted. Although the following arguments cannot be considered conclusive, we believe they may provide some ideas for interpretation.

a. In the grouped area "mathematics, statistics and nonlinear sciences" of our study, the ratio obtained is 64.3%, while for the total amount of papers considered in Larivière et al. (2014) it is less than 50%. As was explained before, this could suggest that mathematicians agree in publishing in standard journals independently of prepublication in arXiv, but mainly of those papers that are considered to be relevant enough to be cited. Alternatively, both facts can be considered as independent, and then this deviation would mean that authors that previously publish in arXiv are more actively involved in diffusion of their work, being also the ones with a bigger rate of standard publication. This publication habit may be specific for some scientific fields, but it seems to be the most general behavior.
b. The opposite trend can be observed with regard to the proportion of publication in our grouped area "physics," which in our case is significantly lower than in the general study of Larivière et al. (2014). This would mean that, to some extent, some authors feel that uploading a manuscript to arXiv is good enough to ensure the visibility of their work, and then prepublication and standard publication are two different tools for diffusion. This would be coherent with the hypothesis that documents in arXiv and papers in standard journals are in fact different enough to make it difficult to link the preprint and the final publication.

Other interesting bibliometric information that can be obtained refers to the average time that is needed for publishing a paper after it is posted in arXiv. As an initial approach, authors are supposed to upload their paper to arXiv when they complete their research work, so that from this point onwards the delay can be interpreted directly as exclusively due to the publication process. The results are shown in Table 2 for the grouped specialties; again the reader can find the complete information in the attached datasets. It should be remarked that the dispersion of the result is very high (high variance).

Table 2. Publication delay for grouped specialties

	Publication delay (year)	Publication ratio (%)
Astrophysics	0.8	74
Computer science	1.1	72
Condensed matter	0.8	75.2
Mathematics, statistics, nonlinear sciences	2.5	64.3
Physics	1.5	51.8

A comparison with other publication delay data that can be found in the literature makes sense. In Larivière et al. (2014), the data computed with the whole of the arXiv database show that the specialties grouped in our case with the label "mathematics, statistics and nonlinear sciences" have a publication delay of more than 1.4 years (see Fig. 5 in the referred paper). The value estimated in the present investigation is however higher (2.5 years). Also the time elapsed for publication in areas of the grouped variable "physics" is shown shorter in the analysis in Larivière et al. (2014) than in ours (1.5), which almost doubles the expected value (0.8). Therefore, a fairly large difference has been found with the previous analysis. A suitable explanation of the reason for the higher delay could be the bias produced by our selection method of arXiv manuscripts. Citations to the arXiv version of a manuscript that will be published later may mean that there is a delay on publication. Otherwise, it seems natural to cite the standard published version if possible, or both versions; it is known that this facilitates citation counts by the WoS—what benefits the authors—and also ensures to the potential reader of the citing paper that its reference have been peer reviewed.

Another aspect that should be mentioned is the relationship observed between the delay in publication and the publication rate in each grouped specialty. Delay in publication increases when the proportion of publications decreases, as can be seen in Table 2, although only a weak correlation can be observed. The topic "quantitative biology," which is found in the original sample, has been eliminated due to the small sample size. Each discipline seems to have its own delay/ratio characteristics.

There is an inverse relationship between the number of articles deposited in arXiv and the length of the publication process: the greater the publication ratio, the shorter the delay in publication. However, this relationship is weak. For example, the area "mathematics, statistics and nonlinear sciences" shows its own particular values. It seems to be again a consequence of the authors' publication policy together with the characteristics of the journals that publish in different scientific fields. At one extreme we find "condensed matter" and "astrophysics," with low delay and high ratio, while at the other extreme we find "mathematics, statistics and nonlinear sciences" and "physics," with different proportions between these terms. There are long delays in publication along with relatively small publication rates, which is consistent with what appear to be different philosophical views on the role of arXiv. For example, it could mean that for mathematicians, the final publication is made even if the results are made available to

the scientific community some years earlier. On the other hand, the standard peer reviewed publication would also be important for astrophysicists, but also the rapid presentation of the results facilitated by arXiv.

3.3. Measuring Impact of the Papers with Non-Standard Tools

The second part of our study is dedicated to analysing how to measure the impact of documents previously deposited in arXiv. Due to the nature of the documents—which are not considered as "citable objects" by WoS for the calculation of their impact—an alternative way of measuring the influence of the article other than the number of citations from WoS has been developed. The number of citations of all the documents has been considered in two different sources, which in a sense are complementary. For all manuscripts, citations were calculated in arXiv (provided by High Energy Physics information system). Then, two different procedures were applied, depending on whether the work was finally published or not.

1. If the preprint was finally published, the number of citations was found in the journal where it was published, or in Google Scholar instead in case the journal did not

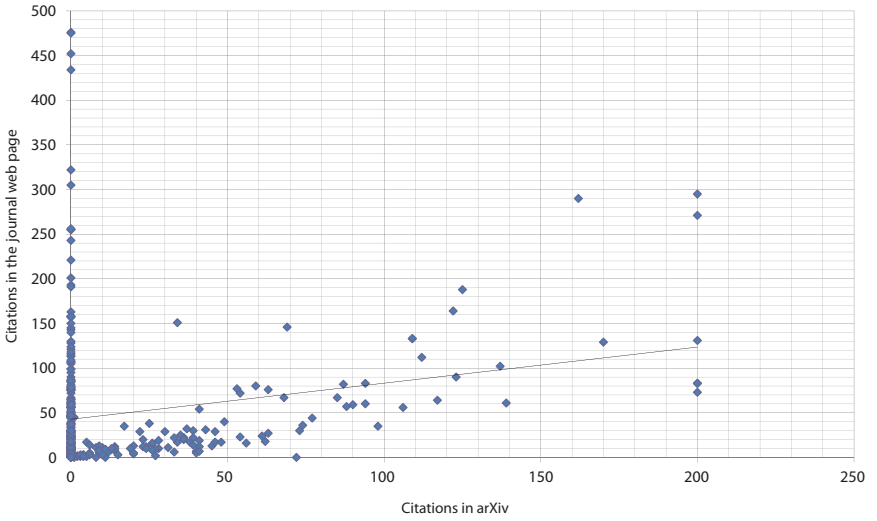


Fig. 1. Citations in arXiv (by documents in arXiv) versus citations registered in the website of the journal where the paper was finally published (Five points out of scale were removed for the representation).

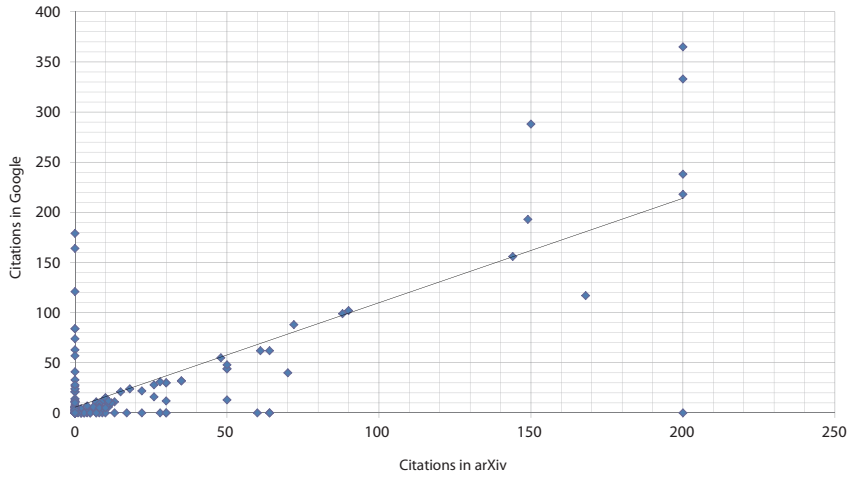


Fig. 2. Citations in arXiv (by documents in arXiv) versus citations in Google Scholar (Two points were removed for the representation).

provide this information. Fig. 1 shows the result.

2. If the article was never published, an external citation measure was used: the number given by Google Scholar. As explained in the methodology section, the reason for choosing this option is to provide an adequate measure of the citations that do not come from WoS, where only the standard published version is considered as a scientific document. The result can be seen in Fig. 2. This procedure differs from previous intensive studies on the subject that we have used as main references for our analysis.

Although a weak correlation may be observed in both cases, the large amount of papers on the axis suggests that the three citation computations are completely different. For example, there are many papers without citations inside arXiv that have a big citation rate when they are published in a standard journal, or in Google Scholar (see the axis OY in both Figs. 1 and 2). This reinforces our hypothesis about citation behaviors by the three methods considered—arXiv, journals' websites, and Google Scholar—and somewhat independent, and the reasons why authors use arXiv differ in each case.

4. DISCUSSION: HOW AND WHY RESEARCHERS USE arXiv

The standard interpretation of why researchers use arXiv is that it follows the essential aspects of a traditional form of research dissemination and scientific information exchange: e-preprints are the current version of the classic manuscript that was shared with colleagues to communicate research results (see for example the paper by Confrey, 1996; see also the references in the paper by Larivière et al., 2014). Material of this type (electronic preprints) should be understood as the same type, and contained in the broad category sometimes referred to as research manuscripts.

arXiv is mainly devoted to physics, mathematics, and computer science. In recent work (Larivière et al., 2014) it is shown that 64% of all arXiv papers are finally published in WoS. Previous studies have shown that the ratio of deposited manuscripts in arXiv with respect to total published papers in mathematics was 81% in 2010 (Fowler, 2011). Also 75% of the papers on physics of condensed matter are deposited in arXiv (Moed, 2007). A complete review of the existing literature on the high ratio of deposition in arXiv by physicists and mathematicians can be found in Li et al. (2015). Summing up, it can be said that arXiv provides a standard tool for prepublication and post-publication in those fields in which rapid communication is at least as important as the fact that the publication is peer reviewed. It should also be noted that in mathematics the backlog associated with the process of editing a paper is generally quite large, so the academic community is committed to disseminating its results in this way, although peer review is also seen as fundamental.

However, it seems that scientists in these areas know that citing an arXiv manuscript is not the best way to refer to a published paper, as it is supposed to have some sort of temporary value only until the last peer reviewed version is published. This would be supported mainly by the small rate of references to arXiv preprints that we have found in WoS. Also, some editors request that these references be updated in the latest versions of the paper if possible, and some even reject the original submission of papers containing such citations. From the authors' point of view, there is a conflict of interest regarding the balance between the rapid dissemination of a manuscript and the "quality" of citations to this manuscript.

Let us explain a suitable scheme of authors' motivations in a more detailed way, when the problem of the possible dispersion of cites is taken into account. It is based mainly on the analysis of the studies quoted above and our bibliometric data.

a. In the first scenario, the manuscript is supposed to provide a rapid communication of research results and no further publication is expected, or in case publication is done it is of secondary importance. Then, a big rate of citations in arXiv was expected if the citing papers follow the same rule, that is, if the scientific group interested in the topic considers arXiv as a primary and reputable source of relevant information. As already announced in Haustein (2016), a parallel system of scholarly communication is supposed to exist in this case, based on arXiv type documents. Since we have shown that the total amount of documents of this kind referred to in WoS is small, the system would work independently of the standard publication. The extreme cases of this expected behavior are the papers appearing in the OX axis of Figs. 1 and 2. From our personal experience as researchers, it must be said that some publishers refuse—explicitly or implicitly—to publish articles that refer to unreviewed documents.
b. In a second scenario, the paper is deposited in arXiv to ensure authorship of the research or rapid presentation of results, but this version is not assumed to be the final support for the investigation presented in it. Again

the extreme case would give no citations in arXiv and many citations in other sources—a journal website or Google Scholar. These are the manuscripts that appear in the OY axis of Figs. 1 and 2.

The final picture would be given by all the intermediate cases between the two extreme situations mentioned above. When an author considers depositing a paper in arXiv, the arguments supporting the decision may be, in a sense, a mixture of those explained. On the one hand, he or she wants to offer the result of his scientific work to the community as soon as possible, while ensuring his authorship. This would provide a long-term benefit—prestige, but could be dangerous in terms of the citation count. It has been shown how this prepublication would affect this count by producing "poor quality" citations to the arXiv document—from standard bibliometric measuring tools. On the other hand, the author may consider having arXiv as a permanent support for his results. Depending on the scientific area and the uses in each research community, each of these arguments becomes the main reason. But in many selection/evaluation processes, a paper in arXiv is not a paper—even if it has a hundred citations—and so prepublication could damage the professional career of the researcher since he cannot put it in his list of publications.

5. THE CURRENT SITUATION: AN INCREASE IN THE NUMBER OF NEW REPOSITORIES AND PREPRINT UPLOADS

We have focused our attention on arXiv because of its recognized position in the world of scientific publishing. However, a long list of new platforms for preprints has appeared, which have been consolidated in recent years. The reader can find in the 'researchpreprints' platform a list of repositories, in which it is easy to see that there are many new records (e.g., AgriXiv, ChemArxiv, ChinaXiv, LIA Scholarship Archive, and OSF preprints). Preprints publishing has also grown very rapidly in the last two years, making it increasingly convenient to analyze the role of scientific manuscript prepublication (see Lin, 2018). It seems clear that this practice benefits the authors in terms of dissemination of their work, but as we have observed in the present study, it also produces some dispersion of citations, against the interest of authors to the extent that this may harm their careers as researchers.

In any case, in view of the growing tendency to deposit preprints in repositories, it seems that authors are increasingly concerned about this alternative form of distribution of their research results. For this practice to be consolidated and useful, applying the main conclusions of our analysis seems urgent: A standard citation method and regulated bibliometric rules must be imposed for the evaluation of the research. Along with the crisis of the peer review system, the recognition of the value of all kinds of scientific material (including data, preprints, projects, etc.) seems to be the main current problem of the global scientific information system. We cannot expect these changes to be promoted by large publishing companies, as preprint publication could affect their business. Bearing in mind that the same companies that own the publishing houses are sometimes also owners of the bibliometric platforms, it does not seem that the changes will come from this part. This will probably be done by national research evaluation agencies or international bodies.

However, it seems that the consequences of prepublication for authors depend to a large extent on the field of research, and researchers generally know very well how this may affect their scientific activity. Therefore, the regulation of prepublication seems to be an issue that will depend on each particular field, although some standard rules should be imposed.

6. CONCLUSIONS

We have added some bibliometric explanation regarding the citation-based interaction between the standard publication world and preprint publication to the existing ones. Our aim was to understand the behavior of authors with regard to the prepublication of their scientific results. It seems that there are no universal trends that can explain this behavior, which seems to depend on each scientific specialty. This may be a consequence of: 1) the existence of prepublication rules implicitly accepted by all researchers only in local communities associated with specific scientific areas, as well as 2) the result of the lack of reliable specific tools for measuring preprint citations. The second aspect critically affects the evaluation of authors and conflicts with the benefits of preprint publication. We must say that these benefits are solid and have been proved in various works, and probably counteracts the loss-citation-problem analyzed in this paper. The small amount of references to arXiv papers in WoS that we have found supports the idea that the damage caused to authors by citation to arXiv preprints is, in any case, small.

Our work also confirms—although with information

collected from a small sample—the previous analysis proving the benefit of preprint publication. The direct relationship between the upload of manuscripts and rapid communication—which is not covered by the standard publication—would fit in with the results that we have found.

The existing literature on the topic shows that many researchers consider arXiv to be an autonomous network for scientific dissemination in some disciplines. We therefore believe that a rapid development of recognized tools to measure citations in the world of prepublishing is necessary to facilitate impact assessment. Although the evaluation of research is done using bibliometric indicators, measuring the impact of preprints seems to be the only way to support and reinforce the prepublication of manuscripts, especially if this is going to be the final form of publication of a relevant part of these articles. This issue is not only about the dissemination of science, but also about open science initiatives, as preprint publication is often free of charge. A new paradigm including preprints and other "non-standard" sources of information as valuable scientific documents for research evaluation is needed. This is already done by some national agencies for research assessment, but in other cases (such as Spain or Poland, for instance) standard bibliometric tools still play a fundamental role. A researcher's career is developed through a sequence of evaluation processes, at all levels. Some new tools—such as downloads of associated electronic files and other altmetrics—are beginning to be considered in these evaluations, but even so, citation count still plays a relevant role. Therefore, it seems natural to think that in the near future the consideration of non-peer-reviewed articles will also enter evaluation systems. In fact, the need for these new rules is evident, as preprint publication seems to be increasing exponentially, as we mentioned in the previous section. The impact of preprints can be measured in terms of, for example, citations or downloads, but our analysis suggests that some of the existing tools are not adequate yet. For example, both Google Scholar and arXiv have been shown to provide a citation counter of all documents appearing in any standard search, but the results obtained are somewhat random (Figs. 1 and 2). No uniformity is observed when different scientific areas or even individual works are considered.

ACKNOWLEDGMENTS

The work of the first, second, and third author was supported by Ministerio de Economía, Industria y Competitividad, Spain, under Research Grant CSO2015-65594-C2-1R Y 2R (MINECO/FEDER, UE). The work of the fourth author was supported by Ministerio de Economía, Industria y Competitividad, Spain, and FEDER, under Research Grant MTM2016-77054-C2-1-P. The authors would also like to thank the referees for their useful comments and references, which helped them to improve the work, especially in Section 5.

REFERENCES

Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management*, 7(1), 64-77.

Bar-Ilan, J. (2014). Astrophysics publications on arXiv, Scopus and Mendeley: A case study. *Scientometrics*, 100(1), 217-225.

Bohlin, I. (2004). Communication regimes in competition: The current transition in scholarly communication seen through the lens of the sociology of technology. *Social Studies of Science*, 34(3), 365-391.

Confrey, E. A. (1996). The information exchange groups experiment. *Publishing Research Quarterly*, 12(3), 37-39.

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2), 203-215.

Fowler, K. K. (2011). Mathematicians' views on current publishing issues: A survey of researchers. *Issues in Science and Technology Librarianship*, 67. Retrieved November 2, 2018 from https://doi.org/10.5062/F4QN64NM.

Haque, A. U., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 60(11), 2203-2218.

Haque, A. U., & Ginsparg, P. (2010). Last but not least: Additional positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 61(12), 2381-2388.

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Sciento metrics*, 108(1), 413-423.

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., & Murray, S. S. (2006). Effect of e-printing on citation rates in astronomy and

physics. *Journal of Electronic Publishing*, 9(2). Retrieved November 2, 2018 from https://quod.lib.umich.edu/j/jep/3336451.0009.202/--effect-of-e-printing-on-citation-rates-in-astronomy?rgn=main;view=fulltext.

Kim, J. (2011). Motivations of faculty self-archiving in institutional repositories. *Journal of Academic Librarianship*, 37(3), 246-254.

Klein, M., Broadwell, P., Farb, S. E., & Grappone, T. (2016). Comparing published scientific journal articles to their pre-print versions. In N. R. Adam, B. Cassel, & Y. Yesha (Eds.). *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (pp. 153-162). New York: ACM.

Kling, R., Spector, L., & McKim, G. (2002). Locally controlled scholarly publishing via the Internet: The Guild Model. *Proceedings of the American Society for Information Science and Technology*, 39(1), 228-238.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. *Information Processing & Management*, 41(6), 1395-1402.

Kurtz, M. J., & Henneken, E. A. (2007). Open Access does not increase citations for research articles from The Astrophysical Journal. Retrieved November 2, 2018 from https://arxiv.org/ftp/arxiv/papers/0709/0709.0896.pdf.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6), 1157-1169.

Li, X., Thelwall, M., & Kousha, K. (2015). The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication. *Aslib Journal of Information Management*, 67(6), 614-635.

Lin, J. (2018, May 31). Preprints growth rate ten times higher than journal articles. Retrieved November 2, 2018 from https://www.crossref.org/blog/preprints-growth-rate-ten-times-higher-than-journal-articles/.

Manuel, K. (2001). The place of e-prints in the publication patterns of physical scientists. *Science & Technology Libraries*, 20(1), 59-85.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.

Neuhaus, C., & Daniel, H. D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation*, 64(2), 193-210.

Pinfield, S. (2005). Self-archiving publications. In G. E. Gorman, & F. Rowland (Eds.). *International yearbook of library and information management 2004/2005: Scholarly publishing in an electronic era* (pp. 118-145). London: Facet Publishing.

Swan, A. (2010). *The Open Access citation advantage: Studies and results to date*. Southampton: University of Southampton Institutional Repository.

Youngen, G. K. (1998). Citation patterns to traditional and electronic preprints in the published literature. *College & Research Libraries*, 59(5), 448-456.

Zha, X., Li, J., & Yan, Y. (2013). Understanding preprint sharing on Sciencepaper Online from the perspectives of motivation and trust. *Information Development*, 29(1), 81-95.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 17-27, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.2

Psychological Aspects of Job Satisfaction Among Library and Information Science Professionals

Ramesh Pandita*
Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India
BGSB University, Rajouri, Jammu & Kashmir, India
E-mail: rameshpandita90@gmail.com

Dr. J. Dominic
Karunya University, Karunyanagar Coimbatore, Tamil Nadu, India
E-mail: jdom16@gmail.com

ABSTRACT
This study assesses the psychological aspects which influence job satisfaction among library and information science professionals. The study is based on primary data collected from the library and information science professionals working in the higher education institutions of Jammu and Kashmir, India. In all 264 responses were collected, comprising 44.3% male respondents and 55.7% females. The majority, 74.2% of respondents, are under 45 years of age, while 67.4% of respondents have a master's degree in library and information science. Of the total respondents, 7.6% conceded to being incompetent, while 13.3% viewed their peers as incompetent. The majority, 25% of respondents, replied that the library profession is a thankless job and 70.8% of respondents viewed that they are emotionally attached to their profession, while at the gender level, compared to 75.5% females, 65% of male respondents admitted to being emotionally attached to their profession. The encouraging part is that 26.5% of respondents replied that they love doing their job beyond office hours and 75.8% of respondents replied that they would not seek voluntary retirement, while 41.7% of respondents showed willingness to continue working as library and information science professionals post-retirement, if engaged.

Keywords: job satisfaction, library professionals, individual psyche, Jammu and Kashmir, higher education, guidance and counselling

Open Access

Accepted date: August 31, 2018
Received date: June 22, 2018

***Corresponding Author:** Ramesh Pandita
Research Scholar
Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu 641046, India
E-mail: rameshpandita90@gmail.com

1. INTRODUCTION

Studying the cognitive capacities of individuals is one of the most important subjects studied extensively by researchers all across the world, especially while seeking solutions to the problems, which may directly or indirectly emanate from differences in the behavioural patterns of an individual or a group. Job satisfaction and dissatisfaction among a group of employees, working under similar or different conditions apart from being influenced by extrinsic conditions, is equally influenced by intrinsic factors. These factors can be both physical as well as psychological in nature, what we generally owe to the cognitive capacities of an individual.

To live a wholesome and fulfilling life, it is always imperative that an individual should be satisfied on various fronts, and job satisfaction is one of the foremost aspects associated with the wholesome satisfaction of an individual. Salary, work environment, job security, interpersonal relationships, recognition, advancement, and so on are some of the important variables associated with the job satisfaction of employees and the absence of any of these variables may lead to job dissatisfaction. Apart from these and various other variables associated with job satisfaction and dissatisfaction, the psychology of an individual employee is equally important, which plays a very significant role in the overall job satisfaction of an employee.

Till the recent past, especially in the absence of the application of information technology (IT), the job of a library professional was more or less monotonous in nature, less interactive and less attractive, with no spur to inspire an individual or motivate a professional to do something new and significant. With the result, the majority of library professionals used to show lower levels of job satisfaction. The application of information and communication technology in libraries has somewhat helped a great deal in making the library profession more attractive and interactive (Pandita & Dominic, 2018); still, it is the individual psychology of an employee which comes into play for reaping wholesome job satisfaction.

The undergoing study has been undertaken with a view to assess individual psychological perceptions among library and information science (LIS) professionals about their job, which contributes towards their job satisfaction or dissatisfaction. To undertake the study, data were collected from the LIS professionals working in the higher education institutions of Jammu and Kashmir, India. The study is based on the 264 responses generated from respondents working in both government and private institutions. The respondents represent academic, technical, and professional institutions, ranging from universities and colleges to autonomous institutions.

2. BRIEF INTRODUCTION ABOUT THE HIGHER EDUCATION SECTOR OF JAMMU AND KASHMIR

The state of Jammu and Kashmir is in the extreme north of the Union of India, having a population of 12.55 million, with a literacy rate of 68.70% (Government of India, 2011). As of date, apart from seven state universities the state has two central universities, one national institute of technology, one Indian institute of technology, one Indian institute of management studies, two Indian institutes of integrative medicine, over 100 government degree colleges, and nearly 170 private colleges.

3. PROBLEM STATEMENT

The scenario of job satisfaction among LIS professionals in India is not that encouraging, as the majority of LIS professionals in the country are more or less dissatisfied with their employers one or the other way (Pandita, 2017). Salary, though important, is not the lone criterion which can be associated with the job satisfaction of an employee or for that matter to a library professional. There are some psychological aspects as well which significantly affect the job satisfaction of an employee and so holds true about LIS professionals. Among various psychological aspects, the individual psychology of an employee towards his/her job plays a very significant role in his/her overall job satisfaction. Given the fact, it was conceived to assess how far the various psychological aspects, both personal and those associated with the job affect the job satisfaction of library professionals, and to undertake this pilot study the LIS professionals working in the higher education institutions of Jammu and Kashmir were surveyed.

4. OBJECTIVES OF THE STUDY

- To understand the psychological aspects associated with the job satisfaction of working LIS professionals
- To find out how far the individual psychology of an employee influences his/her levels of job satisfaction or dissatisfaction

- To explore the gender based psychological perceptions of library professionals towards their job, their job satisfaction, and the need thereof for their guidance and counseling, if any

5. REVIEW OF LITERATURE

Extensive research has been undertaken in the areas of assessing the influence of cognitive capacities of individuals on their job satisfaction. Job satisfaction in fact reflects the cognitive assessment of an employee towards his/her employer, keeping in view the work conditions and other related aspects which are taken care of by the employer (Weiss & Cropanzano, 1996). Hence this cognitive capacity of an individual is always at work, sometimes accepting things and at others rejecting them in more or less in a silent way. This is what was assessed as a personality disposition by Staw and Ross (1985), who viewed it as an individual approach of an employee towards his/her work, which is generally quite apparent and known to peers and other colleagues.

In a study, Dormann and Zapf (2001) observed 35% variation in the job satisfaction of employees when they were offered a more stable work environment in their changeable work environment. The individual personality of an employee is seen as a building block of individual traits, which reflects both the positive and negative aspects of job satisfaction (Brief & Roberson, 1989). Self-esteem, efficacy, control, and non-neuroticism significantly affects the job satisfaction among individuals (Judge, Locke, Durham, & Kluger, 1998).

Although the behaviour of an individual is generally governed by one's individual precepts, one cannot rule out the impact of other social aspects which directly or indirectly influence an individual's behaviour to a significant degree. In the same way, the behaviour of an employee gets equally influenced by his/her peers and by the institutional management. Thereby, it is always an added responsibility of management to govern the behaviour of their employees (Robbins, 2003). Robbins views that transformational leadership, apart from helping in making the maximum out of the employees by helping them to fulfil their professional commitments, also promotes the individual growth of employees and allows less resignation among employees towards their work.

In a study undertaken on 1,300 information professionals, Williamson, Pemberton, and Loundsbury (2005) found a significant correlation between personality variables and job satisfaction. The researchers studied variables like emotional stability, teamwork, work style, work drive, optimism, and emotional resilience under three different sets and found 20%, 19%, and 18% difference in job satisfaction among employees. The researchers also found that library professionals who are more optimistic towards their work, have more emotional resilience, believe in team work, have a vision, and believe in hard work, as such are more satisfied with their jobs than their counterparts who lag in such areas.

In a similar study, Judge, Heller, and Mount (2002) studied a five-factor personality model, which included variables like neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness, by analysing 163 independent samples, and found that neuroticism and extraversion are correlated with job satisfaction. While opting for a conceptual approach, Locke (1976) studied concepts like satisfaction, dissatisfaction, value, emotions, and appraisal by using Rand's theory of emotions and viewed that job satisfaction and dissatisfaction has not advanced following the adoption of the conception of causality having a correlation with it. While studying the different variables associated with job satisfaction among employees, with special reference to LIS professionals, Pandita and Dominic (2016) viewed the individual psyche of a professional as an important aspect of job satisfaction.

While reviewing sociologists' perceptions towards library professionals, Satija (2003) observed that the majority of such studies recommended the need for treating librarians at par with regular teachers in all matters, including faculty status. The library professionals who actively participate in the decision making process of their institution, and communicate freely and openly, have better opportunities for growth and have been found to be more committed and satisfied with their job (Burd, 2003). Library professionals should always feel challenged by their work (Reid, 2005) and so should they embrace the changes which are inevitable, especially in this fast and ever changing service oriented world. Most of the changes are aimed towards the betterment of both the profession and professionals, and any resistance to any such positive change hampers both professional and individual growth, which by no means goes well with the overall benefit of society at large. In a study, Pandita (2017) observed that in India every third library professional enters into the profession by chance and not by choice. This in itself speaks about the psychological state of library professionals in India. It should not come to us as a surprise if these one-third of library professionals show dissatisfaction with their job.

6. RESEARCH METHODOLOGY

This study is based on primary data collected from 264 LIS professionals working in the higher education institutions of Jammu and Kashmir, India. The data were collected from the respondents by circulating the questionnaire, specially designed for the purpose. The respondents include the library professionals working in both state and central universities, private and government colleges, medical colleges, nursing colleges, engineering colleges, and other institutes of national importance. Questions were framed keeping in view the objectives of the study. The data were collected by circulating the online link of the questionnaire among the respondents through emails and other social media platforms. Questionnaires were also sent by post to potential respondents at their official postal addresses along with self addressed envelopes so as to facilitate the return of the filled in questionnaire, and thirdly the questionnaires were also circulated and responses were collected by visiting the respondents in person.

7. RESULTS

The responses collected from respondents were compiled and tabulated using MS Excel for easy filtration of data. Computations like addition, subtraction, multiplication,

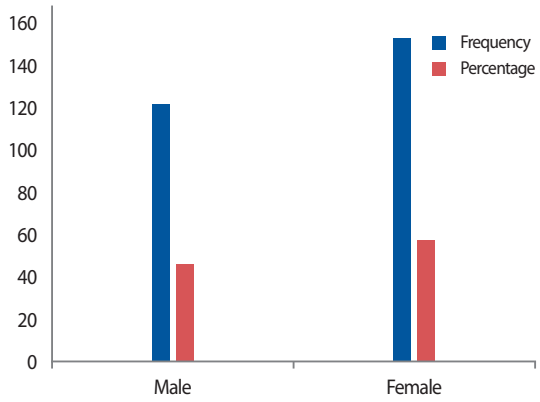


Fig. 1. Gender based response representation of respondents.

Table 1. Biographical information of respondents

Response variable		Gender				Total		
		Male		Female				
		Freq	%age	Freq	%age	Freq	%age	CF%
Total responses		117	44.3	147	55.7	264	100	100
Age group (yr)	Under 25	2	0.8	5	1.9	7	2.7	2.7
	26–35	48	18.3	53	20.2	99	37.5	40.2
	36–45	50	19.1	47	17.9	97	36.7	76.9
	46–55	13	4.9	30	11.5	43	16.3	93.2
	Above 56	4	1.5	10	3.8	14	5.3	98.5
	Prefer not to say	-	-	2	0.8	2	0.8	100
Work experience (yr)	Less than 1	4	1.5	6	2.3	10	3.8	3.8
	2–5	32	12.1	33	12.5	65	24.6	28.4
	6–10	28	10.6	36	13.6	64	24.2	52.6
	11–15	35	13.3	23	8.7	58	22.0	74.8
	16–20	11	4.2	16	6.1	27	10.2	85.0
	Above 20	7	2.7	30	11.4	37	14.0	99.0
	Prefer not to say	-	-	3	1.1	3	1.1	100
Professional qualifications	Certificate course	3	1.1	14	5.3	17	6.4	6.4
	B.Lib.Sc/BLISc	10	3.8	16	6.1	26	9.8	16.2
	M.Lib.Sc/MLISc	74	28.0	104	39.4	178	67.4	83.6
	M.Phill	7	2.7	5	1.9	12	4.5	88.1
	Ph.D	13	4.9	7	2.7	20	7.6	95.7
	Prefer not to say	10	3.8	1	0.4	11	4.2	100
Designation	University librarian/equivalent	-	-	-	-	-	-	-
	University deputy librarian/equivalent	2	0.8	-	-	2	0.8	0.8
	University senior assistant librarian/ equivalent-2	12	4.5	18	6.8	30	11.4	12.2
	University assistant librarian/equivalent	35	13.3	46	17.4	81	30.7	42.9
	Information scientist	1	0.4	1	0.4	2	0.8	43.7
	Professional assistant/senior library assistant	6	2.3	11	4.2	17	6.4	50.1
	Semi professional assistant/library assistant	41	15.5	53	20.1	94	35.6	85.7
	Junior library assistant	10	3.8	12	4.5	22	8.3	94.0
	Library attendant	7	2.7	4	1.5	11	4.2	98.2
	Prefer not to say	3	1.1	2	0.8	5	1.9	100

and drawing percentage were performed using MS Excel. Percentage at all places has been drawn to one decimal place and has been rounded off to 100% figure.

Of the total responses generated, 44.3% are male and 55.7% female respondents (Fig. 1). Over 37.5% of respondents are in the age group of 26 to 35 years, followed by 36.7% in the age group of 36 to 45 years. The majority, 24.6% of respondents, have 2 to 5 years of work experience, followed by 24.2% of respondents having 6 to 10 years of work experience. Similarly, in terms of academic qualifications the majority, 67.4% of respondents, have a master's degree in LIS, followed by 9.4% having a bachelor's degree in LIS, while a majority 35.6% of respondents are working as semi-professional assistants, followed by 30.7% as assistant librarians and equivalent positions (Table 1).

Gender is not a constraint in the library profession in general and for those working in the state of Jammu and Kashmir in particular, as the majority of the respondents in the present survey are female library professionals and are academically well qualified. Since the majority of respondents are in the 26 to 45 years of age group, as such, it can be emphatically inferred that professionals in the state are quite young, hence they are expected to be tech savvy, and as a result application of information and communication technology in the library services and activities may not be a constraint for these young professionals. LIS in India is believed to be growing more gender specific, as more and more females prefer to join the library profession than males, which is also evident from the response percentage of the present survey.

Table 2. Respondents' level of agreement on some psychological aspects associated with job satisfaction

Statement		Agree (%age)	Partly agree (%age)	Can't say (%age)	Partly disagree (%age)	Disagree (%age)	No response (%age)	Total (%age)
You are an incompetent professional								
	Male	9 (7.7)	31 (26.5)	14 (12.0)	6 (5.1)	51 (43.6)	6 (5.1)	117 (100)
	Female	11 (7.5)	25 (17.0)	18 (12.2)	4 (2.7)	83 (56.5)	6 (4.1)	147 (100)
	Total	20 (7.6)	56 (21.2)	32 (12.1)	10 (3.8)	134 (50.8)	12 (4.5)	264 (100)
Your peers are least competent								
	Male	17 (14.5)	24 (20.5)	19 (16.2)	15 (12.8)	31 (26.5)	11 (9.4)	
	Female	18 (12.2)	29 (19.7)	31 (21.1)	14 (9.5)	46 (31.3)	9 (6.1)	147 (100)
	Total	35 (13.3)	53 (20.1)	50 (18.9)	29 (11.0)	77 (29.2)	20 (7.6)	
Your work & contribution is not being recognized by your institutional administration								
	Male	26 (22.2)	31 (26.5)	18 (15.4)	12 (10.3)	22 (18.8)	8 (6.8)	117 (100)
	Female	22 (15.0)	48 (32.7)	24 (16.3)	13 (8.8)	34 (23.1)	6 (4.1)	147 (100)
	Total	48 (18.2)	79 (29.9)	42 (15.9)	25 (9.5)	56 (21.2)	14 (5.3)	264 (100)
The library profession is a thankless job								
	Male	29 (24.8)	32 (27.4)	17 (14.5)	12 (10.3)	20 (17.1)	7 (6.0)	117 (100)
	Female	37 (25.2)	39 (26.5)	20 (13.6)	12 (8.2)	34 (23.1)	5 (3.4)	147 (100)
	Total	66 (25.0)	71 (26.9)	37 (14.0)	24 (9.1)	54 (20.5)	12 (4.5)	264 (100)
You don't like to take new initiatives								
	Male	20 (17.1)	16 (13.7)	14 (12.0)	14 (12.0)	45 (38.5)	8 (6.8)	117 (100)
	Female	18 (12.2)	33 (22.4)	11 (7.5)	15 (10.2)	64 (43.5)	6 (4.1)	147 (100)
	Total	38 (14.4)	49 (18.6)	25 (9.5)	29 (11.0)	109 (41.3)	14 (5.3)	264 (100)
You don't like to accept new challenges								
	Male	15 (12.8)	19 (16.2)	9 (7.7)	11 (9.4)	52 (44.4)	11 (9.4)	117 (100)
	Female	16 (10.9)	25 (17.0)	19 (12.9)	13 (8.8)	64 (43.5)	10 (6.8)	147 (100)
	Total	31 (11.7)	44 (16.7)	28 (10.6)	24 (9.1)	116 (43.9)	21 (8.0)	264 (100)
You love doing your job beyond office hours								
	Male	32 (27.4)	30 (25.6)	22 (18.8)	9 (7.7)	17 (14.5)	7 (6.0)	117 (100)
	Female	37 (25.2)	40 (27.2)	13 (8.8)	18 (12.2)	24 (16.3)	15 (10.2)	147 (100)
	Total	69 (26.1)	70 (26.5)	35 (13.3)	27 (10.2)	41 (15.5)	22 (8.3)	264 (100)
You are least satisfied with your job								
	Male	17 (14.5)	31 (26.5)	16 (13.7)	16 (13.7)	29 (24.8)	8 (6.8)	117 (100)
	Female	30 (20.4)	20 (13.6)	22 (15.0)	18 (12.2)	46 (31.3)	11 (7.5)	147 (100)
	Total	47 (17.8)	51 (19.3)	38 (14.4)	34 (12.9)	75 (28.4)	19 (7.2)	264 (100)

Normally a professional would never admit of being incompetent, no matter even if one is, but contrary to it, 7.6% of respondents agreed to being incompetent professionals, while 21.2% partly agreed with the statement. An employee can judge his/her competence by assessing the performance of his/her colleagues. If an employee believes that he/she is not able to perform to the extent to which his/her colleagues do, he/she can easily rate himself/herself as incompetent. This self-assessment can help an employee to look for the reasons which act as an impediment in his/her job performance and ultimately affect his/her job satisfaction. Mal-employment is seen as one of the reasons which results in the underperformance of employees. The majority, 50.8% of respondents, disagreed with the statement, while 12.1% of respondents were not sure and 4.5% of respondents preferred not to reply to this particular question (Table 2).

In continuation to the above question, the respondents were asked regarding their peers being least competent; the majority at 29.2% disagreed and 11% partly disagreed, while 13.3% agreed with the statement and 20.1% partly agreed to it. Compared to 26.5% of males, 31.3% of females showed disagreement with the statement, compared to 14.5% of males and 12.2% of females who agreed to it. Here again the reasons can be the same but the other way round. A competent professional is always in a position to judge the job performance capability of his/her colleagues. Incompetent professionals mostly lean against competent professionals for professional help and other advice more often (Table 2).

Undermining the importance of libraries and library professionals is a very common problem faced by library professional across India and is the most compelling reason which leads to job dissatisfaction among LIS professionals. Accordingly, 18.2% of respondents of the present survey agreed and 29.9% of respondents partly agreed with the statement that their contribution is not recognized by their institutional administration. It is quite obvious that library professionals working in institutions where their contribution, role, and importance are not recognized by their institutional administration are bound to show signs of job dissatisfaction. Contrarily, the institutions which recognize the role of library professionals and their contribution in the teaching and research activities show higher levels of job satisfaction. On the gender level, compared to 18.8% of males, 23.1% of females disagree with the statement, while compared to 22.2% of males, 15% of females agree to it. Compared to females, more males agree with the statement and this again corroborated the fact that compared to males more females disagree with the statement.

Appreciating an employee while on the job is perhaps the greatest reward a person gets. Respect and gratitude are the instant rewards a professional receives while serving clientele, and if professionals are not being paid gratitude or respect the way they expect, their morale is bound to come down, which ultimately influences their overall job satisfaction. While trying to ascertain from the professionals how far people pay gratitude to them while serving them, one fourth of the respondents (25%) replied that the library profession is a thankless job, followed by the majority, 26.9% of respondents, who partly agreed to it. However, 20.5% of respondents totally disagreed with it. Here it is equally desired that library professionals should be thorough professionals, as they should never expect gratitude from their clientele, so that it may not lower down the moral and enthusiasm in a professional.

The majority, 41.3% of respondents, disagreed with the statement that they do not like to take new initiatives, while 14.4% deemed the statement right. It is quite encouraging to see the LIS professionals refuting the statement and emphatically saying that they are open to new initiatives. At the gender level, compared to 43.5% of females, 38.5% of males believe that library professionals like to take new initiatives, compared to 17.1% of males and 12.2% of female library professionals who believe that library professionals do not like to take new initiatives. From the analysis it is evident that female library professionals are more proactive in taking new initiatives than their male counterparts. The more a professional is satisfied which his/her job, the more he/she is going to involve himself/herself, which is bound to result in taking new initiatives, and this enthusiasm towards new initiatives remains missing in those employees who are generally dissatisfied with their job (Table 2).

Library professionals are quite open to accepting new challenges, as 43.9% of respondents have refuted the statement that library professionals do not like to accept new challenges. However, 11.7% of respondents viewed that library professionals do not like to accept new challenges. At the gender level, both male and the female respondents have shown almost similar levels of agreement with the statement.

Commitment towards one's job gets reflected through different ways and means and doing one's job beyond office hours somewhat signifies the passion of an employee towards his/her job. Accordingly, respondents were asked whether they love doing their job beyond office hours. The majority, 26.5% of respondents, partly agreed and 26.1% of respondents totally agreed with the statement, while 15.5% disagreed and 10.2% partly disagreed with the statement.

At the gender level, it can be inferred from the response percentage that compared to female library professionals, male library professionals love to do their job beyond office hours more.

With a view to evaluate overall job satisfaction among LIS professionals, the respondents were asked if they are least satisfied with their job. Contrary to the statement, the majority at 28.4% disagreed and 12.9% partly disagreed with the statement. However, 17.8% of respondents agreed with the statement, backed by 19.3% of respondents, who partly agreed with the statement. At the gender level, compared to 24.8% of males, 31.3% of females disagreed with the statement and compared to 14.5% of males, 20.4% of females agreed to it (Table 2). From the figures it is evident that a mixed trend can be seen in the overall job satisfaction of library professionals working in the state; however, at the gender level, compared to male library professionals, female library professionals are more satisfied with their job.

In the present-day scientific world, if an employee is asked how far he or she is emotionally attached to his/her profession may sound a bit unusual, but the harsh reality is, how so objective we may reflect outlook, but in the heart of hearts there is always space for emotions in each individual. Accordingly, the respondents were asked about their emotional attachment to their job and a whopping 70.8% of respondents admitted to being emotionally attached to

Table 3. Respondents' opinions on some other psychological aspects associated with job satisfaction

Statement		Yes (%age)	No (%age)	To some extent (%age)	Can't say (%age)	No response (%age)	Total (%age)
You are emotionally attached to your profession							
	Male	76 (65.0)	16 (13.7)	16 (13.7)	4 (3.4)	5 (4.3)	117 (100)
	Female	111 (75.5)	13 (8.8)	18 (12.2)	1 (0.7)	4 (2.7)	147 (100)
	Total	187 (70.8)	29 (11.0)	34 (12.9)	5 (1.9)	9 (3.4)	264 (100)
Would you like to seek voluntarily retirement							
	Male	15 (12.8)	89 (76.1)	5 (4.3)	4 (3.4)	4 (3.4)	117 (100)
	Female	18 (12.2)	111 (75.5)	4 (2.7)	10 (6.8)	4 (2.7)	147 (100)
	Total	33 (12.5)	200 (75.8)	9 (3.4)	14 (5.3)	8 (3.0)	264 (100)
Would you like to seek post retirement engagement in LIS profession							
	Male	54 (46.2)	37 (31.6)	17 (14.5)	4 (3.4)	5 (4.3)	117 (100)
	Female	56 (38.1)	56 (38.1)	14 (9.5)	12 (8.2)	9 (6.1)	147 (100)
	Total	110 (41.7)	93 (35.2)	31 (11.7)	16 (6.1)	14 (5.3)	264 (100)
Are you able to strike a balance between your work, family, and personal life							
	Male	63 (53.8)	18 (15.4)	32 (27.4)	-	4 (3.4)	117 (100)
	Female	80 (54.4)	20 (13.6)	37 (25.2)	2 (1.4)	8 (5.4)	147 (100)
	Total	143 (54.2)	38 (14.4)	69 (26.1)	2 (0.8)	12 (4.5)	264 (100)
Would you have opted for LIS profession had you been familiar with it?							
	Male	36 (30.8)	50 (42.7)	22 (18.8)	4 (3.4)	5 (4.3)	117 (100)
	Female	28 (19.0)	70 (47.6)	33 (22.4)	5 (3.4)	11 (7.5)	147 (100)
	Total	64 (24.2)	120 (45.5)	55 (20.8)	9 (3.4)	16 (6.1)	264 (100)
LIS professionals have less work related stress							
	Male	20 (17.1)	49 (41.9)	36 (30.8)	9 (7.7)	3 (2.6)	117 (100)
	Female	17 (11.6)	85 (57.8)	34 (23.1)	6 (4.1)	5 (3.4)	147 (100)
	Total	37 (14.0)	134 (50.8)	70 (26.5)	15 (5.7)	8 (3.0)	264 (100)
Would you encourage your child to pursue his/her career in LIS profession?							
	Male	36 (30.8)	55 (47.0)	15 (12.8)	9 (7.7)	2 (1.7)	117 (100)
	Female	45 (30.6)	53 (36.1)	27 (18.4)	16 (10.9)	6 (4.1)	147 (100)
	Total	81 (30.7)	108 (40.9)	42 (15.9)	25 (9.5)	8 (3.0)	264 (100)
If given a chance, would you like to be a library professional again?							
	Male	54 (46.2)	34 (29.1)	16 (13.7)	9 (7.7)	4 (3.4)	117 (100)
	Female	74 (50.3)	39 (26.5)	14 (9.5)	16 (10.9)	4 (2.7)	147 (100)
	Total	128 (48.5)	73 (27.7)	30 (11.4)	25 (9.5)	8 (3.0)	264 (100)

LIS, library and information science.

their profession. Compared to 75.5% of females, 65% of male respondents admitted to being emotionally attached to their profession. However, 11% of respondents did not show any emotional attachment to their job, while 12.9% rated it to some extent. 1.9% of respondents were not sure and 3.4% of respondents did not reply to this particular question (Table 3).

Upon asking respondents as whether they would like to seek voluntary retirement from the library profession, a whopping 75.8% of respondents replied no, while 12.5% replied yes. This somewhat corroborates that the majority of LIS professionals are willing to carry on with their profession and can be deemed as somewhat satisfied with their job. There is no significant difference between male and female respondents, as professionals of both genders have reflected almost similar opinions.

With a view to corroborate the above statement, the respondents were asked whether they would like to seek post retirement engagement in the LIS profession; the majority at 41.7% replied yes, and 11.7% replied maybe. Compared to 38.1% of females, 46.2% of male respondents replied that if engaged, they will take a post retirement job in the LIS profession. This again corroborates that the majority of LIS professionals find their job meaningful and fulfilling hence they have shown willingness to carry on with the LIS profession if engaged after retirement. Contrarily, 35.2% of respondents replied that they will not take post retirement engagement in the LIS profession. This also somewhat means that nearly one-third of the library professionals working in the higher education institutions in Jammu and Kashmir are not fully satisfied with their job, hence they do not want to continue with it as their post retirement engagement (Table 3).

Table 4. Respondents other preferred choices of employment apart from library profession

Total responses	Gender		Total (%age)
	Female (%age)	Male (%age)	
Teacher	44 (29.9)	29 (24.8)	73 (27.7)
Doctor	5 (3.4)	4 (3.4)	9 (3.4)
Administrative officer	5 (3.4)	2 (1.7)	7 (2.7)
Banker	4 (2.7)	2 (1.7)	6 (2.3)
Businessman/ entrepreneur	2 (1.4)	4 (3.4)	6 (2.3)
IT professional/engineer	1 (0.7)	4 (3.4)	5 (1.9)
Army/police, etc.	-	3 (2.6)	3 (1.1)
Auditor/economist	2 (1.4)	1 (0.9)	3 (1.1)
Scientist	2 (1.4)	1 (0.9)	3 (1.1)
Social worker	2 (1.4)	1 (0.9)	3 (1.1)
Writer	2 (1.4)	1 (0.9)	3 (1.1)
Lawyer	2 (1.4)	-	2 (0.8)
Sports person	-	1 (0.9)	1 (0.4)
Can not say	5 (3.4)	3 (2.6)	8 (3.0)
Other	2 (1.4)	7 (6.0)	9 (3.4)
No response	69 (46.9)	54 (46.2)	123 (46.6)
Total	147 (100)	117 (100)	264 (100)

The majority, 54.2% of respondents, are able to strike a balance between their work, family, and personal life, while 26.1% are able to do this balance to some extent. However, 14.4% find it difficult to strike a balance between their work, family, and personal life. It is pertinent to note that if an employee is not able to strike a balance between work and family life, he/she is bound to show signs of dissatisfaction with his/her job. No considerable difference was found at the gender level, as both male and female respondents equally weighed the question.

There is a perception among LIS circles that the majority of LIS professionals entering into the profession are not generally familiar with the nature of the job and other allied aspects of the LIS profession. Accordingly, the respondents were asked whether they would have opted for the LIS profession had they been familiar with it; 24.2% replied yes, while 20.8% rated to some extent. However, the majority at 45.5% of respondents replied no, which is quite encouraging and reflects that nearly 50% of library professionals entering into the library profession are familiar with the library profession. Here the need can be pressed for drawing awareness about the library profession among the potential library professionals, as a sizable portion of potential library professionals are not aware of the nature of the LIS profession and the types of services they are to render to their clientele.

Upon asking the respondents whether LIS professionals have less work-related stress, 14% agreed while 26.5% rated to some extent. However, the majority at 50.8% of respondents disagreed with it and believe that LIS professionals do have their own lot of job-related stress. Compared to 41.9% of males, 57.8% of female respondents viewed that library professionals have work related stress. This refutes the notion that library professionals have much less work-related stress (Table 3).

The respondents were asked, would they encourage their children to pursue their careers in the library and information profession? The majority, 40.9% of respondents, replied no, and 30.7% replied yes. Here again, it can be inferred that the majority of respondents do not find the LIS profession that rewarding and fulfilling, hence they do not

want to encourage their children to pursue their career in LIS, which is quite worrisome. The significant difference can be found at the gender level as well, wherein, against 36.1% of females, 47% of males replied that they will not encourage their children for the library profession. This somewhat also corroborates the fact that compared to male library professionals, female library professionals are more satisfied with their jobs, hence they will not hesitate to encourage their children to take up the library profession. Another 15.9% of respondents were not sure, but replied that maybe they will encourage.

The respondents were asked, if given the chance, would they like to be library professionals again? The majority at 48.5% replied yes, while 11.4% replied to some extent. Compared to 46.2% of males, 50.3% of females reflected their willingness to take up the library profession once again if given a chance. However, 27.7% of respondents categorically rejected the idea.

The general notion is that if one does not get the job of his/her choice, he/she contends with the job which serves as a substitute to his/her choice. In the same way 27.7% of respondents replied that had they not been LIS professionals, they would have been teachers. Compared to 24.8% of males, 29.9% of female respondents replied that they would have been a teacher had they not been a library professional. This confirms that the library and the teaching profession are two faces of the same coin, and as such, those who may have aspired to being a teacher would have definitely felt contented at landing in the library profession. It is quite good to have LIS professionals who had aspired to be doctors, administrative officers, bankers, entrepreneurs, IT professionals, scientists, lawyers, sports persons, and more. What is more surprising to know is that a majority 123 (46.6%) respondents did not reply to this particular question (Table 4).

Table 5. Respondents were asked about their preferred places of work, where a professional can drag more job satisfaction

Options	Gender		Total (%)
	Female (%)	Male (%)	
Working in the home state	71 (48.3)	51 (43.6)	122 (46.2)
Working outside the home state	6 (4.1)	8 (6.8)	14 (5.3)
Working abroad	9 (6.1)	11 (9.4)	20 (7.6)
Hardly makes any difference	51 (34.7)	45 (38.5)	96 (36.4)
No response	10 (6.8)	2 (1.7)	12 (4.5)
Total	147 (100)	117 (100)	264 (100)

The majority, 46.2% of respondents, replied that library professionals working in their home state can find more job satisfaction than those working outside their home state. The scenario is no way different at the gender level, as against 43.6% of males, 48.3% of females uphold this view. Besides, it is an understood fact that females generally tend to work within their home state than their male counterparts. Similarly, 36.4% of respondents viewed that working within the home state or outside the home state makes hardly any difference. Still more, 5.3% of respondents believe that working outside the home state gives more job satisfaction, 7.6% believe that working abroad helps in getting more job satisfaction, while 4.5% of respondents are not sure, hence they did not reply to this particular question (Table 5).

8. FINDINGS AND DISCUSSION

One can understand the cognitive state of a person who himself admits to being an incompetent professional. It is not about being straightforward in reply, but about assessing whether a professional is satisfied with his/her job who believes that he/she is not a competent person. If 7.6% of respondents of the present study admit to being incompetent, in a way it also means that they are aware of the fact that they are not delivering the services the way and to the extent they are expected, hence they can never be satisfied with their job. Contrary to this fact, if the respondent deems his/her peers as incompetent, this also impliedly means that the respondent is quite confident of what the profession demands and how far his/her peers are able to deliver on the expected lines. The professionals who rate their peers as incompetent can be inferred to be those who yearn for and feel happy if their peers are competent and result oriented, hence a peer's work efficiency may also serve as a sense of satisfaction to an employee.

Working beyond office hours in the library can be for two reasons, either the person is overburdened or the person loves his/her job to such an extent that working beyond office hours does not bother him/her at all. If a person loves to do his/her job beyond the office hours it means the person is satisfied with his/her profession. From the data analysis it is evident that a fair percentage of library professionals working in the higher education institutions of the state love to do their job beyond office hours. However, compared to female library professionals, male library professionals are more open to working beyond office hours. By and large, the library professionals working in the higher education sector of Jammu and Kashmir are satisfied with their job

and those who have shown dissatisfaction with their job can be for different reasons and the foremost being that a good number of respondents are working on a temporary basis. Similarly, advancement, work environment, and other factors may be contributing to their dissatisfaction.

Every job of whatsoever nature involves stresses of its own kind, and so holds true of the library profession and this gets corroborated by the fact that over 50% of respondents replied that they have work related stress. Compared to male library professionals, female professionals feel more stressed. Interestingly, one third of the library professionals viewed that they will encourage their children to opt for the LIS profession. This clearly indicates the fact that parents who are willing to encourage their children toward the library profession are satisfied with their job. Compared to males, female library professionals have shown more interest in encouraging their children to pursue a career as a library professional, which clearly reflects that female LIS professionals find the library profession more fulfilling and satisfying than their male counterparts.

The majority of library professionals see the LIS profession parallel to that of the teaching profession and are content with the way they contribute to the teaching and research activities of their respective institution. Teaching appears to be the most preferred profession among the females given the way the majority of the female respondents have replied that they had aspired for the teaching profession. The majority of library professionals working in the higher education institutions of Jammu and Kashmir are of the view that a professional can find more job satisfaction by working in his/her home state rather than outside or abroad. But at the same time nearly one-third of the respondents believe that working within one's home state or outside the home state hardly makes any difference on the job satisfaction of an employee.

It is an accepted fact that if there is someone who can uplift the image of the library profession among the masses, it is the library professionals themselves. The integration of technology is one of the key aspects of modern-day library services and thereon harnessing these IT applications to their optimum can definitely help a great deal in overcoming the misconceptions of the public towards the profession. Still more, the time has come when library professionals need to specialize in their own way. A library professional, if apart from having information handling specialization has the subject specialization, can definitely change the recourse of information handling and its provision. Such an information specialist, apart from helping in retrieving the relevant information, can also help the information seeker in its best exploitation. Similarly, some better means can be adopted whereby the library profession and the psyche of library professionals can be changed by turning their job more demanding, hence more fulfilling, with far greater acceptance in society. Besides, the individual psyche of library professionals should help them to accept this fact that no job is small or big; it is the mindset of people which rates one job better over the other. The fact remains that every menial job acts as a foundation of a bigger job; besides, it is said that work is worship, so a library professional should not feel offended if not paid gratitude by the library clientele or if undermined for not being important. The greatest job satisfaction and the greatest reward an employee gets for work done is in the form of pleasure one obtains by rendering service to information seekers.

More than half of respondents viewed that the library profession is a thankless job; this in itself means that library professionals believe that people in general do not find the library profession important, when compared to other professions. It is quite obvious that the library professionals who receive gratitude from their clientele for rendering services will feel elated, hence will serve a sense of satisfaction to them, which in fact is a reward in its own way.

9. CONCLUSION

The individual perception of an employee towards his/her job and other intrinsic and extrinsic aspects associated with his/her job play a very significant role in his/her overall job satisfaction. It is not always the external factors, like work environment, salary, interpersonal relationship, recognition, advancement, or other administrative aspects which solely influence job satisfaction, but the psyche of an individual plays a very vital role in the overall job satisfaction of an employee. Societies have moved a long way, whereby jobs have become unisex and are no more gender specific. In the same way, keeping in view the response percentage of this particular study, it can be argued that the library profession is preferred by both males and females equally; nevertheless, a majority of the respondents of this particular survey are females. However, on the gender level, the male and female library professionals have different psychological perceptions about the profession and so do vary their degree of job satisfaction on different grounds.

The foremost aspect which emerges from the analysis is that the library profession as a whole is somewhat undermined by the public at large, which results in the public in general and the library clientele in particular

failing to recognize the contribution of libraries and library professionals in their routine information searching and handling. Though it is customary all across the world to pay gratitude to the person who supports or lends a helping hand in any manner, even if as a part of his/her duty, the irony is a majority of people pay gratitude only if they find that the help rendered is important. Still more, the harsh reality is that the majority of library clientele somewhat undermine the importance of the library profession; hence they hardly bother to pay gratitude to the library professionals serving them at the desk. Given this fact, there is a need to orient working library professionals through guidance and counseling mechanisms about public attitudes towards libraries and library professionals, which by no means should deter LIS professionals from rendering their duty with the utmost sincerity and integrity.

There is also a need to draw awareness about the library profession among potential library professionals, as a good number of potential professionals are not aware of the nature of the job and its other allied aspects. The foremost aspect is that librarianship is an allied aspect of teaching whereby a teacher is actively involved with the teaching and research activities of his/her institution, so is a library professional passively involved with all such activities.

REFERENCES

Brief, A. P., & Roberson, L. (1989). Job attitude organization: An exploratory study. *Journal of Applied Social Psychology*, 19(9), 717-727.

Burd, B. (2003). Work values of academic librarians: Exploring the relationships between values, job satisfaction, commitment and intent to leave. *ACRL Eleventh National Conference*. Retrieved October 1, 2018 from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/pdf/burd.PDF.

Dormann, C., & Zapf, D. (2001). Job satisfaction: A meta-analysis of stabilities. *Journal of Organizational Behavior*, 22(5), 483-504.

Government of India (2011). *Census of India: Provisional population totals India. Paper 1: Census 2011*. Retrieved October 1, 2018 from http://censusindia.gov.in/2011-prov-results/prov_results_paper1_india.html.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 87(3), 530-541.

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology*, 83(1), 17-34.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297-1349). Chicago: Rand McNally.

Pandita, R. (2017). Job satisfaction among library and information science professionals in India: A case study. *Journal of Information Science Theory and Practice*, 5(1), 47-64.

Pandita, R., & Dominic, J. (2016). Variables of job satisfaction: A review study with special reference to LIS professionals. *International Journal of Information Dissemination and Technology*, 6(4), 258-267.

Pandita, R., & Dominic, J. (2018). Impact of information technology on the job satisfaction of LIS professionals: A case study of Jammu & Kashmir. *DESIDOC Journal of Library and Information Technology*, 38(2), 75-81.

Reid, B. (2005). What do I do now: Suggestions for the frustrated mid-career professional. *IEEE Antennas and Propagation Magazine*, 47(5), 159-163.

Robbins, S. P. (2003). *Organizational behavior* (10th ed.). New Delhi: Prentice Hall.

Satija, M. P. (2003). Academic status of librarians as perceived by sociologists. *ILA Bulletin*, 39(200), 5-10.

Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, 70(3), 469-480.

Weiss, H. M., & Cropanzano, R. (1996). Affective events theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior: An annual series of analytical essays and critical reviews, Vol. 18* (pp. 1-74). New York: Elsevier Science/JAI Press.

Williamson, J. M., Pemberton, A. E., & Lounsbury, J. W. (2005). An investigation of career and job satisfaction in relation to personality traits of information professionals. *The Library Quarterly*, 75(2), 122-141.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 28-38, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.3

Minimally Supervised Relation Identification from Wikipedia Articles

Heung-Seon Oh
Korea University of Technology and Education, Cheonan, Korea
E-mail: ohhs@koreatech.ac.kr

Yuchul Jung*
Kumoh National Institute of Technology, Gumi, Korea
E-mail: jyc@kumoh.ac.kr

ABSTRACT
Wikipedia is composed of millions of articles, each of which explains a particular entity with various languages in the real world. Since the articles are contributed and edited by a large population of diverse experts with no specific authority, Wikipedia can be seen as a naturally occurring body of human knowledge. In this paper, we propose a method to automatically identify key entities and relations in Wikipedia articles, which can be used for automatic ontology construction. Compared to previous approaches to entity and relation extraction and/or identification from text, our goal is to capture naturally occurring entities and relations from Wikipedia while minimizing artificiality often introduced at the stages of constructing training and testing data. The titles of the articles and anchored phrases in their text are regarded as entities, and their types are automatically classified with minimal training. We attempt to automatically detect and identify possible relations among the entities based on clustering without training data, as opposed to the relation extraction approach that focuses on improvement of accuracy in selecting one of the several target relations for a given pair of entities. While the relation extraction approach with supervised learning requires a significant amount of annotation efforts for a predefined set of relations, our approach attempts to discover relations as they occur naturally. Unlike other unsupervised relation identification work where evaluation of automatically identified relations is done with the correct relations determined a priori by human judges, we attempted to evaluate appropriateness of the naturally occurring clusters of relations involving person-artifact and person-organization entities and their relation names.

Keywords: relation identification, Wikipedia mining, unsupervised clustering

Open Access

Accepted date: August 08, 2018
Received date: December 08, 2017

***Corresponding Author:** Yuchul Jung
Assistant Professor
Kumoh National Institute of Technology, 61 Daehak-ro, Gumi 39177, Korea
E-mail: jyc@kumoh.ac.kr

1. INTRODUCTION

Wikipedia, the largest online encyclopedia, is composed of millions of articles, each of which explains an entity with various languages in the real world. Since the articles are contributed and edited by a large population of diverse experts with no specific authority, Wikipedia can be seen as a naturally occurring body of human knowledge. This characteristic attracts researchers to focus on mining structured knowledge from Wikipedia.

Relation extraction (RE) often refers to the task of extracting relations between named entities. Most past RE research has focused on development of supervised learning methods for the task of identifying a predefined set of relations from a known corpus, e.g., the ACE corpus. Supervised learning tasks, however, require heavy human annotation efforts to build training data for different domains. To alleviate the problem, semi-supervised methods using a search engine were developed (Etzioni et al., 2005; Pantel & Pennacchiotti, 2006), which start with initial seeds and go through a bootstrapping process using a search engine. Unlike the RE task, recent work on unsupervised relation identification (Hasegawa, Sekine, & Grishman, 2004; Rosenfeld & Feldman, 2006; Rozenfeld & Feldman, 2007; Y. Yan, Okazaki, Matsuo, Yang, & Ishizuka 2009) does not assume a predefined set of target relations, attempting to discover meaningful relations from a given corpus using a clustering algorithm.

As Wikipedia becomes a major knowledge resource, there have been some attempts to extract relations with Wikipedia structural characteristics. A research work (Wu & Weld, 2008) focused on extracting an "infobox" which describes attribute-value pairs of an entity of an article as a way of constructing ontology. A conditional random fields (CRFs) model is automatically trained with sentences related to infobox entries. In Nguyen, Matsuo, and Ishizuka (2007), a system is proposed to extract relations among entity pairs. Rather than using a named entity (NER) tagger to determine the semantic type of an entity, an entity type classifier is trained with features generated from category structures of Wikipedia. Then, relations are extracted with a support vector machines (SVMs) classifier trained by sub-tree features from the dependency structure of entity pairs. Compared to the methods above limited to a set of predefined relations, a method (Y. Yan et al., 2009) was proposed based on unsupervised relation identification framework by incorporating two context types of an entity pair: surface patterns from search results of an entity pair and dependency patterns from parsing the structure of a sentence of an entity pair in Wikipedia. Even though it shows the feasibility of identifying relations in combination with the Web, we thought that considering Wikipedia characteristics to identify relations is much more important.

In this paper, we propose a method to identify meaningful relations from Wikipedia articles with minimal human effort. Our method first detects entity pairs by utilizing the characteristics of Wikipedia articles. Similar to Nguyen et al. (2007), human effort only is required to prepare training data for an entity type classifier. Then, a set of entity pairs not associated in a grammar structure is filtered out. Then, context patterns are generated over sentences with respect to the remaining entity pairs. Based on them, entity pairs are clustered automatically. At last, a cluster label is chosen by selecting a representative word for each cluster. Experimental results show that our method produces many relation clusters with high precision. In previous work (Nguyen et al., 2007; Y. Yan et al., 2009), analysis of utilizing the characteristics of Wikipedia was not reported in detail even though the importance of the characteristics is not addressed. This paper reports our deep investigation.

The rest of this paper is organized as follows. Section 2 briefly introduces relevant research. The details of our method are described in Section 3. Section 4 delivers experimental results. Finally, we conclude in Section 5 with a suggestion for future work.

2. RELATED WORK

Wikipedia has been utilized for other purposes. Semantic relatedness (Gabrilovich & Markovitch, 2007; Strube & Ponzetto, 2006) is measuring the relatedness of two words or phrases utilizing characteristics such as the unique names of the articles and category hierarchy. Text classification (Gabrilovich & Markovitch, 2006) also utilizes the unique names of Wikipedia articles. Rather than using a bag of words approach, it utilizes the names of Wikipedia articles as semantic concepts for input text. When two input texts are entered, they are mapped to articles including each text and get the names of the articles as semantic concepts. The concepts are used as features for text categorization. Wikipedia also was used in taxonomy or ontology generation (Strube & Ponzetto, 2006; Wu & Weld, 2008). Due to the various usages of Wikipedia, the tasks of extracting entities and relations from Wikipedia are quite meaningful.

There have been some attempts to extract entities and relations from Wikipedia. One research work (Culotta,

McCallum, & Betz, 2006) regards RE as a sequential labeling task like NER and applies a CRFs model with conventional words and patterns as features for learning a classifier. In Nguyen et al. (2007) an entity detector and SVMs classifier were built using the characteristics of Wikipedia articles. Then, relations among the detected entities were determined by using another SVMs classifier trained with sub-trees mined from the syntactic structure of text. Unlike our approach, these approaches restrict target relations and require a significant amount of human labor for building the training data. KYLIN (Wu & Weld, 2007) automatically generates training data using infoboxes of Wikipedia articles to learn a CRFs model and extracts attribute-value pairs from the articles that have incomplete or no infoboxes.

Open information extraction (OpenIE) is a research area aiming to extract a large set of verb-based triples (or propositions) from text without restrictions of target entities and relations. Reverb (Fader, Soderland, & Etzioni, 2011) and ClauseIE (Corro & Gemulla, 2013) are representative projects to pursue OpenIE. Due to the no restrictions, OpenIE systems tried to consider all possible entities and relations in text of interest and thus produces many meaningless extractions. Unlike OpenIE, we are interested in somewhat normalized entities and relations existing in Wikipedia.

For the task of unsupervised relation identification, a research work (Hasegawa et al., 2004) shows a successful result of applying clustering to relation discovery from large corpora. It detects named entities using a NER tagger and considers entity pairs that often co-occur in a corpus for relation discovery. Entity pairs with intervening words between them are clustered using a hierarchical clustering technique. For each cluster, a representative word is chosen as the relation name based on word frequency. Instead of using intervening words, other systems (Rosenfeld & Feldman, 2006; Rozenfeld & Feldman, 2007) adopted a context pattern extraction and selection methods that uses dynamic programming and an entropy-based measure among the extracted patterns, respectively. The relation identification method in our system resembles the aforementioned method but with some unique technical details for a different resource, namely, Wikipedia.

In recent work (D. Zeng, Liu, Lai, Zhou, & Zhao, 2014), neural networks are employed to train an extraction model. D. Zeng et al. (2014) utilized the convolutional neural network to automatically extract features that are not dependant on traditional natural language processing tools and evade the error propagation problem. Although other approaches based on deep learning adopted long short-term memory networks along the shortest dependency path (X. Yan et al., 2015) and proposed an attention mechanism with bidirectional long short-term memory networks (Zhou et al., 2016), all of these models require sufficient training data and time to generate a high-performing model.

To alleviate the difficulties of producing training examples for RE, distant supervision has been used (Craven & Kumlien, 1999; Mintz, Bills, Snow, & Jurafsky, 2009). There exist two major research directions for the distant supervision. One direction is to use it for directly enriching knowledge bases from unstructured text, as well as leveraging the knowledge bases to generate the distant supervision labels (Poon, Toutanova, & Quirk, 2015; Parikh, Poon, & Toutanova, 2015). The other direction, so called Socratic learning (Varma et al., 2016), uses the differences in the predictions of the generative model to reduce the noise in distant supervision labels. Meanwhile, those approaches require multiple sources of weak supervision. More recently, a reinforcement learning approach was proposed to conduct large scale RE by learning a sentence relation extractor with distant supervised datasets (X. Zeng, He, Liu, & Zhao, 2018).

3. PROPOSED METHOD

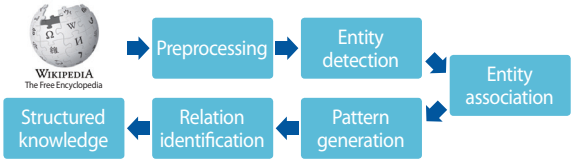


Fig. 1. Overview of proposed method.

Fig. 1 shows the overview of our method. From Wikipedia articles as input, structured knowledge is identified with minimal human effort. At first, preprocessing is performed, such as tokenization, Part-of Speech tagging, and chunking to Wikipedia articles. For each sentence, entities are detected and associated to make entity pairs. Discriminative patterns for entity pairs are retained. Entity pairs are clustered based on the patterns with hierarchical clustering method. Then, for each cluster, a representative word is selected as a name of the cluster.

3.1. Preprocessing

Several preprocessing stages are performed on Wikipedia articles. We first retain the raw text of an article by filtering out markup tags. Then, several miscellaneous parts not related to the main text such as *See Also* and *References* are

discarded. The remaining text parts of the articles undergo tokenization, sentence splitting, Part-of Speech tagging, and chunking steps in turn via OpenNLP tools.[1]

To ensure that a sufficient amount of contextual information exists surrounding entities, we discarded sentences having less than five words, and articles consisting of less than 25 sentences. Sentences with more than 30 words were also discarded to avoid potential errors due to the complexity involved in sentence processing.

3.2. Entity Detection

In Culotta et al. (2006), two types of entities are defined in Wikipedia articles: a principal entity and secondary entity. A principal entity refers to an instance of the name (title) of the article which is being described. A secondary entity refers to mentioned entities anchored in the same article which is linked to another Wikipedia article. A principal entity is often expressed in a different way with an anaphor. This is a natural phenomenon of English. For example, "Bruce Willis," a famous movie star, can be mentioned with "Willis," "he," or "an American actor" in the corresponding article. Definitely, we may miss many mentions of a principal entity without considering anaphors. There are various methods to resolve anaphora and co-references (Sukthanker, Poria, Cambria, & Thirunavukarasu, 2018). We adopted the heuristic method in Nguyen et al. (2007) for resolving anaphors referring to principal entities. Secondary entities linked to other Wikipedia articles are identified in a straightforward manner as they are tagged as such. Entities ending with a proper noun are only considered since our current focus is on named entities. The above step results in sentences with a principal and secondary entity pair.

To retain meaningful relations, the semantic classes of entities should be considered. For example, a chairman relation only occurs between person and organization. In our work, four semantic classes of entities are considered: person, organization, location, and artifact. An article does not belong to any of four semantic classes because they do not cover all Wikipedia articles. For that reason, we add other types for undefined classes.

As each entity corresponds to a Wikipedia article, entity classification can be regarded as text classification aiming at classifying an article to one of five classes. Unlike a common text classification, we assumed that all parts of an article are not effective to classify among five semantic classes. Similar to Nguyen et al. (2007), the SVMs classifier is trained with five features incorporating Wikipedia's structural characteristics: 1) category feature (categories collected by tracking back from the article up to k parent levels of the Wikipedia category hierarchy), 2) category term feature (the terms in the category feature), 3) category headword feature (the headwords of categories in the category feature), 4) first sentence term feature (terms in the first sentence in the article), and 5) title term feature (terms consisting of the article title). In this step, human effort is required to prepare an annotated dataset. Fortunately, this is cheap and easy because our task is just to assign a semantic class to an article, not a label sequence of a word sequence, for common NER tasks.

3.3. Entity Association

A major goal of our research is to identify relations between principal and secondary entities in a Wikipedia article. To satisfy the goal, we should find potentially useful entity pairs that can have a certain relation. Two approaches are possible: based on co-occurrence or a grammatical relation between two entities. The first approach as used in Hasegawa et al. (2004) selects entity pairs that occur more frequently than a threshold. A pair of entities that occur together very rarely would not possess a relation of sufficient interest. The second approach selects entity pairs involved in a grammatical relation, like a subject-object or object-subject relation, as in Shinyama and Sekine (2006).

Unlike more frequently used data for relation extraction, such as news data, however, there are few co-occurring entity pairs in Wikipedia because of the nature of

(a) Entity pair information with corresponding sentences

ID	First Entity	Second Entity
1	(0, 0, P, PER, He/PRP)	(2, 3, S, PER, Dan/NNP Rather.NNP)
2	(0, 0, P, PER, He/PRP)	(6, 7, S, ORG, Planet/NNP Hollywood/NNP)

ID	Sentences
1	He/PRP interviewed/VBD Dan/NNP Rather/NNP in/IN what/WP he/PRP would/MD later/PB call/VB the/DT most/RBS serious/JJ conversation/NN of/IN my/PRP$ entire/NN life/NN *J*.
2	He/PRP is/VBZ also/RB a/DT co-founder/NN of/IN Planet/NNP Hollywood/NNP *J*.

(b) Slot-marked sentences

ID	Sentences
1	<PER>/ENT interviwed/VBD <PER>/ENT in/IN what/WP he/PRP would/MD later/RB call/VB the/DT most/RBS serious/JJ conversation/NN of/IN my/PRP$ entire/NN life/NN *J*.
2	<PER>/ENT is/VBZ also/RB a/DT co-founder/NN of/IN <ORG>/ENT *J*.

Fig. 2. Entity pair information for the article on "Bruce Willis" (a) and sentences after slot-marking (b).

[1] https://opennlp.apache.org/

encyclopedia articles. For that reason, we parsed sentences and retained predicate-argument structures. Based on the structure, sentences with entity pair matched to subject-object pair are assumed to have a relation. Fig. 2 shows an example "Bruce Willis" article. The entry (0, 0, P, PER, He/PRP) indicates the start token, end token, principal entity, person type, and the entity text, respectively.

For further processing, entities in sentences are generalized by being slot-marked with a corresponding entity type and ENT indicating an entity tag. In addition, numbers are normalized to "#NUM#". This generalization process makes it easier to find common patterns for clustering. An example for a slot-marked sentence is shown in Fig. 2(b).

3.4. Pattern Generation

To identify relations, each entity pair is encoded as a feature vector representation. A feature vector should consist of discriminative features and values. To satisfy two conditions, feature vectors are constructed through a pattern extraction and selection (Fradkin & Mörchen, 2015).

The aim of pattern extraction is to provide necessary data for clustering entity pairs. In order to provide sufficient context information of entities, we applied Smith-Waterman (SW) algorithm (Smith & Waterman, 1981), which is one of the dynamic programming methods for a local alignment of molecular subsequences, for context pattern extraction.

The SW algorithm starts with constructing a score matrix D for two different input sentences using the scoring scheme shown below. The two input sentences are represented as $s = s_0 s_1 \cdots s_i$ and $t = t_0 t_1 \cdots t_j$ where s_i and t_j indicates i-th and j-the words in the two input sentences, respectively.

$$D(i,j) = \max \begin{pmatrix} 0 \\ D(i-1,j-1) + D(s_i, t_j) \\ D(i-1,j) - gap \\ D(i,j-1) - gap \end{pmatrix} \quad (1)$$

Here D(i, j) is a cost function for i-th and j-th words and gap is a penalty cost for a gap. We set gap to 1 and defined the cost function below.

$$D(i,j) - \begin{cases} 2 & \text{if } s_i = t_j \\ -1 & \text{if } s_i \neq t_j \end{cases} \quad (2)$$

Initially, all positions of the score matrix are initialized with 0. By comparing s_i and t_j, the score matrix is filled with D(i, j). After constructing the score matrix, backtracking is carried out for finding the best local alignment starting from the position assigned a maximum score on the matrix following the policies in turn.

$$D(i,j) - \begin{cases} D(i-1,j-1) & \text{if } D(i,j) = D(i-1,j-1) + D(s_i, t_j) \\ D(i-1,j) & \text{if } D(i,j) = D(i-1,j) - gap \\ D(i,j-1) & \text{if } D(i,j) = D(i,j-1) - gap \end{cases} \quad (3)$$

Fig. 3 shows an example for computing the alignment matrix and the resulting alignment between two input sentences.

An alignment is converted to a pattern after replacing mismatching and similar words with a wild card character that allows for any word sequence. In the example, the alignment is converted to a pattern "<ORG> * located in * <LOC>".

(a) Computing an alignment matrix

	<ORG>	wes	located	in	near	<LOC>	.
<ORG>	2	1	0	0	0	0	0
is	1	1	0	0	0	0	0
located	0	0	3	2	1	0	0
in	0	0	2	5	4	3	2
<LOC>	0	0	1	4	4	6	5
.	0	0	0	3	3	5	8

(b) An alignment between two sentences

<ORG>	is	located	in	GAP	<LOC>	.
|	:	|	|	.	|	|
<ORG>	wes	0	3	3	5	5

Symbol	Meaning
|	Match
.	Mismatch
:	similar

Fig. 3. Example of computing an alignment matrix (a) with the resulting alignment (b).

Even though pattern extraction aims at reflecting common contextual information of entity pairs, not all of the patterns are helpful for identifying relations. Many patterns are not discriminative because they are too specific or too general to certain contexts. In the clustering phase, such patterns may introduce noise and result in unexpected entity pair clusters. As such, selecting patterns with sufficient entity revealing contextual information is critical.

There are several feature selection methods such as information gain and x^2 that work with labeled data (Forman, 2003). However, they are not applicable because we do not have labeled data for relations. For that reason, an unsupervised feature selection method is adopted for selecting useful patterns (Jinxiu, Donghong, Lim, & Zhengyu, 2005; Rosenfeld & Feldman, 2007). The intuition behind the method is that good clustering features should

improve the separability of the dataset, making points that are close together still closer, and points that are far from each other still farther apart.

Let $C = \{c_0, c_1, \ldots, c_n\}$ be a set of examples where an example consists of patterns as features. Then, cosine similarity between two examples is defined:

$$S(c_i, c_j) = S_{ij} = \frac{c_i \cdot c_j}{|c_i||c_j|} \quad (4)$$

Using the similarity, scoring function for a feature f is defined:

$$Score(f) = E - E_{-f} \quad (5)$$

Where

$$E = -\sum_{c_i, c_j \in C} S_{ij} \log S_{ij} + (1 - S_{ij}) \log(1 - S_{ij}) \quad (6)$$

$$E_{-f} = -\sum_{c_i, c_j \in C} S_{ij}^{f} \log S_{ij}^{f} + (1 - S_{ij}^{f}) \log(1 - S_{ij}^{f}) \quad (7)$$

and S_{ij}^{f} is the similarity between c_i and c_j after removing the feature f.

Performing the feature selection for full feature space over all examples is very time-consuming. To reduce the feature space with retaining patterns directly related to entities, we discard patterns which do not have entity slots and content words such as noun, verb, and adjective before feature selection. For example, "* located in *" is discarded because no entity slot occurs.

3.5. Relation Identification

Our goal is to discover relations from all entity pairs represented as a set of discriminative patterns. For that reason, a hierarchical agglomerative clustering (HAC) algorithm which is not concerned with the number of clusters in advance is a natural choice. As reported in Rosenfeld and Feldman (2007), we opted for single link HAC because it outperforms average and complete link HACs for relation identification tasks.

In single link HAC, initially, each of the data points is regarded as a single cluster. When the similarity distance of two clusters is within a threshold, two clusters merge. As a result, determining the threshold affects the clustering results. In our case, we utilized cosine similarity and set the threshold to 0.3. Since clusters without a sufficient number of instances cannot have a representative for the identified relation, those with less than five instances were not considered for further processing.

Since entity pairs are clustered based on the similarities of context patterns, we can assume that instances in each cluster have a common meaning for the context patterns, i.e., a relation between entities in our case. Instead of classifying the meaning to one of the existing relation names as in RE tasks, we opted for naming it with a representative word found in the cluster. The terms between the two entities in a cluster are candidates and evaluated with the TF*IDF scheme where TF is the term frequency in the cluster and IDF is the inverse document frequency of the term over entity paired sentences. The identified relations are shown in the last of this paper.

4. EXPERIMENTS

For experiments, we downloaded English Wikipedia articles and randomly selected a total of 32,355 articles after filtering, where an article was filtered if it did not represent a real-world entity. For example, entity Forrest Gump was discarded because he is not an actual person but the main character of a movie, while Tom Hanks, an actor who played the character, was kept because he is a real world entity. After going through entity detection and association explained in subsections 3.2 and 3.3, 103,526 sentences with principal and secondary entity pairs were retained. For example, let us see the sentence "Hanks has collaborated with film director Steven Spielberg on five films to date." Hanks is a principal entity while Steven Spielberg, a famous movie director, is a secondary entity in the article "Tom Hanks."

To assign the semantic classes of each entity, we built an entity classifier with LIBSVM (Chang & Lin, 2011). 4,123 and 415 articles were manually annotated and tested. Table 1 shows the results of the entity classifier. We obtained the best result performance when all features such as 1) category feature, 2) category term feature, 3) category headword feature, 4) first sentence term feature, and 5) title term feature were used with up to four parents in category structure.

Table 1. Performance of entity classifier

Features	Parent levels	Accuracy
All features	3	0.8364
	4	0.8571
	5	0.8475
	6	0.8356

Table 2. Results of pattern extraction and selection (# of instances)

Domain	Sentence	Extracted pattern	Selected pattern
PER-ART	4,567	14,856	3,782
PER-ORG	6,703	29,800	9,603

Table 3. Results of clustering with entity pairs

Domain	Relation cluster (Entity Pair)	Garbage cluster (Entity Pair)
PER-ART	115 (1,549)	1,548 (2,229)
PER-ORG	160 (2,180)	3,304 (4,015)

Table 4. Performances of anaphor identification and entity classification

Criterion	Total	Correct	Precision
Anaphor (PER-ART)	1,549	1,467	0.947
Anaphor (PER-ORG)	2,180	2,125	0.975
Entity (PER-ART)	3,098	2,382	0.769
Entity (PER-ORG)	4360	4280	0.982

To analyze the results in detail, we focused on two domains, person-organization (PER-ORG) and person-artifact (PER-ART). Table 2 shows simple statistics resulting from pattern extraction and selection for the two different cases. It can be seen that the number of surviving patterns after the selection process is only one third of the extracted patterns.

For clustering of entity pairs, we utilized LingPipe,[2] freely usable natural language tools, for single link HAC. Clusters that contain less than five entity pairs are considered a garbage cluster. Table 3 shows the results of clustering. In the case of PER-ART, for example, a total of 1,549 entity pairs form 115 relation clusters, indicating that 1,549 entity-relation-entity triples with 115 relations can be generated.

In entity detection, a heuristic method is adapted for identifying anaphors of principal entities. The effects of anaphor identification should be investigated because many entity pairs include anaphors and are processed further.

Table 4 shows the performances of anaphor identification and entity classification. It shows promising results in both domains. However, the precision of entity classification in the PER-ART domain is surprisingly lower than that of the PER-ORG domain, indicating that entity classification for

[2] http://alias-i.com/lingpipe/

Table 5. Precisions on PER-ART domain

Case	Total	Correct	Precision
1	1,549	1,093	0.706
2	1,467	1,059	0.722
3	836	626	0.749
4	798	604	0.757

Table 6. Precisions on PER-ORG domain

Case	Total	Correct	Precision
1	2,180	1,841	0.844
2	2,125	1,796	0.845
3	2,099	1,782	0.849
4	2,056	1,745	0.849

ART is more difficult than that of ORG. We have found two reasons resulting in the performance drop. The first is that entity classification is conducted for each article, not for each sentence. As a result, every entity receives the same entity type regardless of the context of an entity pair in a sentence. For example, in the following two sentences, Singapore General Hospital is supposed to have two different entity types: organization for the first and artifact for the second.

1. Ratnam began his career as a houseman at the Singapore General Hospital in 1959.
2. Singapore General Hospital was built in 1920.

The second reason is the insufficient coverage of training data. For example, many historical war names such as American Civil War are classified as artifacts. However, they should be classified as other categories and filtered out for further processing. It turns out that those incorrectly classified entities share the same category hierarchy information from Wikipedia, which is a key feature for our classifier, with those correctly classified. We evaluated the appropriateness between an entity pair and a relation by determining whether or not a relation is a representative word for an entity pair. For that, an entity pair and a relation are represented as a relation triple like entity-relation-entity. As a result, a precision indicates the overall appropriateness with respect to all of the relation triples. In order to avoid biased subjectivity, we counted a relation triple for precision when two evaluators (i.e., two authors of this paper) both agree with a relation triple as being appropriate.

Tables 5 and 6 show the results for each domain. To

Table 7. Example of cluster name errors

Relation	Example
German	Hattie McDaniel was the first performer of African descent to win an Academy Award.
enrolled	Bruce R. McConkie enrolled in Army ROTC while at the University of Utah.

analyze the effect of erroneous results of entity detection, we conducted four different evaluations for each domain. Case 1 includes all of the incorrect results from anaphor identification and entity classification. Case 2 excludes the incorrect results from anaphor identification. Case 3 excludes the incorrect results from entity classification. Case 4 excludes all of the incorrect results from anaphor identification and entity classification. The results show that excluding the incorrect results in the earlier phases improves the precision .051 and 0.005 in PER-ART and PER-ORG domains, respectively.

Considering only case 4 of both domains, we found two error types. The first error type shown in Table 7 is that the identified relation is not appropriate to represent the relation among entities. The second error type is caused by incorrect subject-object entity pairing. In the second example of Table 7, University of Utah is not an object of Bruce R. McConkie. This error type entirely depends on the results of parsing predicate-argument structure.

5. CONCLUSION

In this paper, we presented a method that identifies naturally occurring relations between entities in Wikipedia articles with an aim to minimize human annotation efforts. The manual annotations are required to construct training data for an entity classifier in general. However, the efforts should be minimized because it is a simple task of assigning a class to a Wikipedia article. Using the entity classifier, entity pairs which may have a meaningful relation are kept for relation identification. Relations are identified in an unsupervised way based on hierarchical clustering and pattern generation and selection. Our experimental results showed promising results for both entity classification and relation identification. From the analysis of experiments, we found that error propagation from entity classifier and heuristic anaphora detection is a critical issue for improving performance, but hard to avoid since our method heavily relies on unsupervised learning.

As Wikipedia grows and evolves via the contributions of the general public, this kind of automatic identification of relations among key entities that reflect the real world would be very useful for a variety of applications such as ontology and knowledge base construction, guided searching and browsing, and question answering. More specifically, in aspects of ontology construction, our proposed methods can be effectively used for building core (basic) ontology of specific domains. After that, the core ontology can be populated further by combining with domain-specific patterns, knowledge-based approaches, and other state-of-the-art supervised/unsupervised approaches.

Our future work includes the following extensions: expansion of the entity type pairs, more thorough and larger scale evaluation of the relation identification task, and more direct evaluation of the value of the entity and relation identification for ontology construction.

ACKNOWLEDGMENTS

This work was supported by the research fund for a newly appointed professor of Korea University of Technology & Education in 2018. This work was also supported b y the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF -2018R1C1B5031408).

REFERENCES

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Corro, L. D., & Gemulla, R. (2013). ClausIE: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 355-365). New York: ACM.

Craven, M., & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (pp. 77-86). Menlo Park: AAAI Press.

Culotta, A., McCallum, A., & Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 296-303). Stroudsburg: Association for Computational Linguistics.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S.,…Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1), 91-134.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Stroudsburg: Association for Computational Linguistics.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Fradkin, D., & Mörchen, F. (2015). Mining sequential patterns for classification. *Knowledge and Information Systems*, 45(3), 731-749.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1301-1306). Menlo Park: AAAI Press.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (pp. 1606-1611). San Francisco: Morgan Kaufmann Publishers.

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 415-422). Stroudsburg: Association for Computational Linguistics.

Jinxiu, C., Donghong, J., Lim, T. C., & Zhengyu, N. (2005). Unsupervised feature selection for relation extraction. In R. Dale, K. F. Wong, J. Su, & O.Y. Kwong (Eds.), *Natural Language Processing: IJCNLP 2005* (pp. 390-401). Berlin: Springer.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1003-1011). Stroudsburg: Association for Computational Linguistics.

Nguyen, D. P. T., Matsuo, Y., & Ishizuka, M. (2007). Relation extraction from Wikipedia using subtree mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence* (pp. 1414-1420). Menlo Park: AAAI Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 113-120). Stroudsburg: Association for Computational Linguistics.

Parikh, A. P., Poon, H., & Toutanova, K. (2015). Grounded semantic parsing for complex knowledge extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 756-766). Stroudsburg: Association for Computational Linguistics.

Poon, H., Toutanova, K., & Quirk, C. (2015). Distant supervision for cancer pathway extraction from text. In *Pacific Symposium on Biocomputing Co-Chairs* (pp. 120-131). Singapore: World Scientific.

Rozenfeld, B., & Feldman, R. (2006). High-performance unsupervised relation extraction from large corpora. In *Proceedings of Sixth International Conference on Data Mining* (ICDM'06) (pp. 1032-1037). Piscataway: IEEE.

Rosenfeld, B., & Feldman, R. (2007). Clustering for unsupervised relation identification. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (pp. 411-418). New York: Association for Computing Machinery.

Shinyama, Y., & Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 304-311). Stroudsburg: Association for Computational Linguistics.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195-197.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1419-1424). Menlo Park: AAAI Press.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2018). *Anaphora and coreference resolution: A review*. Retrieved September 2, 2018 from https://arxiv.org/pdf/1805.11824.pdf.

Varma, P., He, B., Iter, D., Xu, P., Yu, R., De Sa, C., & Ré,

C. (2016). *Socratic learning: Augmenting generative models to incorporate latent subsets in training data*. Retrieved September 2, 2018 from https://arxiv.org/abs/1610.08123.

Wu, F., & Weld, D. S. (2007). Autonomously semantifying Wikipedia. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (pp. 41-50). New York: Association for Computing Machinery.

Wu, F., & Weld, D. S. (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 634-644). New York: Association for Computing Machinery.

Yan, X., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z. (2015). Classifying relations via long short term memory networks along shortest dependency path. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1785-1794). Stroudsburg: Association for Computational Linguistics.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1021-1029). Stroudsburg: Association for Computational Linguistics.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics* (pp. 2335-2344). Sheffield: International Committee on Computational Linguistics.

Zeng, X., He, S., Liu, K., & Zhao, J. (2018). Large scaled relation extraction with reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 5658-5665). Palo Alto: Association for the Advancement of Artificial Intelligence.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 207-212). Stroudsburg: Association for Computational Linguistics.

APPENDIX. SAMPLE CLUSTER DETAILS

Table A1. Sample cluster details in PER-ART domain

Relation	Entity pair	Context pattern	Entity paired sentence	True	False
appear	Tyra Banks- Felicity Tamera Mowry- Smart Guy	<PER> appeared on <ART> <PER> appeared in	<PER> also appeared on <ART>. <PER> appeared in <ART >.	35	2
role	Chris Elliott - Cabin Boy Sylvester McCoy- The Cabaret of Dr Caligari	<PER> had * role in * <ART> <PER> played * role of *	<PER> had title role in <ART>. <PER> played the role of Snuff in the macabre BBC Radio 4 comedy series <ART>.	24	4
performed	Kellie Pickler- Red High Heels Alan Autry:Autry- Rudolph the Red Nosed Reindeer	<PER> performed * <ART> <PER> performed * of <ART>	<PER> performed live <ART> <PER> performed his rendition of <ART>	13	1
won	Philip K. Dick- The Man in the High Castle Robert Fuller- Golden Boot Award	<PER> won * <ART> for * in In #NUM# * <PER> won * <ART>	In 1963, <PER> won the Hugo Award for <ART >. In 1989, <PER> won the <ART>.	32	5

Table A2. Sample cluster details in PER-ORG domain

Relation	Entity pair	Context pattern	Entity paired sentence	True	False
educated	Haldane - Dragon School Dick McCreery- Eton College	<PER> was educated at <ORG> * College <PER> was educated at <ORG> * , * and	<PER> was educated at <ORG>, Eton College and at New College, Oxford. <PER> was educated at <ORG > .	48	0
professor	Haushofer - University of Munich Von Laue - University of Zurich	<PER> * professor * at <PER> * professor of * at * <ORG>	In 1919, <PER> would become professor of geography at the <ORG>. <PER> became professor of physics at the <ORG> in 1912.	42	0
attended	Brookings - Bowdoin College Hicks - University of Houston	<PER> attended <ORG> * , <PER> * attended * <ORG>	<PER> attended <ORG> in Brunswick. <PER> also attended the <ORG> for a short time.	140	3
member	Vance Plauche - American Legion Merlin Olsen - Phi Beta Kappa	<PER> was * member of <ORG> <PER> * member of * <ORG> and * in	<PER> was also a member of the <ORG>. <PER> is a member of Sigma Chi fraternity and <ORG > and was a letterman in football as a defensive tackle.	140	0

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 39-53, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.4

Topics and Trends in Metadata Research

Jung Sun Oh

School of Information and Library Science, University of North Carolina at Chapel Hill, NC, USA
E-mail: ohjungsun@gmail.com

Ok Nam Park*

Department of Library and Information Science, Sangmyung University, Seoul, Korea
E-mail: ponda@smu.ac.kr

ABSTRACT

While the body of research on metadata has grown substantially, there has been a lack of systematic analysis of the field of metadata. In this study, we attempt to fill this gap by examining metadata literature spanning the past 20 years. With the combination of a text mining technique, topic modeling, and network analysis, we analyzed 2,713 scholarly papers on metadata published between 1995 and 2014 and identified main topics and trends in metadata research. As the result of topic modeling, 20 topics were discovered and, among those, the most prominent topics were reviewed in detail. In addition, the changes over time in the topic composition, in terms of both the relative topic proportions and the structure of topic networks, were traced to find past and emerging trends in research. The results show that a number of core themes in metadata research have been established over the past decades and the field has advanced, embracing and responding to the dynamic changes in information environments as well as new developments in the professional field.

Keywords: topic modeling, metadata research, research trends, library and information science

Open Access

Accepted date: July 09, 2018
Received date: December 07, 2017

***Corresponding Author:** Ok Nam Park
Associate Professor
Department of Library and Information Science, Sangmyung University, 20 Hongjimun 2-gil, Jongno-gu, Seoul 03016, Korea
E-mail: ponda@smu.ac.kr

1. INTRODUCTION

Metadata lies at the intersection of multiple core areas of information science including knowledge organization and information retrieval. While having its root in traditional bibliographic control in libraries, the area has grown substantially along with the evolution and expansion of the Internet, encompassing principles and practices of resource description for both digital and non-digital materials.

The proliferation of distributed information repositories on the web brought about the need for standardized mechanisms for describing resources, which led to the developments of an array of metadata standards since the 1990s, including the Dublin Core Metadata Element Set (DCMES or DC), Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), and Learning Object Metadata (LOM), to name just a few. Research addressing various issues related to the creation and use of metadata started to appear in scholarly publications in the mid-1990s, and since then the body of literature has grown substantially. Recently, there has been a new wave of advances in web technologies, including Linked Data, which extends the horizon for metadata research even further. Zeng and Qin (2016) noted, in one of the well-known textbooks on the subject of metadata, that "the last two decades of metadata development have witnessed a continual expansion and evolution of metadata research and practices at almost all levels and in almost all disciplines" (p. 18). Yet, to our best knowledge, there has been a lack of systematic analysis of the field of metadata.

In this study, we attempt to examine the body of literature on metadata and address questions regarding the development of the field. While metadata research and practice grow to span multiple disciplines and various areas of interest, as a starting point we set out to trace the growth of the field within library and information science (LIS). More specifically, we apply a text mining method, topic modeling, to the research papers on metadata in LIS literature for the past 20 years to discover main topics addressed in these papers, and the trends in those topics over time. What are the main themes/topics in the discussion of metadata? How has the field changed over time? What topics have gained increasing attention, and what topics have declined over the years? How are these topics interrelated and how have the relationships evolved as individual topics developed? These are questions that we intend to explore. In doing so, we present the topic modeling method combined with network analysis as a promising way for identifying major research topics and studying research trends in literature.

In the following, we will first review the topic modeling method in general and then discuss specifics of our methods. In the result section, the topics identified as a result of topic modeling and our interpretation of the modeling outputs will be presented first, and the analysis of research trends and topic networks will follow. Lastly, the main findings of the study will be discussed to conclude the paper.

2. TOPIC MODELING

Topic modeling has attracted much attention over the past ten years as a tool for computational analysis of large document collections. Based on a probabilistic model, topic modeling uncovers latent 'topics' in a collection of documents or a text corpus (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). It starts from an assumption that each document contains a mixture of topics, and the words in the document reflect those topics. The modeling algorithm then infers the hidden topics in a document collection using the observable data—documents and words therein—through unsupervised learning. A topic is represented as a semantically related cluster of words that are likely to appear together in text discussing the topic, with each word having different probabilities of occurrence with regard to the topic. A document, with the occurrences of words associated with different topics, can in turn be abstracted as a probabilistic mixture of those topics. In this way it is possible not only to discover topics addressed in a collection as a whole, but also to figure out what topics appear in which proportions in each document in the collection.

Topic modeling is in fact a label for a family of probabilistic learning algorithms for discovering topics from text corpora. The first and most common method, called Latent Dirichlet Allocation (LDA), was introduced in 2003 in the seminal paper of Blei, Ng, and Jordan (2003). LDA is a generative probabilistic model for a collection of documents, based on the joint probability distribution of the hidden variables (topics) and the observed variables (words in documents). Given a pre-specified number of topics K and a collection of documents containing a fixed set of words (a vocabulary) V, LDA computes the conditional distribution of the underlying topic structure and derives K topics, each as a multinomial distribution over the vocabulary. At the same time, LDA delivers topic assignments for documents, with each document being described as a multinomial distribution over topics (Blei, 2012).

There are several advantages of using topic modeling for the analysis of a large collection of documents. First, as an unsupervised learning technique, it requires no intervention or supervision during the modeling process once the input parameters, including the number of topics, are set. Therefore, using a tool for topic modeling is relatively simple and does not require sophisticated computational skills. The analyst may adjust the input parameters to, for instance, get a more or less find-grained result, but the rest is taken care of by the tool. Second, the model output is readily interpretable by human analysts. The discovered topics can be presented as a list of words with weights denoting the relative importance of each term in describing the topic. Moreover, an examination of the words associated with a topic together with the documents representative of the topic (i.e., documents of which a large proportion is allocated to the topic) helps clarify the meaning of the topics and verify the results. Third, the topics are known to be robust against the inherent ambiguity in language (e.g., synonymy, polysemy) as the model's reliance on word co-occurrence in effect sorts out different contexts in which a word appears. For instance, the term 'library' may appear in two different topics in a collection—in one topic it appears along with terms like 'catalog,' 'monograph,' or 'lending' while in another topic it comes next to 'java,' 'programing,' or 'functions.' The different meanings of the term in these two topics are apparent thanks to the other associated terms. Finally, one of the primary advantages of topic modeling is in its representation of documents as a mixture of multiple topics, which provides a 'soft' clustering or classification of documents. This sets this method apart from other clustering techniques such as K-Means where a document belongs to a single class. Instead of assigning a single topic to a document, topic modeling finds the proportions of multiple topics (the proportions of words associated with those topics) for each document. This more realistic and flexible representation of documents allows further exploration of relationships between topics and documents (Mimno, McCallum, & Mann, 2006).

Over the past decade, many studies show empirically that topic modeling discovers a semantically meaningful set of topics as well as inducing a sensible decomposition of individual documents in terms of those topics. Among others, those studies applying the method to discover topics in research publications reported its usefulness in finding subfields or topical divisions of a research field. Moreover, it was shown that, using the quantitative measures of topic proportions in research literature, it is possible to track changes in the relative prevalence of topics over time, and thereby trace the overall progression of a research field as well (Griffiths & Steyvers, 2004; Hall, Jurafsky, & Manning, 2008; Daud, 2012).

3. LITERATURE REVIEW

Studies related to research trends have already been conducted in several areas of LIS. This study reviewed previous studies regarding the methodology used in research efforts, research trends in knowledge organization, and research trends by means of topic modeling.

3.1. Methodology Used in Research Trends Analysis

The research methods used to analyze trends in research disciplines are distinguished largely by bibliometrics, content analysis, and social network analysis.

3.1.1. Bibliometrics

Bibliometrics is a methodology to apply quantitative methods for literature analysis, mainly quoted by utilizing the index method. It identifies the most highly cited journals in the areas of research, discipline, author, and author cooperation. Patra, Bhattacharya, and Verma (2006), who leveraged the Library and Information Science Abstracts (LISA) to investigate the tendency for bibliometrics literature, described the mid-range of the relevant literature, the language of literature, and authorship patterns. Blessinger and Hrycaj (2010) analyzed the 10 top Journal Citation Ranking (JCR) journals by impact factors to analyze trends in the LIS field of study. They examined 2,200 articles published from 1996 to 2004 and identified the most highly cited journals, articles, and subject areas.

3.1.2. Content Analysis

Content analysis is also a methodology to interpret documents by text analysis. It analyzes the structure, intention, and characteristics that appear in the text, and can be carried out by quantitative or qualitative methods. Shiri (2003) analyzed articles published from three conferences in 2012 to identify research trends in digital library areas, and the study found standards, architecture, usability, issues, and digital content used as focal issues in digital library areas. Julien, Pecoskie, and Reed (2011) conducted a content analysis to analyze trends in information behavior research. They analyzed 749 articles on information behavior published from 1999 to 2008 according to authorship, types of article, journal type, theoretical framework, user groups, degree of attention to users' cognitive processes,

and interdisciplinarity. They found there was an increase in interdisciplinarity in information behavior. Greifeneder (2014) studied 155 articles that were written in the field of information behavior between 2012 and 2014. They employed publication title, authors, publication years, methods, and topics, and main research topic, and identified information seeking as the main research topics, and qualitative methods as the primary methodology.

3.1.3. Social Network Analysis

Social network analysis (SNA) is a methodology to analyze and visualize the network characteristics of group, organization, and data objects. SNA evaluates the network focusing on frequency of keywords, network size, network centrality, and density. Cho (2013) studied articles published in the Republic of Korea and Japan between 2010 and 2012 that were focused on field knowledge organization. The frequency of keywords and network map of the main keywords were analyzed. Feicheng and Yating (2014) utilized SNA to investigate the co-occurrence of tags. They studied tags from the CiteULike and found centricity and groups of tags. They propose that SNA of online tags can be employed as a visualization tool and recommendation resources.

3.2. Research Trends in Knowledge Organization

Studies to investigate research trends regarding knowledge organization have not been carried out to any great extent. Pattuelli (2010) analyzed 34 courses related to knowledge organization in LIS schools in the United States, and outlined the topics and readings taught in the courses. Cho (2013) examined the knowledge organization literature in Japan and the Republic of Korea by network analysis. In addition, Hunter (2003) studied metadata research trends by survey, and found XML semantic web, metadata harvesting, web services, and so on as the main research areas. Parlmer, Zavalina, and Mustafoff (2007) performed an analysis of Institute of Museum and Library Services digital collections projects conducted in 2003 and 2008. They surveyed the project managers and gained an understanding of the metadata audience, metadata application, decision factors, and problems for metadata schema development.

3.3. Topic Modeling Utilization in Research Trends Analysis

Topic modeling has not been much employed in research trend analysis. Most studies have utilized the literature as a dataset to investigate the applicability of topic modeling algorithms.

Griffiths and Steyvers (2004) analyzed abstracts from *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) published from 1991 to 2001. They employed LDA and a Markov chain Monte Carlo algorithm to infer about topic modeling. Topics that continue to decrease and increase in the dataset were presented along with topic related terms. Mimno and McCallum (2008) investigated 300,000 articles related to artificial intelligence. They argued that topic modeling can be usefully applied to discover main authors, topics, and predict lead authors for topics. Hall et al. (2008) analyzed 12,500 papers from a journal of the Association for Computational Linguistics Anthology. He employed the LDA technique and discovered hot topics that have been emphasized in anthology, cold topics, and the decline of the leading conference in the field. Daud (2012) carried out topic modeling analysis based on DataBase systems and Logic Programming dataset as well as temporal analysis by year for the main topic, and determined the key authors, key topics, and changes in key topics in the literature.

Previous studies have contained a number of flaws. Studies to understand research trends have been conducted in various areas but there is little specific literature on the research trends of metadata, and studies have relied on limited methodology such as surveys and descriptive content analysis. To unlock key areas, changes in key research areas, and interlinking among research areas, more intensive studies of metadata research need to be done. Looking at the research on topic modeling performed to date, topic modeling has employed literature as data set. Through this, it can be seen that topic modeling can be used to analyze research tendencies.

4. METHODS

4.1. Dataset

In order to investigate topics addressed related to metadata in LIS literature, we constructed our dataset by searching three databases commonly used in the LIS field—LISA, Library and Information Science Source (LISS), and Library, Information Science & Technology Abstracts (LISTA). The three databases have comparable search features, allowing us to construct the same or equitable queries. The searches were done in February 2015. From each database, we retrieved all the records for peer-reviewed scholarly papers written in English that have the term 'metadata' in the title, keyword, or subject fields. The initial dataset had a total of 5,473 records including 1,059 records from LISA, 2,065 records from LISTA, and 2,349 records from LISS. From

the initial dataset, we removed those records that did not contain an abstract, either having an empty abstract field or having text other than an actual abstract (e.g., copyright notes, accession number, etc.) in the abstract field. Since the three databases overlap in their coverage of LIS literature, there were numerous duplicate records in the initial set. We detected duplicates by comparing both title and abstract and removing them. We also removed five papers published before 1995 and 19 papers published in January/February, 2015. The resulting dataset included 2,713 records published in 370 journals between 1995 and 2014. Fig. 1 shows the number of papers in the final dataset by year.

As can be seen in the table, the research literature on metadata grew steeply in the first ten years, from only a handful of papers per year in the mid-1990s to almost 200 papers in 2005. The number of yearly publications has stabilized around 200 since then. The growth can also be illustrated by the number of journals represented in our dataset. Until the mid-1990s the papers on metadata appeared in only a few journals, but drastic increases occur in the last few years of the 1990s and early 2000s. For instance, between the year 2000 and 2001, the number of journals in our dataset rises from 24 to 41, coinciding with a notable increase of papers in 2001.

4.2. Topic Modeling

4.2.1. Preprocessing

The combined title and abstract from each record in our dataset constituted a single document for topic modeling. The documents were preprocessed prior to the topic modeling. After removing punctuation, numbers, and special characters, we deleted words belonging to a stop-word list or with a length less than 3 letters. Since most, if not all, documents contain the word 'metadata,' we also decided to remove the term as it has little value in differentiating topics. For the remaining words (tokens), we performed stemming. All data processing and analysis were done in the *R* computing environment (R Core Team, 2014). The *tm* package (Feinerer & Hornik, 2014) was used to preprocess the documents.

4.2.2. LDA Topic Modeling

For topic modeling, we applied LDA to the preprocessed corpus using the *mallet* package (Mimno & McCallum, 2008), which is an *R* wrapper for the Java-based software MALLET (McCallum, 2002), in *conjunction* with the *topicmodels* package (Grün & Hornik, 2011) in *R*.

As mentioned before, LDA requires a specification of the number of topics *K* as an input for running the algorithm. Considering that the actual number of topics or organizing themes in a corpus is usually unknown, finding the optimal number of topics to be discovered is an important and challenging part of the analysis. Often a trial-and-error approach is taken in practice, trying different *K* values (different number of topics) and settling with a number that leads to the most meaningful results. In our analysis, we

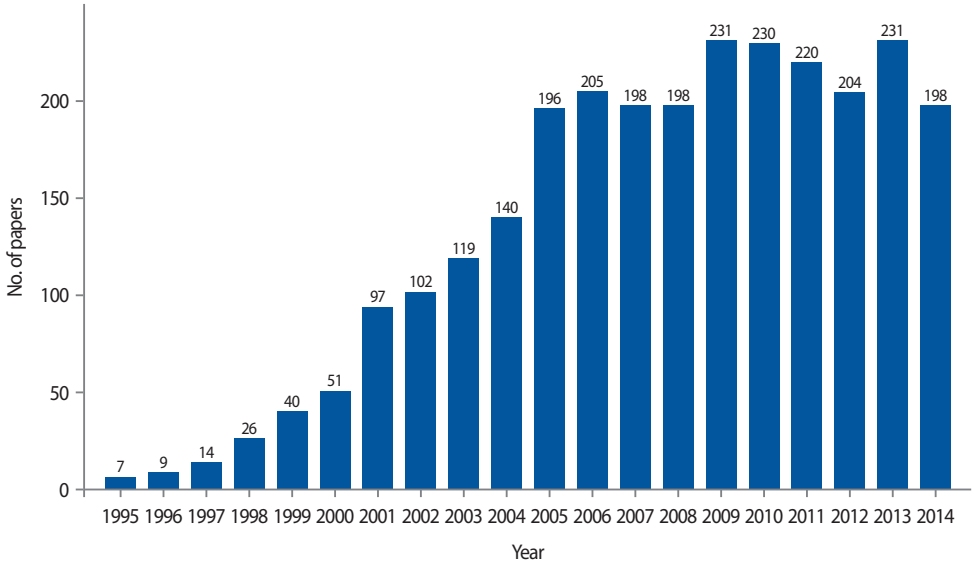


Fig. 1. The final number of papers by years removing duplicates.

tried *K* values of 20, 30, and 40 and chose 20 topics. In the subsequent analysis, we used the model output taken after 5,000 iterations of Gibbs sampling, with *K*=20.

4.2.3. Interpretation

In using topic modeling, assessing the quality of latent topics and interpreting the results is the key analytic task. In interpreting each topic derived from topic modeling, we examined 1) the most probable words (i.e., the words with a high probability of being associated with the topic), 2) the words more distinctive to the topic, and 3) the most representative papers addressing the topic (i.e., the papers where the vast majority or a significantly large proportion of its content words belongs to the topic). A label was then assigned to each topic.

The difference between the probable words and the distinctive words (1 and 2 above, respectively) related to a topic is explained in the following. As mentioned before, the outcome of topic modeling includes the distributions of words over topics. Therefore, a ranked list of words with a high probability of appearing in a topic can be easily created, and in fact, this list is most commonly used for topic interpretation. However, since the probability is affected by the overall frequencies of words, it is often the case that some generic words or common jargon in the domain take place in the top ranked list for multiple topics, making it difficult to differentiate the meanings of these topics. In order to solve this problem, different measures for identifying and ranking topic words were suggested. Among others, we adopt Sievert and Shirley (2014)'s measure of 'relevance' defined as relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w). This measure considers how uniquely or distinctively a word is associated with a topic, and thereby reduces the weight of those generic words occurring frequently in many topics while increasing the weight of most 'relevant' terms for a given topic. We examined the top thirty probable terms and top thirty relevant terms (at λ=0.6) for our interpretation of topics. In addition, as mentioned above, a number of documents where a given topic appears dominantly are also reviewed to verify the topic contents.

4.3. Research Trends and Topic Networks

In addition to finding out metadata-related topics addressed in research literature, we are interested in how these topic areas have evolved over time. As in Griffiths and Steyvers (2004), we conducted a post hoc analysis on the topic modeling outcomes to track changes in topic distributions over time. More specifically, the proportion of topics assigned to documents were aggregated by the publication year of the documents, and each topic's share in yearly publications was plotted to uncover any trends.

In order to look at the relationships between topics and the dynamics of the relationships over time, we employed network analyses of topics, using the topic distribution over documents. In topic modeling, each document is represented as a mixture of topics. In some cases, a document has a single prevalent topic with other topics having marginal proportions. In some other cases, a document comprises multiple topics each with a non-trivial proportion. We suppose that, if two topics are frequently addressed together in substantial proportions in the same documents, it indicates possible relationships between those topics. That is, relationships between topics are established based on the documents in which the given topics appear together.

5. RESULTS

As a result of performing the topic modeling of datasets, twenty latent topics were discovered. Table 1 shows the twenty topics with assigned labels, in the order of their prevalence in the entire dataset, along with the most probable words next to each topic.

As explained in the method section, in order to interpret the topic content and assign the label for each topic, we examined 1) the most probable words, 2) the most relevant words (words that are most distinctive to the given topic), and 3) the representative papers of the topic. Due to the space limitation, Table 1 includes only seven most probable terms for each topic. We will mention any notable differences between probable words and relevant words in our discussion of derived topics in the following.

Note that the topic modeling outcome included words in their stemmed form since we did stemming of words in the preprocessing phase, but for better readability, stems were changed back to complete words in Table 1. Note also that we removed the term 'metadata' before topic modeling because most, if not all, of the documents in our collection include the term, and therefore it does not have a value for identifying distinct topics in the collection. However, when we interpret the results presented in Table 1, it would be reasonable to presume that the top terms may be used in conjunction with the term 'metadata' or in a broad context of metadata related issues. Therefore, we used the term 'metadata' in topic labels where it seems appropriate.

Table 1. Discovered topics

No	Topic name	Probable terms
1	Digital library projects	digital, collection, library, project, preservation, archive, access
2	Development or management of information systems/services	system, manage, service, base, develop, implement, integrate
3	Role of metadata or metadata librarians	library, resource, librarian, service, develop, technology, digital
4	Evaluation of metadata quality	study, paper, quality, analysis, research, result, data
5	Semantic web and ontology	semantic, web, ontology, model, knowledge, base, paper
6	Metadata standard development	project, standard, develop, resource, access, nation, work
7	Record management	manage, record, paper, system, research, knowledge, context
8	Cataloging	catalog, library, record, bibliographic, catalogue, author, marc
9	Social tag and folksonomy	tag, user, social, subject, term, index, folksonomy
10	Automatic extraction methods	document, base, method, extract, automatic, system, index
11	Search and retrieval	search, user, retrieve, image, music, query, result
12	Linked Data	data, link, research, map, science, scientific, geographic
13	Metadata harvesting	repository, institute, open, oai, harvest, research, protocol
14	Publishing and access	publish, article, journal, book, access, public, scholar
15	XML and encoding standards	standard, description, xml, archive, encode, article, document
16	Dublin Core	core, element, dublin, resource, standard, set, develop
17	Search engine and web sites	web, site, page, engine, meta, search, description
18	Domain metadata: education and heath	learn, education, object, student, health, resource, medical
19	Metadata for multimedia and social media	content, network, media, user, video, social, multimedia
20	Conference and meeting reports	library, conference, present, report, association, meet, discuss

5.1. Prominent Topics

In this section, we review topics 1 to 8 in detail, as they turned out to be the most prominent topics in metadata literature. Collectively, these topics account for 57.6 % of entire words (tokens) in the corpus, with each topic assuming more than 5%.

Topic 1 is about digital collection or digital library projects, as can be seen in the words falling in this topic with high probability. Included in the top third probable or relevant words are a set of words referring to the organizations in charge of such projects, including *library*, *archive*, *cultural*, *heritage*, and *institution*, as well as a group of words for information objects, such as *object*, *material*, and *image*. In addition, the words related to some key functions of metadata, such as *describe*, *access*, or *preservation*, also appeared high in the lists. The words associated with this topic occupy more than 10% of the corpus, making it the most prominent topic in the metadata literature in our dataset.

As expected, many of the representative papers of this topic report and share the experiences of building a digital collection or setting up a digital library, often discussing metadata related issues or challenges encountered in the process (e.g., Boyd & King, 2006; Woodley, 2002). There are also papers addressing specific aspects of a project, such as the implications of technical choices for digitization or the need for documenting metadata decisions (e.g., Lalitha, 2009; Symonds & May, 2009).

Topic 2 concerns the development or management of information systems/services. The main keywords for this topic include nouns such as *system*, *service*, *software*, and *application* as well as verbs like *develop*, *manage*, *implement*, and *integrate*. Also ranked high on the list of probable/relevant words are *design*, *architecture*, *model*, and *framework*. Combinations of these terms give a fairly good idea as to the topic area.

The papers in this topic discuss metadata based approaches or solutions for specific problems at hand in relation to information systems or services: for instance, access control (Yagüe, Maña, & Lopez, 2005), content management (Yeh, Chen, Sie, & Liu, 2014), or a federation of distributed resources (Aktas, Fox, & Pierce, 2010). Issues concerning the design and implementation of such solutions, or the proposed models or architectures, are commonly found in those papers.

Topic 3 deals with role of metadata or metadata librarians. The most probable words for this topic are *library*, *resource*,

librarian, *service*, *develop*, *technology*, and *digital*, but this topic can be better understood when the relevant words distinctive to this topic, such as *role*, *skill*, *future*, and *profession* are considered.

The papers in this topic appear to divide broadly into two groups, while commonly noting the challenges that the proliferation of electronic resources have brought to library services. One group tackles the problem of organizing electronic resources and discusses the increasingly important role of metadata (e.g., Medeiros, 2003; Emery, 2007). The other group centers on the discussion of professional roles of librarians in the digital era, reflecting on the influence of technologies on library and information services. Many state the need for reconfiguring technical services or cataloging practices to better meet current and future challenges, and call for attention to the changing roles and competencies of librarians (e.g., Schottlaender, 2003; Han & Hswe, 2011).

Topic 4 covers evaluation of metadata quality. The top ten words with the highest probabilities of being associated with this topic are *study*, *paper*, *quality*, *analysis*, *research*, *result*, *data*, *find*, *survey*, and *evaluate*. The rank order of relevant words differs slightly, with *quality*, *analysis*, *survey*, and *evaluate* placed higher.

The most representative papers of this topic report the results of studies on metadata quality. Some analyze a number of metadata records and identify patterns of problems or errors therein. Often certain criteria such as accuracy, consistency, and completeness are used to evaluate the quality of metadata (e.g., Chuttur, 2012). Some further suggest and/or test mechanisms for quality assurance (e.g., Park & Tosaka, 2010; Chuttur, 2014). Many papers present empirical studies where a variety of methods including experiments, focus group interviews, and surveys are employed. Yet, some discuss quality issues based on a review and an analysis of research and practice in the field (e.g., Park, 2009).

Topic 5 is focused on semantic web and ontology. Top keywords for this topic are *semantic*, *web*, *ontology*, *model*, *knowledge*, *base*, *paper*, *relation*, *concept*, *structure*, *domain*, and so on. As can be seen in the above list of keywords, where all except for one rather general term *paper* represent a coherent theme, there is little to no ambiguity about this topic area.

The papers in this topic review various semantic web technologies (e.g., Kanellopoulos & Kotsiantis, 2007), or discuss ontology modeling or implementation, including a conversion of existing controlled vocabulary or metadata (e.g., Qin & Paling, 2001).

Topic 6 is about metadata standard development. The first glance at the list of probable words in this topic does not give a clear idea as to its topic content, as it includes rather generic terms that appear in multiple topics, such as *project*, *develop*, *resource*, and *access*. Only the term *standard* is relatively unique to this topic. However, when we examine further down the list of the terms relevant to this topic, along with its representative papers, it becomes evident that this topic centers on metadata standards, especially the development of various national or international standards. The names of standards organizations, such as *NISO* and *ISO*, are often mentioned in the papers and related terms like *initiative*, *committee*, (*working*) *group*, or *programme* appear high on the list of relevant terms.

Some papers in this group discuss the importance of developing and adopting metadata standards (Lagace, Breeding, Romano Reynolds, & Han, 2013), some introduce a newly developed standard or provide updates on a standard under development (e.g., Feick, Henderson, & England, 2011), and some provide a review of an existing standard, often with a discussion of emerging issues (e.g., Mullen, 2001).

Topic 7 mainly concerns record management. The top keywords in this topic include *manage*, *record*, *paper*, and *system*. Additional words particularly relevant to this topic are *recordkeeping*, *context*, and *scheme*.

A variety of questions regarding metadata for record management or record keeping have been addressed, including the role, purpose, or capacities of metadata in the context of electronic record management, the specifications for records metadata, the methods for acquiring or capturing record keeping metadata, the need for standards and tools, and so on (e.g., Evans & Rouche, 2004; Evans, 2007; Cumming, 2007).

Topic 8 is related to cataloging or bibliographic description of resources, as clearly shown in the lists of probable and relevant topic words including *catalog*(*ue*), *record*, *library*, *bibliographic*, *MARC*, *RDA*, *access*, *control*, *description*, *resource*, and so on.

Although sharing the central theme, cataloging, research problems addressed in the papers range broadly from the pertinence of specific aspects of cataloging rules and/or principles in modern catalogs (e.g., Conners, 2008), to the applicability of cataloging tools to organization of Internet resources (e.g., Ferris, 2002), and to future directions for library catalog or cataloging rules (e.g., Wakimoto, 2009). Many of the recent papers report testing and implementation of the new cataloging standard, Resource Description and Access (RDA) (e.g., Danskin, 2014).

5.2. Research Trends

The topics of prominence we looked at in the previous section are based on the proportion of words associated with a given topic within the entire corpus. That is, the result shows the distribution of the twenty topics across all the papers in the corpus that are published in the span of 20 years. In order to see the dynamics of these topics, we now turn our attention to the changes in the topic proportions in the corpus over time.

Since topics are assigned to each document with their respective proportions, when we aggregate the proportions of a given topic in the documents published in a year, we can obtain the share of the topic in that year's literature. This provides quantitative measures of rise and/or fall of topics in popularity over time, as well as their relative prevalence.

For the overall trend analysis, we plotted topic shares from 2000 to 2014. The publications from 1995 to 1999 were not included in the plots, since the number of publications during the period was too small for this analysis. The plots for all twenty topics were first drawn, and among them, we identified topics with an upward trend, a downward trend, and those reaching a peak at different points in time. In the following, we present those topics.

5.2.1. Emerging Topics

Topic 4 (evaluation of metadata quality), Topic 10 (document extraction methods), Topic 12 (Linked Data) have increased in volume since 2000 as shown in Fig. 2.

Topic 4 (evaluation of metadata quality) shows a steep increase from 2000 to 2010, with its topic share being more than tripled. This trend demonstrates that growing attention has been given to quality issues as metadata has become a pillar of information services.

Topic 10 (document extraction methods) is related to automatic extraction of metadata from documents. Noting both the need for and the expenses of creating metadata, the papers in this topic explore various strategies and methods for automatically generating metadata using different parts of documents per se. This topic shows a good deal of fluctuation, yet the overall topic share has increased considerably over time.

The papers associated with Topic 12 (Linked Data) show a rapid increase since 2006, the year when Tim Berners-Lee coined the term and outlined the key principles for publishing and connecting structured data on the web. Also included in this topic are papers on data mining, or on curation or retrieval of special data including scientific data and geographic data. This explains why the topic was present before 2006 and has a spike in 2002. The vast majority of the representative papers (24 out of the top 30 papers) in this topic, however, were published after 2006, many touching on the application or adoption of Linked Data concepts, technologies, and practices to the creation, transformation, and use of Libraries, Archives, Museums (LAM) metadata.

5.2.2. Declining Topics

Interestingly, topics showing a downward trend (Fig. 3) are related to metadata standards: topic 6 (standard development), topic 15 (XML and encoding standards), and topic 16 (Dublin Core). It appears that these topics attracted the most interest at the early stage of metadata research, before or around the beginning of the 2000s, but began to fade at least from the research front.

As described in the previous section, topic 6 has to do with the development of metadata standards, and many

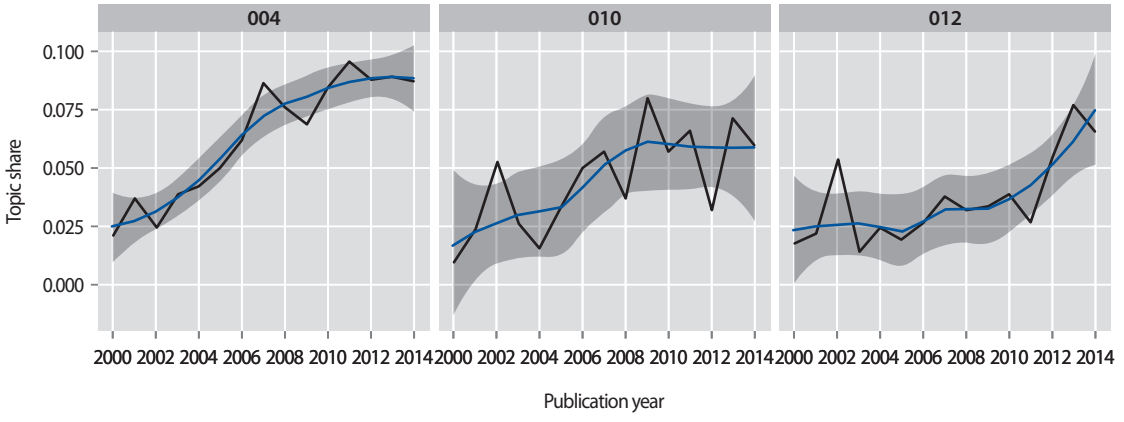


Fig. 2. Emerging topics showing upward trends.

papers in this topic introduced a then new standard or reported updates on the development or implementation of a standard.

Topic 15 is about XML and encoding standards. There is a surge of interest in this topic in the metadata community from 2000 to 2002, demonstrated by a special section on XML in *Library Hi Tech* (volume 19 in 2001). Many of the papers in this period provide an introduction to the suite of XML specifications and technologies, along with a discussion of their implications for metadata sharing. In addition, the release of the XML-compliant EAD version 1 in 1998 seems to kindle the interests of archivists in this topic, resulting in a series of papers on XML encoding of archival resources in following years.

Topic 16 about Dublin Core has declined steadily. As one of the first metadata standards that received international recognition and enjoyed wide adoption, Dublin Core took a central place in the early phase of metadata research, but has given its share to other emerging topics as the research horizon expands. In fact, this topic reached its peak in 1997 (not shown in the plot), where its topic share amounted to 16.8% of the entire corpus. The topic share continued to decrease to 8% in 2000, 3.8% in 2006, and finally to 2% in 2014.

5.2.3. Topics with Peaks and Valleys

The dynamic changes in the area are also observed in a set of topics that reached their peak (or valley) in different points in time at different pace (Fig. 4). Topic 9 (social tag and folksonomy) and topic 13 (metadata harvesting) gained sizable attention at one point then have gone down, and topic 8 (cataloging) showed an opposite pattern.

Topic 9 is about social tag and folksonomy. The steep increase of this topic in literature is closely related to the development of social bookmarking or tagging tools. Del.icio.us was founded in 2003, Flickr in 2004, and Ma.gnolia

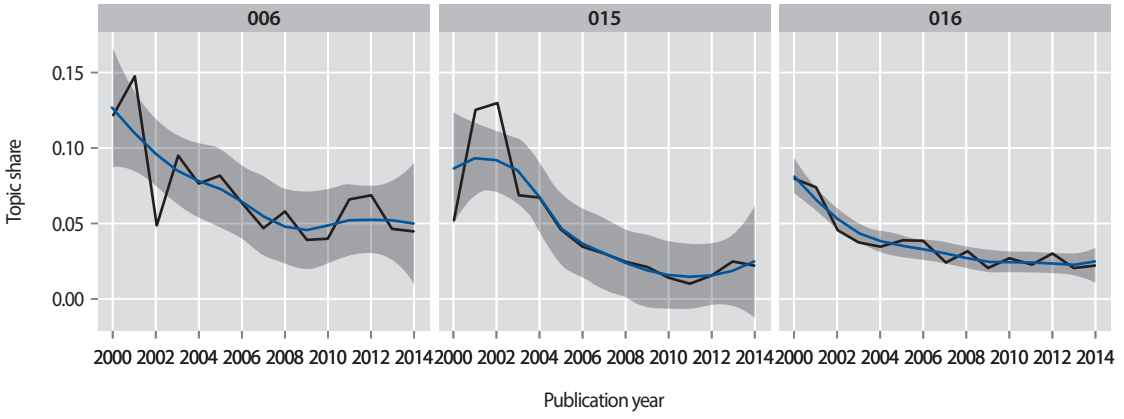


Fig. 3. Declining topics showing downward trends.

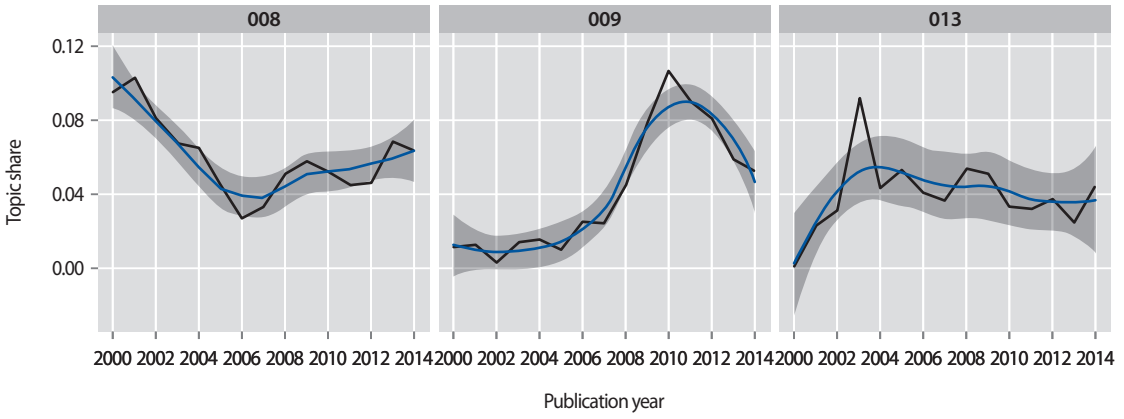


Fig. 4. Topics with peaks and valleys showing dynamic changes.

in 2006. The idea of building up a bottom-up taxonomy, dubbed as *folksonomy*, using user-generated tags was booming as those tools gained enormous popularity. Many libraries have also started incorporating tagging features into their catalogs. The published work on this topic peaked in 2010 but has rapidly decreased since then, following the downturn of tagging services in general.

Topic 13 represents a coherent set of discussions on metadata harvesting and issues related to constructing and maintaining metadata repositories. The research on metadata harvesting took off shortly after the introduction of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 1.0 in early 2001, and the sudden burst of this topic between 2002 and 2003 coincides with the release of OAI-PMH version 2.0 specification. While the high level of attention to this topic per se in literature did not sustain itself after its peak in 2003, the breath of discussions appear to have expanded, as shown in its increased relationships with other topics (presented in the next section).

Topic 8 is an interesting case. It showed a clear downward trend until 2006, but bucked up the trend since then. This change appears to be a response to the substantial developments undergoing in the cataloging field, including the work on the new cataloging rule, RDA, and the continuing discussion on improving or substituting Machine Readable Cataloging (MARC).

5.3. Topic Networks

Having identified the main themes of metadata discourse and having looked at the changes in prominence of such topics over time, we now are interested in how these topics are interrelated and how the strengths of the connections between topics have changed during the fifteen year period.

As described in the method section, in order to look into the relationships between topics, we adopted a network analytic method. Using topic proportions in each paper, we derived connections between topics based on the papers addressing two or more topics together. In order to determine those papers spanning multiple topics, we set the threshold at 30%. That is, if a paper's topic composition consists of two or more topics each in a proportion of 30% or more, the paper constitutes a potential link between the topics. Between 2000 and 2014, about 32% of the papers (838 out of 2,617) fell under this category.

Fig. 5 shows the topic network considering all papers published between 2000 and 2014. The size of a node is proportional to its degree, and the thickness of a link between two nodes denotes the strength of their connections

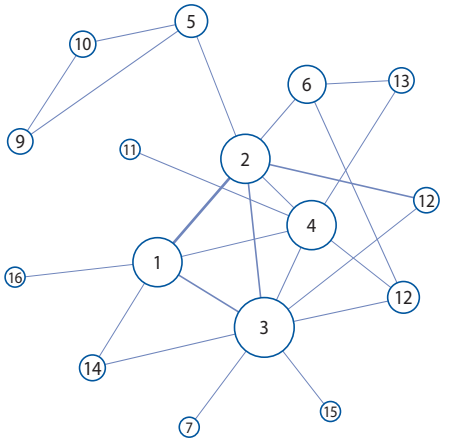


Fig. 5. Topic network considering all papers published between 2000 and 2014.

determined by the number of papers addressing the two together. Each link in this network has a minimum value of 10, which means ten or more papers addressed the topics on each side of the link together. The isolated nodes, topics with no connections, are removed from the figure.

Given the condition of ten or more shared documents to establish connections, 16 out of 20 topics have at least one connection with other topic(s). The node with the highest degree, the most well-connected topic, is topic 3, having links to eight other topics. This is not surprising since the role of metadata or metadata librarians is a topic that can be discussed in various contexts. Most of the prominent topics assuming a large proportion in the corpus (topics 1 through 8) have a relatively high degree ranging from three to eight, except for topic 7 (record management). Topics 1 through 4 clearly constitute the tightly-knit core of the network, to which smaller topics are connected with varying strengths. It is notable that the more technically oriented topics, topics 5, 9, and 10, form a clique somewhat separated from the core.

In order to look at how the relationships between topics have evolved over time, as well as to get more insights into their composition, we divide the 15-year period into three sub-periods (period 1: 2000–2004, period 2: 2005–2009, period 3: 2010–2014) and draw a topic network for each sub-period. Considering the relative sizes of the document sets in these periods, we adjusted the number of shared documents to create links between topics—a minimum of four shared documents was used for period 1, and a minimum of five for period 2 and period 3. Fig. 6 presents the three sub-period networks.

As can be seen in Fig. 6, the networks show considerable differences not only in the volume of connections but also

in their structure. Overall, it is clear that increasingly more connections among topics have arisen as time passes, but at the same time there are notable changes in topics assuming central positions.

In the first five year period (2000–2004), the most noticeable difference compared to subsequent periods is the prominence of topic 6 (metadata standard development) and topic 15 (XML and encoding standards) both in terms of their degree and the strengths of connections that they have with other topics. Topic 16 (Dublin Core) is present only in period 1, being connected to topic 6 and topic 15. Note that these topics all showed a downtrend from around 2000, as described in the previous section. The fact that they remained salient in the network suggests that, while declining in volume, these topics relating to base standards still had an important place in discussion of other topics. These topics, however, become peripheral in subsequent periods.

In period 2, topics 1 (digital library projects) and topic 2 (development or management of information systems/services) moved to the center of the network. Topic 1's central position, with its links to a variety of topics, indicates that various metadata research topics were often introduced and discussed in a context of digital projects. Topic 2, on the other hand, showed a tendency of having connections with more tech-oriented topics, including topic 5 (semantic web and ontology) and topic 12 (Linked Data). It is also notable that topic 4 (evaluation of metadata quality) started to appear on the network in this period.

In period 3, topic 3 (role of metadata or metadata librarians) emerged as a center of the network, with a degree of nine. It indicates that a discussion of the function of metadata or the professional role of metadata librarians takes place in a variety of topics in more recent literature. In addition, topic 12 (Linked Data) and topic 14 (publishing and access) show a considerable growth in links, reflecting new trends in research.

The composition of networks and the changes therein portray how topics develop in relation to other topics. For instance, looking at topic 13 (metadata harvesting), the topic was first connected to topic 6 in period 1, topics 1 and 2 in period 2, and finally topics 3 and 4 in period 3. That is, papers introducing the ideas and mechanisms of metadata harvesting, including the OAI protocol standard itself, first formed a link between topic 13 and topic 6. Then papers describing a project creating regional or national repositories using OAI-PHM and papers concentrating on design and implementation of services building upon metadata repositories appeared, connecting this topic to topic 1 and topic 2, respectively. Finally, papers focusing on the evaluation of consistency of metadata elements and values across harvested records constitute the link between topic 13 and topic 4 in period 3. In addition, discussions on academic libraries' role in managing and promoting OAI compliant institutional repositories spanned both topic 13 and topic 3.

6. DISCUSSION AND CONCLUSION

In this study, we collected 20 years of metadata literature, and identified and labeled 20 topics present in metadata literature. We found a set of sensible and coherent research topics from the literature spanning 20 years, and also traced a number of research trends. Among those, more prominent topics, in terms of their proportion in literature,

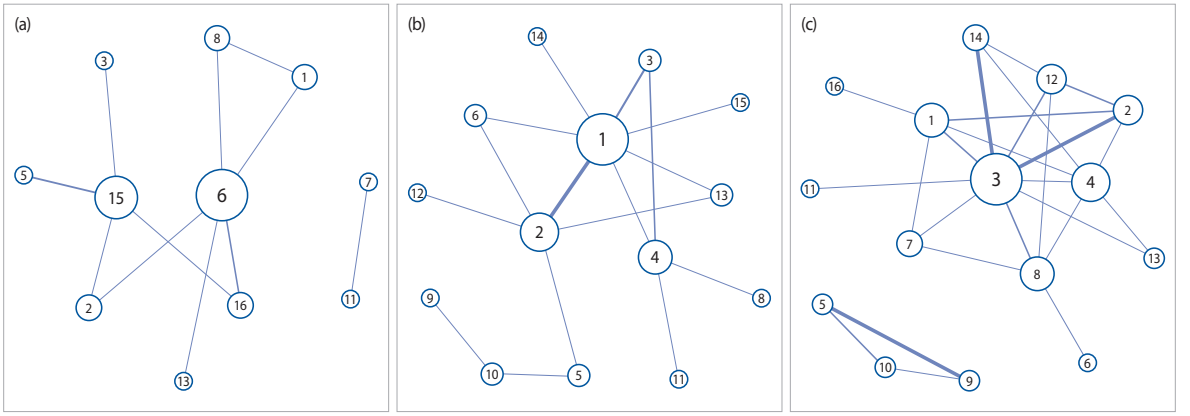


Fig. 6. Evolution of topic networks. (a) Period 1 (2000–2004), (b) period 2 (2005–2009), and (c) period 3 (2010–2014).

were examined in detail. Some overarching topics such as digital library projects, development of information systems/services, the role of metadata and metadata librarians, and evaluation of metadata quality turned out to be prevalent in metadata research, while more specifically focused topics such as semantic web and ontology, and record management are also found to have a significant share.

We also looked at research trends by comparing relative proportions of topics by year. In addition, as a means of gaining a better insight into how the topics have developed over time in connection with one another, we conducted a network analysis based on the distribution of topics over documents. Overall, the results show that, while some core topics more or less have retained their relatively large proportions in metadata literature, many topics exhibit considerable changes in popularity over time. At the same time, connections between topics continue to grow in volume and in diversity. A variety of factors may affect the rise and/or fall of a topic as well as the formation of relationships among topics, but some trends appear to be closely tied to the overall development of the field.

In the early days of metadata research, discussions related to building the infrastructure and some basic mechanisms for discovery and access in digital environments prevailed. Needless to say, development of international or national metadata standards as well as those proposed by communities in different domains constituted a large part of such groundwork, and metadata literature was once flooded with papers examining various aspects of standard development and deployment. Topic 6 (initiatives for standard development), topic 16 (Dublin Core), and topic 15 (XML and encoding standard) fall under this category. Although their importance in the field of metadata is hardly diminished, as the horizon of research has expanded, the proportions of these topics all show a decline later on.

Practitioners and researchers in metadata fields are keen on new technologies or developments in information environment, as shown in the topics peaking at different points in time. For instance, responding to the surge of social tagging tools, papers addressing various approaches to harnessing user generated tags for creating metadata or enhancing metadata-based services appeared soon after. Once a technology or an innovation is widely adopted in the field and its application is tested and reported in the context of metadata, words referring directly to the technology tend to dwindle in research papers. However, the decline in the relative share of a topic may reflect shifts in focus or integrations with other topics. Our network analysis of topic relations demonstrated this point, as explained with the case of topic 13 (metadata harvesting). The focus of discussion on this topic has moved from the harvesting standard itself to the development of repository services and to evaluation of the qualities of harvested metadata.

Metadata research has also been tightly connected to movements in the professional field of LIS. The turn of the trend of topic 8 (cataloging) is related to the development of the new model and standard for cataloging, which triggered the resurge of research interest in the topic in the context of broader metadata issues. Not only has the share of this topic increased, the topic appeared together with various other topics in recent literature, including topic 12 (Linked Data), indicating that the discourse surrounding this topic has extended beyond the traditional boundaries of library catalogs.

As the field has matured, empirical studies assessing the quality of metadata or examining the efficacy of current approaches to creation and use of metadata have emerged. The strong rise of topic 4 (evaluation of metadata quality) and the increased interest in automatic extraction methods (topic 10) showcase this trend. In addition, the recent stream of papers on Linked Data indicate that metadata research is continuously evolving in response to new developments in the global information infrastructure.

Finally, the increasingly dense and diversified connections among topics, as shown in the three sub-period topic networks, testify that topics emerged and evolved over time, not in isolation but in connection with one another. Note that the networks are constructed based on those papers addressing multiple topics together. The extended connections among topics therefore indicate that researchers become more attentive to related topics and attempt to incorporate relevant ideas and discourse into their work. Spawning connections to topic 3 and topic 4 in the recent topic network suggest that the discussions on the role of metadata for various purposes as well as the considerations of quality issues are becoming a cornerstone of metadata research.

Since the dataset in this study was constructed by searching well-known databases in the LIS field, the findings may reflect more the perspective of researchers and practitioners of the LIS field rather than a broader metadata research community. Metadata is an interdisciplinary field, and the current study has included only LIS journals. Therefore, it is necessary to analyze research trends by analyzing metadata research in various fields in future research. In addition, since the research data were analyzed by 2014, it is necessary to carry out the time series analysis continuously including the studies conducted thereafter.

Following up this initial effort to grasp the development of metadata research, we plan to expand the scope of the dataset in a future study to encompass research streams in other related fields.

REFERENCES

Aktas, M. S., Fox, G. C., & Pierce, M. (2010). A federated approach to information management in grids. *International Journal of Web Services Research*, 7(1), 65-98.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blessinger, K., & Hrycaj, P. (2010). Highly cited articles in library and information science: An analysis of content and authorship trends. *Library & Information Science Research*, 32(2), 156-162.

Boyd, K., & King, D. (2006). South Carolina goes digital: The creation and development of the University of South Carolina's Digital Activities Department. *OCLC Systems & Services*, 22(3), 179-191.

Cho, J. (2013). The recent trends of information organization research in Japan and Korea. *Library Collections, Acquisitions, and Technical Services*, 37(3-4), 107-117.

Chuttur, M. Y. (2012). An experimental study of metadata training effectiveness on errors in metadata records. *Journal of Library Metadata*, 12(4), 372-395.

Chuttur, M. Y. (2014). Investigating the effect of definitions and best practice guidelines on errors in Dublin Core metadata records. *Journal of Information Science*, 40(1), 28-37.

Conners, D. (2008). A ghost in the catalog: The gradual obsolescence of the main entry. *The Serials Librarian*, 55(1-2), 85-97.

Cumming, K. (2007). Purposeful data: The roles and purposes of recordkeeping metadata. *Records Management Journal*, 17(3), 186-200.

Danskin, A. (2014). Implementing RDA at the British Library. *CILIP Update*, 40-41.

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems*, 26, 154-163.

Emery, J. (2007). Ghosts in the machine: The promise of electronic resource management tools. *The Serials Librarian*, 51(3-4), 201-208.

Evans, J. (2007). Evaluating the recordkeeping capabilities of metadata schemas. *Archives and Manuscripts*, 35(2), 56-84.

Evans, J., & Rouche, N. (2004). Utilizing systems development methods in archival systems research: Building a metadata schema registry. *Archival Science*, 4(3-4), 315-334.

Feicheng, M., & Yating, L. (2014). Utilising social network analysis to study the characteristics and functions of the co-occurrence network of online tags. *Online Information Review*, 38(2), 232-247.

Feick, T., Henderson, H., & England, D. (2011). One identifier: Find your oasis with NISO's I2 (institutional identifiers) standard. *The Serials Librarian*, 60(1-4), 213-222.

Feinerer, I., & Hornik, K. (2014). tm: Text Mining Package: A framework for text mining applications within R. Retrieved September 22, 2018 from http://CRAN.R-project.org/package=tm.

Ferris, A. M. (2002). Cataloging internet resources using MARC21 and AACR2: Online training for working catalogers. *Cataloging & Classification Quarterly*, 34(3), 339-353.

Greifeneder, E. (2014, Semptember). *Trends in information behaviour research*. Paper presented at ISIC: the information behaviour conference (part 1), Leeds, United Kingdom.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1-30.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363-371). Hawaii: Association for Computational Linguistics.

Han, M. J., & Hswe, P. (2011). The evolving role of the metadata librarian. *Library Resources & Technical Services*, 54(3), 129-141.

Hunter, J. (2003). Working towards MetaUtopia: A survey of current metadata research. *Library Trends*, 52(2), 318-344.

Julien, H., Pecoskie, J. L., & Reed, K. (2011). Trends in information behavior research, 1999-2008: A content analysis. *Library & Information Science Research*, 33(1), 19-24.

Kanellopoulos, D. N., & Kotsiantis, S. B. (2007). Semantic

web: A state of the art survey. *International Review on Computer and Software*, 2(5), 428-442.

Lagace, N., Breeding, M., Romano Reynolds, R., & Han, N. (2013). Everyone's a player: Creation of standards in a fast-paced shared world. *The Serials Librarian*, 64(1-4), 158-166.

Lalitha, P. (2009). Importance of digitization of cultural and heritage materials. *SRELS Journal of Information Management*, 46(3), 249-266.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved September 22, 2018 from http://mallet.cs.umass.edu.

Medeiros, N. (2003). A pioneering spirit: Using administrative metadata to manage electronic resources. *OCLC Systems and Services*, 19(3), 86-88.

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of 24th Conference on Uncertainty in Artificial 1 Intelligence* (pp. 411-418). Arlington: AUAI Press.

Mimno, D., McCallum, A., & Mann, G. S. (2006). Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries* (*JCDL '06*) (pp. 65-74). New York: ACM.

Mullen, A. (2001). GILS metadata initiatives at the state level. *Government Information Quarterly*, 18(3), 167-180.

Palmer, C. L., Zavalina, O. L., & Mustafoff, M. (2007, June). Trends in metadata practices: A longitudinal study of collection federation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 386-395). New York: ACM.

Park, J. R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3-4), 213-228.

Park, J. R., & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696-715.

Patra, S. K., Bhattacharya, P., & Verma, N. (2006). Bibliometric study of literature on bibliometrics. *DESIDOC Journal of Library & Information Technology*, 26(1), 27-32.

Pattuelli, M. C. (2010). Knowledge organization landscape: A content analysis of introductory courses. *Journal of Information Science*, 36(6), 812-822.

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: The case of GEM. *Information Research: An International Electronic Journal*, 6(2). Retrieved September 22, 2018 from http://www.informationr.net/ir/6-2/paper94.html.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.

Schottlaender, B. E. C. (2003). Why metadata? Why me? Why now? *Cataloging and Classification Quarterly*, 36(3-4), 19-29.

Shiri, A. (2003). Digital library research: Current developments and trends. *Library Review*, 52(5), 198-202.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63-70). Baltimore: Association for Computational Linguistics.

Symonds, E., & May, C. (2009). Documenting local procedures: The development of standard digitization processes through the Dear Comrade project. *Journal of Library Metadata*, 9(3-4), 305-323.

Wakimoto, J. C. (2009). Scope of the library catalog in times of transition. *Cataloging & Classification Quarterly*, 47(5), 409-426.

Woodley, M. S. (2002). A digital library project on a shoestring. *Library Collections, Acquisitions, and Technical Services*, 26(3), 199-206.

Yagüe, M. I., Maña, A., & Lopez, J. (2005). A metadata-based access control model for web services. *Internet Research*, 15(1), 99-116.

Yeh, J., Chen, C., Sie, S., & Liu, C. (2014). X-System: An extensible digital library system for flexible and multi-purpose contents management. *International Journal of Digital Library Systems*, 4(1), 25-40.

Zeng, M. L., & Qin, J. (2016). Metadata (2nd ed.). Chicago: American Library Association.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 54-64, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.5

Information Needs of Korean Immigrant Mothers in the United States for Their Children's College Preparation

JungWon Yoon
School of Information, University of South Florida, Tampa, FL, USA
E-mail: jyoon@usf.edu

Natalie Taylor
School of Information, University of South Florida, Tampa, FL, USA
E-mail: ngtaylor@usf.edu

Soojung Kim*
Department of Library and Information Science, Chonbuk National University, Jeonju, Korea
E-mail: kimsoojung@jbnu.ac.kr

ABSTRACT
This study aims to understand the information needs of Korean immigrant mothers in the United States for their high school children's college preparation. A content analysis was conducted for the messages posted to a "motherhood" forum on the MissyUSA website. In total, 754 posts were analyzed in terms of a child's grade, college preparation stage, type of post, and topic of post. The study found that there is a range of information needed at different stages in a child's education. Many of the demonstrated information needs showed similarities to those of other immigrant groups, but there were also community-specific themes, such as an emphasis on STEM (science, technology, engineering, and math) and standardized tests. The forum was mainly used for factual questions, not emotional support. We concluded that the findings of the study would help researchers in understanding immigrant information needs for the college application process and how information professionals and educators could combine the needs of different ethnic groups to create customized services for them.

Keywords: information needs, information-seeking behavior, online forum, Korean immigrants, immigrant mothers, college preparation

Open Access

Accepted date: September 14, 2018
Received date: September 06, 2018

***Corresponding Author:** Soojung Kim
Professor
Department of Library and Information Science, Chonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju 54896, Korea
E-mail: kimsoojung@jbnu.ac.kr

1. INTRODUCTION

Preparing for college is a complex process that requires extensive information gathering. A lack of information about college and financial aid (FA) can be a barrier that results in even students with high education aspirations being prevented from attending a college (Kao & Tienda, 1998). The information regarding college preparation needs to be obtained in a timely manner, but finding information can be a daunting task, especially for immigrant parents who lack understanding of or experience with the United States education system.

Previous studies have been conducted on the college choice process of students of different ethnicities/races and parental involvement (Teranishi, Ceja, Antonio, Allen, & McDonough, 2004; Ceja, 2006; Kim, 2014), and these studies provide insight into the role of parents in the college choice process. Immigrant parents not only have linguistic and socio-cultural barriers but also lack knowledge about the United States education system. Therefore, for information professionals and educators to best facilitate immigrant parents, it is necessary to understand what they need to know to prepare for their children's post-secondary education. However, there is a lack of research on what information immigrant parents seek, particularly for certain specific immigrant groups. To fill this gap, the current study aims to understand Korean immigrant mothers' information needs for their high school children's college preparation.

Koreans are well-known for education fervor. Education is one of the central motives for Korean families' immigration to the United States (Choi, Cranley, & Nichols, 2001). For these families, getting into a good college is essential to a student's future success. That the culture puts a strong value on college education is evident in Korean immigrants' high education attainment. In 2015, 53% of Korean immigrants aged 25 and over had a bachelor's degree or higher, compared to 29% of the total United States foreign-born population and 31% of the native-born population (Zong & Batalova, 2017). Knowing Korean parents' high interest in college education, the researchers expect that they actively seek information during the college choice process of their children. In particular, maternal support for children's education in Korea has been well documented (Park, 2007). Considering Korean mothers' critical role in children's education, this study seeks to demonstrate what information Korean immigrant mothers need to know for their children's college preparation. Knowing what information they seek will help information professionals, educators, school library media specialists, and school counselors to provide appropriate information services to user groups who are not familiar with the process of United States college preparation.

The purpose of this study is to understand Korean immigrant mothers' information needs for their high school children's college preparation through the analysis of an online forum for Korean women in the United States. This study will address the following specific research question: How can Korean immigrant mothers' information needs for their high school children's college preparation be characterized in terms of types of posts, topics, college preparation stages, and grades?

2. LITERATURE REVIEW

The following literature review will focus on two major aspects of this study: 1) a general overview of studies on mothers' participation in online forums and 2) studies on immigrant parents' involvement in college preparation and choice, and, related to this, the College Choice Model.

2.1. Parenting Online Forums

Many of the studies focusing on mothers' participation in online forums have focused on those with young children. For example, Evans, Donelle, and Hume-Loveland (2012) conducted a content analysis of messages in an online forum for mothers battling postpartum depression, finding that participants' postings gave emotional support—giving hope, honesty, and affection and empathy, as well as information—reassurance and validation, peer expertise, medical advice, and instrumental support, such as help with daily activities. Porter and Ispa (2013) found through an ethnographic study of messages posted by mothers of children under two on the websites of two best-selling parenting magazines in the United States that mothers had many questions about their children's sleeping and eating habits. The authors also highlighted themes in the posts, such as parenting stress, questions about advice given by others, and concerns over their children's development.

Overall, these studies and others point to several factors as reasons for mothers participating in parenting-oriented online forums. Valtchanov, Parry, Glover, and Mulcahy (2014) found that "[t]he distinctive ease, convenience, and speed of online connectivity, combined with the respectfulness cultivated within this particular online community, facilitated mothers' access to essential peer support in the forms of emotional sustenance, 'appraisal assistance,' and informational resources" (p. 187). Drentea

and Moren-Cross (2005) found three main types of communication present in their study of a motherhood parenting board—emotional support, instrumental support, and community building/protection (also highlighted in the Evans et al. article). One literature review of international research on parents' participation in online forums found that anonymity and availability are two central factors for why parents participate (Doty & Dworkin, 2014). Other websites seem to be more of a space for venting frustrations or otherwise engaging in anonymity by saying things one would not say in real life—the YouBeMom message board studied by Schoenebeck (2013) is one example.

Other research has examined the potential effects of income, age, education, and comfort on participation. Doty, Dworkin, and Connell (2012) surveyed 1,518 parents about their online activities and found that although income had a slight impact on information seeking behaviors, the factor with the most impact on information seeking and online activities was the participant's comfort with technology. Age and education had no effects. However, there has been less study of mothers of older children participating in online forums, as well as the potential differences that the immigrant experience might make to both participation in and needs of online forums. Our study aims to address some of these gaps.

2.2. Immigrant Parents' Involvement in College Preparation and Choice

The process of high school students' college choice is complicated, but academic studies have attempted to explain the various stages. One example is Hossler and Gallagher's (1987) model. This model consists of three phases—predisposition, search, and choice. The predisposition phase is where students make a decision whether they would like to attend a college. The search phase is the second phase where students begin seeking information about colleges. The final phase is that of choice where students decide which college to attempt to attend. During this phase, students' families decide to accept any offers and apply for FA. This model has been tested with Asian Pacific Americans with findings that show

> the college decision-making processes varied by the ethnic and socioeconomic class backgrounds of students. This general finding was true among such specific factors as the influence of social networks, the impact of cost and FA availability, numbers of college applications submitted, and perceptions of the prestige and reputation of different colleges (Teranishi et al., 2004, p. 546).

This study demonstrates the multiple "inputs" into the college decision making process, including socioeconomic status, ethnicity, and social characteristics.

Litten (1982)'s five-stage model is another example. This model is composed of college aspirations, decision to start process, information gathering, applications and enrollment, and identified college actions (admit/deny and FA) between the applications and enrollment stages. For each stage, they also identified influencing factors such as background (race, income, socioeconomic status, parents' education), personal attributes (academic ability, self-image), environment, high school attributes, student performance, public policy (FA), college actions (recruitment, activities, academic/admission policies), and college characteristics. The College Choice Model, whether it is a three-stage or a five-stage model, is an important first step in determining the various points at which immigrant parents have information needs in the college choice process.

The literature on college preparation and choice is clear that the role of the college applicant's support network is powerful. For example, Ma and Yeh (2010) studied 265 Chinese immigrants at a New York high school and found that "career-related support from parents" (along with other factors) positively predicted the youths' plans to attend college. Carolan-Silva and Reyes (2013) found in their study of first-generation Latino high school students that "parents influenced their children's aspirations indirectly through the values and qualities they modeled for their children, and directly through communicating their desire for them to attend college" (p. 341). Cabrera and La Nasa (2000) found that "[b]arriers to college access and completion occur along the K-16 pipeline when students are not supported in their development of college aspirations; their enrollment in college-bound classes; their access to sufficient information about college; and their completion of the necessary applications to enroll in college and receive financial aid" (cited in Carolan-Silva & Reyes, 2013, p. 335).

There have been some studies about parents' information needs with regard to their children's preparation for and choice of higher education, particularly in Latino communities. In 2002, Tornatzky, Cutler, and Lee conducted a phone survey of over 1,000 Latino parents in three major United States cities, as well as follow-up interviews with 41 of these parents and found that these parents' knowledge of college was low, with over 60% of the parents missing more than half of the eight informational questions, ranging from topics about tuition and fees of different types of schools to when students should begin college preparation courses. At the same time, almost all of the parents expected

their children to attend college, so clearly there is a need for education. Fann, McClafferty Jarsky, and McDonough (2009) conducted an action study wherein Spanish-language workshops were held with Latino parents on college knowledge. They found that parents were generally looking for information on FA, general information on the system of higher education, application processes, academic requirements, and tests. Researchers found that these immigrant parents differed from many native parents, as they had little prior knowledge of the formal higher education structure.

Less is known about Korean-American communities. As mentioned in the introduction, parents of Asian American students place a heavy emphasis on education, often engaging their children in "shadow education," such as SAT preparation tests and other tutoring (Byun & Park, 2012), suggesting a high amount of interest in higher education and a need for information about the college application process. Our study aims to address these gaps in knowledge.

3. METHODOLOGY

This study is based on a content analysis of messages posted to a "motherhood" forum on the MissyUSA website. MissyUSA is one of the largest online communities among Korean immigrants in the United States, which was launched in 2002. The members of the site are married Korean women living in America and they seek and share information about their lives in their adopted country. The site contains many forums for discussion on various topics. On the forums, the members of the site can post messages or reply to posted messages.

The Motherhood forum is divided into three categories—Pregnancy, Baby, and Schooling, the last of which has six subgroups—Daycare and Preschool, Kinder and Elementary school, Middle and High school, College Admissions, College and Graduate school, and Advertisements. The College Admissions subgroup forum is for sharing questions and information on college preparation and application. Since this study attempts to determine Korean mothers' information needs regarding their high school students' college preparation processes, messages from the College Admissions forum were analyzed. The number of posts posted in the College Admissions forum from July 1, 2016 to June 30, 2017 was 8,319. Out of these 8,319 posts, every 10th post was collected. After discarding 84 posts as they were unfit for analysis (e.g., posts without messages, advertisements, posts not related to college preparation, etc.), a total of 747 posts were analyzed by two of the authors. For posts that had multiple queries, each query was treated as a separate post, so in total, 754 posts were analyzed.

The collected data were coded in four areas: grade, stage, type of post, and topic of post. First, grade was determined based on a student's current grade, if it was identifiable in the post or if it could be directly inferred from the post. For example, if a mother asked when a specific college's admission results were released, it was inferred that her child was a high school senior. Second, a coding scheme for stage was determined as Preparation, Information gathering, Application, Admission/Acceptance, and Enrollment. The authors defined a preliminary scheme for stages based on previous studies on college models (Cabrera & La Nasa, 2000; Chapman, 1981; Hossler & Gallagher, 1987; Litten, 1982), and the five stages were finalized by reflecting unique characteristics of the current dataset. Third, regarding type of post, a previous study (Kim & Yoon, 2012) that conducted content analysis of online forums was considered to establish the precoding scheme, and the scheme was revised based on the unique characteristics of the current dataset. Fourth, due to lack of previous studies in this area, the topic of post coding scheme was developed in vivo and refined several times throughout an iterative coding process. Two of the authors coded 10% of the randomly selected posts and established precoding schemes. Then they coded another 10% of the randomly selected posts using the precoding schemes, and necessary revisions were made to the precoding schemes. The two authors next coded another 10% of the randomly selected posts each, and the percentages of agreement were calculated to check intercoder reliability. Using Holsti's (1969) reliability formula, the percentage of intercoder agreement was 88% on average (grade 89%; stage 87%; type of posting 89%; topic of posting 88%). The final coding system is as follows:

- Grade
  - Middle schooler
  - Freshman
  - Sophomore
  - Junior
  - Senior
  - Not specific

- Stage:
  - Preparation: a stage in which students prepare themselves through high school academic performance and extracurricular activities
  - Information gathering: a stage in which students gather

information for their college choice and application process
  - Application: a stage in which college application and related documents are prepared and submitted
  - Admission/Acceptance: a stage in which students get admission results, decide where they want to go, and accept an offer from a college
  - Enrollment: a stage after making the decision about which college to attend in which students work on related paperwork and prepare for their college life
  - Uncertain: a stage cannot be clearly identified from a post

- Type of post:
  - Seeking information: posts seeking factual and objective information
  - Asking opinions: posts asking for advice, opinions, and subjective information
  - Sharing information: posts sharing factual and objective information
  - Sharing experiences: posts sharing personal experiences
  - Sharing emotions: posts mainly expressing emotions and feelings

- Topic of post:
  - High school academic performance: posts related to high schoolers' academic performance
  - High school curriculum: posts related to high school curricula, including Advanced Placement (AP) courses and the International Baccalaureate (IB) program
  - High school extracurricular and camps: posts related to extracurricular and other activities, including camps not hosted by a college (camps hosted by a college fall under the 'campus visits' category)
  - Standardized tests: posts related to standardized tests such as Scholastic Assessment Test (SAT), Preliminary SAT (PSAT) and American College Test (ACT). This category includes posts about comparison of different tests, test preparation, and test-taking process.
  - Other high school academic related: posts related to high school academics but not categorized in the above categories, such as posts related to high school academic environments and finding tutors
  - Specific college information: posts related to a specific college, including educational or living environment of the college, information about a certain program, qualifications needed for entrance, and so on. When a post asks about what SAT scores are needed to get admitted into a college, the post is categorized in 'specific college information,' not 'standardized tests.'
  - General college information: posts related to college in general, including college rankings and asking about college recommendations
  - Major and future career: posts asking about/sharing information regarding majors or future careers. Posts requesting college recommendation for a specific major fall under the 'general college information' category.
  - College expenses/FA: posts related to college expenses (tuition, living costs), FA, student loans, and scholarships
  - FA application: posts related to FA application forms and the FA process
  - College application: posts related to college application forms, the college application process, and preparing application documents, including recommendation letters and essays
  - College admission results: posts asking about/sharing college admission results and related information (e.g., notifying dates)
  - Accept decision and process: posts related to deciding to attend and accepting a college's admission offer and the acceptance process
  - College visits: posts related to college campus visits, including summer camps for high schoolers and college orientations
  - College life preparation: posts related to college life preparation after deciding on a college. Posts related to preparation for a specific college fall under the 'specific college information' category.
  - Immigration status: posts related to any issues caused by immigration status. When a post has two topics (e.g., immigration status and FA application), the post is categorized in the 'immigration status' category.
  - Others: posts that do not fit into the above categories

4. RESULTS

4.1. Topics and Types of Posts

As shown in Table 1, the most frequently posted topics were: specific college information (15%), standardized tests (13%), college application forms and processes (11%), college admission results (10%), college expenses and FA (8%), and general college information (8%). Some examples of posts in the most popular topic category, 'specific college information,' are "Please tell me about Brandeis University. I can't find much information about that university"; "If your kid got admitted into UVA, please share his/her grades,

test scores, and extracurricular activities. My son's wish is attending UVA"; "Do you know about the pre-med school at the University of Richmond?"

Regarding the type of posts, the number of posts asking for advice and opinions accounted for 57%, while 31% of posts asked for factual or objective information. Only a small portion of the posts intended to share information (5.8%), personal experiences (3.3%), or emotions (2.9%).

Several relations between topics and post types were observed. Questions seeking others' advice and opinions were related to various topics, including high school academic performance (e.g., What can be done if a junior student has a low GPA?), curriculum (e.g., Which class should be selected between AP Biology and AP Chemistry?), extracurricular activities (e.g., Would it be useful to participate in a specific extracurricular activity for college admission?), standardized tests (e.g., Are there disadvantages if a student takes SAT multiple times?), specific college information (e.g., How is the living environment of a certain college?), general college information (e.g., Which colleges would admit a student with certain standardized test scores?), major/future career (e.g., Which major would be best for a certain future career?), accept decision process (e.g., How can I deposit money to the college?), and campus life preparation (e.g., How should I prepare bedding for my child going to a cold area?).

Questions containing requests for factual and objective information were often found in topics such as standardized tests (e.g., How long does it take to get ACT scores back?), FA application process (e.g., What should be recorded in a specific field in the FAFSA form?), college application process (e.g., How should students report AP exam scores?), college admission results (e.g., Where can I find admission results for a specific college?), accept decision process (e.g., How can students deny an acceptance offer from a college that they do not want to attend?), and immigrant status (e.g., How can a student indicate their immigration status if he/she is pending in the permanent resident process?).

The most frequently shared information was general college information, such as college rankings. The mothers shared personal experiences and emotions when discussing college admission results—for example, they shared the specs of a student who was admitted to a specific college and shared their emotions after they learned their child had been admitted.

Table 1. Number of posts by topic and type of posts

	Seeking information	Asking opinion	Sharing information	Sharing experience	Sharing emotion	Total
HS academic performance	3	18	4	-	-	25 (3.3%)
HS curriculum	8	37	-	1	-	46 (6.1%)
HS extracurricular and camps	4	17	1	-	1	23 (3.1%)
Standard tests	35	60	1	-	2	98 (13.0%)
Other HS academic related	2	10	1	-	-	13 (1.7%)
Specific college information	18	92	4	1	-	115 (15.3%)
General college information	5	41	12	1	1	60 (8.0%)
Major & future career	-	17	5	1	1	24 (3.2%)
College expenses/FA	27	27	4	1	2	61 (8.1%)
FA application	26	16	-	-	-	42 (5.6%)
College application	37	37	2	3	2	81 (10.7%)
College admission results	38	12	6	15	7	78 (10.3%)
Accept decision & process	12	16	-	2	3	33 (4.4%)
College visits	7	10	-	-	-	17 (2.3%)
Campus life preparation	5	10	1	-	-	16 (2.1%)
Immigration status	6	3	2	-	-	11 (1.5%)
Others	1	6	1	-	3	11 (1.5%)
Total	234 (31%)	429 (56.8%)	44 (5.8%)	25 (3.3%)	22 (2.9%)	(100%)

HS, high school; FA, financial aid.

4.2. College Preparation Stages and College Information Needs

In terms of stages, the posts were relatively evenly distributed, although the preparation (26%) stage garnered the most posts by a few percentage points, followed by the admission/acceptance stage (23%), the information gathering stage (21%), and the application stage (20%) (Table 2). The patterns between topics and stages were obvious. In the preparation stage, posts regarding high school academic performance, curriculum, extracurricular activities, and standardized tests were frequent. Particularly, the ratio of standardized test related posts was notably high: 44% of posts in the preparation stage were in this category. During the information gathering stage, specific college information, general college information, major recommendation, and college expenses and FA were the main topics mentioned in the posts. In the application stage, questions on colleges as well as FA application forms and processes were often asked. During the stage when students are admitted and decide on what college to attend, questions on specific college information, college expenses and FA information, admission results, and acceptance decision and process were frequently posted. After students had decided on a college to attend (enrollment stage), questions on specific college and campus life preparations were often posted. Posts related to college expenses and FA and posts related to specific college appeared throughout the stages from information gathering and enrollment. Whereas college models include a predisposition stage (Cabrera & La Nasa, 2000) or college aspiration stage (Litten, 1982) in which aspirations for higher education are developed for high school students, the current dataset showed that Korean immigrant mothers who go to the online forum have already decided to have their children go to college. The appendix includes sample questions with the categories of topic, type of posts, and stage.

4.3. Grades and College Information Needs

As shown in Table 3, over a half of the questions were asked by parents whose children were seniors, followed by those whose children were juniors. Depending on a student's grade, topics of the questions varied. Mothers of freshmen high school students asked mainly about high school curricula, and mothers of sophomore high school students

Table 2. Number of posts by topic and stage

	Preparation	Information gathering	Application	Admission/ acceptance	Enrollment/ college life	Uncertain	Total
HS academic performance	21	4	-	-	-	-	25 (3.3%)
HS curriculum	40	3	-	2	-	1	46 (6.1%)
HS extracurricular and camps	22	-	-	1	-	-	23 (3.1%)
Standard tests	87	3	6	1	-	1	98 (13.0%)
Other HS academic related	11	2	-	-	-	-	13 (1.7%)
Specific college information	2	54	7	39	11	2	115 (15.3%)
General college information	3	51	1	2	-	3	60 (8.0%)
Major & future career	1	14	2	4	-	3	24 (3.2%)
College expenses/FA	-	11	21	17	7	5	61 (8.1%)
FA application	-	-	39	3	-	-	42 (5.6%)
College application	4	5	67	4	1	-	81 (10.7%)
College admission results	-	1	2	75	-	-	78 (10.3%)
Accept decision & process	-	-	2	24	5	2	33 (4.4%)
College visits	4	5	-	1	7	-	17 (2.3%)
Campus life preparation	-	3	-	-	12	1	16 (2.1%)
Immigration status	1	4	4	1	-	1	11 (1.5%)
Others	1	1	2	1	-	6	11 (1.5%)
Total	197 (26.1%)	161 (21.4%)	153 (20.3%)	175 (23.2%)	43 (5.7%)	25 (3.3%)	754 (100%)

HS, high school; FA, financial aid.

asked about standardized tests. Some questions on college information (specific/general) were also noticed. Mothers of junior high school students asked mostly about high school academic performance, curricula, and standardized tests. These mothers also looked for specific and general college information. Mothers of senior high school students still asked about high school curricula and standardized tests, but they more actively gathered college information, application process information, and admission/accept process information. As Fig. 1 shows, the preparation stage starts as early as middle school, and information gathering for college stage starts in the sophomore year of high school.

Table 3. Number of posts by topic and grade

	Middle schooler	Freshman	Sophomore	Junior	Senior	Not specific	Total
HS academic performance	-	1	5	9	3	7	25 (3.3%)
HS curriculum	2	8	3	12	7	14	46 (6.1%)
HS extracurricular and camps	-	4	3	6	2	8	23 (3.1%)
Standard tests	2	2	12	28	18	36	98 (13.0%)
Other HS academic related	2	1	-	2	1	7	13 (1.7%)
Specific college information	-	-	2	11	72	30	115 (15.3%)
General college information	1	1	3	9	15	31	60 (8.0%)
Major & future career	-	-	1	2	6	15	24 (3.2%)
College expenses/FA	-	-	-	3	46	12	61 (8.1%)
FA application	-	-	-	2	37	3	42 (5.6%)
College application	-	-	1	2	68	10	81 (10.7%)
College admission results	-	-	-	-	75	3	78 (10.3%)
Accept decision & process	-	-	-	1	31	1	33 (4.4%)
College visits	-	1	2	2	11	1	17 (2.3%)
Campus life preparation	-	-	-	-	12	4	16 (2.1%)
Immigration status	-	-	-	1	6	4	11 (1.5%)
Others	-	-	-	-	3	8	11 (1.5%)
Total	7 (0.9%)	18 (2.4%)	32 (4.2%)	90 (11.9%	413 (54.8%)	194 (25.7%)	754 (100%)

HS, high school; FA, financial aid.

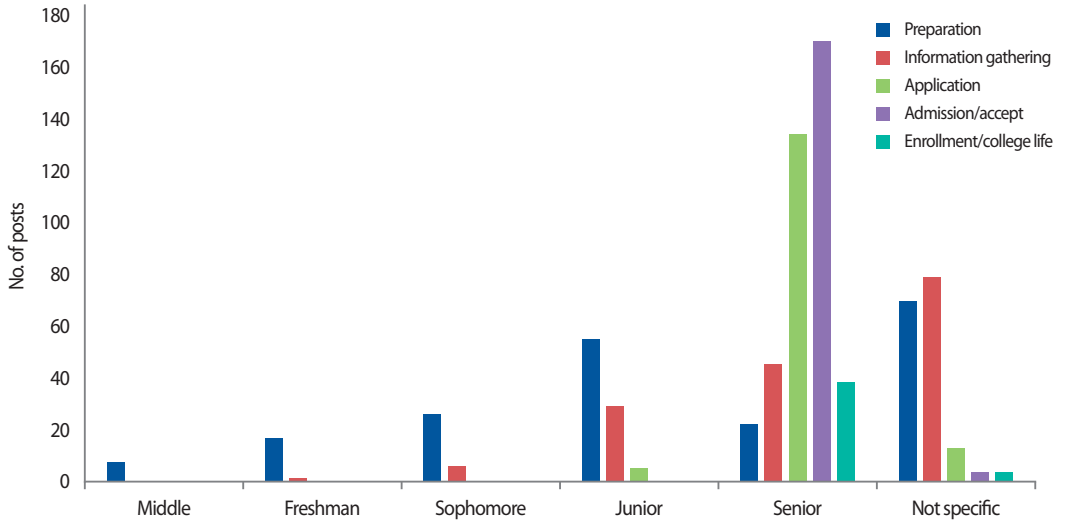


Fig. 1. Number of posts per grade by stage.

5. DISCUSSION

The findings of the current study highlight four key issues essential to understanding Korean immigrant mothers' information needs to assist their children along the path to college.

First, Korean immigrant mothers have a wide range of information needs from high school academics through the college application process to FA. At a glance, these information needs are common to any ethnic/racial groups of parents in the United States. On looking further, however, Korean immigrant mothers who are not familiar with the United States. education system lack basic knowledge of United States colleges and the application procedures including FA. This finding supports earlier research indicating that information on the application process and FA are of utmost importance for Latino parents of first-generation students (Fann et al., 2009). Applying for FA is especially challenging for Korean immigrant mothers because of different procedures between Korea and the United States. In Korea, the FA application process begins after a student is fully accepted to college and the procedures are relatively simple. In the United States, the process is not only a parallel track of the college preparation process, but is very complicated. Those parents with limited English abilities face significant difficulties in acquiring and understanding the FAFSA, CSS, and other FA-related information. Therefore, Korean mothers, like other immigrant parents, are in need of a base knowledge of the overall college application process including the FA application process, as well as knowledge of the United States education system to help their children prepare for college.

Second, in addition to general information needs, immigrant community-specific information needs call for special attention. Among the most frequently sought information by Korean immigrant mothers was information on standardized tests such as the SAT. It is known that Asian parents put a strong emphasis on standardized tests and thus East Asian students including Koreans are more likely to take SAT test prep courses than White and Hispanic students (Byun & Park, 2012). Whether such emphasis is outsized to their importance is debatable, but what is certain is that Korean mothers are anxious to acquire substantial information about standardized tests from basics (e.g., how to cancel SAT scores) to more strategic guidance (e.g., choices of subjects in SAT2). Also, a large portion of the questions related to college information were towards science, engineering, and medical fields. This finding confirms Kim's study (2014), which revealed that it is a norm for Korean immigrant parents to guide their children to science, engineering, and business fields as they are perceived to guarantee high payoff and job security. Other community-specific inquiries were concerned with visa/immigration status. These community-specific information needs reflect cultural values and concerns Korean immigrant mothers have and should be adequately addressed by counselors, teachers, and information professionals.

Third, although many online forums are used for emotional sharing, MissyUSA is used mainly for informative purposes for the topic of college preparation. In terms of type of information requested, almost 60% of the mothers sought advice while 30% looked for factual information. Advice-seeking questions are, for example, "What would be the best choices of courses for the 10th grader who hasn't decided on a major?", "Which college would be better between Duke and Johns Hopkins for an 11th grade girl interested in the education or medical field?" These questions cannot be answered by college-planning resources that only provide generic information. MissyUSA is appealing because the mothers can receive customized advice/suggestion for an individual child anonymously. Furthermore, since Korean immigrant mothers lack English skills and confidence in interacting with teachers/counselors (Kao, 2014), they may prefer the online forum where they can easily communicate with their Korean peers who have common understandings.

Fourth, the questions in this study were almost evenly distributed across the college preparation stages. This implies that the mothers seek information constantly as their children go through the college-preparation process.

6. CONCLUSION

Through an exploration of posts on a popular online forum for Korean immigrant mothers, this study explored Korean immigrant mothers' information needs for their high school children's college preparation. By analyzing posts made over a one-year period, researchers found that many information needs, such as questions about the FA process, demonstrated similarities to those of other immigrant groups. At the same time, there were also community-specific themes that emerged, such as an emphasis on science, technology, engineering, and math (STEM) and standardized tests. The forum was mainly used for factual questions, not emotional support. Finally, there is clearly a

need for informal support networks, perhaps because of a lack of confidence in English skills and in interacting with teachers and counselors.

This study demonstrates that there are many ways that educators, counselors, and information professionals can better meet immigrant Korean mothers' information needs regarding their children's higher education admissions. The findings show that there are both general types of information that can be offered that will be useful to the general population of parents of high school-aged students, as well as customized types of information for this specific immigrant group. In general, parents need to know information on the details of the application process, particularly FA information. This is particularly important for parents who might not have gone through the process themselves, such as in the case of adult immigrants or parents who did not attend college. This also presupposes a knowledge of the higher education system in America, which, again, should not be assumed for those parents who have not been exposed to the system first-hand. Many immigrant groups, or parents with little exposure to the higher education system, may be nervous about speaking to educators about these issues, so outreach is critical. An emphasis on providing access for those with limited English language skills and those who are not used to interacting with information professionals is also important.

More specific to the Korean immigrant experience, this study shows that mothers would like community-specific information, such as issues that may arise around visas or immigration. Additionally, there are cultural factors that are important, such as a general emphasis on STEM and standardized tests, which could be used as means of entry to the community. If educators know that these mothers are already concerned about these specific issues, starting with those topics and then leading into other parts of the process about which they are less aware might work well.

Finally, the findings show that starting parental education when children are younger—as early as middle school—would be beneficial. The results demonstrate that there is a range of information needed at different stages in a child's education, confirming earlier studies on college application information-seeking processes. This is again an opportunity that would benefit not only this specific immigrant community, but all parents interested in helping their parents navigate the complicated college admissions process.

This study has limitations like other contents analysis studies of online community forums. Since this study examined the messages posted to a single online community, although it is a major online community among Korean immigrant women, the results may not represent the entire Korean immigrant population. In future studies, it will be critical to conduct further investigations on Korean immigrant parents' needs and barriers through interviews to get a better understanding of what type of information they are seeking outside of online forums—and why they choose various venues for pursuing this information. It would also be beneficial to understand where emotional support is coming from, given the emphasis on factual information-seeking that was present in the online discussions. Additionally, exploring other immigrant groups, particularly for the groups who have lower college attendance, would be a useful point of comparison. This would aid researchers in creating a general understanding of immigrant information needs for the college application process and how information professionals and educators could combine the needs of different groups to create services that are better able to reach all those who need this type of information.

REFERENCES

Byun, S. Y., & Park, H. (2012). The academic success of East Asian American youth: The role of shadow education. *Sociology of Education*, 85(1), 40-60.

Cabrera, A. F., & La Nasa, S. M. (2000). Three critical tasks America's disadvantaged face on their path to college. *New Directions for Institutional Research*, 2000(107), 23-29.

Carolan-Silva, A., & Reyes, J. R. (2013). Navigating the path to college: Latino students' social networks and access to college. *Educational Studies*, 49(4), 334-359.

Ceja, M. (2006). Understanding the role of parents and siblings as information sources in the college choice process of Chicana students. *Journal of College Student Development*, 47(1), 87-104.

Chapman, D. W. (1981). A model of student college choice. *The Journal of Higher Education*, 52(5), 490-505.

Choi, S., Cranley, M. E., & Nichols, J. D. (2001). Coming to America, becoming American: Narration of Korean immigrant young men. *International Education Journal*, 2(5), 47-60.

Doty, J. L., & Dworkin, J. (2014). Online social support for parents: A critical review. *Marriage & Family Review*, 50(2), 174-198.

Doty, J. L., Dworkin, J., & Connell, J. H. (2012). Examining

digital differences: Parents' online activities. *Family Science Review*, 17(2), 18-39.

Drentea, P., & Moren-Cross, J. L. (2005). Social capital and social support on the web: The case of an Internet mother site. *Sociology of Health & Illness*, 27(7), 920-943.

Evans, M., Donelle, L., & Hume-Loveland, L. (2012). Social support and online postpartum depression discussion groups: A content analysis. *Patient Education and Counseling*, 87(3), 405-410.

Fann, A., McClafferty Jarsky, K., & McDonough, P. M. (2009). Parent involvement in the college planning process: A case study of P-20 collaboration. *Journal of Hispanic Higher Education*, 8(4), 374-393.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading: Addison-Wesley.

Hossler, D., & Gallagher, K. S. (1987). Studying student college choice: A three-phase model and the implications for policymakers. *College and University*, 62(3), 207-221.

Kao, G. (2014). Parental influences on the educational outcomes of immigrant youth. *International Migration Review*, 38(2), 427-449.

Kao, G., & Tienda, M. (1998). Educational aspirations of minority youth. *American Journal of Education*, 106(3), 349-384.

Kim, E. (2014). When social class meets ethnicity: College-going experiences of Chinese and Korean immigrant students. *The Review of Higher Education*, 37(3), 321-348.

Kim, S., & Yoon, J. (2012). The use of an online forum for health information by married Korean women in the United States. *Information Research*, 17(2), 1-18.

Litten, L. H. (1982). Different strokes in the applicant pool: Some refinements in a model of student college choice. *The Journal of Higher Education*, 53(4), 383-402.

Ma, P.-W. W., & Yeh, C. J. (2010). Individual and familial factors influencing the educational and career plans of Chinese immigrant youths. *The Career Development Quarterly*, 58(3), 230-245.

Park, S. J. (2007). Educational manager mothers: South Korea's neoliberal transformation. *Korea Journal*, 47(3), 186-213.

Porter, N., & Ispa, J. M. (2013). Mothers' online message board questions about parenting infants and toddlers. *Journal of Advanced Nursing*, 69(3), 559-568.

Schoenebeck, S. Y. (2013). The secret life of online moms: Anonymity and disinhibition on YouBeMom.com. *Proceedings of the seventh international AAAI Conference on Weblogs and Social Media* (pp. 555-562). Palo Alto: The AAAI Press.

Teranishi, R. T., Ceja, M., Antonio, A. L., Allen, W. R., & McDonough, P. M. (2004). The college-choice process for Asian Pacific Americans: Ethnicity and socioeconomic class in context. *The Review of Higher Education*, 27(4), 527-551.

Tornatzky, L.G., Cutler, R., & Lee, J. (2002). *College knowledge: What Latino parents need to know and why they don't know it*. Retrieved September 20, 2018 from http://trpi.org/wp-content/uploads/archives/College_Knowledge.pdf.

Valtchanov, B. L., Parry, D. C., Glover, T. D., & Mulcahy, C. M. (2014). Neighborhood at your fingertips: Transforming community online through a Canadian social networking site for mothers. *Gender, Technology and Development*, 18(2), 187-217.

Zong, J., & Batalova, J. (2017). *Korean immigrants in the United States*. Retrieved September 20, 2018 from https://www.migrationpolicy.org/article/korean-immigrants-united-states.

Call for Paper

Journal of Information Science Theory and Practice (JISTaP)

We would like to invite you to submit or recommend papers to **Journal of Information Science Theory and Practice** (JISTaP, eISSN: 2287-4577, pISSN: 2287-9099), a fast track peer-reviewed and no-fee open access academic journal published by Korea Institute of Science and Technology Information (KISTI), which is a government-funded research institute providing STI services to support high-tech R&D for researchers in Korea. JISTaP marks a transition from Journal of Information Management to an English-language international journal in the area of library and information science.

JISTaP aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics.

We welcome materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. Topics covered by the journal include: information processing and management; information policy; library management; knowledge organization; metadata and classification; information seeking; information retrieval; information systems; scientific and technical information service; human-computer interaction; social media design; analytics; scholarly communication and bibliometrics. Above all, we encourage submissions of catalytic nature that explore the question of how theory can be applied to solve real world problems in the broad discipline of information science.

Co-Editors in Chief : Gary Marchionini & Dong-Geun Oh

Please click the "Online Submission" link in the JISTaP website (http://www.jistap.org), which will take you to a login/ account creation page. Please consult the "Author's Guide" page to prepare your manuscript according to the JISTaP manuscript guidelines.

Any question? Suhyeon Yoo (managing editor) : jistap@kisti.re.kr

Information for Authors

The Journal of Information Science Theory and Practice (JISTaP), which is published quarterly by the Korea Institute of Science and Technology (KISTI), welcomes materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. JISTaP is an open access journal run under the Open Access Policy. See the section on Open Access for detailed information on the Open Access Policy.

A. Originality and Copyright

All submissions must be original, unpublished, and not under consideration for publication elsewhere. Once an article is accepted for publication, all papers are accessible to all users at no cost. If used for other researches, its source should be indicated in an appropriate manner and the content can only be used for uncommercial purpose under Creative Commons license.

B. Peer Review

All submitted manuscripts undergo a single-blind peer review process in which the identities of the reviewers are withheld from the authors.

C. Manuscript Submission

Authors should submit their manuscripts online via Article Contribution Management System (ACOMS). Online submission facilitates processing and reviewing of submitted articles, thereby substantially shortening the paper lifecycle from submission to publication. After checking the manuscript's compliance to the Manuscript Guidelines, please follow the "Online Submission" hyperlink in the top navigation menu to begin the online manuscript submission process.

D. Open Access

With the KISTI's Open Access Policy, authors can choose open access and retain their copyright or opt for the normal publication process with a copyright transfer. If authors choose open access, their manuscripts become freely available to public under Creative Commons license. Open access articles are automatically archived in the KISTI's open access repository (KPubS, www.kpubs.org). If authors do not choose open access, access to their articles will be restricted to journal users.

E. Manuscript Guidelines

Manuscripts that do not adhere to the guidelines outlined below will be returned for correction. Please read the guidelines carefully and make sure the manuscript follows the guidelines as specified. We strongly recommend that authors download and use the manuscript template in preparing their submissions.

1. Page Layout :
All articles should be submitted in single column text on standard Letter Size paper (21.59 × 27.94 cm) with normal margins.

2. Length :
Manuscripts should normally be between 4,500 and 9,000 words (10 to 20 pages).

3. File Type :
Articles should be submitted in Microsoft Word format. To facilitate the manuscript preparation process and speed up the publication process, please use the manuscript template.

4. Text Style :
- Use a standard font (e.g., Times New Roman) no smaller than size 10.
- Use single line spacing for paragraphs.
- Use footnotes to provide additional information peripheral to the text. Footnotes to tables should be marked by superscript lowercase letters or asterisks.

5. Title Page :
The title page should start with a concise but descriptive title and the full names of authors along with their affiliations and contact information (i.e., postal and email addresses). An abstract of 150 to 250 words should appear below the title and authors, followed by keywords (4 to 6).

Author1
Affilliation, Postal Address. E-mail

Author2
Affilliation, Postal Address. E-mail

ABSTRACT
A brief summary (150-250 words) of the paper goes here.

Keywords : 4 to 6 Keywords, separated by commas.

6. Numbered Type :

1. INTRODUCTION
All articles should be submitted in single column text on standard letter size paper (21.59 × 27.94 cm) with normal margins[1 . Text should be in 11-point standard font (e.g., Times New Roman) with single line spacing.
[1 Normal margin dimensions are 3 cm from the top and 2.54 cm from the bottom and sides.

2. SECTIONS

The top-level section heading should be in 14-point bold all uppercase letters.

2.1. Subsection Heading 1

The first-level subsection heading should be in 12-point bold with the first letter of each word capitalized.

2.1.1. Subsection Heading 2

The second-level subsection heading should be in 11-point italic with the first letter of each word capitalized.

7. Figures and Tables :

All figures and tables should be placed at the end of the manuscript after the reference list. To note the placement of figures and tables in text, "Insert Table (or Figure) # here" should be inserted in appropriate places. Please use high resolution graphics whenever possible and make sure figures and tables can be easily resized and moved.

Figure

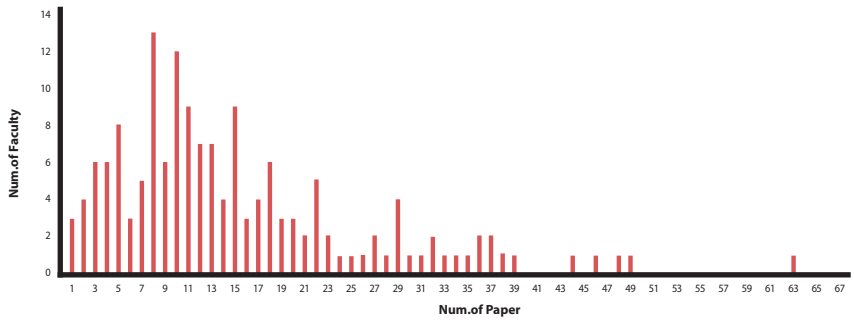


Fig. 1. Distribution of authors over publication count.

Table

Table 1. The title of table goes here

Study	Time period study	Data
Smith Wesson (1996)	**1970 - 1995**	684 papers in 4 SSCI journals
Reeves [a (2002)	**1997 - 2001**	597 papers in 3 SSCI journals
Jones Wilson [b (2011)	**2000 - 2009**	2,166 papers in 4 SSCI journals

[a Table footnote a goes here
[b Table footnote b goes here

8. Acknowledgements :

Acknowledgements should appear in a separate section before the reference list.

9. Citations :

Citations in text should follow the author-date method (authors' surname followed by publication year).

- Several studies found... (Barakat et al., 1995; Garfield, 1955; Meho & Yang, 2007).
- In a recent study (Smith & Jones, 2011)...
- Smith and Jones (2011) investigated...

10. Reference List :

Reference list, formatted in accordance with the American Psychological Association (APA) style, should be alpha-betized by the first authors last name.

Journal article

- Author, A., Author, B. & Author, C. (Year). Article title. *Journal Title*, volume(issue), start page-end page.
- Smith, K., Jones, L. J., & Brown, M. (2012). Effect of Asian citation databases on the impact factor. *Journal of Information Science Practice and Theory*, 1(2), 21-34.

Book

- Author, A., & Author, B. (Year). *Book title*. Publisher Location: Publisher Name.
- Smith, K., Jones, L. J., & Brown, M. (2012). *Citation patterns of Asian scholars*. London: Sage.

Book chapter

- Author, A., & Author, B. (Year). Chapter title. In A. Editor, B. Editor, & C. Editor (Eds.), *Book title* (pp. xx-xx). Publisher Location: Publisher Name.
- Smith, K. & Brown, M. (2012). Author impact factor by weighted citation counts. In G. Martin (Ed.), *Bibliometric approach to quality assessment* (pp. 101-121). New York: Springer.

Conference paper

- Author, A., & Author, B. (Year). Article title. In A. Editor & B. Editor (Eds.), *Conference title* (pp. xx-xx). Publisher Location: Publisher Name.
- Smith, K. & Brown, M. (2012). Digital curation of scientific data. In G. Martin & L. J. Jones (Eds.), *Proceedings of the 12th International Conference on Digital Curation* (pp. 41-53). New York: Springer.

Online document

- Author, A., & Author, B. (Year). Article title. Retrieved *month day*, *year* from URL.
- Smith, K. & Brown, M. (2010). The future of digital library in Asia. *Digital Libraries*, 7,111-119. Retrieved *May 5, 2010*, from http://www.diglib.org/publist.htm.

JISTaP

Journal of Information Science
Theory and Practice

http://www.jistap.org

KiSTi
www.kisti.re.kr
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea (ZIP code: 02456)
Tel. +82-2-3299-6102 Fax. +82-2-3299-6067 http://www.jistap.org