Natural Language Processing Applications in Library and Information Science

Preprint *in* Online Information Review · January 2019

DOI: 10.1108/OIR-07-2018-0217

Zehra Taskin
Adam Mickiewicz University

Umut Al
Hacettepe University

Natural Language Processing Applications in Library and Information Science

Zehra Taşkın[*] & Umut Al[**]

Abstract

Purpose: With the recent developments in information technologies, natural language processing (NLP) practices have made tasks in many areas easier and more practical. Nowadays, especially when big data are used in most research, NLP provides fast and easy methods for processing these data. The main objective of this paper is to identify subfields of library and information science (LIS) where NLP can be used and to provide a guide based on bibliometrics and social network analyses for researchers who intend to study this subject.

Design/methodology/approach: Within the scope of this study, 6,607 publications, including NLP methods published in the field of LIS, are examined and visualized by social network analysis methods.

Findings: After evaluating the obtained results, the subject categories of publications, frequently used keywords in these publications, and the relationships between these words are revealed. Finally, the core journals and articles are classified thematically for researchers working in the field of LIS and planning to apply NLP in their research.

Originality/value: The results of this study draws a general framework for LIS field and guides researchers on new techniques that may be useful in the field.

Introduction

Natural language processing (NLP) is a process of understanding how texts, speeches, and similar materials are used by computerized systems and how they are operated on computers (Chowdhury, 2003, p. 51). The Oxford Dictionary defines NLP as “the application of computational techniques to the analysis and synthesis of natural language and speech” (Natural Language Processing, 2017). The main goal of these applications is to realize a human-like language processing for several tasks or applications and to analyze the generated texts with computational techniques (Liddy, 2010, p. 3864). With similar applications, detailed linguistic analyses are possible, and large texts can easily be analyzed.

Although NLP approaches were initially applied for endangered languages to prevent their extinction, these approaches have been recently used in many studies to organize and make sense of big data. Currently, it would be more difficult and time-consuming to work without NLP many fields, including marketing, information validation, and information retrieval or visualization. The main aim of this study is to provide

[*] Hacettepe University, Department of Information Management (iSchool), Turkey.
https://orcid.org/0000-0001-7102-493X | http://www.bby.hacettepe.edu.tr/akademik/zehrataskin/
[**] Hacettepe University, Department of Information Management (iSchool), Turkey.
http://yunus.hacettepe.edu.tr/~umutal/

information about the main stages of NLP applications and to provide a detailed bibliometric approach on the literature about NLP published in the field of library and information science (LIS). In this respect, potential subjects, which can employ different NLP methods, can be revealed, and ideas about future trends can be obtained. The main research questions are as follows:

- What are the main NLP applications used in LIS literature? What is the historical evolution of NLP subjects?
- Which journals should be followed by researchers new to NLP, and how is the distribution of these journals according to the subtopics of NLP applications?
- What are the prominent publications on NLP for LIS?

In this context, basic features of NLP applications are first presented, and then works of the literature are gathered and evaluated to reveal the main characteristics of the field.

Natural Language Processing

Levels of NLP Tasks

Natural language processing is divided into seven basic levels (Fig.1), from the simplest to the most difficult (Feldman, 1999, p. 62-64; Liddy, 2010, p. 3867-3868). In this manner, NLP tasks are carried out for many studies, and large texts in any language can easily be analyzed.

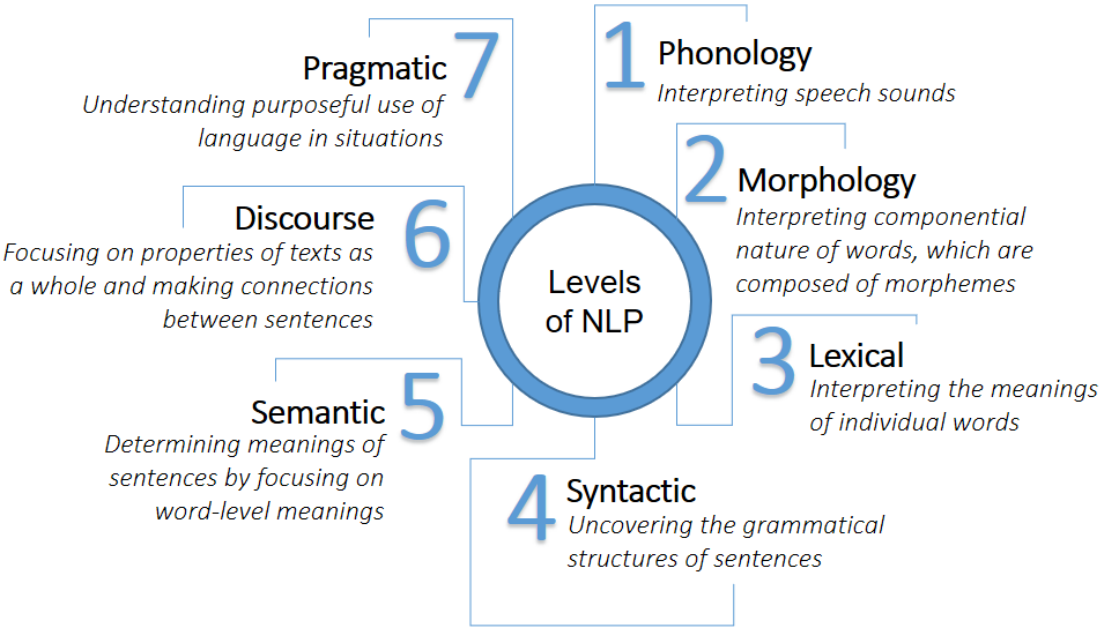

Levels of NLP
1 Phonology
Interpreting speech sounds
2 Morphology
Interpreting componential nature of words, which are composed of morphemes
3 Lexical
Interpreting the meanings of individual words
4 Syntactic
Uncovering the grammatical structures of sentences
5 Semantic
Determining meanings of sentences by focusing on word-level meanings
6 Discourse
Focusing on properties of texts as a whole and making connections between sentences
7 Pragmatic
Understanding purposeful use of language in situations

Figure 1. Seven levels of NLP (Feldman, 1999, p. 62-64; Liddy, 2010, p. 3867-3868).

Studies carried out at a phonetic level are designed to understand spoken languages or pronunciations. In the literature, various studies have been conducted on languages to prevent their extinction (Çarkı, Geutner, & Schultz, 2000; Schultz & Waibel, 2001; Shen, Wu, & Tsai, 2015). The next level, morphology, deals with the smallest parts of words, morphemes. Using morphology, the main features of languages can be revealed, and the roots and suffixes of words can be discriminated. For example, similar applications developed for Turkish language, which is one of the agglutinative languages in the world, are included in the literature (Eryiğit, 2014; Zemberek NLP, 2015). The main aim of the lexical level is to understand the meanings of words. The fourth level (syntactic) uncovers the grammatical structures of sentences, and several studies have been conducted on it (e.g., Arısoy, Saraçlar, Roark, & Shafran, 2010), while the semantic level aims to reveal the meanings of words. These two levels (semantic and syntactic) are combined for discourse level studies. The most complicated level is the pragmatic level, which aims to understand the purposeful use of languages in situations. To achieve NLP tasks successfully, it is necessary to investigate all or a few of these levels. Each level is a continuation of the previous level; that is, while the simplest NLP task is performed at the phonetic level, the most complicated/detailed task is performed at the pragmatic level.

NLP Applications

Since there are various purposes for processing natural language texts, many applications have been developed for each purpose. The most common NLP applications in the literature are as follows:

- *Information Retrieval:* With the recent increase in the amount of information, the application of NLP to access meaningful information has become important. The main purpose of this application, which is defined as “document retrieval” or “text retrieval” in various studies, is to enable individuals’ access any section of a paragraph, book, or large-scale text (Lewis & Jones, 1996, p. 92).
- *Information Extraction:* The application of information extraction is based on the identification, labeling, and extraction of key elements, such as people’s names, institutions, locations, and countries from large texts (Liddy, 2010, p. 3871). Information extraction may be used with other NLP applications because extracted data may form the basis of complex NLP tasks (Blake, 2013, p. 129).
- *Machine Translation:* This application is based on the automatic translation of texts or speeches from one language to another. This practice aims to integrate different cultures, remove language barriers, and preserve ancient languages (Manning & Schütze, 1999, p. 463).
- *Summarization:* Summarization systems are based on using some linguistic or statistical methods to choose the most important words or phrases from sentences or paragraphs in large texts, and creating a meaningful summary to represent the text (Chowdhury, 2003, p. 60).
- *Text Categorization:* These applications are predictive models used for various purposes such as image or pattern recognition, weather forecasting, and disease diagnostics (Silahtaroğlu, 2013, p. 67). Categorization techniques are based on various classes depending on the basic text features (Blake, 2013, p. 136).

NLP and Bibliometrics

Studies in which NLP methods and bibliometric analyses are performed together have been very common in recent years. These studies can be categorized into two groups: studies in which NLP applications are used to increase performance of bibliometric studies and studies in which NLP articles are analyzed by bibliometric methods.

In the literature, the abovementioned NLP applications have all been used to improve the performance of bibliometric studies. Information extraction, which is the most frequently used application, is used to access the names of authors, institutions, and countries of publications, especially when there are standardization problems in citation indexes (Galvez & Moya-Anegón, 2007; Hooper, Neves, & Bordea, 2015; Taşkın & Al, 2013; Taşkın & Al, 2014). In these cases, automated information extraction processes can be carried out with NLP software such as Nooj, Saffron, and Intex. In addition, information extraction algorithms have been developed for detecting citation sentences from full texts (Kim, Le, & Thoma, 2014; Taşkın, Al, & Sezen, 2017). In a recent study, it was also stated that an NLP-based extraction system has been used to analyze the interpretation of a specific term (Web of Science) in a dataset (Li, Rollins, & Yan, 2018).

The approach of pennant diagrams, developed in the center of relevance and bibliometrics and based on information retrieval, provides a convergence between information retrieval applications and bibliometrics. Currently, these diagrams developed by Howard White (2007) are frequently used for bibliometric studies (e.g., Akbulut, 2016; Carevic & Mayr, 2014; White, 2018).

There is a strong relationship between bibliometrics and information retrieval when citation indexes are the main data source for bibliometric studies. The quality of data in citation indexes can increase the success of these studies. This strong connection between the concepts, bibliometrics, and information retrieval has facilitated the organization of new events that address this connection. The intersections of bibliometrics and information retrieval are presented in detail at the International Society for Scientometrics and Informetrics conference in Vienna, 2013, with the theme “Combining Bibliometrics and Information Retrieval.” Its selected papers are published in *Scientometrics* in 2015 (Mayr & Scharnhorst, 2015). Similarly, the Joint Workshop on Bibliometric-enhanced IR and NLP for Digital Libraries (BIRNDL) event, which has been held regularly for three years as part of the Association for Computing Machinery's Special Interest Group on Information Retrieval (ACM SIGIR) conferences, also provides platforms to present works on bibliometric-enhanced information retrieval for authors working in the field (BIRNDL2018, 2018).

Summarization is one of the subjects of some bibliometric studies in the literature. For instance, one study focused on creating a summary of a field by using different publications in the dataset (Nanba & Okumura, 1999). In a similar study, text summarization was performed using co-citation clusters, and as result, the automatic summarization system significantly outperformed the baseline in both metrics (Fiszman, Demner-Fushman, Kılıçoğlu, & Rindflesch, 2009). In another study, domain summarization was successfully performed as a result of co-citation analyses (Chen, Ibekwe-SanJuan, & Hou, 2010).

Semantic applications are often used for analyzing research results or determining the meaning of citations. For instance, the bibliometric and semantic nature of negative results publications was examined in a study, wherein negative expressions were determined using Nooj software (Gumpenberger, Gorraiz, Wieland, Roche, Schiebel, Besagni, & François, 2013). In another study, a citation-weighting proposal was introduced using semantic analysis applications for computational bibliometrics domain (Avram, Velter, & Dumitrache, 2014). Semantic and syntactic analyses are often preferred for content-based citation analyses, which may replace the traditional citation counting. The main aim of studies using semantic and syntactic features of citations is to develop new-generation citation indicators using NLP techniques (e.g., Adedayo, 2015; Catalini, Lacetera, & Oettl, 2015; Jha, Jbara, Qazvinian, & Radev, 2016; Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, 1998; Taşkın & Al, 2018). In addition, the lexical level of NLP is preferred for nano-bibliometric studies in the literature (e.g., Glänzel, Heeffler, & Thijs, 2017).

All the above studies are applications of NLP meant to enhance the quality of bibliometric studies. However, in the literature, papers on NLP have been analyzed using bibliometric methods. One of these works is the bibliometric analysis of NLP studies in health field (Chen, Xie, Wang, Liu, Xu, & Hao, 2018). In this study, sub-themes affecting the field were determined as computational biology, terminology mining, information extraction, text classification, and information retrieval. In another study that examined clinical NLP studies, it was found that the most frequently used text types are patient-authored texts, transcribed audios, and online encyclopedic resources (Névéol & Zweigenbaum, 2016). In another study, papers on data mining published in a conference proceedings book were analyzed (Mikova, 2016). In this study, the emerging words were identified as web scraping, ontology modeling, advanced bibliometrics, semantics, and sentiments analysis. In a review, publications on NLP were examined in depth, and it was found that the future of NLP lies in the biologically and linguistically motivated computational paradigms that enable narrative understanding and hence sense making (Cambria & White, 2014).

Although there are a large number of studies on bibliometric and information science using NLP applications in the literature, no bibliometric analysis has been made on these publications. This work aims to fill this gap and provide a roadmap for new researchers.

Data and Methods

In this study, LIS studies on NLP and its various applications were examined in depth by the Web of Science's core indexes using a data collection process. The first step was to create the dataset using the query below:

> *TS=("natural language processing" OR "text categorization" OR "information retrieval" OR "machine translation" OR "information extraction" OR summarization) OR TI=("natural language processing" OR "text categorization" OR "information retrieval" OR "machine translation" OR "information extraction" OR summarization)) AND WC="Information science & Library Science"*
>
> *Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI*

The query to create the dataset was performed on the Web of Science on April 19, 2018, and we realized 6,607 publications. These publications were downloaded in tab delimited (.txt) format and corrections/unifications were made on author (AU), citation (CR), source (SO), keyword (DE & ID), and abstract (AB) fields (Clarivate Analytics, 2018) to ensure data quality and standardization. The steps carried out in the data standardization process were as follows:

- The authors with two names/surnames and different authors with the same names in the datasets causes consistency problems for social network analysis studies. Therefore, we conducted name standardization process for author names. In this context, the names of the authors were listed and problematic author names were unified.
- In some cases, identifying and analyzing citations of some publications that have not a DOI number can create consistency problems for co-citation analyses. In order to ensure consistency in the most cited publications in our dataset, the entire list of cited papers was obtained and repetitive publications were unified.
- Journal name changes may also affect social network analyses negatively because the journals, which changed their names, can be represented as different journals on the maps. In this context, even if the journals were renamed, it was standardized under the last names of the journals to ensure consistency.
- The most important consistency problem in keyword or co-occurrence maps is conducting analyses without parsing of keywords in titles or abstracts. If the standardization process for keywords is not conducted before analysis, it is possible to encounter many words that are duplicated in keyword maps for reasons such as plural suffixes, word phrases and verb inflections. In this context, all the keywords were simplified to the most basic form for analysis. In addition, word phrases were identified to provide consistency.

Data was made ready for analysis with this standardization process carried out for the dataset. To mention the features of the publications in our dataset, 65% of publications are articles, 18% are proceedings and 9% are book reviews. The publications spread across 17 different document types. The distribution of publications and citations by year is shown on Fig. 2.

Two different kinds of visualization software were used to visualize the evolution of the NLP literature and its changes over time. While VOSviewer[1] software is used for drawing a general framework of the field, CiteSpace[2] software is used for visualizing trends for years. The Vosviewer software works with tab-delimited text and CiteSpace software with field delimited text file formats. Therefore, standardized data files were converted into formats that meets software requirements. The use of both software is different from each other and thus, manuals of them may be examined for detailed information (Chen, 2018; van Eck, & Waltman, 2018).

[1] http://www.vosviewer.com/

[2] http://cluster.cis.drexel.edu/~cchen/citespace/

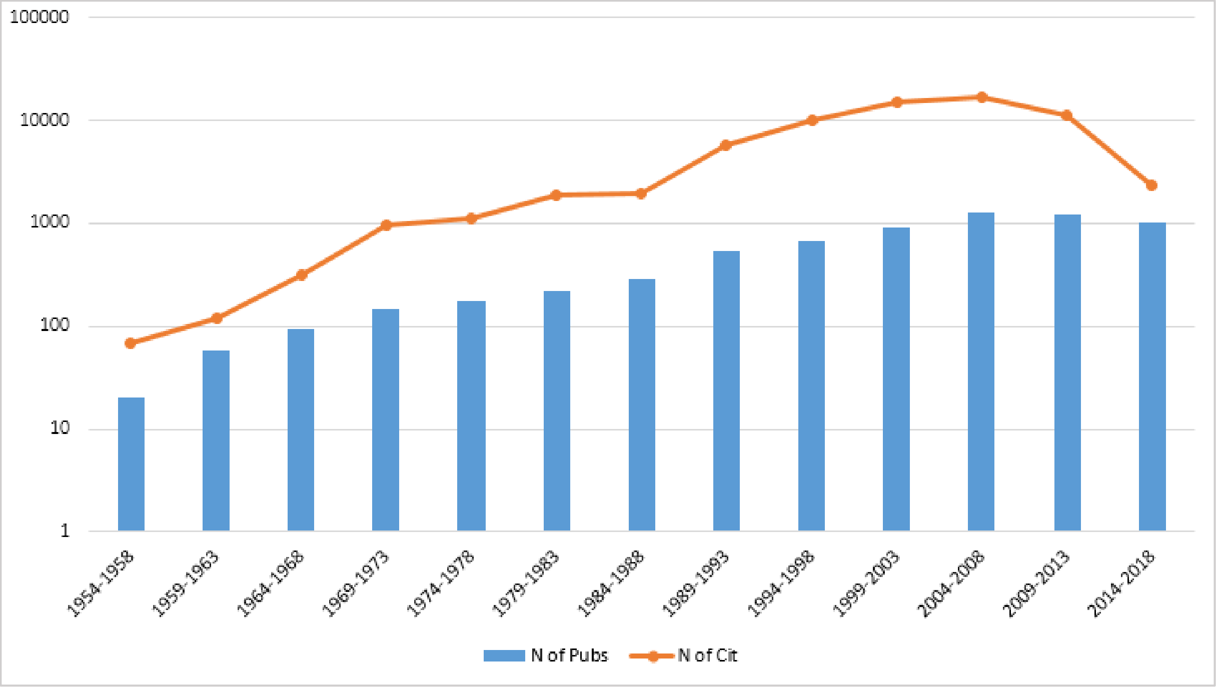


Figure 2. Distribution of articles and their citations by year (logarithmic scale).

Findings

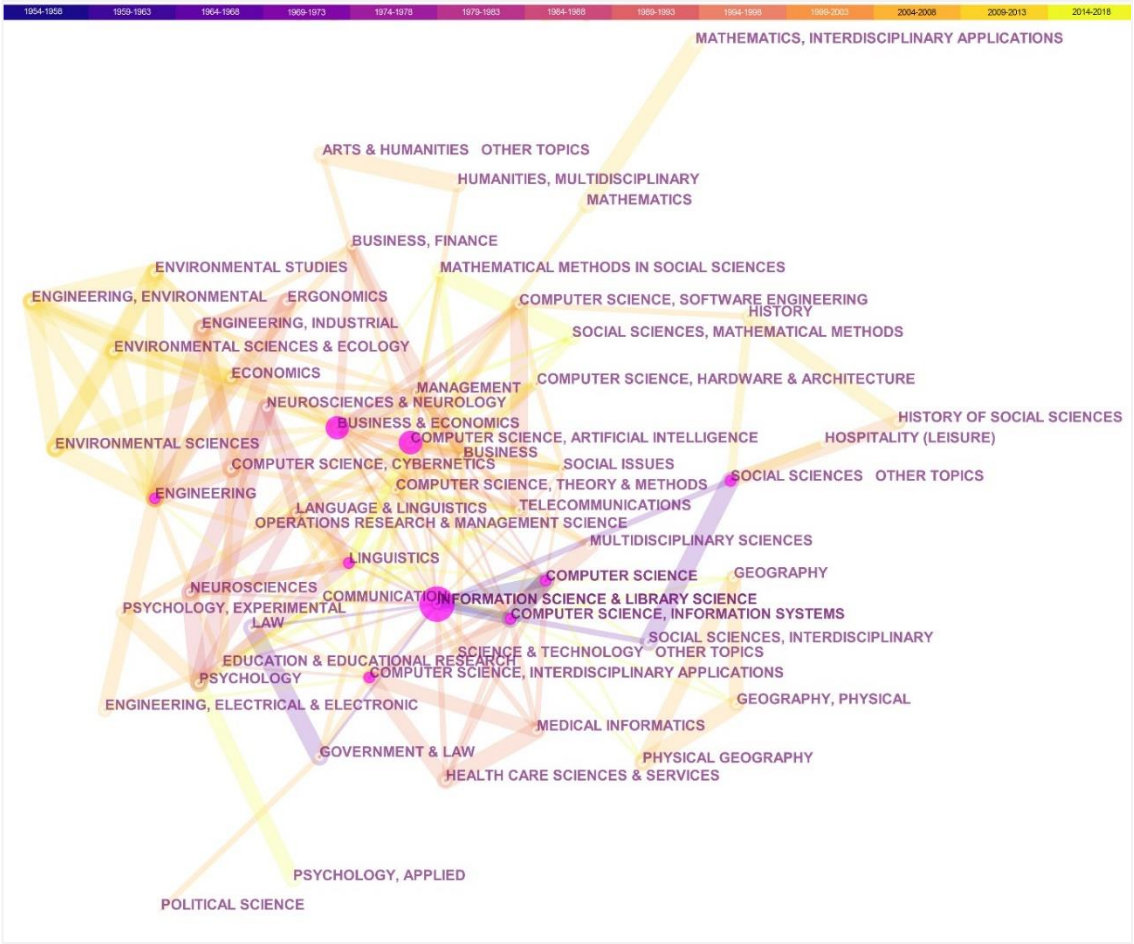
Convergence of Categories

The subject categories of the Web of Science were used to find the distribution of studies in the field of NLP and to show their convergence. Regarding the number of publications, the most centralized and frequent category is naturally "Information Science & Library Science" because the dataset is limited by LIS. According to the centralities, LIS is followed by "Business & Economics" (0.40) and "Computer Science, Artificial Intelligence" (0.35). The category with the second highest frequency of publications is "Computer Science" (4,737 publications), and it is followed by "Computer Science, Information Systems" with 4,576 publications. In Fig. 3, the relationships of each subject categories for years can be observed.[3]

Five-year intervals were used to draw the timeline, and all the items for each slice are shown in Fig. 3. The distribution of the categories can be evaluated in four basic clusters, which are NLP, decision support system, dictionary based systems, and recommender systems. Cluster names were defined by CiteSpace using the most common words in publications. According to the map, studies in the field of computer science and information science began in 1954. In the 1960s, it converged to only the fields of law and social sciences. Until the 1980s, no other subject areas were seen in the map. By the 1980s, it became clear that the applications of NLP began in education, health sciences, management, and linguistics. In addition, NLP techniques have been used in recent years for environmental studies such as environmental risk assessment. It can also be traced from the network map that, in recent years, NLP techniques have been

[3] The extended version of the network, which includes divisions of subject clusters, is available at https://goo.gl/6FaVRk

applied in history, psychology, and art. This proves that the NLP, which is thought to be related to only computer science, actually has an interdisciplinary structure and can be used in many fields from law to history. This also indicates that the transdisciplinary structure of the field will continue to increase in the following years, as studies on NLP have been carried out in 58 of the 252 Web of Science subject categories.



1954-1958
1959-1963
1964-1968
1969-1973
1974-1978
1979-1983
1984-1988
1989-1993
1994-1998
1999-2003
2004-2008
2009-2013
2014-2018
MATHEMATICS, INTERDISCIPLINARY APPLICATIONS
ARTS & HUMANITIES OTHER TOPICS
HUMANITIES, MULTIDISCIPLINARY
MATHEMATICS
BUSINESS, FINANCE
ENVIRONMENTAL STUDIES
MATHEMATICAL METHODS IN SOCIAL SCIENCES
ENGINEERING, ENVIRONMENTAL
ERGONOMICS
COMPUTER SCIENCE, SOFTWARE ENGINEERING
HISTORY
ENGINEERING, INDUSTRIAL
SOCIAL SCIENCES, MATHEMATICAL METHODS
ENVIRONMENTAL SCIENCES & ECOLOGY
ECONOMICS
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE
MANAGEMENT
NEUROSCIENCES & NEUROLOGY
BUSINESS & ECONOMICS
HISTORY OF SOCIAL SCIENCES
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE
HOSPITALITY (LEISURE)
ENVIRONMENTAL SCIENCES
BUSINESS
COMPUTER SCIENCE, CYBERNETICS
SOCIAL ISSUES
SOCIAL SCIENCES OTHER TOPICS
COMPUTER SCIENCE, THEORY & METHODS
ENGINEERING
LANGUAGE & LINGUISTICS
TELECOMMUNICATIONS
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
MULTIDISCIPLINARY SCIENCES
LINGUISTICS
GEOGRAPHY
COMPUTER SCIENCE
NEUROSCIENCES
COMMUNICATION
INFORMATION SCIENCE & LIBRARY SCIENCE
PSYCHOLOGY, EXPERIMENTAL
COMPUTER SCIENCE, INFORMATION SYSTEMS
LAW
SOCIAL SCIENCES, INTERDISCIPLINARY
SCIENCE & TECHNOLOGY OTHER TOPICS
EDUCATION & EDUCATIONAL RESEARCH
PSYCHOLOGY
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
GEOGRAPHY, PHYSICAL
ENGINEERING, ELECTRICAL & ELECTRONIC
MEDICAL INFORMATICS
GOVERNMENT & LAW
PHYSICAL GEOGRAPHY
HEALTH CARE SCIENCES & SERVICES
PSYCHOLOGY, APPLIED
POLITICAL SCIENCE

Figure 3. Distribution of publications to the subject categories and their relations.

Title and Keyword Analysis

By analyzing various words used in the different sections of scientific articles such as keywords, abstracts, titles, and full texts, the word maps of the fields can be revealed, and thematic links between studies can be obtained. Keywords are considered as the main elements for analyses in some studies (Ding, Chowdhury, & Foo, 2001; Sue & Lee, 2010), while the words in the abstracts and titles are used to expand the scope of some studies (Rotto & Morgan, 1997; Sedighi, 2016). Rotto and Morgan (1997, p. 101) stated that co-word analyses should include abstracts because specific research can only be revealed in this way. In this study, co-word analysis was carried out using titles and abstracts of NLP publications. For this analysis, singular/plural words and synonyms were parsed to standardize the words. Then, the abbreviations were unified, to ensure that each word is included only once in the co-word map. The network map creation after these steps is as shown in Fig. 4.[4]

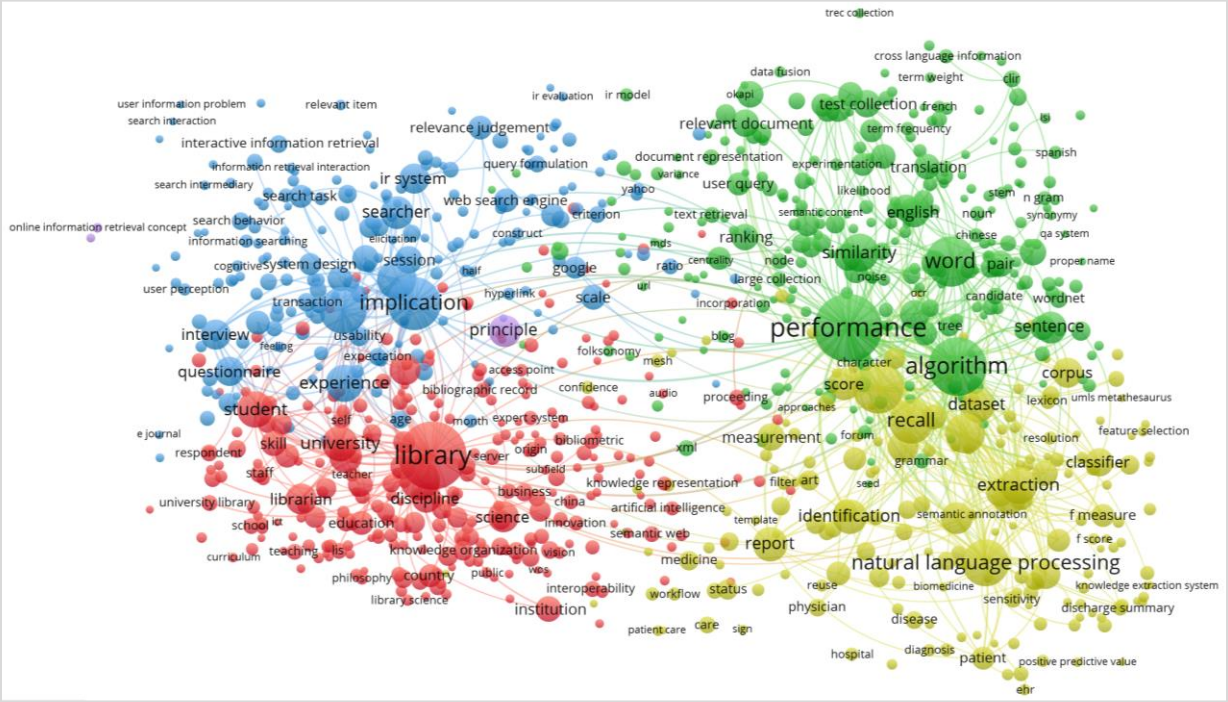


Figure 4. Co-word analysis of title and abstracts (for interactive map, please visit *https://goo.gl/GspS9C*).

VosViewer determined four main clusters[5] in the co-word analysis, and 289 nodes were determined for the first cluster (red), 269 for the second (green), 201 for the third (blue), and 185 for the fourth (yellow). To name these clusters in terms of thematic distribution, the first cluster can be considered to contain the basic/traditional subjects of LIS. The most used and strongest term in the cluster is “Library,” which has strong connections with the other clusters (total link strength: 4,928; links: 808) followed by “student,” “university,” and “information science” terms. The most used words in bibliometric studies such as “citation analysis,” “bibliometrics,” and “citation data” are also in this cluster. In addition, concepts such as bibliography, documentation, and knowledge management, which are traditional LIS subjects, can be traced in this cluster. The use of NLP techniques or applications in traditional subjects is important as it shows that all the areas of the LIS field are harmonized to technological innovations along with their subfields.

Considering the second cluster, the word “performance” is not only the strongest node of its own cluster but of the whole map (total link strength: 5,995; links: 829). “Performance” seems to have strong connections with other clusters according to the importance of performance evaluations in LIS and NLP. “Algorithm,” “word,” and “similarity” follow “performance.”

The third cluster includes keywords on human information behaviors, information retrieval models, and user experiences. The strongest words are “implication,” “behavior,” and

[4] Co-word map is also created by CiteSpace to visualize changes over time. The map is accessible on the link: https://goo.gl/ENep6B

[5] Five clusters are shown on the map, however, pink cluster includes only three keywords and these keywords do not have strong links to others. Therefore, it is excluded from the analysis.

"participant." Many elements in this cluster are often related to the end user experiences, such as user interaction and usability. This cluster converges to the red cluster because information behaviors are one of the fundamental subtopics of the LIS domain.

The last cluster contains key applications, algorithms, and techniques used for performance evaluations in NLP studies. The most powerful keywords of this cluster are "precision," "recall," and "NLP." It can be traced from the map that this cluster is close to the green cluster containing the performance evaluation. Therefore, the green and yellow clusters can be correctly evaluated together and classified under the title "NLP and performance evaluations."

Similar results were obtained in the map produced by CiteSpace[6], which shows the changes of keywords and their relationships with each other over the years. Studies that began with "information retrieval" at the end of 1950s focused only on computer technology until the 1990s; however, after the 90s, NLP studies cover almost all areas of information science. CiteSpace identified seven main clusters. The first cluster is named as "web search engines" by CiteSpace, and it comprises studies covering information retrieval on the web. This cluster is close to the second cluster named "full text documents," and thus, these two clusters are considered as one. The works in these clusters continue from the 50s, and it is expected to continue in the future. The most important observation about this cluster is social media, which started in 2010, and the more recent information literacy issues. It can be inferred that while the previous problem of information on the web is "accessing information," the current problem is "accessing meaningful and correct information."

Keywords, which refer to human behaviors that affect information retrieval, are shown on the "collaborative information behaviors" cluster. Recent words are generally focused on social media just as in the previous cluster.

It is possible to track keywords on online catalogs and databases subjects, which are the results of the use of online systems in libraries, from the "adaptive information retrieval system" cluster. This cluster is convergent to the LIS domain, and it includes issues focused on information retrieval in libraries.

Works on health information retrieval help health subjects to create a cluster of health issues on the map. The keywords related to the methods and techniques for accessing health information can be seen from the "clinical narrative" cluster. Perhaps, the main reason for the formation of a separate cluster for health field in this map is the journals (e.g., *Qualitative Health Research*, *Journal of Health Communication* and *Health Information, and Libraries Journal)* that are indexed in the Web of Science under the LIS topic and are focused on health issues.

In the "linking libraries" cluster, where the NLP techniques are used in bibliometric studies, there are words such as "information representation," "citation analysis," and "bibliometrics." The "partial relevance judgment" cluster, which deals with information problems and is included in the "collaborative information behavior" cluster after a 10-year period, seems to focus on the end user experience and information retrieval processes.

[6] https://goo.gl/MRK9Rd

Journal Co-Citation Analysis

Co-citation analyses aim to reveal similarities to present general structures of research domains. Journal co-citation analyses help to develop core journal lists that are specific to the domains, reveal the expertise of the areas through published literature, and develop effective collections for users (McCain, 1991, p. 291). The aim of this study is to suggest a core journal collection for researchers and librarians working on the field of LIS and using NLP techniques in their works. In this context, the most frequently cited journals[7] and the links between these journals are shown in Fig. 5. All journal titles are standardized so that journal title changes do not affect the consistency of the journal co-citation map.

As seen in Fig. 5, CiteSpace determined six clusters for journal co-citation map, which are "information retrieval," "seeking context," "information retrieval system design," "information retrieval: Health," "authorship attribution," and "library’s role." The “information retrieval” cluster starts with *Information Retrieval Journal*. Association for Computing Machinery has a significant impact in this cluster. Several ACM journals can be traced in the cluster. It is important for researchers conducting investigations on information retrieval to follow ACM publications and ACM SIGIR conferences to be informed about latest developments in their fields. In the cluster #3 (information retrieval), there are journals that focus on information retrieval in the health sciences. Researchers working in the field of health information may benefit from these journals.

The second cluster containing journals on information behaviors starts with *The Journal of the Association for Information Science and Technology*.[8] This journal is also connected with the information retrieval cluster. Other journals in this cluster are *Online Information Review, Aslib Proceedings, Library Quarterly*, and *the Annual Review of Information Science*, which cover key topics in LIS.

Considering cluster #2, researchers seeking publications on information retrieval system design for information centers may follow LIS journals such as *Electronic Library*, *Library Trends*, *Journal of Academic Libraries*, and *Cataloging and Classification Quarterly*.

The “authorship attribution” cluster includes journals on the subject of bibliometrics, such as *Scientometrics* and *Journal of Informetrics* as well as interdisciplinary journals such as *Science* and *Nature*. It seems that cluster #4 is a cluster that includes bibliometrics studies using NLP techniques. Researchers conducting bibliometric studies may follow the journals listed in this cluster.

Although the last cluster is identified by CiteSpace as “library’s role,” this cluster includes journals on computer science and management indexed in LIS subject in the Web of Science. Interdisciplinary researchers who work for computer science and LIS may benefit from the journals listed in the cluster.

[7] Top 1% of cited journals are selected for visualization.

[8] The name of the journal was *American Documentation* between 1956 and 1969.

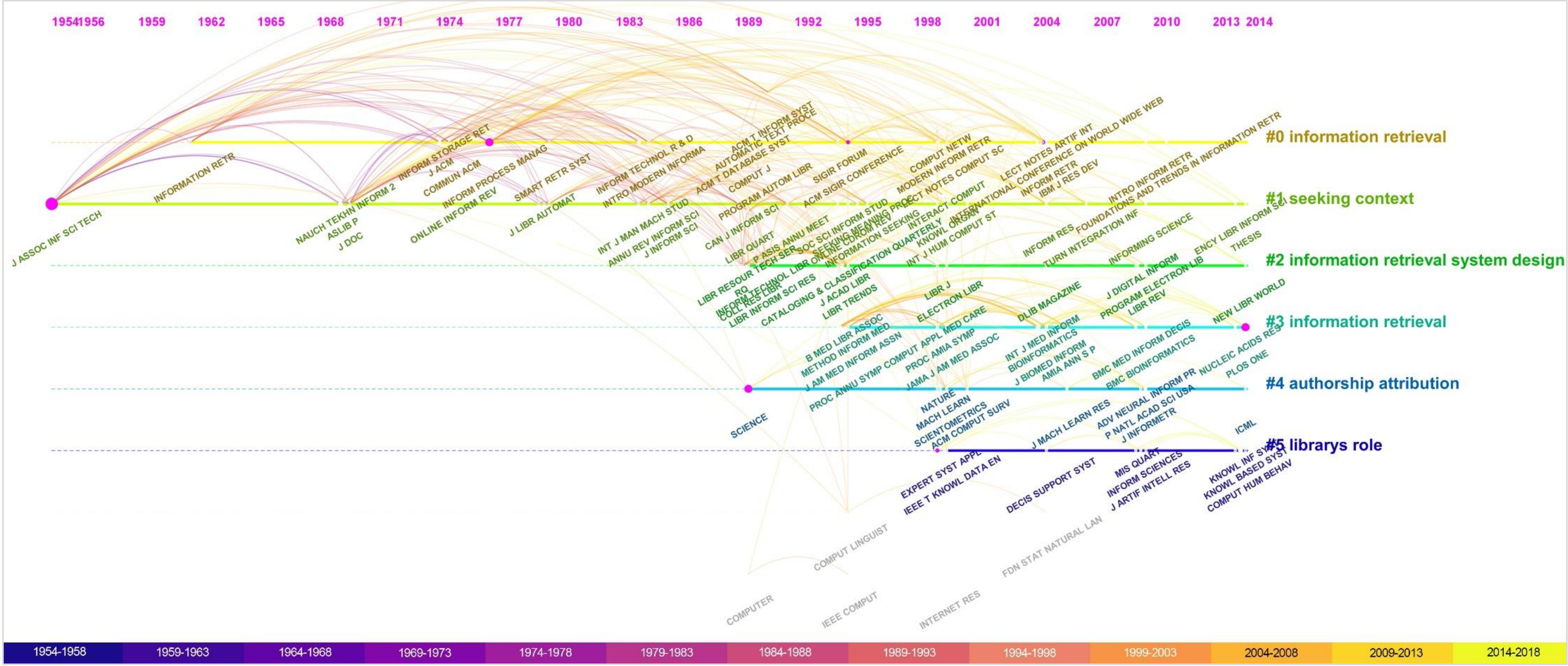


Figure 5. Journal co-citation map of NLP papers (*https://goo.gl/fdULTP*).

Document Co-Citation Analysis

Document co-citation analysis method, which is primarily based on identifying pairs of highly cited papers (Garfield, 2001, p. 2), may be useful for expediting knowledge and building consilience across disciplines (Trujillo & Long, 2018). The most co-cited articles by NLP papers on LIS subject is shown in Fig. 6.

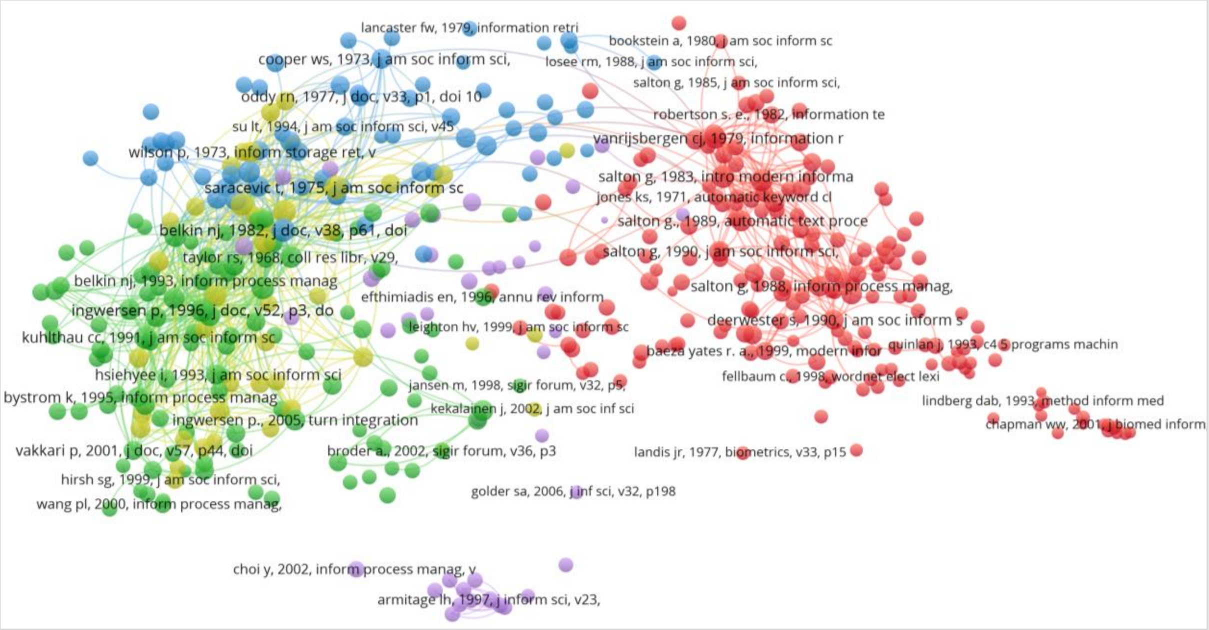

lancaster fw, 1979, information retri
bookstein a, 1980, j am soc inform sc
cooper ws, 1973, j am soc inform sci,
losee rm, 1988, j am soc inform sci,
salton g, 1985, j am soc inform sci,
oddy rn, 1977, j doc, v33, p1, doi 10
robertson s. e., 1982, information te
su lt, 1994, j am soc inform sci, v45
vanrijsbergen cj, 1979, information r
wilson p, 1973, inform storage ret, v
salton g, 1983, intro modern informa
saracevic t, 1975, j am soc inform sc
jones ks, 1971, automatic keyword cl
salton g., 1989, automatic text proce
belkin nj, 1982, j doc, v38, p61, doi
salton g, 1990, j am soc inform sci,
taylor rs, 1968, coll res libr, v29,
belkin nj, 1993, inform process manag
salton g, 1988, inform process manag,
efthimiadis en, 1996, annu rev inform
ingwersen p, 1996, j doc, v52, p3, do
deerwester s, 1990, j am soc inform s
leighton hv, 1999, j am soc inform sc
kuhlthau cc, 1991, j am soc inform sc
quinlan j, 1993, c4 5 programs machin
baeza yates r. a., 1999, modern infor
hsiehyee i, 1993, j am soc inform sci
fellbaum c., 1998, wordnet elect lexi
jansen m, 1998, sigir forum, v32, p5,
bystrom k, 1995, inform process manag
lindberg dab, 1993, method inform med
kekalainen j, 2002, j am soc inf sci
ingwersen p., 2005, turn integration
chapman ww, 2001, j biomed inform
vakkari p, 2001, j doc, v57, p44, doi
broder a., 2002, sigir forum, v36, p3
landis jr, 1977, biometrics, v33, p15
hirsh sg, 1999, j am soc inform sci,
golder sa, 2006, j inf sci, v32, p198
wang pl, 2000, inform process manag,
choi y, 2002, inform process manag, v
armitage lh, 1997, j inform sci, v23,

Figure 6. Document co-citation map of mostly cited papers (for interactive map, please visit https://goo.gl/yiw5ix).

According to the map created by VosViewer, five different clusters were determined. The first cluster (red) includes 179 publications on information retrieval subject. The book entitled "Introduction to Modern Information Retrieval" (Salton, 1983) is the strongest node in this cluster with 304 links and 1695 total link strength. It is followed by "Information Retrieval" (Van Rijsbergen, 1979) and "Relevance Weighting of Search Terms" (Robertson & Sparck-Jones, 1976). Publications in this cluster can serve as reference guides for authors who work in information retrieval domain.

The cluster defined as the green color contains 107 publications on human information behaviors. The most co-cited paper in this cluster is "Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory" with 316 links and 1997 total link strength (Ingwersen, 1996). It is followed by the papers "The Design of Browsing and Berrypicking Techniques for the Online Search Interface" (Bates, 1989) and "Inside the Search Process: Information Seeking from the Users Perspective" (Kuhlthau, 1991).

The blue cluster contains 62 papers published on topics of human information behavior and information retrieval. The most important difference of this cluster from the green one is the similarity of blue and red cluster subjects. The most frequently co-cited paper of this cluster is "Ask for Information Retrieval: Background and Theory" by Belkin, Oddy, and Brooks (1982).

Forty-four publications are shown on the yellow cluster. The cluster includes articles published in the main LIS subfields. The green cluster containing the information behavior issues resembles the yellow one from the point at which it contains similar subjects. The most commonly cited title is “Relevance: Review of and a Framework For Thinking on Notion in Information Science” (Saracevic, 1975).

The last cluster contains 41 publications and is defined with a pink color. The publications in this cluster seem to be focused on the philosophy of librarianship and information retrieval. The most frequently co-cited publication is “Language and Representation” (Blair, 1990).[9]

Discussion

Natural language processing can be used to understand and evaluate information effectively, and it became especially useful with the continuous rise in the amount of information. The main purpose of this study is to reveal opportunities to use NLP techniques in LIS field as well as to draw a framework for early career researchers who intend to conduct studies on LIS field using NLP techniques. In accordance of this purpose, this study focused on finding answers to three basic research questions, which were discussed below.

The first research question was designed to determine which natural language processing applications were generally used by LIS field. According to the findings, the first NLP studies in the field of LIS were based on the subject of information retrieval and information retrieval systems. Besides, it was found that NLP applications were used in many subjects in LIS from traditional librarianship to next-generation practices. The main research topics of NLP applications in LIS are user experiences, information behaviors, bibliometrics and information systems design.

The second research question aimed to determine the most frequently cited journals in articles that used NLP applications in LIS field. It was found that preferred journals for citation varied according to the subjects of the studies. According to the results, ACM is the most important organization for people working for information retrieval subject in LIS. *JASIS&T* and *Online Information Review* journals may be followed for information behavior subjects; *Electronic Library* and *Library Trends* for designing effective information retrieval systems; *Scientometrics* and *Journal of Informetrics* for bibliometric studies; and journals of *IEEE* and *MIS Quarterly* for interdisciplinary issues.

In our study, it was aimed to determine the core publications in order to provide guideline for researchers who are new to the field. As result, the core publications were classified in certain subject categories such as information retrieval and philosophy of librarianship. Through the publications in each subject classes focusing on information retrieval, information behaviors, information seeking and philosophy of LIS, it is possible to follow the reference publications of the field.

[9] For a complete list of co-cited papers and their distribution to the clusters, please follow this link: https://goo.gl/hPKF7W

Conclusion

In this study, we aimed to reveal main characteristics of NLP-based LIS publications by using the benefits of social network analysis and bibliometrics. In recent years, the increase in the amount of information and difficulties in information processing have led to the use of NLP practices in LIS, which is also an information-based field. Although it is thought to be a subfield of computer science, NLP techniques have been widely used in LIS for the past 10 years. In addition, NLP techniques have made it possible to perform more tasks with less human power in the field of LIS. Because of their usefulness, it is important to provide ideas about next-generation methods to people working in LIS field. This study also revealed that NLP is a transdisciplinary subject for LIS studies. NLP is not only used in technical studies, but also in studies on library philosophy. Moreover, it was revealed that NLP methods have been used in many LIS works in the literature, and through this study, the general framework of these studies in the field of LIS has been drawn.

The most important benefit of this study is the guidance it provides to early-career researchers who work in LIS field. Through the obtained results, core journals and the most influential publications were identified and relations of these publications were revealed. In addition, this study provides a detailed literature review on which NLP applications can be used in the future studies, which journals can be followed and which core publications can be read.

Acknowledgement

This article was supported in part by a research grant from the Turkish Scientific and Technological Research Center (115K440).

References

Adedayo, A.V. (2015). Citations in introduction and literature review sections should not count for quality. *Performance Measurement and Metrics, 16*(3), 303-306.

Akbulut, M. (2016). *Atıf klasiklerinin etkisinin ve ilgililik sıralamalarının pennant diyagramları ile analizi [The analysis of the impact of citation classics and relevance rankings using pennant diagrams].* Unpublished master’s thesis, Hacettepe University, Ankara.

Arısoy, E., Saraçlar, M., Roark, B. & Shafran, I. (2010). Syntactic and sub-lexical features for Turkish discriminative language models. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010 (pp. 5538-5541). Dallas, TX: IEEE.

Avram, S., Velter, V. & Dumitrache, I. (2014). Semantic analysis applications in computational bibliometrics. *Control Engineering and Applied Informatics, 16*(1), 62-69.

Bates, M.J. (1989). The Design of Browsing and Berrypicking Techniques for the online search interface. *Online Review, 13*(5), 407-424.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982). Ask for information-retrieval .1. Background and theory. *Journal of Documentation, 38*(2), 61-71.

BIRNDL2018. (2018). 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018). http://wing.comp.nus.edu.sg/~birndl-sigir2018/

Blair, D.C. (1990). Language and representation. *Annual Review of Information Science and Technology, 44*(1), 159-200.

Blake, C. (2013). Text mining. *Annual Review of Information Science and Technology, 45*(1), 121-125*.*

Cambria, E. & White, B. (2014). Jumping NLP curves: a review of natural language processing research. *IEEE Computational Intelligence Magazine*. doi: 10.1109/MCI.2014.2307227

Carevic, Z. & Mayr, P. (2014). Recommender systems using pennant diagrams in digital libraries. *13th European Networked Knowledge Organization Systems (NKOS) Workshop.* https://arxiv.org/ftp/arxiv/papers/1407/1407.7276.pdf

Catalini, C., Lacetera, N. & Oettl, A. (2015). The incidence and role of negative citations in science. *PNAS, 112*(45), 13823-13826.

Chen, C. (2018) *The CiteSpace manual*. https://leanpub.com/howtousecitespace

Chen, C., Ibekwe-SanJuan, F. & Hou, J. (2010). The structure and dynamics of cocitation clusters: a multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386-1409.

Chen, X., Xie, H., Wang, F.L., Liu, Z., Xu, J. & Hao, T. (2018). A bibliometric analysis of natural language processing in medical research. *BMC Medical Informatics, 18*(1).

Chowdhury, G.G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89.

Clarivate Analytics. (2018). *Web of Science core collection field tags.* https://images.webofknowledge.com/images/help/WOS/hs_wos_fieldtags.html

Çarkı, K., Geutner, P. & Schultz, T. (2000). Turkish LVCSR: Towards better speech recognition for agglutinative languages. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings* (pp. 1563-1566). İstanbul: IEEE.

Ding, Y., Chowdhury, G.G. & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*, 817-842.

Eryiğit, G. (2014). *ITU Turkish natural language processing pipeline*. http://tools.nlp.itu.edu.tr/MorphAnalyzer

Feldman, S. (1999). NLP meets the jabberwocky: Natural language processing in information retrieval. *Online, 23*, 62-72.

Fiszman, M., Demner-Fushman, D., Kılıçoğlu, H. & Rindflesch, T.C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. *Journal of Biomedical Informatics, 42*, 801-813.

Galvez, C. & Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics, 70*(1), 3-26.

Garfield, E. (2001). From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography. http://garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf

Glänzel, W., Heeffler, S. & Thijs, B. (2017). Lexical analysis of scientific publications for nano-level scientometrics. *Scientometrics, 111*, 1897-1906.

Gumpenberger, C., Gorraiz, J., Wieland, M., Roche, I., Schiebel, E., Besagni, D. & François, C. (2013). Exploring the bibliometric and semantic nature of negative results. *Scientometrics, 95*, 277-297.

Hooper, C.J., Neves, B. & Bordea, G. (2015). A disciplinary analysis of internet science. In Tiropanis, T., Vakali, A. Sartori, L & Burnap, P. (eds) *Internet Science. INSCI 2015. Lecture Notes in Computer Science*, vol 9089 (pp. 63-77). Switzerland: Springer Cham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50.

Jha, R., Jbara, A-A., Qazvinian, V. & Radev, D.R. (2016). NLP-driven citation analysis for scientometrics*. Natural Language Engineering, 23*(1), 93-130.

Kim, I.C., Le, D.X. & Thoma, G.R. (2014). Automated model for extracting citation sentences from online biomedical articles using SVM-based text summarization technique. In *Proceedings 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA* (pp. 1991-1996). San Diego: IEEE.

Kuhlthau, C.C. (1991). Inside the search process - information seeking from the users perspective. *Journal of the American Society for Information Science, 42*(5), 361-371.

Lewis, D.D. & Jones, K.S. (1996). Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92-101.

Li, K., Rollins, J. & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics, 115*, 1-20.

Liddy, E.D. (2010). Natural language processing. In *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 3864-3873). New York: Taylor and Francis.

Manning, C.D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. London: MIT Pres.

Maričić, S., Spaventi, J., Pavičić, L. & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science*, *49*(6), 530-540.

Mayr, P. & Scharnhorst, A. (2015). Combining bibliometrics and information retrieval: preface. *Scientometrics, 102*(3), 2191-2192.

McCain, K.W. (1991). Mapping economics through the journal literature: an experiment in journal cocitation analysis. *Journal of the American Society For Information Science, 42*(4), 290-296.

Mikova, N. (2016). Recent trends in technology mining approaches: quantitative analysis of GTM Conference Proceedings. In Daim, T.U., Chiavetta, D., Porter, A.L. & Sarıtaş, O. (Eds) *Anticipating Future Innovation Pathways through Large Data Analysis* (pp. 59-70). Switzerland: Springer Nature.

Nanba, H. & Okumura, M. (1999). Towards multi-paper summarization reference information. *In IJCAI'99 Proceedings of the 16th International Joint Conference on Artificial intelligence - Volume 2,* (pp. 926-931). Stockholm: Margan Kaufmann Publishers Inc.

Natural Language Processing. (2017). *Oxford Living Dictionaries*. Erişim adresi: https://en.oxforddictionaries.com/definition/natural_language_processing

Névéol, A. & Zweigenbaum, P. (2016). Clinical natural language processing in 2015: leveraging the variety of texts of clinical interest. *IMIA Yearbook of Medical Informatics 2016*, 234-239. Doi: 10.15265/IY-2016-049

Robertson, S.E. & Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129-146.

Rotto, E. & Morgan, R.P. (1997). An exploration of expert-based text analysis techniques for assessing industrial relevance in U.S. engineering dissertation abstracts. *Scientometrics, 40*, 83-102.

Salton, G. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Saracevic, T. (1975). Relevance - review of and a framework for thinking on notion in information-science. *Journal of the American Society for Information Science, 26*(6), 321-343.

Schultz, T. & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication, 35*(1-2), 31-51.

Sedighi, M. (2016). Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics). *Library Review, 65*(1-2), 52-64.

Shen, H-P., Wu, C-H. & Tsai, P-S. (2015). Model generation of accented speech using model transformation and verification for bilingual speech recognition. *ACM Transactions on Asian and Low-Resource Language Information Processing, 14*(2).

Silahtaroğlu, G. (2013). *Veri madenciliği: kavram ve algoritmaları [Data mining: conceps and algorithms]*. İstanbul: Papatya Yayıncılık Eğitim A.Ş.

Sue, H-N. & Lee, P-C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics, 85*, 65-79.

Taşkın, Z. & Al, U. (2013). Institutional name confusion on citation indexes: The example of the names of Turkish hospitals. *Procedia - Social and Behavioral Sciences, 73*, 544-550.

Taşkın, Z. & Al, U. (2014). Standardization problem of author affiliations in citation indexes. *Scientometrics, 98*(1), 347-368.

Taşkın, Z. & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics, 114*(1), 335-357.

Taşkın, Z., Al, U. & Sezen, U. (2017). First stage of an automated content-based citation analysis study: detection of citation sentences. In *STI2017, open indicators: innovation, participation and actor-based STI indicators, Paris, 2017*. https://goo.gl/hdbnt3

Trujillo, C.M. & Long, T.M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances, 4*(1). doi: 10.1126/sciadv.1701130

van Eck, N.J. & Waltman, L. (2018). VOSviewer manual. http://www.vosviewer.com/download/f-z2x2.pdf

van Rijsbergen, C.J. (1979). *Information retrieval*. London: Butterworth-Heinemann Newton.

White, H.D. (2007). Combining bibliometrics, information retrieval, and relevance theory. Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology, 58*, 536-559.

White, H.D. (2018). Pennants for Garfield: bibliometrics and document retrieval. *Scientometrics, 114*, 757-778.

Zemberek NLP. (2015). http://zembereknlp.blogspot.com.tr/