# Information Retrieval as a Domain: Visualizations Based on Two Data Sets†

K. S. Raghavan*, K. H. Apoorva** and Aarti Jivrajani***

Centre for Knowledge Analytics & Ontological Engineering (KAnOE),
PES University, Bangalore 560085, India,
*<ksragav@hotmail.com>, **<apoorva.khp@gmail.com>, ***<aartij17@gmail.com>

S. Raghavan works as a visiting scientist at the Centre for Knowledge Analytics & Ontological Engineering, PES University, Bangalore. He is currently a visiting professor at the Wee Kim Wee School of Communication & Information of the Nanyang Technological University, Singapore. His research interests are knowledge organization with emphasis on lateral semantic relations, knowledge organization in culture-specific domains and multilingual thesauri.

K. H. Apoorva is an undergraduate computer science and engineering student at the PES Institute of Technology, Bangalore. She is currently working as a research assistant at the Centre for Knowledge Analytics and Ontological Engineering, PES University, Bangalore. Her research interests are natural language processing, semantic web, knowledge organization, big data analytics and green cloud computing.

Aarti Jivrajani is an undergraduate computer science and engineering student at the PES Institute of Technology, Bangalore. She is currently working as a research assistant at the Centre for Knowledge Analytics and Ontological Engineering, PES University, Bangalore. Her research interests are natural language processing, semantic web, knowledge organization, big data analytics and green cloud computing.

Raghavan, K. S., Apoorva, K. H. and Jivrajani, Aarti. **Information Retrieval as a Domain: Visualizations Based on Two Data Sets.** *Knowledge Organization. 42(8)*, 591-601. 2 references.

**Abstract:** A domain is seen as a subject of discourse whose extensions and intensions are defined by the scope and content of its research literature. Mapping the contours of a domain over a period helps in visualizing the changes in the research frontiers of the domain thus indicating the transformations as well as trends in research in the domain. In this paper research literature in information retrieval from IEEE and EBSCO databases were used as data sets to map the contours of the research literature in the area of information retrieval over the last 14 years. The two data sets suggest differing perspectives and emphasis between the two research communities.

Received: 5 November 2015; Accepted: 6 November 2015

Keywords: information retrieval, domain, domain analysis

† This work was supported in part by the World Bank/Government of India research grant under the TEQIP Programme (Subcomponent 1.2.1) to the Centre for Knowledge Analytics & Ontological Engineering (KAnOE) at the PES Institute of Technology, PES University, Bangalore, India. The authors wish to thank Ms. Srilakshmi of KAnOE for her assistance, and also the anonymous reviewer for helpful comments.

## 1.0 Introduction

Domain analysis—a technique proposed by Hjørland and employed by Smiraglia in the area of knowledge organization—uses a range of empirical techniques to comprehend and map the facets of a domain that help in visualizing the contours of the domain and its transformation over time. Smiraglia provides an operational definition of a domain as: "a group with an ontological base that reveals an underlying teleology, a set of common hypotheses, epistemo-

logical consensus on methodological approaches, and social semantics" (Smiraglia 2012, 114). In this paper we view a domain as a subject of discourse whose extensions and intensions are primarily represented by the research literature generated by its scholar community. Domain analysis is seen as the process of mapping the contours of a domain with a view to study its evolution and transformation over time. Domain analysis is an effective method for visualizing the wide-ranging array of themes and sub-themes that constitute the topics being studied within the domain and helps identify the trends and directions in research. In knowledge organization, domain analysis helps build an ontology of a domain. The extension (facets that constitute the core of the domain) and intension (areas of applications) of a domain at different points in time during its evolution and development can be brought out using domain analytic techniques.

## 2.0 The present study

The term "information retrieval" (IR) was coined by Calvin Mooers in 1950, although as a process it is as old as the early bibliographic tools. Information Retrieval has always been at the core of information science as a research theme and is the single most important contributing factor leading to the acceptance of information science as a discipline. A volume edited by Tefko Saracevic entitled *Introduction to Information Science* and published in the 1970s had nearly 40% of its pages devoted to the subject of information retrieval. The first few major experimental studies in information science were also in the area of information retrieval, including the UNITERM-ASTIA, Aslib-Cranfield I and II, and MEDLARS evaluation studies. IR has also been the area in which there has been a large interface between information science and computer science since the early 1950s. Professional bodies such as the Association for Information Science and Technology (ASIST) as well as the Institute of Electrical and Electronics Engineers (IEEE), and the Association for Computing Machinery (ACM) have special interest groups (SIGs) devoted to information retrieval. Information retrieval can be seen as a process, as an expression of information-seeking behavior, as an information seeking and access activity related to a specific problem-solving task involving human beings interacting with information sources. The emergence of the World Wide Web as an important channel of communication and publishing has only made the task more complex and even crucial. As a result in more recent years IR has evolved into a highly inter-disciplinary area of research and today there is a substantial body of research literature on both the theoretical aspects of information retrieval as well as findings of experimental studies on information retrieval. There is no question, therefore, that information retrieval meets all the requirements of a domain. These were the major factors that prompted the choice of information retrieval as the domain for this study. In this paper we seek to map the contours of information retrieval and explore its evolution and transformation since 2001.

## 3.0 Methodology

The present study is based on an analysis of the metadata of research literature in information retrieval published in the last 14 years (i.e. from 2001 to 2014). There are four major sources that collectively cover research literature on IR. While there is a certain degree of overlap, none probably is an exhaustive record of the research literature in IR. These are:

- Bibliographic databases in LIS such as LISA, LISTA, etc.
- IEEE database
- ACM SIGIR conferences, and
- Presentations at TREC.

In this study two different data sets were used:

- IEEE database (http://ieeexplore.ieee.org/Xplore/home.jsp)
- Library, Information Science & Technology Abstracts (LISTA) from EBSCO http://search.ebscohost.com)

Metadata for the papers presented at the ACM SIGIR conferences and presentations at TREC were not readily accessible; a more comprehensive picture of the contours of the domain will be reported in another paper that will take into consideration the research reported in ACM SIGIR conferences as also presentations made at TREC.

Searches were carried out in IEEExplore and LISTA using 'advanced search' technique (papers published in scholarly academic journals in the English language) for papers on IR; only records which contained the term "information retrieval" in the INSPEC controlled terms field (in the IEEE database) and in the "subject terms" field in LISTA were downloaded. The details of the number of items retrieved for further analysis are as in Table 1.

A brief explanation of the factors that influenced the choice of the two source databases should be in order. IEEE and LISTA reflect slightly different perspectives and viewpoints from which they look at the domain of information retrieval. Information retrieval has traditionally been a forte of the library and information science (LIS) community as early retrieval tools such as library catalogues, bibliographies and indexes were built as part of li-

| Name of the database | Number of records |
|---|---|
| IEEE | 11614 |
| LISTA (EBSCO) | 7677 |

*Table 1.* Source databases.

brary services. The techniques and tools required for developing such information products have been the focus of research of the LIS community at least since the time of Panizzi and Cutter. The LIS community has focused on the development of techniques and tools for indexing such as schemes of classification, thesauri, other vocabulary control devices, and other metadata standards to enhance the quality of information retrieval. The computer science community that worked on problems of information retrieval, on the other hand, focused more on automatic extraction of metadata, search techniques such as enhanced Boolean approaches, weighted term searches, proximity operators, etc. It was thought that an analysis of the data sets from EBSCO and IEEExplore should reveal whether this continues to be the case; and what are the principal aspects of research as reflected by research covered in the two databases.

### *3.1 Tools of data analysis and visualization*

The records retrieved for every year from the two databases were separately examined as it was one of the objectives of the study to see the differences, if any, between the two. At the first level, the terms that co-occurred with "information retrieval" were identified. In other words all the other terms that were assigned to each document (*INSPEC Control Terms* in the case of IEEE and *Subject Terms* in the case of LISTA) were noted. The list of terms co-occurring with "information retrieval" for each year since 2001 in the two databases were prepared along with frequency of occurrence. This way the topics that occurred in a year and all the years in which a topic occurred were identified. The frequency of occurrence of a topic in a year was taken to be an indication of its importance ("degree") within the domain of IR. For graphically visualizing the major topics which figured in the domain the NetworkX Python package along with the open graph visualization platform Gephi were used. Based on the aforementioned criteria graphs were constructed for the IEEE and EBSCO data sets separately.

## 4.0 Analysis

### *4.1 The dominant themes*

In the following paragraphs we present an analysis of the two data sets with a view to see the similarities as well as differences in the contours of the domain as revealed by the two data sets. In making the analysis we have considered the contributors to resources covered by the IEEExplore and those covered by EBSCO LISTA as two different research communities. While there may be areas of common interest to the two communities, it is hypothesized here that the emphasis and focus differ. The IEEE data indicated that as many as 2,080 unique terms co-occurred with "information retrieval" during the period 2000-2014. The data set from EBSCO LISTA, on the other hand, had over 8,500 terms co-occurring with "information retrieval" during the period. An examination of the terms suggested four different kinds of terms associated with "information retrieval:"

– Terms indicating topics representing specific subthemes of information retrieval;
– Terms indicative of application areas (e.g., medicine, knowledge management (KM), music, etc.);
– Terms indicative of broad context specifying areas (e.g., library science, information science, academic libraries, etc.); and,
– Terms that did not fall into any of the categories above but merely indicated an aspect of the resource such as the nature of content (e.g., research)

For further analysis, the top 20 terms co-occurring with "information retrieval" were identified for each year for the two data sets. Figure 1 indicates the dominant themes in IR research based on the two data sets. Only a few areas are of common interest to the two research communities.

Figures 2 and 3 are visualizations of IR research as represented by the two data sets; more particularly the IEEE data set is clearly leading the domain into new territories.

In terms of percentage of total research output, the top 20 research themes in the EBSCO data set accounted for about 25% or more of the total research output between 2001 and 2007; However, there has been a decline since 2008 and in 2012 the top 20 terms accounted for just over 15% of the total research output. The figures have been below 20% for all the years since 2008. This suggests that the IR research literature as covered by the EBSCO database has been experiencing a much higher degree of distribution and scatter in terms of research themes.

Evidently the core research literature is distributed over a larger number of themes. In contrast to this the top 20 research themes in the IEEE data set account consistently for 30% or more of the total IR research output in all the years.

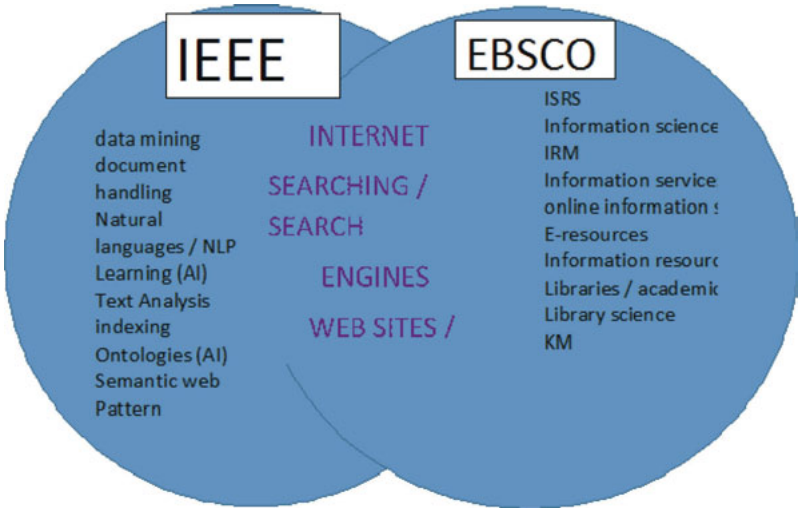


*Figure 1*. Dominant research themes in IR.

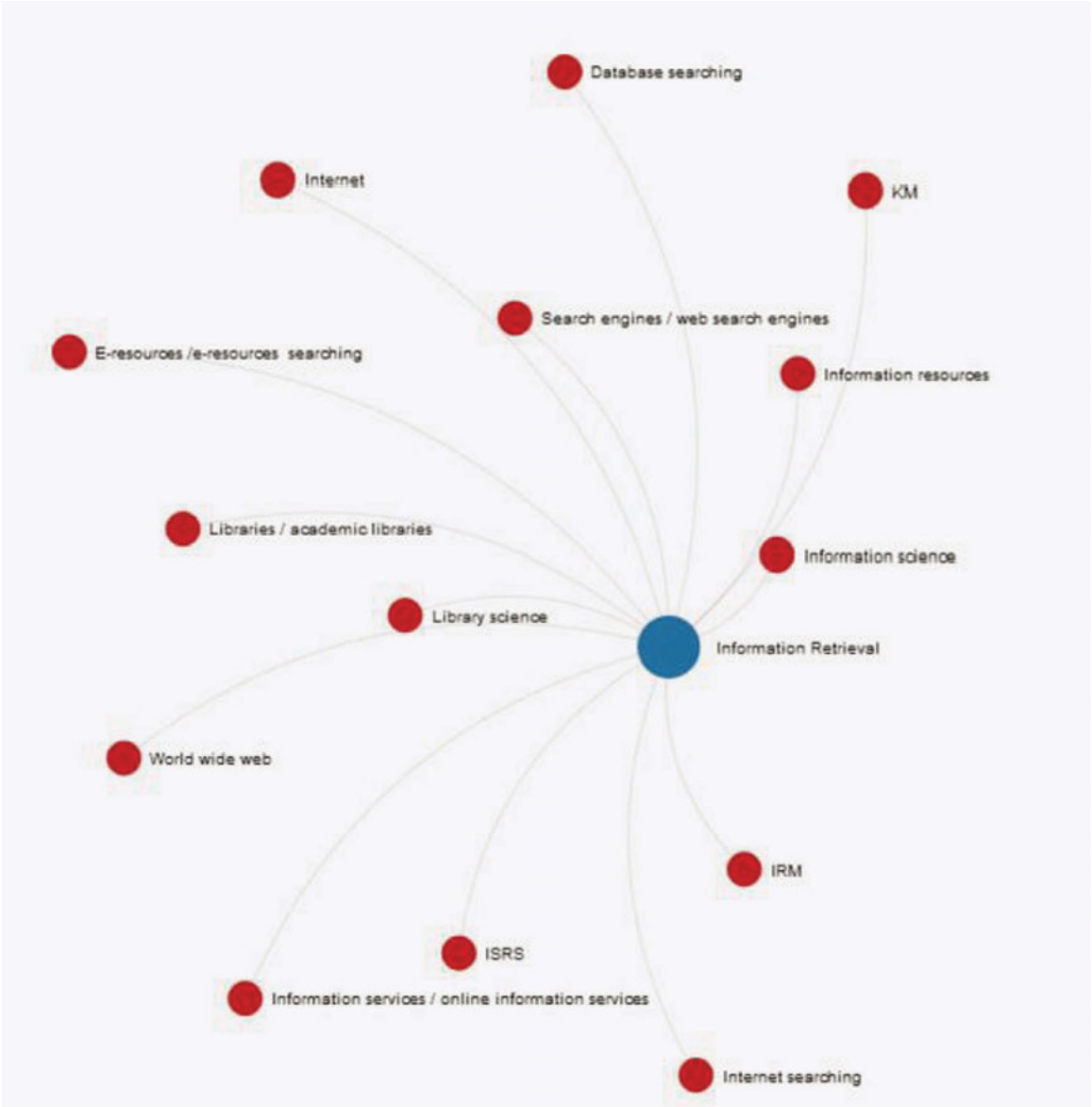


*Figure 2*. Dominant research themes (EBSCO data set).


*Figure 3*. Dominant research themes (IEEE data set).

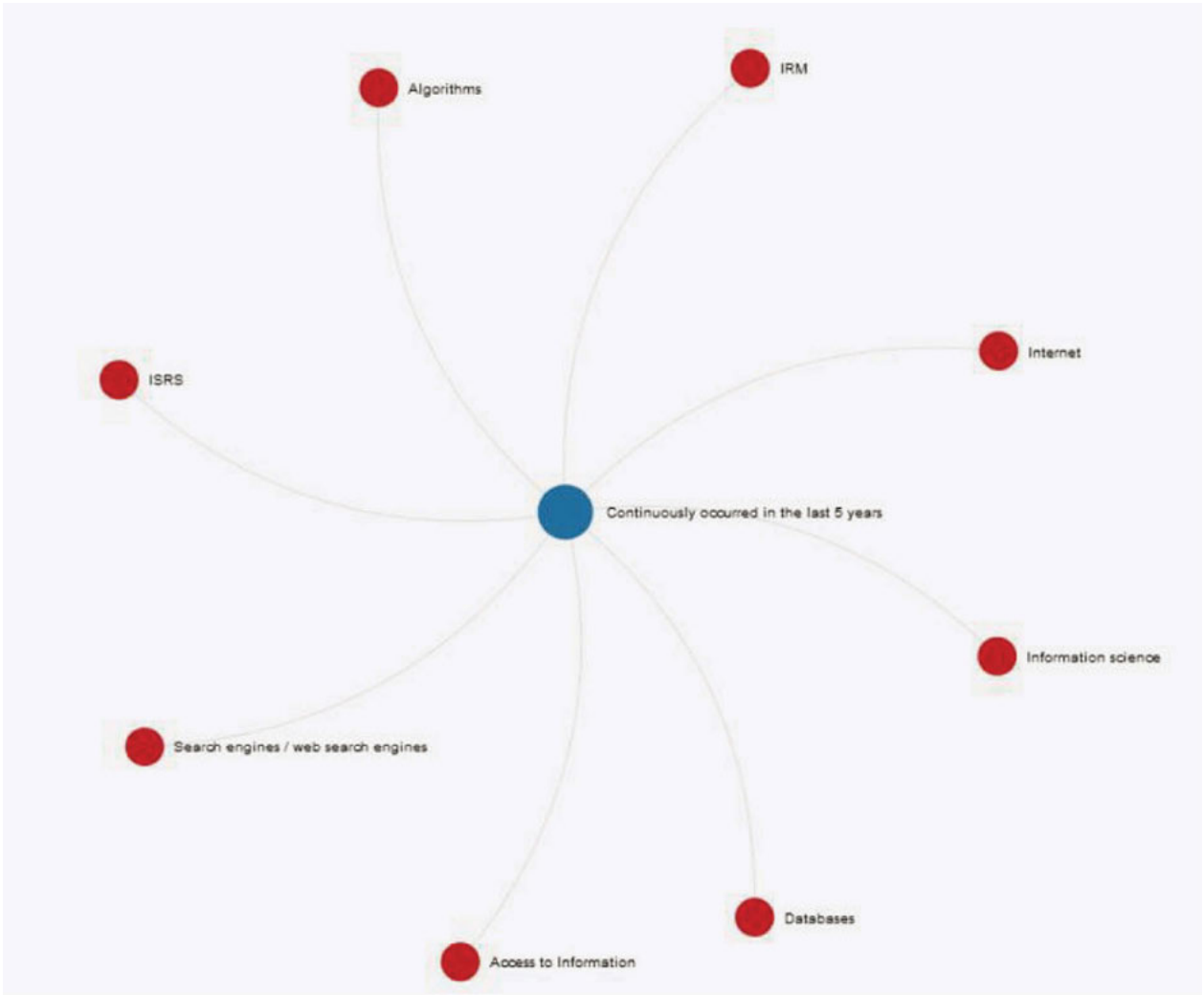


*Figure 4*. Research themes in the last 5 years (EBSCO data set).


*Figure 5.* Research themes in the last 5 years (IEEE data set).

*4.2 The top 20 research themes*

That the domain of IR is charting new territories should also be evident from the top subject terms that represent the predominant themes of research in the two data sets. Tables 4 and 5 present details of the top twenty research themes over the years since 2001 for the two data sets.

*4.3 The most productive journals*

A large number of journals and serials appear to be publishing papers on IR. Based on the EBSCO data for 2014 the most productive journals are as shown in Table 1. *Journal of the Association for Information Science & Technology* (*JASIST*) tops the list of most productive journals in the EBSCO dataset. Besides *JASIST,* only *Scientometrics* and *Information Processing & Management* are the other two journals that figure in the list of ten most productive journals in which the LIS community has been a major player. In the IEEE data set, however, the proceedings of the ACM/IEEE Joint Conference on Digital Libraries tops the list as the most productive serial. The differences in the keywords contained in the titles of the two lists (for the two data sets) is also significant from the point of view of understanding the contours of the domain. The titles of the most productive serials in the IEEE dataset suggest that the intension and extension of the domain of information retrieval have changed considerably in the last decade and more and are clearly indicative of the emerging trends and directions in research in the domain. Clearly IR is today a domain with contours very different from what they were when Calvin Mooers coined the term in the 1950s.

| | |
|---|---|
| *Journal of the Association for Information Science & Technology* | 41 |
| *Journal of the American Medical Informatics Association* | 27 |
| *Information Systems* | 26 |
| *Scientometrics* | 20 |
| *Information Processing & Management* | 19 |
| *International Journal of Database Theory & Application* | 17 |
| *Journal of Information Systems Education* | 17 |
| *Journal of Medical Internet Research* | 17 |
| *International Journal of Information Management* | 16 |
| *ACM Transactions on Information Systems* | 15 |

*Table 2*. Most productive journals (2014)—EBSCO Data Set.

| | |
|---|---|
| *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on* | 20 |
| *ICASSP (2014 IEEE International Conference)* | 18 |
| *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on* | 14 |
| *Big Data (Big Data), 2014 IEEE International Conference on* | 11 |
| *ICSME, (2014 IEEE International Conference)* | 11 |
| *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference* | 11 |
| *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on* | 10 |
| *Information Communication and Embedded Systems (ICICES), 2014 International Conference on* | 10 |
| *Knowledge and Data Engineering, IEEE Transactions on* | 10 |
| *Data Engineering (ICDE), 2014 IEEE 30th International Conference on* | 9 |

*Table 3*. Most productive serials/journals (2014) —IEEE data set.

The changing dimensions of IR as a domain is even more clearly visible when we examine the top 20 research themes for the years since 2001. Clearly there are perceptible differences between the two data sets. However, the two data sets also suggest that major transformation has taken place in the last 14 years. The growing research in the areas of natural language processing, ontologies, learning (AI), and search engines, read along with the near complete disappearance of such themes as cataloging, classification, indexing, etc., are pointers to the shape of things to come. The developments suggest not merely a change in the emphasis but may also mean major terminological changes. Clearly the Web has been the major influencing factor in determining the direction of research in IR.

## 5.0 Conclusion

The study brings out the changes and transformation in the contours of the domain of information retrieval. However, it should be understood that the analysis and visualization presented here are based on two data sets. Clearly there are at least two other sources which are quite representative of certain schools of thought in so far as research in IR is concerned. These are the papers presented to the ACM SIGIR meetings and presentations made at the annual TREC conferences. The work of examining the research reported in these two is in progress and a more comprehensive picture of the domain of information retrieval will be presented in a subsequent paper.

| Terms | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Access to Information | | | | | | | | x | x | x | x | x | x | x |
| Algorithms | | | | | | | | | | x | x | x | x | x |
| Bioinformatics | x | | | | | | | | | | | | | |
| Cataloguing | | | | | x | | | | | | | | | |
| Citation Analysis | | | | | | | | | | | | | | x |
| Classification | | | | | | | | | | | x | | | |
| Computer s/w | | x | | | | | | | | | | | x | x |
| Computers in medicine | x | | | | | | | | | | | | | |
| Cross-language IR | | | | x | | | | | | | | | | |
| Data Analysis | | | | | | | | | | | | | x | x |
| Data mining | | x | | | | | x | x | | | | x | | |
| Database searching | x | x | x | x | x | x | x | x | | x | | x | | x |
| Databases | | x | | | | x | | | | x | x | x | x | x |
| Digital libraries | | x | | | x | x | x | | x | | | | | |
| Documentation | | x | x | x | x | x | x | | x | | | | | |
| Electronic data processing | | | x | | | | | | | | | | | |
| Electronic publications | | | | | | | | | | | x | x | x | x |
| E-resources /e-resources searching | x | x | x | x | x | x | x | x | x | x | | x | | |
| Image processing | x | | | | | | | | | | | | | |
| Image retrieval | x | | | | | | | | | | | | | |
| Indexing | | | | x | | | | | x | x | | | | |
| Information Literacy | | | | | | | | x | | | | | | |
| Information resources | x | | x | x | x | x | x | x | x | x | x | | | |
| Information science | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Information Seeking Strategies / Behavior | | | | | | | | x | x | x | | x | | |
| Information services / online information services | x | x | x | x | x | x | x | x | x | x | x | x | | |
| Internet | x | | x | x | x | x | | x | | x | x | x | x | x |
| Internet searching | x | x | x | x | x | x | x | x | x | x | | | | |
| Information Resources Management | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Information Storage & Retrieval Systems | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Information Technology | | | x | | x | x | x | | x | | | x | | |
| Knowledge Management | x | | x | | | | x | x | x | x | x | | x | |
| Librarians | | | | x | | | | | | | | | | |
| Libraries / academic libraries | | x | x | | x | x | x | | x | | x | x | x | x |
| Library science | | | x | x | x | x | x | x | x | x | | | | |
| Medical informatics | x | x | | | | | | | | | | | | |
| Medical records | x | | | | | | | | | | | | | |
| Metadata | | | | | x | | | | | x | x | | x | x |
| Multimedia systems | x | x | | | | | | | | | | | | |
| Public institutions | | x | | | | | | | | | | | | |
| Querying (Computer Science) | | | | | | | | | | | x | | | |
| Research | | | | | | | x | x | x | | x | x | x | x |
| Search algorithms | | | | | | | | | | | | x | | |
| Search engines / web search engines | x | x | x | x | | x | x | x | x | x | x | x | x | x |
| Semantics | | | | | | | | | | | x | | x | x |
| Slavic literature | | | | | x | | | | | | | | | |
| User Interfaces (Computer Systems) | | | | | | | | | | | x | | x | x |
| Websites | x | | x | | | | | | | | | | | |
| World wide web | x | x | x | x | x | x | x | x | | | x | | | |

*Table 4*. Top 20 research themes—EBSCO data set.

| Terms | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Big data | | | | | | | | | | | | | | x |
| Classification | x | x | | | | x | x | x | | | | | | |
| client-server systems | | | | | | | | | | | | | | |
| Cloud Computing | | | | | | | | | | | x | x | x | x |
| Computer-aided instruction | | | | | | x | | | x | x | x | x | | |
| DBMS | | | | | | | x | | | | | | | |
| Data Analysis | | | | | | | | | | | | | x | x |
| data mining | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| data visualization | | x | x | x | | | | | | | x | x | | |
| digital libraries | | | x | x | x | x | | | | | | | | |
| Distributed databases | x | | | | | | | | | | | | | |
| document handling | x | | x | x | x | x | x | x | x | x | x | x | x | x |
| E-Commerce | | x | | | | | | | | | | | | |
| Feature Extraction | | x | | x | | | | x | | x | x | x | x | |
| GIS | | | | | | | | | | x | | | | |
| Grid Computing | | | | x | x | x | | | | | | | | |
| Groupware | x | | | | | | x | | | | | | | |
| Human factors | x | | | | | | | | | | | | | |
| hypermedia mark-up languages | x | x | x | | | | | | | | | | | |
| indexing | x | x | x | x | x | x | x | x | x | x | x | | x | x |
| Information filters | | | x | | | | | | | | | | | |
| information needs | | x | | | | | | | | | | | | |
| information resources | x | x | x | | | | | | | | | | | |
| Interactive Systems | x | | | | | | | | | | | | | |
| Internet | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Knowledge-based systems | | | x | | | | | | | | | | | |
| Learning (AI) | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Medical Computing | | | | | | | | | | | | | x | |
| Medical information systems | | | | | | x | x | | | | | | x | |
| meta data | x | x | x | x | x | x | | x | | | | | | |
| mobile computing | | | | | x | | x | | x | | x | x | x | x |
| Multi-agent systems | | | | | x | | | | | | | | | |
| multimedia databases | x | x | | | | | | | | | | | | |
| Natural languages / NLP | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| online front-ends | | x | x | x | | | | | | | | | | |
| Ontologies (AI) | | | | x | x | x | x | x | x | x | x | x | x | x |
| Pattern Classification | | | | | | | | x | | x | x | x | x | x |
| Pattern Clustering | | | | x | | x | x | | x | x | x | x | x | x |
| Peer-to-peer computing | | | | | x | | x | | x | | | | | |
| Recommender systems | | | | | | | | | | | | x | x | x |
| search engines | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Security of Data | | | | | x | | | | | | | | | |
| Semantic web | | | | x | x | x | x | x | x | x | x | x | | x |
| Social Networking (Online) | | | | | | | | | x | x | x | x | x | x |
| software agents | x | x | x | | | | | | | | | | | |
| speech recognition | x | | | | | | | | | | | | | |
| Support Vector Machines | | | | | | | | | x | | | | | |
| Text Analysis | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| User interfaces | x | x | x | x | x | | | | | x | | | | |
| Web Services | | | | | | | x | x | x | x | | x | | |
| Websites | | x | x | x | x | x | x | x | x | x | x | x | x | x |
| XML | | | x | x | x | x | x | x | x | x | | | | |

*Table 5. Top 20 research themes—IEEE data set.*

**References**

Saracevic, Tefko, comp. and ed. 1970. *Introduction to Information Science*. New York: Bowker.

Smiraglia, Richard P. 2012. "Epistemology of Domain Analysis." In *Cultural Frames of Knowledge*, edited by Richard P. Smiraglia and Hur-li Lee. Wurzbürg: Ergon Verlag, 111-24.