Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization

Richard P. Smiraglia* and Xin Cai**

Knowledge Organization Research Group, School of Information Studies, University of Wisconsin, Milwaukee, NWQB2569, 2025 E. Newport St., Milwaukee WI 53211
*<smiragli@uwm.edu>, **<xincai@uwm.edu>

Richard P. Smiraglia is a professor and member of the Knowledge Organization Research Group in the iSchool at the University of Wisconsin-Milwaukee. He has explored domain analysis for evolution of knowledge organization, epistemological analysis of the role of authorship in bibliographic tradition, the evolution of knowledge and its representation in knowledge organization systems, and the phenomenon of instantiation among information objects. He is Editor-in-Chief of this journal.

Xin Cai is a PhD candidate in the iSchool at the University of Wisconsin-Milwaukee. He holds a master's degree in information science from Central China Normal University, China, and a bachelor's degree in computer science from Shenyang Normal University, China. His research interests include information retrieval and systems, data mining, and domain analysis.

Smiraglia, Richard P. and Xin Cai. 2017. "Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization." *Knowledge Organization* 44(3): 215-233. 12 references.

Abstract: A very important extension of the traditional domain of knowledge organization (KO) arises from attempts to incorporate techniques devised in the computer science domain for automatic concept extraction and for grouping, categorizing, clustering and otherwise organizing knowledge using mechanical means. Four specific terms have emerged to identify the most prevalent techniques: machine learning, clustering, automatic indexing, and automatic classification. Our study presents three domain analytical case analyses in search of answers. The first case relies on citations located using the ISKO-supported "Knowledge Organization Bibliography." The second case relies on works in both Web of Science and SCOPUS. Case three applies co-word analysis and citation analysis to the contents of the papers in the present special issue. We observe scholars involved in "clustering" and "automatic classification" who share common thematic emphases. But we have found no coherence, no common activity and no social semantics. We have not found a research front, or a common teleology within the KO domain. We also have found a lively group of authors who have succeeded in submitting papers to this special issue, and their work quite interestingly aligns with the case studies we report. There is an emphasis on KO for information retrieval; there is much work on clustering (which involves conceptual points within texts) and automatic classification (which involves semantic groupings at the meta-document level).

Received: 30 January 2017; Revised: 8 March 2017; Accepted: 10 March 2017.

Keywords: automatic indexing, clustering, automatic classification, machine learning, knowledge organization, information

1.0 Extending knowledge organization

A very important extension of the traditional domain of knowledge organization (KO) arises from attempts to incorporate techniques devised in the computer science domain for automatic concept extraction and for grouping, categorizing, clustering and otherwise organizing knowledge using mechanical means. Four specific terms have emerged to identify the most prevalent techniques: machine learning, clustering, automatic indexing, and automatic classification. Dumais et al. wrote (1998, 148): "As the volume of information available on the Internet and corporate intranets continues to increase, there is growing interest in helping people better find, filter, and manage these resources." Techniques in machine learning are usually recognized as helpful for organizing huge amounts of resources. "Machine learning" is the subfield of computer science that has been defined as the "field of study that concentrates on induction algorithms and on other algorithms that can be said to 'learn'" (Ron and Foster 1998, 273). Another term, "data mining," is often conflated with machine learning. However, the algorithms from machine learning are not only used to summarize the data and discover hidden patterns like data mining, but they also can serve as tools for discovery and for making predictions (Rajaraman and Ullman 2011). As a result, we chose to use the term "machine learning" rather than "data mining" in this paper because the goal of knowledge organization is not only organizing existing knowledge but also discovering and organizing future knowledge. Tasks of machine learning could be roughly classified as "supervised learning" and "unsupervised learning." The major difference between them is that supervised learning requires a labeled dataset but unsupervised learning does not.

Clustering, one of the most important unsupervised learning methods is (Rajaraman and Ullman 2011, 241): "The process of examining a collection of "points," and grouping the points into "clusters" according to some distance measure. The goal is that points in the same cluster have small distance from one another." From this point of view, the documents in a collection are viewed as points in a space, and they are categorized as members of groups according to relative distance.

Automatic classification is an application of supervised learning that has been described as (Salles et al. 2016, 2) "creating models that associate documents with semantically meaningful categories," and (Golub et al. 2016, 4), "the assignment of a class or category from a pre-existing scheme … or [discovering] a scheme suitable for the collection at hand and simultaneously assign[ing] to a document one (or more) of the classes discovered." We note that the definition only describes the objection and procedures of automatic classification, but does not specify how to achieve it. The implications beneath the definition are that algorithms from supervised machine learning are not necessarily needed, algorithms like matching terms with controlled vocabulary are also eligible in the deployment of automatic classification.

Finally, automatic indexing has been defined by Gil Leiva (2017, 140) as derived from three perspectives:

> a) Computer programs that assist in the process of storing indexing terms, once obtained intellectually (i.e., computer aided indexing during storage);
> b) Systems that analyze documents automatically, but the indexing terms proposed are validated and published, if necessary, by a professional (semi-automatic indexing); and,
> c) programs without any further validation programs, (i.e., the proposed terms are stored directly as descriptors of that document—automatic indexing).

In summary, automatic indexing, which incorporates traditional term matching techniques, reveals how the early KO domain adopted computer science techniques. Machine learning, a broad term that includes many new techniques devised from the domain of computer science, shows how modern techniques influence the domain of KO. In order to display the picture more clearly, two specific terms from machine learning, "clustering" and "automatic classification," are selected to represent unsupervised learning and supervised learning respectively. Automatic classification, on the other hand, serves as a bridge to track the evolution of techniques devised in computer science for automatic knowledge organization, since it could be realized not only by using traditional term matching techniques, but also algorithms of machine learning. There is a fair amount of overlap among the definitions of the four terms over time; an early article by Soergel (1974) describes automatic and semi-automatic methods for thesaurus construction, document classification and clustering. At present, the four techniques can be seen to form a matrix of sorts, illustrated in Figure 1.

An interesting question for research is: to what extent does an identifiable community in KO embrace these techniques? A corollary question is: to what extent can a community of interest broader than but including the KO domain be identified? Our study presents three domain analytical case analyses in search of answers. The first case relies on citations located using the ISKO-supported "Knowledge Organization Bibliography" (http://www.isko.org/lit.html) using each of the four terms above. The second case relies on works located using each the four terms combined with the term "knowledge organization" in both Web of Science (WoS) and SCOPUS. The online searching represented in cases 1 and 2 took place in July 2016. Case three applies co-word analysis and citation analysis to the contents of the papers in the present special issue, acting as a control group for the emergence of computer science techniques into the science of knowledge organization.

2.0 Case 1: machine learning, automatic indexing, automatic classification and clustering in "KO Literature"

Rather than rely on research databases that have general sciences as their focus, the International Society for Knowledge Organization has, since its inception in 1979, maintained a database (earlier published in segments as "Knowledge Organization Bibliography" in the journal *Knowledge Organization*) carefully monitored by ISKO mem-

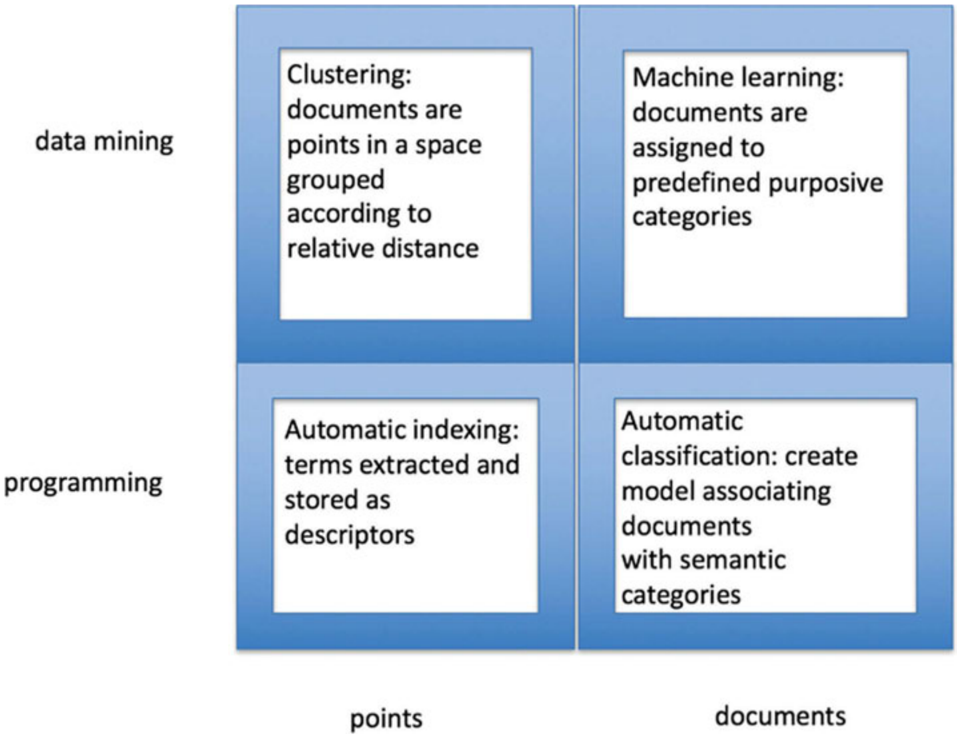
data mining
programming
Clustering: documents are points in a space grouped according to relative distance
Machine learning: documents are assigned to predefined purposive categories
Automatic indexing: terms extracted and stored as descriptors
Automatic classification: create model associating documents with semantic categories
points
documents

Figure 1. Clustering, machine learning, automatic indexing, automatic classification.

ber indexers. Now called "Knowledge Organization Literature" the database is available on the ISKO website (http://www.isko.org/lit.html). Since 2009 the quarterly updates have appeared exclusively online; these may be viewed as individual files, or the database may be searched by keywords, authors, dates, or classification numbers (the database entries bear classification symbols from the "Classification System for Knowledge Organization Literature" described on the website). Using each of the four terms above, we located 237 citations: "machine learning" (12 citations were located), "automatic indexing" (78 citations), "automatic classification" (34 citations) and "clustering" (113 citations). Results were downloaded for analysis following some minor data-cleaning.

Citations located under the terms "clustering," "automatic classification," and "automatic indexing" were dated from 1990-2014. Minor differences can be seen in Figures 2-4; the bulk of "automatic indexing" preceded 2000, most of "automatic classification" occurred after 2005, most "clustering" occurred after 2009.

Citations located under the term "machine learning" were dated from 1996 to 2011, creating the impression that this newer technology is trending. The distribution is shown in Figure 5.

Taken together the distributions indicate that the term "automatic indexing" had much of its usage in the early 1990s and has slowly faded from use in this database. "Automatic classification" is used consistently from 1990 onward, but with low frequency. The highest frequency usages are "clustering" and "machine learning," which post-date 1996 and which seem to be increasing in occurrence. The beginning in 1990 is likely an artifact of the database. It is not clear whether the absence of post-2011 data for "machine-learning" or post-2014 data for "clustering" represent shifting research emphases, or whether it is a reflection of editorial changes in the database.

Indexers assign one or more keywords in conjunction with the classification symbols for each citation in the "KO Literature" database. Analysis of these keywords can suggest how the indexers perceived the topical content of each set. Figures 6-9 show the terms associated with each result.

The purpose of this analysis was to see whether the indexing keyword assignments appeared to be stable *vis a vis* the search terms we used. The analysis is mixed. The richest group is the "clustering" result, which had 160 instances of 53 terms assigned; the most frequently occurring term assigned was "cluster analysis." "Automatic indexing" occurs in all four clusters. "Automatic classification" occurs in all but the "automatic indexing" cluster. "Machine learning" is not used in the "machine learning" cluster, which is dominated by "automatic classification."

There was no overlap among the lists. Analysis of dates of publication indicated growth in all but "auto-

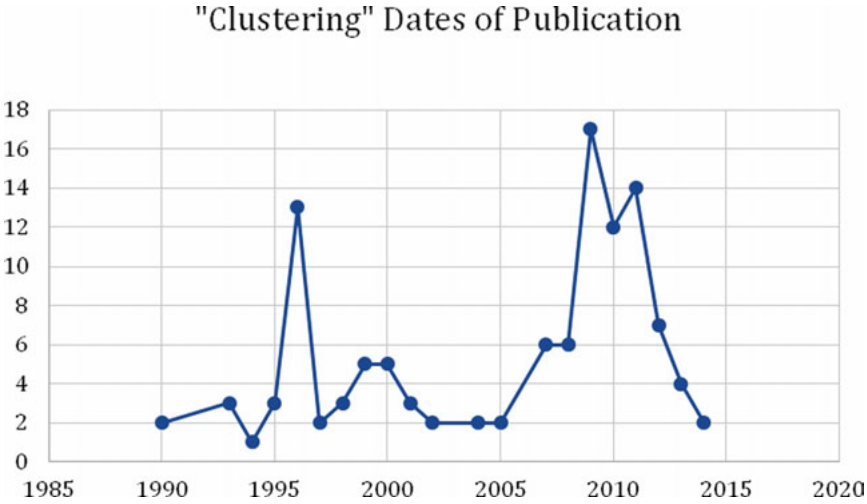

Figure 2. "Clustering" 113 citations published 1990-2014.

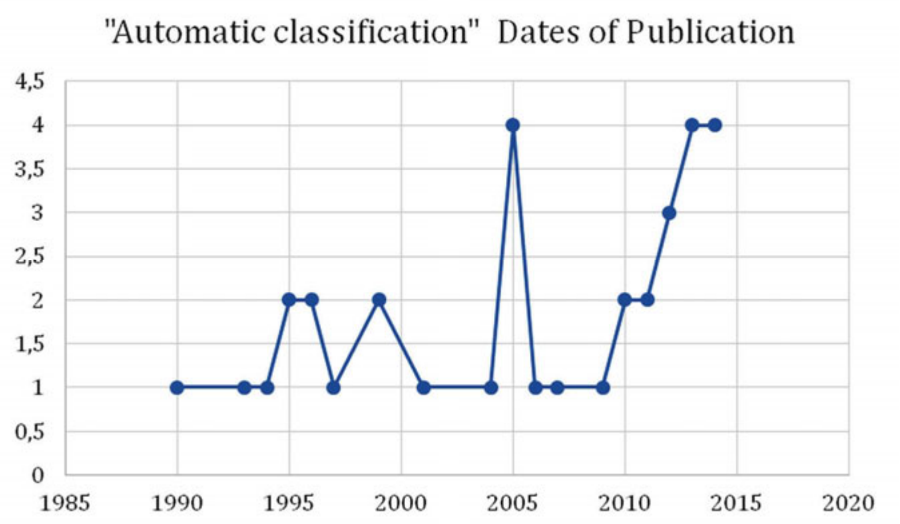

Figure 3. "Automatic classification" 34 citations published 1990-2014.

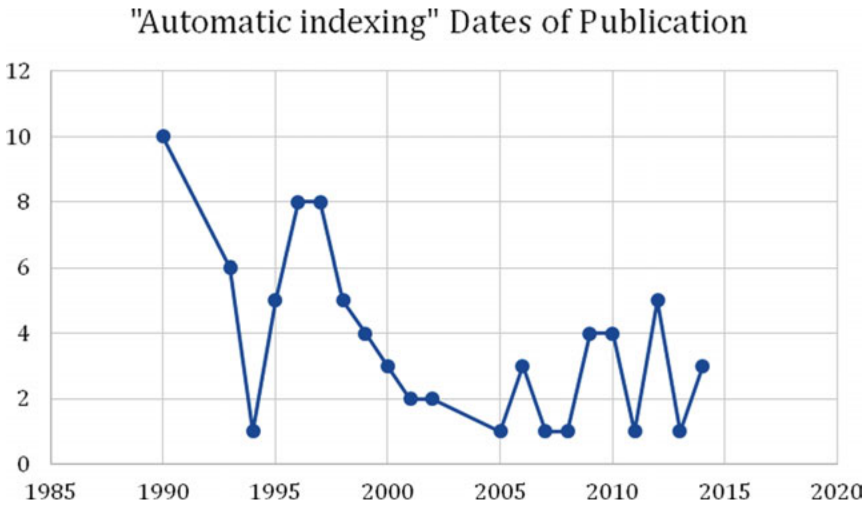

Figure 4. "Automatic indexing" 78 citations published 1990-2014.

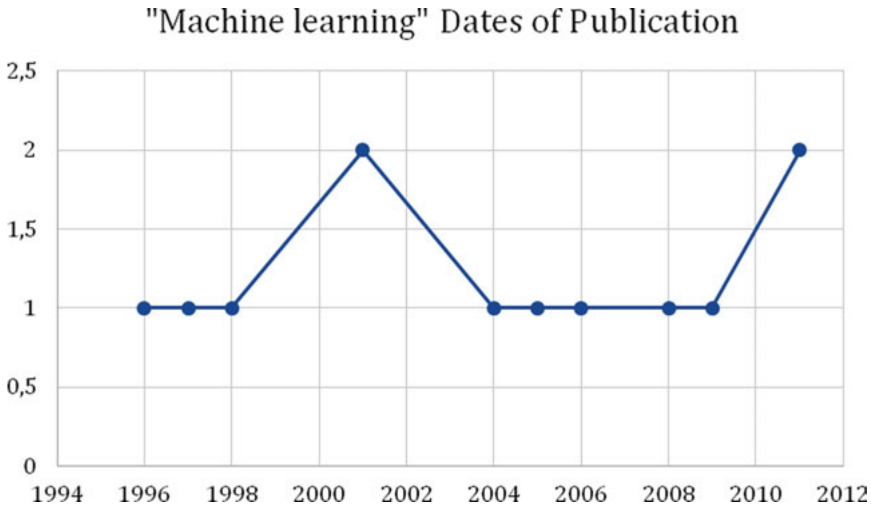

Figure 5. "Machine learning" 12 citations published 1996-2011.

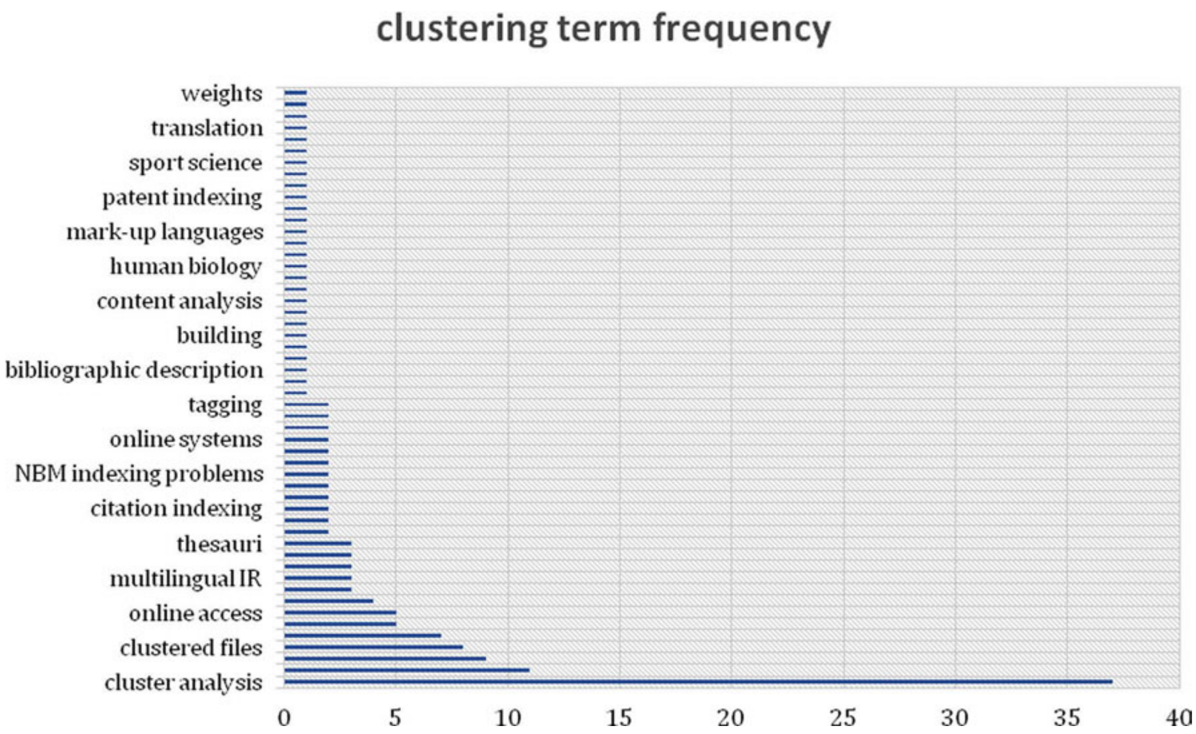

Figure 6. Keywords associated with clustering.

matic indexing," which seems to have peaked about 1996. There was no visible research front; a small group of 13 authors had more than one paper, only 3 had papers in more than one category, these are shown in Table 1.

There is no apparent research front that crosses the boundaries of the four lists. Oberhauser, Zhang and Moens are the only authors whose work appeared in more than one list, thus identified with more than one of the approaches.

Co-word analysis of title keywords also revealed little coherence. The richest group was the "clustering" result, which had 160 instances of 53 terms assigned; the most frequently occurring term assigned was "cluster analysis." "Automatic indexing" occurred in all four clusters. "Automatic classification" occurred in all but the "automatic indexing" cluster. "Machine learning" is not used in the "machine learning" cluster, which is dominated by "automatic classification." Phrases were extracted from all titles, and a frequency distribution of these is shown in Table 2.

https://doi.org/10.5771/0943-7444-2017-3-215

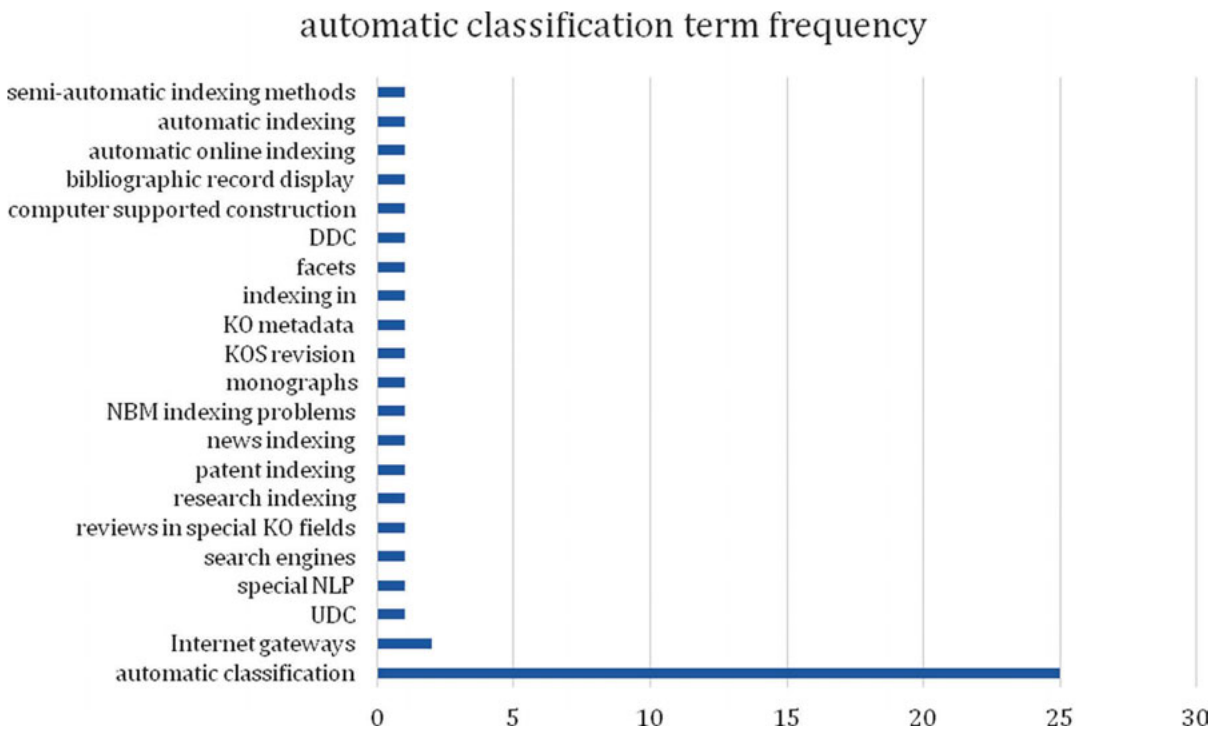

Figure 7. Keywords associated with automatic classification.

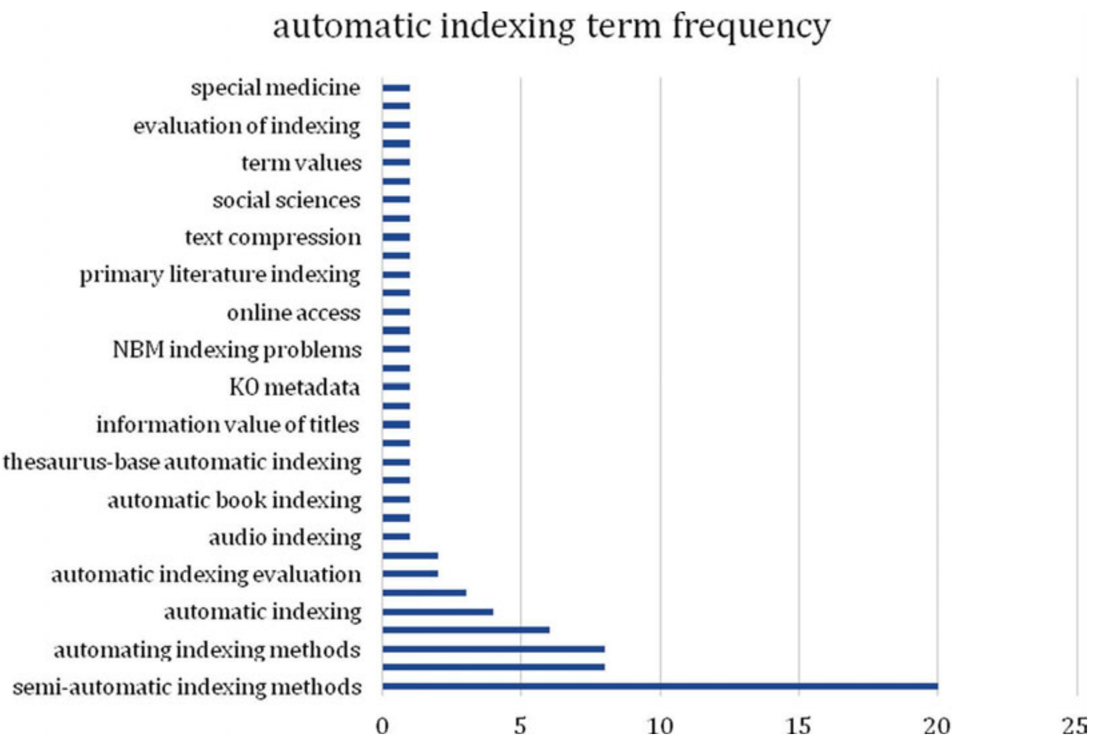

Figure 8. Keywords associated with automatic indexing.

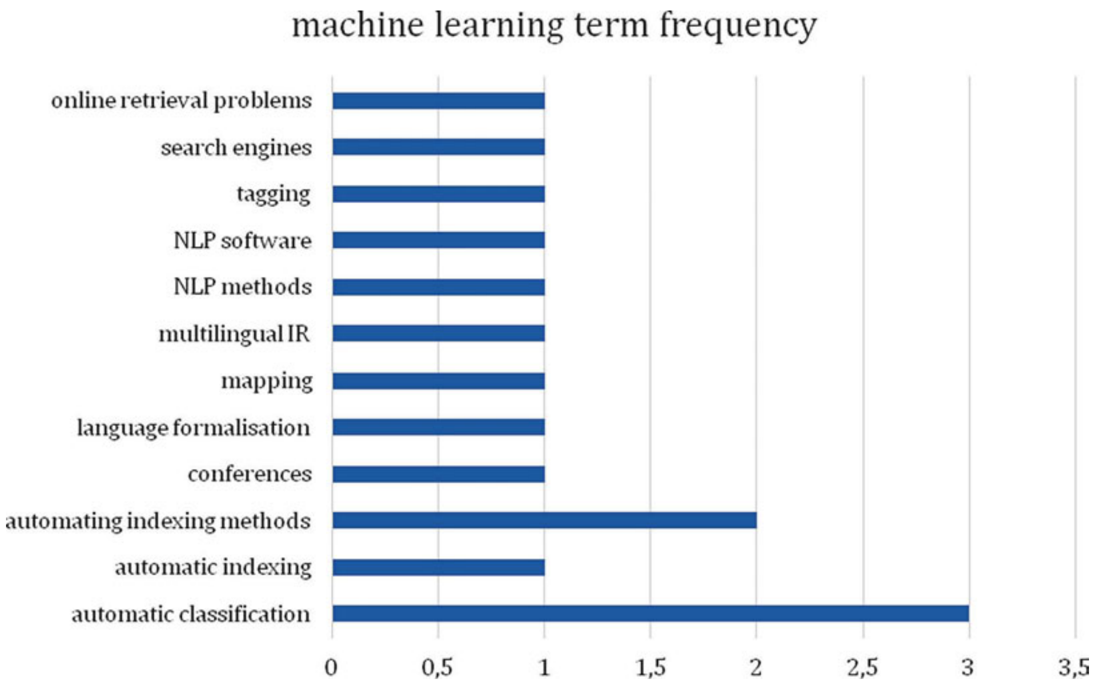

Figure 9. Keywords associated with machine learning.

Authors	Title	Result
Oberhauser, Otto	1 3	automatic indexing automatic classification
Zhang, Lin	2 1	clustering machine learning
Moens, Marie-Francine	1 1	automatic indexing clustering
Carlyle, Allyson	2	clustering
Corrêa, Carlos Alberto	2	automatic indexing
Gödert, Winfried	2	automatic indexing
Khoo, Christopher S.G.	2	clustering
Kishida, Kazuaki	2	clustering
Krauth, Joachim	2	clustering
Gil-Leiva, Isidoro	2	automatic indexing
Lepsky, Klaus	5	automatic indexing
Salton, Gerard	2	automatic indexing
Slivnitsina, N.A.	2	automatic indexing

Table 1. Most productive authors in Case 1.

All Titles	Frequency
automatic indexing	76
automatic classification	34
machine learning	12
document clustering	11
information retrieval	10
automatic indexing system	7
search results	6
automatische indexierung	5
automatic classification system	4
text categorization	4
clustering based	4
hybrid clustering	4
clustering algorithm	4
map for clustering of text	3
automatic indexing of library	3
organizing map for clustering	3
indexation automatique de fonds	3
clustering of text documents	3
indexing and clustering	3
available at http	3
indexing to improve	3
latent semantic indexing	3
zur maschinellen	3
data clustering	3
university library	3
subject access	3
automatisches klassifizieren	3
clustering technique	3
based approach	3
clustering analysis	3
indexing and abstracting	3
relational database	3

Table 2. Phrases from all titles.

"Automatic indexing," "automatic classification," and "machine learning" had little phrase differentiation except for linguistic variants (e.g., Automatisches Klassifizieren). The richest source of meaningful phrases was the clustering set. A visualization produced with WordStat™ software incorporating the most frequently occurring phrases and keywords appears in Figure 10.

The prominence of "systems" tells us that this group of papers is primarily focused on the design of systems. We also see the influence of "information retrieval" and its close relative "index[ing]." The impression formed from this first case study is that although all four approaches are well-represented in literature associated with KO, and the most activity seems to emanate from the clustering group, there is not a coherent domain visible in these data.

3.0 Case 2

Case two involves discovery of works located using each the four terms combined with the term "knowledge organization" in both Thomson-Reuters™' *Web of Science*™ (*WoS*) and Elsevier's *SCOPUS*® citation indexing services. This search was intended to serve as a control for the data in case one. That is, we were interested to see how the major indexing services represented these terms and whether it was similar to or different from the coverage in the KO Literature database. In fact, we found very little available in either source; 19 works were cited in *WoS* and 57 in *SCOPUS*.

The least coverage was in the *WoS*. Table 3 shows the first-named authors, sources and dates of publication for each of the four terms. The citations are given for each term in reverse chronological order. Only one citation appears in two groups—a paper from *Knowledge-Based Systems* with Alex Lopez-Suarez as first author (Lopez-Suarez and Kamel 1994).

None of the authors in Table 3 are represented in the results from the KO Literature database in Case 1. The representation is sparse over all, but only the "clustering" result, which is the largest, shows representation from 2016, the year in which the search was conducted.

There is some overlap between them. Co-word analysis by term was not possible because of the small size of the datasets for three of the terms. An MDS visualization was created using WordStat™ to show the most frequently occurring phrases in the titles and abstracts of the 19 papers found in *WoS* (Figure 11).

The picture of the research front becomes more clear; we see expert systems associated with domain knowledge and conceptual clustering, KOSs associated with patent documents and the technique known as topic-modeling, and KOS construction associated with means clustering and document sets. An attempt to analyze the "clustering" data separately revealed that only one phrase "knowledge organization" was present, and only those two keywords occurred more than twice. To the extent that KO is associated with the techniques in this group, then, it appears to be associated with "clustering," although Table 3 shows the only two approaches represented in the journal *Knowledge Organization* are automatic indexing and machine learning. The data are too sparse to suggest any conclusions.

There were more works cited in *SCOPUS*®. There were 57 works cited; first authors, sources and dates for these are shown in Table 4. These results are more robust.

The largest representation is in "clustering" with the second largest in "automatic classification." Chronologi-

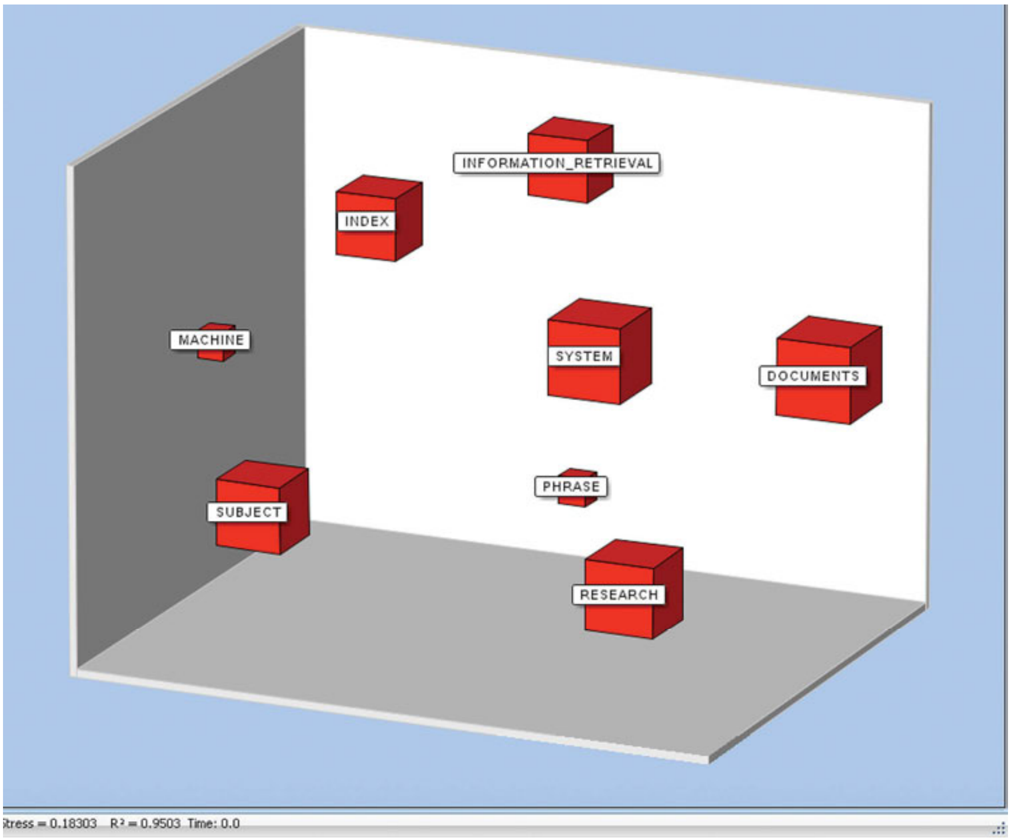

Figure 10. MDS plot of keywords and phrases in Case 1 titles.

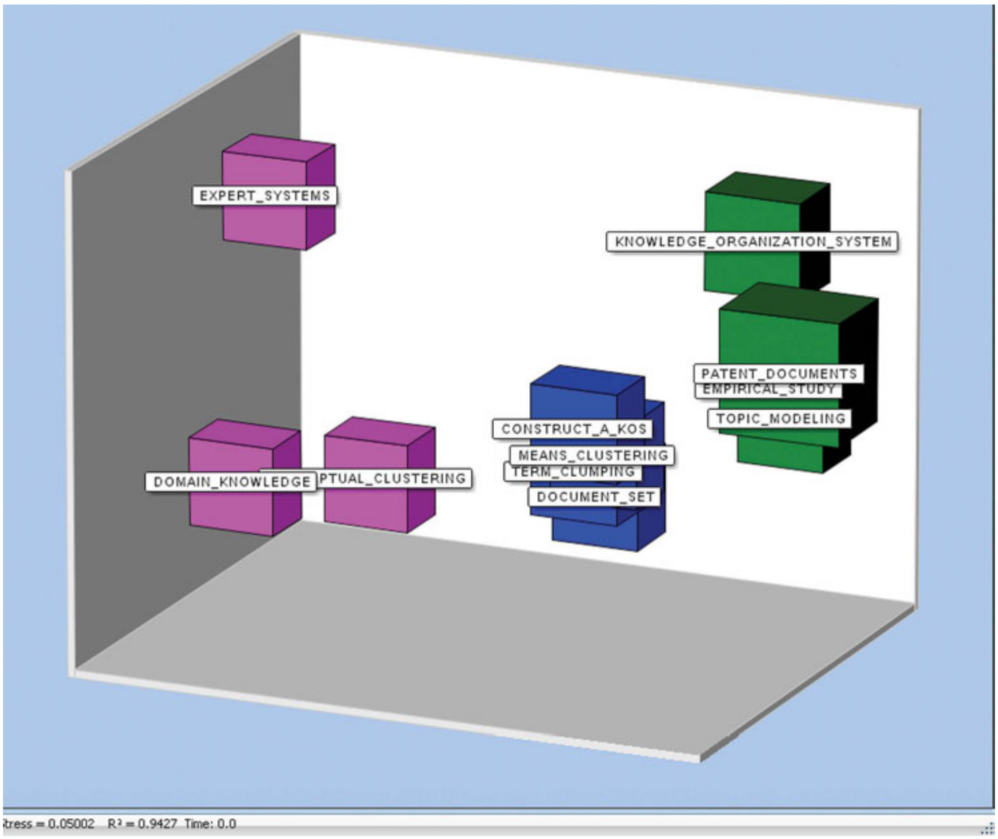

Figure 11. MDS plot of phrases from *WoS* in Case 2 (stress = 0.05002 R^2=0.9427).

Authors	**Journal**	**Year**
Clustering		
Namdar, Bahadir	*International Journal of Science Education*	2016
Majidi, Sharareh	*International Journal of Science and Mathematics Education*	2014
Hu, Zhengyin	*Scientometrics*	2014
Suchecki, Krzysztof	*Advances in Complex Systems*	2012
Koponen, Ismo T.	*Entropy*	2010
Perner, Petra	*Engineering Applications of Artificial Intelligence*	2006
Ibekwe-SanJuan, F	*Journal of Documentation*	2006
Theilhaber, J	*Bioinformatics*	2004
Cooper, L Z	*Journal of The American Society for Information Science and Technology*	2002
Bournaud, I	*Knowledge Engineering and Knowledge Management, Proceedings*	2000
Lopezsuarez, A	*Knowledge-Based Systems*	1994
Cheng, CS	*Computers & Industrial Engineering*	1992
Automatic Classification		
Hu, Zhengyin	*Scientometrics*	2014
Mackenzie, ML	*Library & Information Science Research*	2000
Ingwersen, P	*Libri*	1992
Automatic Indexing		
Sidhom, S	*Knowledge Organization*	2002
Machine Learning		
Cleverley, Paul H.	*Knowledge Organization*	2015
Stanojevic, Mladen	*Expert Systems with Applications*	2007
Lopezsuarez, A	*Knowledge-Based Systems*	1994

Table 3. WoS results by first author, source and date.

cally, "clustering" stretches from 1985 to 2016, "machine learning" from 1992 to 2014, "automatic indexing" from 2001-2015, and "automatic classification" from 1994 to 2016. Many more first authors are named; most of those from the *WoS* results are included but most do not match the most productive authors from the KO Literature database. Among the sources we see *Knowledge Organization* and *Journal of Documentation*, but notably not *Journal of the Association for Information Science and Technology* and we see many conference proceedings (including one international ISKO conference).

Co-word analysis of "clustering" titles and abstracts reveals 15 phrases that occur 4 or more times and that include all of the most frequently occurring keywords. These phrases are visualized in Figure 12.

Similarly Figure 13 is a visualization of phrases from titles and abstracts in "automatic classification."

Authors	**Journal**	**Year**
Clustering		
Namdar, B.A.	*International Journal of Science Education*	2016
Frost, S.	*Studies in Classification, Data Analysis, and Knowledge Organization*	2015
Gao, J.A.	*International Journal of Multimedia and Ubiquitous Engineering*	2015
Hu, Z.A.	*Scientometrics*	2014
Majidi, S.	*International Journal of Science and Mathematics Education*	2014
Matsui, Y.A.	*Studies in Classification, Data Analysis, and Knowledge Organization*	2014
Scaturro, I.	*Knowledge Organization*	2013
Zhu, L.	*ICIC Express Letters, Part B: Applications*	2012
Bedford, D.A.D.	*Proceedings of the 10th Terminology and Knowledge Engineering Conference: New Frontiers in the Constructive Symbiosis of Terminology and Knowledge Engineering, TKE 2012*	2012
Mitrelis, A.A.	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*	2012
Suchecki, K.	*Advances in Complex Systems*	2012
Nousiainen, M.	*Journal of Baltic Science Education*	2011
Wang, H.	*International Conference on Management and Service Science, MASS 2011*	2011
Li, G.	*Communications in Computer and Information Science, 233 CCIS (PART 3)*	2011
Wang, H.	*Advanced Materials Research*	2011
Wang, Y.-H.	*Jisuanji Jicheng Zhizao Xitong/Computer Integrated Manufacturing Systems, CIMS*	2010
Koponen, I.T.	*Entropy*	2010
Lu, H.A.	*Journal of Computational Information Systems*	2010
Sandström, U.	*12th International Conference on Scientometrics and Informetrics*	2009
Perner, P.	*Engineering Applications of Artificial Intelligence*	1006
Ibekwe-Sanjuan, F.	*Journal of Documentation*	2006
Hoskinson, A.	*Computer*	2005
Theilhaber, J.	*BMC Bioinformatics*	2004
Cooper, L.Z.	*Journal of the American Society for Information Science and Technology*	2002
Siddiqui, K.J.	*Proceedings of SPIE - The International Society for Optical Engineering*	2001
Bournaud, I.	*Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*	2000
Bournaud, I.	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*	2000
Lopez-Suarez, A.	*Knowledge-Based Systems*	1994
Cheng, C.-S.	*Computers and Industrial Engineering*	1992
Krishna, M.H.	*Proceedings of SPIE - The International Society for Optical Engineering*	1986
Cheng, Y.	*IEEE Transactions on Pattern Analysis and Machine Intelligence*	1985

Table 4. SCOPUS® results by first author, source and date. (Continued on page 226)

Authors	**Journal**	**Year**
Clustering		
Automatic Classification		
Zhu, Y.	*Scientometrics*	2016
Dong, H.	*Proceedings - 2015 IEEE International Conference on Smart City, SmartCity 2015, Held Jointly with 8th IEEE International Conference on Social Computing and Networking, SocialCom 2015, 5th IEEE International Conference on Sustainable Computing and Communications, SustainCom 2015, 2015 International Conference on Big Data Intelligence and Computing, DataCom 2015, 5th International Symposium on Cloud and Service Computing, SC2 2015*	2015
Diallo, G.	*Journal of Biomedical Semantics*	2014
Nevlud, P.	*Advances in Electrical and Electronic Engineering*	2013
Freire, N.	*Proceedings of the International Conference on Dublin Core and Metadata Applications*	2011
Mynarz, J.	*Grey Journal*	2011
Freire, N.	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*	2011
Mynarz, J.	*GL-Conference Series: Conference Proceedings*	2011
Brank, J.	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*	2008
Hu, K.	*Proceedings of the 6th IEEE International Conference on Cognitive Informatics*	2007
Sharma, A.	*Proceedings of the 3rd Indian International Conference on Artificial Intelligence*	2007
Stanojević, M.	*Expert Systems with Applications*	2007
Yu, P.	*Proceedings of the International Symposium on Test and Measurement*	2001
Lopez-Suarez, A.	*Knowledge-Based Systems*	1994
Automatic Indexing		
Albrechtsen, H.	*Knowledge Organization*	2015
Lima, G.A.B.	*Advances in Knowledge Organization*	2014
Mynarz, J.	*Grey Journal*	2011
Mynarz, J.	*GL-Conference Series: Conference Proceedings*	2011
Sidhom, S.	*Knowledge Organization*	2001
Machine Learning		
Mahesh, K.	*Advances in Knowledge Organization*	2014
Hu, Z.	*Scientometrics*	2014
Huang, T.	*WIT Transactions on Information and Communication Technologies*	2014
Chicaiza, J.	*Communications in Computer and Information Science*	2014
Sharma, A.	*Proceedings of the 3rd Indian International Conference on Artificial Intelligence*	2007
Mackenzie, M.L.	*Library and Information Science Research*	2000
Ingwersen, P.	*Libri*	1992

Table 4. SCOPUS® results by first author, source and date. (Continuation of page 225)

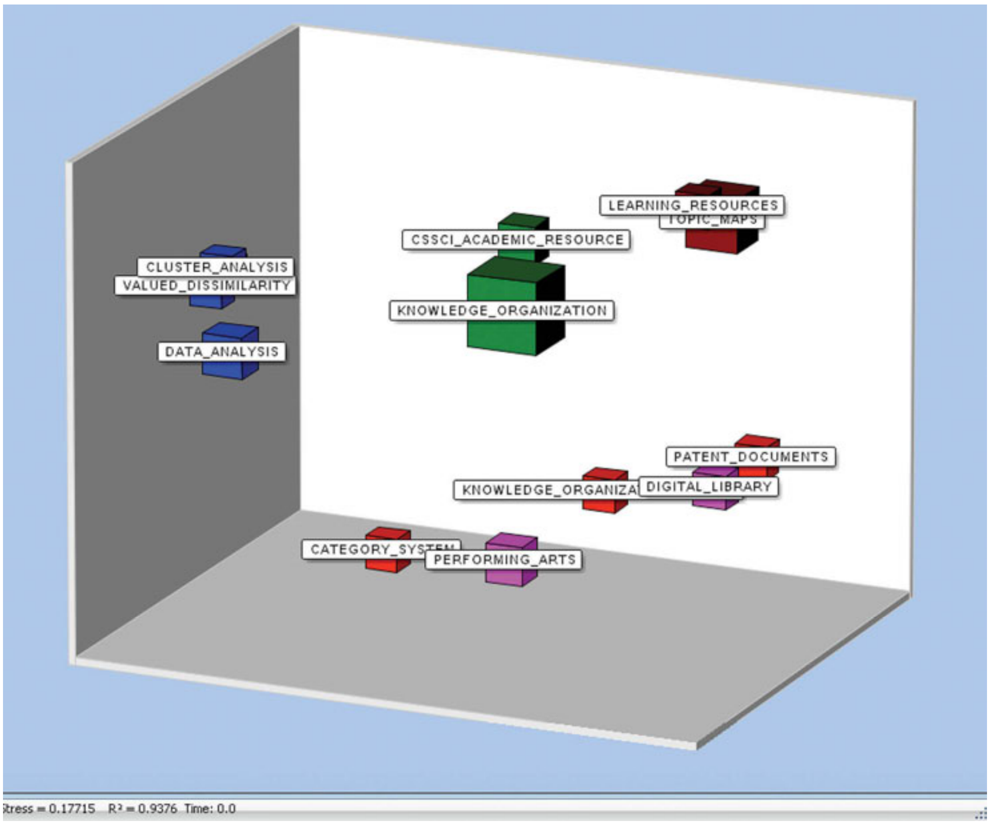

Figure 12. MDS plot of "clustering" phrases from *SCOPUS* in Case 2 (stress = 0.17715 R²=0.9376).

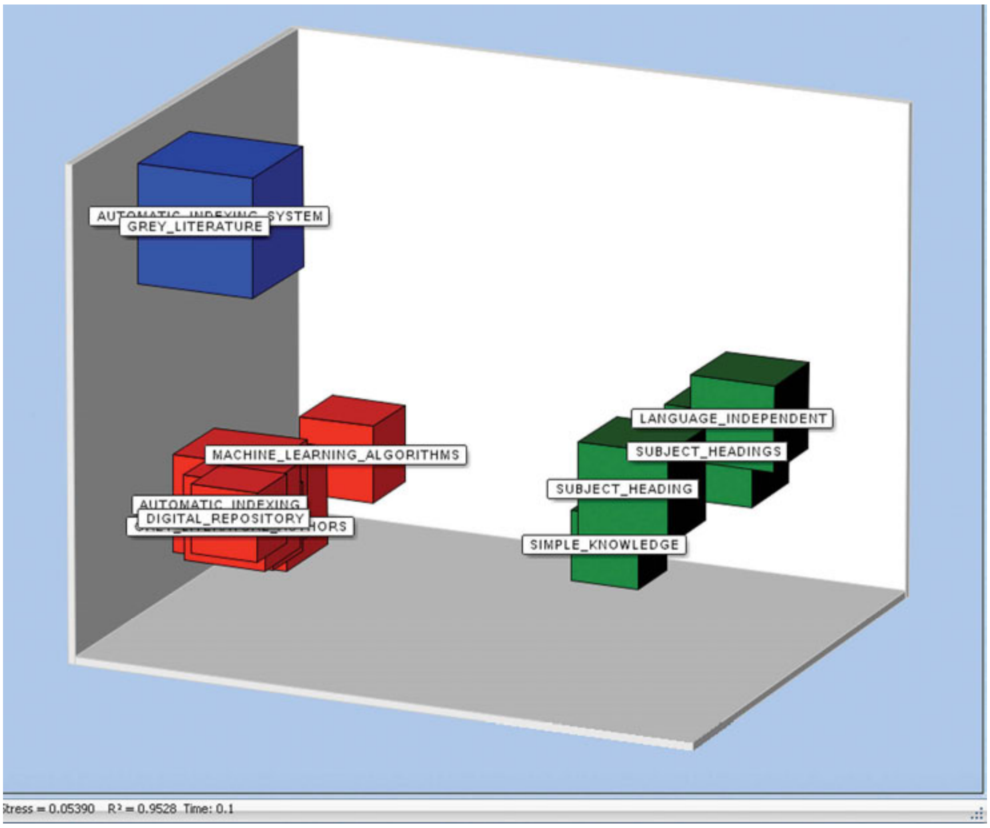

Figure 13. MDS plot of "automatic classification" phrases from *SCOPUS* in Case 2 (stress = 0.17715 R²=0.9376).

Together these two datasets contain most of the result from *SCOPUS*. In "clustering" we see "knowledge organization," "data analysis," "cluster analysis" and "digital libraries." In "automatic classification" we see "machine learning algorithms," "automatic indexing," "subject headings" and "simple knowledge [organization systems]." Collectively the thematic content of this data-set covers the main topics of "automatic classification," "automatic indexing" and "machine learning." Thus in case 2, we see that the authors and sources represented in the citation indexes are different from those in the KO Literature seen in case 1, but the chronological span and thematic content is quite similar. It is obvious that *WoS* does not begin to provide access to the literature of these four areas of emergent research in the KO domain.

5.0 Case 3: Discourse visible in the papers in this special issue

Finally, case 3 provides data triangulation by applying co-word analysis and citation analysis to the works cited in the papers in the present special issue. The five papers are identified in Table 5.

Sandra Collovini, Sandra and Renata Vieira	RelP: Portuguese Open Relation Extraction
Gil-Leiva, Isadoro	SISA: Automatic indexing system for scientific articles. Experiments with location heuristics rules versus TF-IDF rules.
Campos, Maria Luiza and Hagar Espanja Gomes	Ontology as Knowledge Organization System: role of definitions and relations in a domain conceptual modeling
Café, Ligia Maria Arruda and Renato Rocha Souza	Sentiment analysis and knowledge organization: an overview of the international literature
Ibekwe SanJuan, Fidelia and Geoffrey Bowker	Big Data. What can it mean for Knowledge Organization systems and research?

Table 5. Papers in this Special Issue on "New Trends" in KO.

These papers represent self-selection by their authors, who responded to an international call for papers on new and emergent trends in KO. Thus to some extent they constitute a different control group from within the KO domain for analyzing the emergence of computer science techniques into the science of knowledge organization. It is immediately clear that although there is thematic overlap between this group of papers and the contents of the cases described above, we still are not working with a coherent research front or a domain. Rather, we can analyze the works cited by the authors of these five papers for clues to the discourse that might be influencing the inclusion of computer science approaches in KO.

There were 272 works cited in the five papers. Research cited ranged chronologically from 1949 to the present. The mean age of cited work was 15.7 years; the median was 10 years and the mode was 3 years. The midpoint of the distribution was 2008 and the majority of works were published in the most recent three years. The mean suggests the social scientific content that is typical of KO research (Smiraglia 2015); it also suggests a fair amount of reliance on core content alongside recent research.

The largest source was journals; 76 journals were cited, of which 21 were cited more than once. The most cited journals are shown in Table 6.

Journal	Frequency
Journal of the Association for Information Science and Technology	22
Information Processing & Management	11
Knowledge Organization	10
Journal of the China Society for Scientific and Technical Information	6
Journal of Documentation	5
Journal of Information Science	5
Canadian Journal of Information and Library Science	4
Revista Española de Documentación Científica	4
Information Retrieval	3
International Classification	3
International Journal of Communication	3
Knowledge-Based Systems	3

Table 6. Most cited journals in Case 3.

This distribution is dominated by *JASIST*, with an equal number of citations shared by *IP&M* and *KO*. *Knowledge-Based Systems* is shared with the works cited in Case 2. Citations to works in conference proceedings comprise the next largest group of sources. The most-cited proceedings were from conferences shown in Table 7.

The table shows conference series, not individual conferences. We know from other research (e.g., Smiraglia 2015) that KO research typically is split half-and-half between journal articles and conference papers because KO is a domain that relies on frequent international sharing of current research results. We see that mirrored here. We also see some reliance on Portuguese sources, which is obviously related to the work by Brazilian and Portuguese

Proceedings	Frequency
ISKO International Conferences	5
HAREM [Named Entity Recognition for Portuguese]	5
LREC [Language Resources and Evaluation]	3
Computational Processing of the Portuguese Language [PROPOR]	3
IADIS International Conference on Applied Computing	2
ASIST	2
AMIA Symposium, 609–613	2
SIGIR	2

Table 7. Most cited conference proceedings in Case 3.

collaborative teams, a currently growing trend in KO (e.g., Smiraglia 2017). The conferences in Table 7 show a sort of interdisciplinarity that takes into account concept theory (ISKO), alongside natural language processing and related approaches, and (of course) information retrieval. Four theses were cited, and the rest of the citations were divided among monographs and websites, none of which was cited more than once.

As for an author research front, 32 authors were cited more than once; 17 were cited three or more times and these names are given in Table 8.

On this list we find Bowker, Gil-Leiva, Souza and Ibekwe-SanJuan, indicating a high proportion of self-citation among the authors of the papers. We also see Salton and Gil-Leiva, whose names occurred in the author list in Case 1. Additionally, we find Hjørland and Dahlberg, which is typical of KO literature in general.

The most frequently used keywords in these five papers are "indexing," "automatic" and "information." A co-word MDS plot produced using WordStat™ software visualizes the thematic material represented in the titles of these papers (Figure 14).

Here we see information retrieval, language processing, automatic indexing, and ontology, alongside the special interests of the authors, all connected distantly but distinctly to KO.

We suggested earlier that we thought case 3 would allow us to comment on discourse regarding new trends in KO. The first and most obvious observation is that there is a strong Brazilian and Portuguese component present among the five submitted papers. This might be an artifact of the guest editor, a Brazilian KO scholar; it might be an artifact of the 2016 ISKO International Conference in Brazil. It might reflect the possibility that new approaches to KO are coming from the emergent global leader in ISKO-Brazil. But it might also signal that there are now new scholars interested in the intersection of concept theory and information retrieval, and thus there now is a resurgence of interest in automatic methods for classification, especially with regard to machine-learning. A glimpse of Table 8 points to the discourse governing this work. Salton, a pioneer in information retrieval and automatic indexing, tops the list. Dahlberg, founder of KO and pioneer in every aspect of concept theory is highly cited, as is Hjørland, who typically is most-cited in KO today. But the other most highly cited authors in this table are the authors of these papers. This suggests (not, ironically, self interest) that this group of self-selected leaders in new trends are working from their own scientific bases to build theory on their own scientific achievements. The usual proportions in KO between journal articles and conference papers are observed here, but the journal distribution is skewed more toward global research and information retrieval than we usually see in

Cited author	Frequency
Salton, Gerard	8
Collovini, Sandra	5
Dahlberg, Ingetraut	5
Hjørland, Birger	5
Bowker, Geoffrey C.	4
Gil-Leiva, Isidoro	4
Na, Jin-Cheon	4
Smith, Barry	4
Souza, Renato Rocha	4
Auerbach, David	3
Bick, Eckhard	3
Evans, David A	3
Hersh, William R.	3
Humphrey, Susanne M	3
Ibekwe-SanJuan, Fidelia	3
Lancaster, Frederick W	3
Riggs, Fred W	3

Table 8. Most frequently cited authors in Case 3.

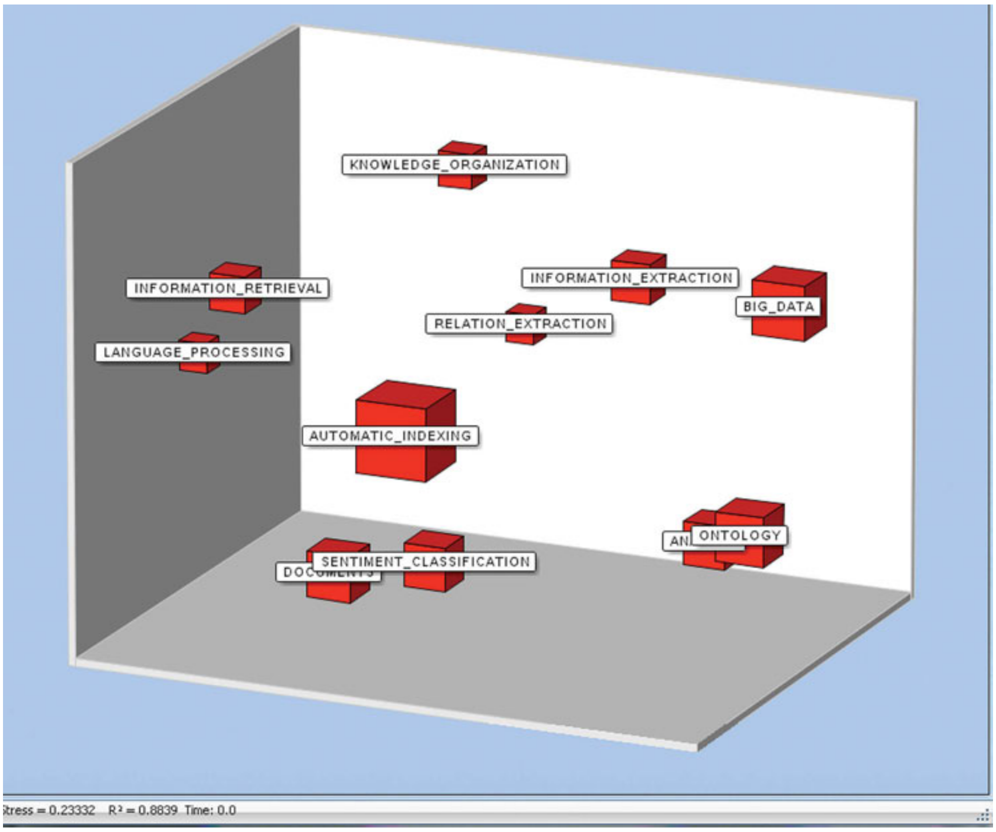

Figure 14. MDS plot of phrases and keywords from five papers in Case 3 (stress = 0.2332 R²=0.8839).

KO. Among the conferences cited by these authors are ISKO international conferences and some Portuguese language processing conferences. Despite the reliance on information retrieval, ASIST is not at the top of the conference list. This would suggest that our authors are drawing research from the information community, but are not finding its conferences to be as productive in applied research as those in KO and HAREM. Case 3, interestingly takes a tangential turn away from the four applied computer science areas in cases 1 and 2, and demonstrates new ways in which the KO community can bring these theories into practice in KO today.

6.0 There might be new trends but there is no domain in KO

We began this study with the idea in mind that we would discover a new subdomain within KO that was devoted to bringing digital solutions such as machine-learning and automatic classification to bear on the problems of the entire domain. By that we mean both the problems of defining and structuring appropriate KOSs for specific domains and the problems of subsequently using those KOSs to index documents. We have discovered a lively group of scholars centered around the use of what is called "clustering" and also around what is called "automatic classification." We have discovered that "automatic indexing" is often thought to be the same as "automatic classification" (although we acknowledge that it is quite different), and that "machine learning" has become a computer science paradigm that is larger than the problems of KO. In other words, we have demonstrated the fact that there are scholars involved in "clustering" and "automatic classification," and that they have a rich series of precedents over two decades, and that they share common thematic emphases. We were able to verify our data by generating Google N-grams for "machine learning," "automatic indexing," and "automatic classification." These are shown in Figures 15-17.

The Google N-gram Viewer compiles chronological visualizations of the frequency of occurrence of terms in books digitized by the Google Books project (http://books.google.com/ngrams). We can see that "machine learning" originated in this corpus as a term allied with today's understanding between 1935 and 1938, occurring together with knowledge representation and robotics, and systems theory. It seems to have entered the corpus substantially about 1960 and its usage is still growing. "Automatic classification" appeared as early as 1851 with reference to the notion of automatically assigning diagnoses

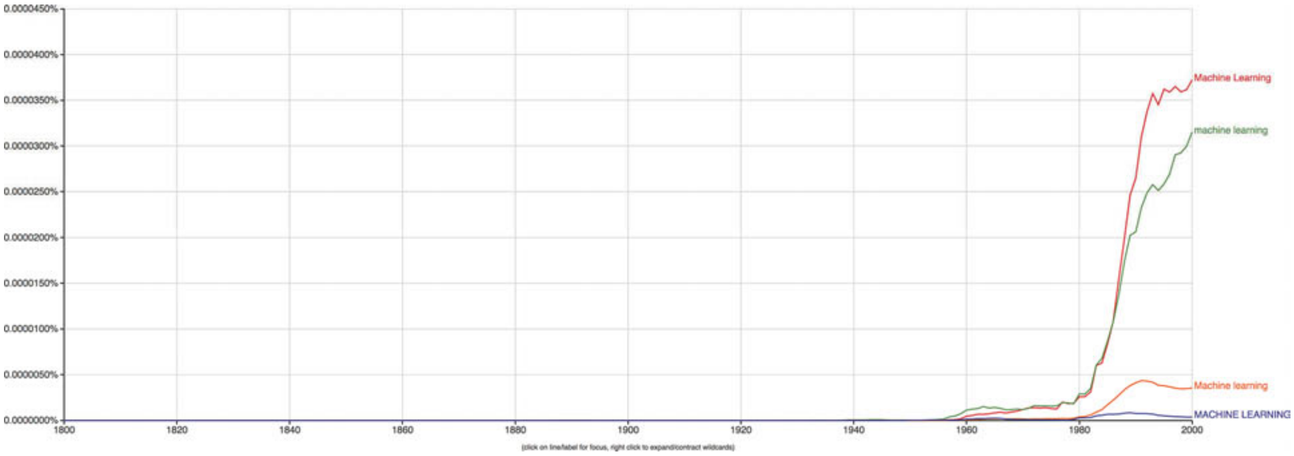

Figure 15. Google N-*gram* for "Machine learning."

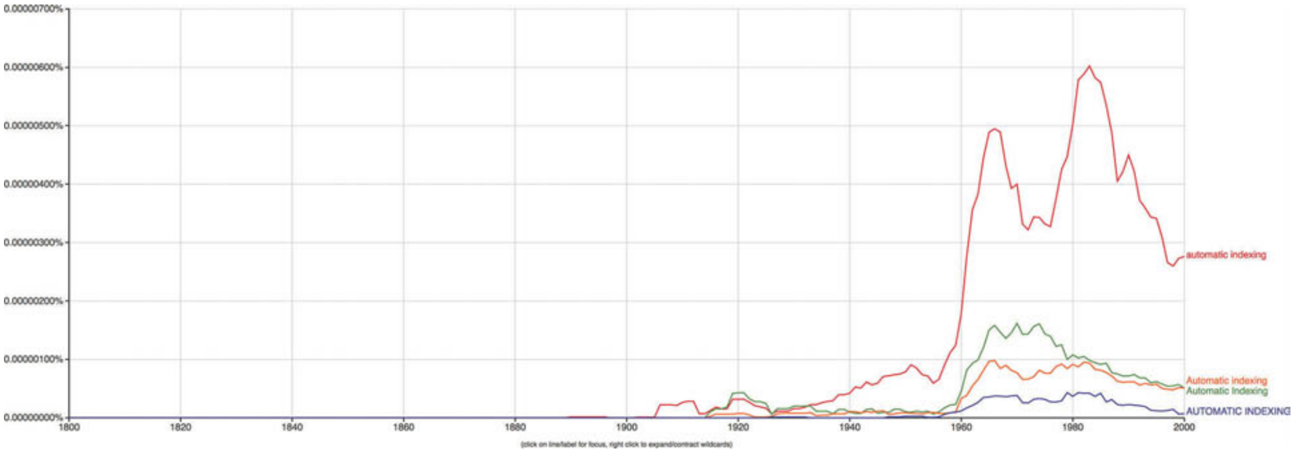

Figure 16. Google N-gram for "Automatic indexing."

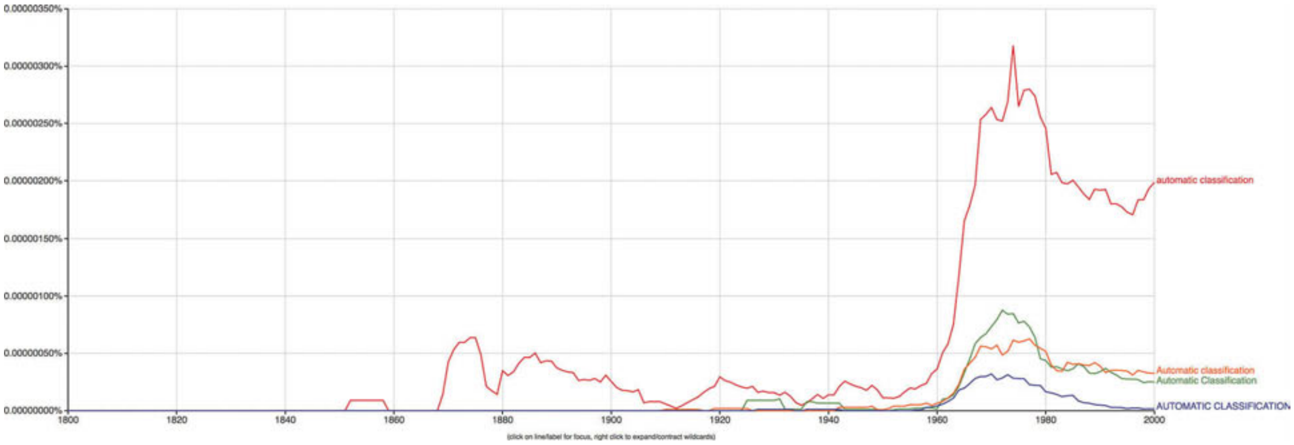

Figure 17. Google N-gram for "Automatic classification."

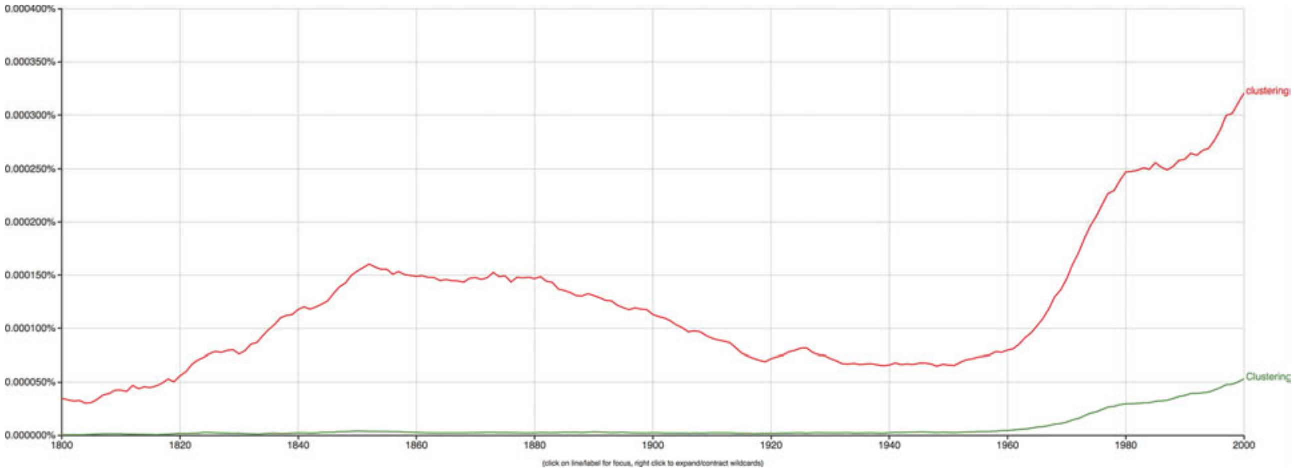

Figure 18. Google N-gram for "Clustering."

such as "healthy" or "abnormal," and by 1940 was associated with automatic character recognition; by 1965 it was associated with keyword classification for information retrieval, and this usage peaks in 1973. "Automatic indexing" appeared in 1889 with regard to mechanisms developed to refine gear rotation in several kinds of mechanical applications. It seems to have first been used in association with indexing and abstracting about 1934, and this meaning supplants the mechanization application about 1959 to reach a peak in 1983. "Clustering" is shown in Figure 18.

It is more difficult to analyze the N-gram for "clustering," because the term appears in the eighteenth century and peaks about 1850, then drops off again until 1960. The nineteenth century usage appears to be associated with normal semantic usage of the term, such as in the sense of clustering grapes. After 1965 usage is similar to that discovered in "automatic classification" and is aligned with information retrieval and computer applications. Nonetheless, we see that "automatic" approaches to knowledge organization have a long history, but gave way to "machine learning" in the end of the twentieth century. The Google N-grams offer data triangulation for the chronological results from our own datasets, which suggested that "automatic indexing," and "clustering" peaked in about 1996, "automatic classification" peaked in 2005, and all of it is now subsumed under "machine learning," of which "clustering" is an accepted component, which is still growing. But, none of it has critical mass among KO researchers.

We also have seen how this scholarship resides in proximity with information retrieval; many of the core phrases visualized here match those in a recent analysis of IR (see Raghavan, Apoorva and Jivrajani 2015). But, measuring according to the most recent definition of a domain (Smiraglia 2012, 114), we have found no coherence, no common activity and no social semantics. There is a common teleology among authors involved in "clustering" and "automatic classification."

What we have not found is a research front, or a common teleology within the KO domain. We see across our three cases that "clustering" and "automatic classification" are terms used by many scholars in KO, increasingly, to describe their work. We also have seen that those terms are used by scholars who might not have considered themselves to have been working in the KO domain, but who have contributed nonetheless to the KO domain. We have found some confusion concerning "automatic indexing" and "automatic classification," that might in the end influence how the research we are studying here is analyzed. We have not found a formal group of scholars in KO devoted to "machine learning," which might in the end be the approach most needed to keep the KO domain at the cutting edge of the research front as we approach the reality of a semantic web.

We have found in the end, a lively group of authors who have succeeded in submitting papers to this special issue, and their work quite interestingly aligns with the case studies we report. There is an emphasis on KO for information retrieval; there is much work on clustering (which involves conceptual points within texts) and automatic classification (which involves semantic groupings at the meta-document level). We have demonstrated the dearth of support for the KO domain coming from Thomson-Reuters *WoS* and the detailed indexing our domain receives from the *SCOPUS* product circle. Still, however, we have seen that the ISKO sponsored KO Literature Database is the most reliable source of citation data about these specific new trends in KO.

References

Dumais, Susan, John Platt, David Heckerman and Mehran Sahami. 1998. "Inductive Learning Algorithms and Representations for Text Categorization." In *Proceedings of the Seventh International Conference on Information and Knowledge Management, Bethesda, MD, USA, November 02 - 07, 1998*, ed. Kia Makki and Luc Bouganim. New York: ACM, 148-55. doi:10.1145/288627.288651

Gil-Leiva, Isidoro. 2017. SISA: Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules versus TF-IDF Rules. *Knowledge Organization* 44(1): xx-xx.

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke and Debra Hiom. 2016. "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval." *Journal of the Association for Information Science and Technology* 67: 3-16. doi:10.1002/asi.23600

Kohavi, Ron and Foster Provost. 1998. "Glossary of Terms." In *Applications of Machine Learning and the Knowledge Discovery Process*, ed. Thomas G. Dietterich. *Machine Learning* 30: 271-4.

Lopez-Suarez, Alex and Mohamed S. Kamel. 1994. "DYKOR: A Method for Generating the Content of Explanations in Knowledge Systems." *Knowledge-Based Systems* 7: 177-88. doi:10.1016/0950-7051(94)90004-3

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

Rajaraman, Anand and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*: Cambridge: Cambridge University Press.

Salles, Thiago, Leonardo Rocha, Marcos André Gonçalves, Jussara M. Almeida, Fernando Mourão, Wagner Meira Jr. and Felipe Viegas. 2016. "A Quantitative Analysis of the Temporal Effects on Automatic Text Classification." *Journal of the Association for Information Science and Technology* 67: 1639-67. doi:10.1002/asi.23452

Smiraglia, Richard P. 2012. *Cultural Frames of Knowledge*. Ed. Richard P. Smiraglia and Hur-Li Lee. Würzburg: Ergon Verlag.

Smiraglia, Richard P. 2015. "Domain Analysis of Domain Analysis for Knowledge Organization: Observations on an Emergent Methodological Cluster." *Knowledge Organization* 42: 602-11.

Smiraglia, Richard P. 2017. "ISKO 14's Bookshelf: Discourse and Nomenclature—An Editorial." *Knowledge Organization* 44: 3-12.

Soergel, Dagobert. 1974. "Automatic and Semi-Automatic Methods as an Aid in the Construction of Indexing Languages and Thesauri." *International Classification* 1: 34-9.