Maria da Graça Simões, Luís Miguel Machado, Renato Rocha Souza, Maurício Barcellos Almeida, António Tavares Lopes

Automatic indexing and ontologies: The consistency of the research chronology and authoring in the scope of Information Science.

Abstract

An alternative to minimize the semantic gap in automatic indexing systems is to make use of knowledge organization systems, like ontologies. Given that these two operational concepts are subjects of study in Information Science (IS), one can deem relevant to identify and analyze the continuity, chronological and authorial consistency of studies between ontologies and automatic indexing. We use as methodology an exploratory / descriptive study based on a systematic review. We conclude that a direct relationship (citation and co-authorship) between the articles analyzed is practically non-existent. Regarding exhaustiveness, specificity, precision and recall rates, we concluded that there are similarities among three of the bases used: LISS, LISTA and ISTA. LSD, the fourth base used, was the one that presented a lower performance in the retrieval, leading to low rates of exhaustiveness and specificity of the indexing process.

Keywords: Automatic Indexing; Ontologies; Collaboration

Introduction

The current society is no longer imaginable without considering the access to information and content in digital formats (Foskett, 1997; Shera & Cleveland, 1977). The digital content has been growing exponentially and is estimated to reach the ratio of five terabytes for each human being by 2020 (EMC, 2017). In this context, information retrieval becomes an increasingly complex task that justifies the importance of describing and identifying the content to be represented by terms (Lancaster, 2003; Stevens & Urban, 1965).

Considering the phenomenon of Big Data, dealing with large digital collections as currently available on the web have made unfeasible if one uses manual indexing. This situation has leveraged the automatic indexing, which traces back its origins to the decade of 1950 when there was the first spike in the availability of electronic texts (Baxendale, 1958; Luhn, 1957; Maron, 1961). Even though information retrieval systems have evolved over the past decades, they are still not efficient, as one would expect for searching based on themes or concepts. Issues like polysemy and synonymy still hinder automatic indexing, and consequently the document retrieval.

One alternative to minimize the semantic gap in automatic indexing systems is to make use of knowledge organization systems, like ontologies. Ontologies are defined here as engineering artifacts consisting of an intensional vocabulary used to describe a certain reality, along with explicit assumptions organized in a logical theory that represents concepts and relations in both clear and unambiguous ways (Gruber, 1992; Guarino, 1998). Although the typification of ontologies is not consensual, three types often emerge: top-level ontologies; domain ontologies; and application ontologies. The use of ontologies in automatic indexing systems, in particular, the use of domain ontologies, have already been explored. One can find these systems in an experimental stage, but presenting very promising results.

Given that these two operational concepts (automatic indexing and ontologies) are subjects of study with a long history in the field of Information Science (IS), the study of the relationship between them is still, to a certain extent, less known. The authors already took a step in this direction in a previous study (Simões, Machado, Souza, & Lopes, 2017), in which we have used a corpus extracted from two databases (*Library & Information Science Source* (LISS) and *Library and Information Science & Technology Abstracts* (LISTA)). We found in this work that there are many potentialities relating ontologies and automatic indexing, many of them reported in the IS field. However, taking into account the authorship analysis (25 articles accounting for 72 different authors where only one is present in two papers), one can deem relevant to identify and analyse the continuity, chronological and authorial consistency of studies between ontologies and automatic indexing. In this context, we extended the study, by including two more databases (Information Science & Technology Abstracts (ISTA), and Library Science Database (LSD)). We aim:

(i) identify and account for scientific papers dealing with these topics (automatic indexing and ontologies) in the databases LISS, LISTA, ISTA, and LSD, and to explore their temporal distribution and departmental affiliations of their authors;

(ii) to verify the overlap of the retrieved articles in the referred databases;

(iii) and to assess the relationship between the studies by mapping the direct citations between them, and the co-authorship in the articles cited in the corpora.

We use as methodology an exploratory / descriptive study based on a systematic review.

Methodology

To meet our goals, we designed an exploratory and qualitative study based on a systematic literature review, along with content analysis techniques (Bardin, 2011; Bernard & Ryan, 2010; Gil, 2008). To assess the relevance of the topics of automatic indexing and ontology, we adopted the same procedures as in the previous study (Simões et al., 2017). To collect the research sample, we queried the four databases using the expression "index* AND ontolog*" in the fields title and subjects. And, to assess the thematic proximity, we worked with four categorical variables that were expressed on a negative scale of intensity (Bardin, 2011), (see Table 1).

Table 1: Variables used to assess the thematic proximity of the papers

Thematic proximity of the paper to the concepts of automatic indexing and ontology	**Degree of proximity**	**Value**
The concept is a core part of the study.	Central	0
The concept is addressed because of its intrinsic relationship with the object of study.	Inherent	-0,5
The concept is addressed because of a secondary relation with the object of study.	Peripheral	-1,5
The concept is not addressed, but we infer a thematic connection.	Inferred	-3

We gathered the collection on October 10, 2017, from the four databases listed. In the comparison between the four, we applied the formulas presented by Ribeiro (1996) to the rates of precision[1] and recall[2]. The judgment of the relevance for the corpus is given by the criteria adopted and previously presented.

Concerning the comparison of the indexing terms observed in each database, we assessed the exhaustiveness and specificity of each database according to the following criteria:

- for the first concept, we computed the mean value of assigned terms;
- and, for the second, we considered the percentage of records that contained the terms ontology (s) and/or automatic indexing in the subject field, (half value if only one of the terms is present).

We want to note that, for the latter two concepts, we used the records of the relevant documents recovered in all four databases.

For the coincidence rate we used the Jaccard index ($| A \cap B | / | A \cup B |$), that is expressed by dividing the cardinality of the intersections of the two groups by the cardinality of their unions.

Regarding the direct relations of citation and co-authorship between the studies that constitute the corpus, we used techniques for analysing social networks and co-occurrence (Alvarenga, 1998; González-Teruel, González-Alcaide, Barrios, & Abad-García, 2015; Scott, 1991; Sugimoto & McCain, 2010).

Results and discussion

As a first result, we can say that the extension of the study to two other databases (ISTA and LSD) did not add new works to the corpus collected in the previous study (see table 2), so that their temporal distribution remains between the years 2003 and 2016 (see Figure 1).

Table 2 – Corpus of the study

Ref	Article	[a]Degree
01	Gilchrist, A. (2003) *Thesauri, taxonomies and ontologies - an etymological note.*	-1,5
02	Kabel, S. et al. (2004) *The added value of task and ontology-based markup for information…*	-0,5
03	Bayer, O. et al. (2005) *Evaluation of an Ontology-based Knowledge-Management-System…*	-2
04	Ciravegna, F. et al. (2006) *Annotating document content: a knowledge-management…*	0
05	Köhler, J. et al. (2006) *Ontology based text indexing and querying for the semantic web*	0
06	Tsinaraki, C. et al. (2007) *Interoperability Support between MPEG-7/21 and OWL in…*	0
07	Hernandez, N. et al. (2007) *Modeling context through domain ontologies*	0
08	Pirrò, G. et al. (2008) *LOM: a linguistic ontology matcher based on information retrieval*	-1,5
09	Allampalli-Nagaraj, G. et al. (2009) *Automatic semantic indexing of medical images using a…*	0
10	Moura, M.A. (2009) *Informação, ferramentas ontológicas e redes sociais Ad Hoc…*	-2
11	Good, B.M. et al. (2009) *Term based comparison metrics for controlled and uncontrolled…*	-2
12	Solskinnsbakk, G. et al. (2010) *Combining ontological profiles with context in information…*	-1,5
13	Bouramoul, A. (2011) *The Semantic Dimension in Information Retrieval, from Document…*	-1
14	Kara, S. et al. (2012) *An ontology-based retrieval system using semantic indexing*	-0,5
15	De Maio, C. et al. (2012) *Hierarchical web resources retrieval by exploiting Fuzzy Formal…*	-1,5
16	Chiaravalloti, M.T. et al. (2012) *URT "Indexing and Classification Systems" Projects and…*	-1,5
17	Willis, C. et al. (2013) *A Random Walk on an Ontology: Using Thesaurus Structure for…*	-0,5
18	Qiu, J. et al. (2014) *Constructing an information science resource ontology based on the…*	-1,5
19	Thenmalar, S. et al. (2014) *Enhanced ontology-based indexing and searching*	0
20	Bendib, I. et al. (2014) *Semantic ontologies for multimedia indexing (SOMI)…*	0
21	Gödert, W. (2014) *Ein Ontologie-basiertes Modell für Indexierung und Retrieval*	0

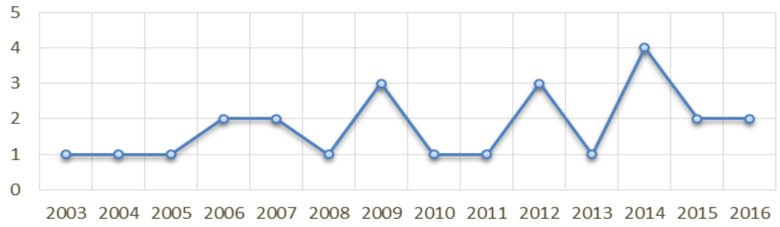
[1] The fraction of relevant instances among the retrieved instances.

[2] The fraction of relevant instances that have been retrieved over the total amount of relevant instances.

22	Pang, C. et al. (2015) *BiobankConnect: software to rapidly connect data elements for pooled …*	-2
23	Du Preez, M. (2015) *Taxonomies, folksonomies, ontologies: what are they and how do they …*	-1,5
24	Vlachidis, A.H. et al. (2016) *A Knowledge-Based Approach to Information Extraction for …*	0
25	Gödert, W. (2016) *An Ontology-Based Model for Indexing and Retrieval*	0

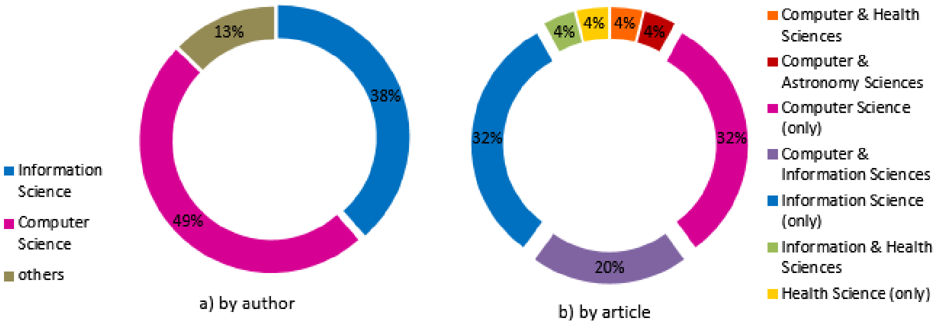
[a] Degree of thematic approximation determined by the sum of the value attributed to the proximity of the approach to each of the two concepts (see table 1), which must be greater than -3.

Figure 1: Temporal distribution of the constituent articles of the analysis corpus



Regarding the distribution of the affiliation of authors by area, we must note that the stated total (78 = 100%) is higher than the number of actual authors (72). This deviation originated from the fact that we considered the Department of Computer and Information Science (two occurrences) and the Department of Social Science Informatics (four) as both Computer Science (CS) and Information Science (IS). This decision set the difference of 11 percentage points in favour of CS (49%) in comparison to IS (38%), while the remaining 13% affiliated to other areas (Astronomy and Health Science, respectively with 1 and 9 authors) (see figure 2, chart a)).

Figure 2: Departmental affiliation of authors



The difference between Computer Science and Information Science coverage disappears, however, when the accounting is carried on the number of papers (see figure 2, chart b). There is an equal distribution of IS and CS papers (eight for each, representing 64% of the total of 25). There are also five articles (20%) that combine the two and, finally, each one of the other papers represents a single remaining area or combination (4%, each).

On the comparison of the four databases, there is a significant difference between LSD and the other three (see Figure 3). The perceptual amplitude of the difference between the LSD and the other three bases vary in the precision rate (between 9 and 21 points) and in recall (between 26 and 40). In the indexing terms aspects, the difference is, for exhaustiveness 16 to 27 percentage points and for specificity, from 4 to 14. In the same matter, LSD values were affected by the existence of two registers with no assigned term (Ref 01 and 02), a situation that, in LISS and in the LIST, occurs in another register (Ref 21). In this comparison, the LISTA base is the one that presents a greater similarity in the results; in the completeness and precision with the ISTA and in the recall with the LISS.

Figure 3: Results of measurements (recovery and indexing

		LISS	LISTA	ISTA	LSD
Retrieved documents (*Rd*)		31	28	17	9
Relevant Retrieved documents (*RR*)		20	21	13	5
Relevant documents (*Rt*)		21	22	16	9
Precision rate (*RR/Rd*)		65%	75%	76%	56%
Recall rate (*RR/Rt*)		95%	95%	81%	56%
Indexing	Exhaustivity	77%	66%	66%	50%
	Especificidy	48%	41%	38%	33%

The percentual amplitude of the difference between the LSD and the other three databases vary in the precision from 9 to 20 points, and in recall from 25 to 39. Concerning indexing terms aspects, the difference is, in exhaustiveness from 16 to 27 percentage points, and in specificity from ~~4~~5 to 1~~4~~5. The computation of indexing aspects of LSD values was affected by the existence of two records with no assigned term (ref 01 and 02), a situation that, for LISS and LISTA, occurred in another record (ref 21). In this comparison table, the LISTA database is the one that presents similarity to others in the results, although not simultaneously; in the completeness and precision with ISTA, and in the recall with LISS.

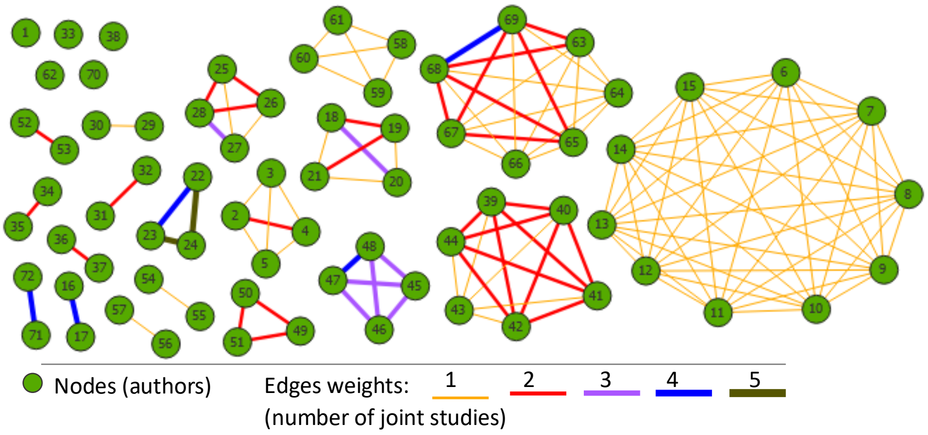
The same pattern, evidencing a difference between LSD and the other three databases, is verifiable in coincidence rates. The three lowest values in the two measurements (Relevant Recovered (*RR*) and Relevant Existing Documents (*Rt*)) report the overlap with the other three (see Figure 4). These values, below 40%, contrast with values above 50% of the other databases. We observed that the LISS database displayed the higher overlap rates in both measurements.

Figure 4: Databases overlap rates

Relevant Retrieved documents (*RR*)	LISS	LISTA	ISTA	LSD	
LSD	25%	24%	38%		LSD
ISTA	65%	55%		39%	ISTA
LISTA	64%		52%	19%	LISTA
LISS		59%	54%	20%	LISS
	LISS	LISTA	ISTA	LSD	Relevant documents (*Rt*)

As far as the relationship between the constituent studies of the corpus is concerned, there are only two direct citations between them (Ref. 19 cites Ref. 05 and Ref. 14 cites Ref. 06). Regarding co-authorship within the universe considered (corpus and its references, that is, 689 works), we verified that the link between the authors of the articles of the corpus is limited to the participating group in the studies themselves (see Figure 5), justifying the low network density (0.04).

Figure 5: Network of co-authorships



In the sociogram of figure 5, which represents the co-authorship relations, 24 unconnected elements are visible, corresponding to the 25 works of the corpus of study, since there is one author who represents two papers (ref 21 and 25). Within some of the elements, we can find authors with other works in common (depicted with lines whose assigned weight is greater than one).

Conclusions

Given the non-increase of the corpus of analysis, with respect to the previous study, the conclusions regarding the temporal distribution are maintained: the joint investigation of the two concepts (ontologies and automatic indexing) is recent in the Information Science area (little more than a decade), having its studies increased in the last five years.

The analysis of the departmental affiliations of the authors confirms the previously mentioned inference regarding the great proximity and interchange between the IS and CS areas in the study of these subjects. It will be possible, as regards to affiliation, to indicate a relationship between these operational concepts and the Health Sciences.

As far as the comparison between the four databases is concerned, we can conclude that there is a close proximity between the LISS, LISTA and ISTA databases, in terms of completeness, specificity, precision and recall rates. We can infer that in such a situation, it is closely related to the fact that the three are included in the EBSCO platform, while the fourth base, the LSD, is inserted in another platform, ProQuest. This last base was the one that presented a lower performance in the retrieval for which, according to the analysis made, the low rates of completeness and specificity of indexation were decisive.

The methodology used in the collection of the corpus emphasized the importance of indexing, particularly its exhaustiveness for a higher recall rate. Although the LISTA database shows a recall equal to the LISS and exhaustiveness 10 percentage points lower than this, this does not contradict the previous statement since LISTA's revocation rate increased due to the retrieval of the title and not the subject of two papers relevant on other bases. We can conclude from the sample that all four databases show little specificity in indexing, with rates below 50% due to lack of use (LISS and LISTA), or even total absence (ISTA and LSD) of terms from automatic indexing.

In the domain of the coincidence rates between the bases, we can infer that the association between the two services (EBSCO and ProQuest) is the reason why we have close values for LISS, LISTA and ISTA bases, as well as a great difference of these rates between the three bases listed and LSD. A comparison with previous studies (Vinson & Welsh, 2014) becomes impractical given the different methodologies adopted and bases used.

Finally, regarding the relationship between the constituent studies of the corpus of analysis, we conclude that the inexists direct relationship between them in the two indicators (direct and co-authoring). The tiny amount of direct citations (two) combined with the lack of other co-authorship between the authors represented in the corpus, in addition to the articles analyzed in the study, points to a weak exchange between the various actors.

This last conclusion and its inference poses some uncertainty in the author's consistency in the study of these matters in the area of IC. This uncertainty is somewhat mitigated by the finding of previous work by authors of the corpus, both individually and in co-authorship, in the bibliographical references analyzed. In contrast, the question is why those papers were not retrieved at the time of the search in the bases. These concerns raise future lines of inquiry, in particular with regard to the mapping of indirect relations, not only in terms of citations, but also the extension of co-authorship to others, in addition to the 72, which may act as bridges or peak (Scott, 1991) increasing the density of the network.

References

Alvarenga, L. (1998). Bibliometria e arqueologia do saber de Michel Foucault: traços de identidade teórico-metodológica. *Ciência Da Informação*, *27*(3), 253–261. https://doi.org/10.1590/S0100-19651998000300002

Bardin, L. (2011). *Análise de conteúdo*. (L. A. R. A. Pinheiro, Trans.). São Paulo: Almedina.

Baxendale, P. B. (1958). Machine-made index for technical literature: An experiment. *IBM Journal of Research and Development*, *2*(4), 354–361.

Bernard, H. R., & Ryan, G. W. (2010). *Analysing qualitative data: Systematic Approaches.* Los Angeles: SAGE.

EMC. (2017). The Digital Universe of Opportunities. Retrieved August 8, 2017, from https://www.emc.com/infographics/digital-universe-2014.htm

Foskett, D. J. (1997). Thesaurus. In *Readings in information retrieval* (pp. 111–134).

Gil, A. C. (2008). *Métodos e Técnicas de Pesquisa Social* (6th ed.). São Paulo: Atlas S.A.

González-Teruel, A., González-Alcaide, G., Barrios, M., & Abad-García, M. F. (2015). Mapping recent information behavior research: an analysis of co-authorship and co-citation networks. *Scientometrics, 103*(2), 687–705.

https://doi.org/10.1007/s11192-015-1548-z

Gruber, T. R. (1992). What is an Ontology? *International Journal Human-Computer Studies*, *43*, 907–928. Retrieved from http://tomgruber.org/writing/ontolingua-kaj-1993.htm

Guarino, N. (1998). Formal ontology and information systems. In N. Guarino (Ed.), *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98)* (pp. 3–15). Amsterdam: IOS Press.

Lancaster, F. W. (2003). Do indexing and abstracting have a future? In *Anales de Documentación*.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309–317.

Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *Journal of the ACM (JACM)*, *8*(3), 404–417.

Ribeiro, F. (1996). *Indexação e controlo de autoridade em arquivos*. Porto: Câmara Municipal do Porto, Arquivo Histórico. Retrieved from http://hdl.handle.net/10216/10721

Scott, J. (1991). *Social network analysis: A handbook*. London: SAGE Publications Ltd.

Shera, J. H., & Cleveland, D. B. (1977). History and foundations of information-science. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 12, pp. 249–275). New York: Knowledge Industry Publications Inc.

Simões, M. da G., Machado, L. M. O., Souza, R. R., & Lopes, A. T. (2017). Indexação automática e ontologias: Identificação dos contributos convergentes na Ciência da Informação. *Ciência Da Informação*, *46*(1), 152–162. Retrieved from http://revista.ibict.br/ciinf/article/view/4020/3459

Stevens, E. M., & Urban, G. H. (1965). Automatic indexing using cited titles. In *Statistical Association Methods for Mechanized Documentation Symposium Proceedings* (Vol. 1964, p. 213).

Sugimoto, C. R., & McCain, K. W. (2010). Visualizing changes over time: A history of information retrieval through the lens of descriptor tri-occurrence mapping. *Journal of Information Science*, *36*(4), 481–493. https://doi.org/10.1177/0165551510369992

Vinson, T. C., & Welsh, T. S. (2014). A Comparison of Three Library and Information Science Databases. *Journal of Electronic Resources Librarianship*, *26*(2), 114–126. https://doi.org/10.1080/1941126X.2014.910407