Vàllez, Mari; Pedraza-Jiménez, Rafael; Blanco, Saúl; Codina, Lluís; Rovira, Cristòfol (2015). Updating controlled vocabularies by analysing query loas. *Online Information Review.* Vol. 39 No. 7.

Updating controlled vocabularies by analysing query logs

Abstract

Purpose – Controlled vocabularies play an important role in information retrieval. Numerous studies have shown that conceptual searches based on vocabularies are more effective than keyword searches, at least in certain contexts. Consequently, new ways must be found to improve controlled vocabularies. This paper presents a semi-automatic model for updating controlled vocabularies through the use of a text corpus and the analysis of query logs.

Design/methodology/approach – An experimental development is presented in which, first, the suitability of a controlled vocabulary to a text corpus is examined. The keywords entered by users to access the text corpus are then compared with the descriptors used to index it. Finally, both the query logs and text corpus are processed to obtain a set of candidate terms to update the controlled vocabulary.

Findings – This paper describes a model applicable both in the context of the text corpus of an online academic journal and to repositories and intranets. The model is able to: a) identify the queries that led users from a search engine to a relevant document; and b) process these queries to identify candidate terms for inclusion in a controlled vocabulary.

Originality/value – The proposed model takes into account the perspective of users by mining queries in order to propose candidate terms for inclusion in a controlled vocabulary.

Research limitations/implications – Ideally, the model should be used in controlled Web environments, such as repositories, intranets or academic journals.

Social implications – The proposed model directly improves the indexing process by facilitating the maintenance and updating of controlled vocabularies. It so doing, it helps to optimise access to information.

Article Type – Research paper.

Keywords – indexing languages, controlled vocabularies, keywords, full-text search, query logs, metadata, information retrieval.

Introduction

The volume of digital data is currently doubling in size every two years (IDC, 2014). In this environment, indexing languages are a key component of information systems, especially in contexts associated with highly complex, high-quality information, such as professional or academic information. In such cases, they are used both to represent the content of documents and to facilitate access to them. In other words, they are used both to index the documents and to specify users' information needs.

Thesauri, taxonomies, ontologies and authority lists are all examples of indexing languages in which the vocabulary is controlled (Pedraza-Jiménez et al., 2009). They are called 'controlled vocabularies' because they use sets of descriptors to prevent the ambiguity of natural language. To this end, in these languages each concept is identified with a unique term. An exhaustive compilation of controlled vocabularies can be found at Taxonomy Warehouse.[i]

However, in the current Google era, users have grown used to simple search systems, in which they need only enter keywords in a search box to be taken to the information they need. Although these systems have proven to be quite effective, they nevertheless have certain shortcomings. For instance, in scientific literature searches, they can make it difficult for researchers to find the exact information required (Bowen et al., 2009; Kajanan et al., 2014). One reason for this is because academic databases, and complex information repositories in general, lack the signals, such as link analysis, that Google uses to return Web pages.

The trend imposed by search engines has rekindled the debate over the usefulness and viability of controlled vocabularies (Beall, 2008; Hjørland, 2012). The debate dates back decades, although it initially focused on the use of controlled vocabularies at libraries. Numerous studies have shown that controlled vocabularies continue to be an essential tool for helping users access information in the aforementioned contexts (Gross et al., 2015; Nowick and Mering, 2003; Rowley, 1994).

As a result, indexing processes based on controlled vocabularies remain necessary in contexts that both make intensive use of information with common features and require effective information retrieval systems. Intranets and subject or institutional repositories are two other clear cases in which the indexing process is effective (Tejeda-Lorente et al., 2014; White, 2013).

In such cases, indexing can be performed manually by human experts or by automatic or semi-automatic systems, which accelerate the process while at the same time ensuring consistency (Moens, 2002; Olson and Wolfram, 2008). The proposal presented in this paper falls into this category. First, it checks for overlap between the keywords entered by users in a search engine to access a set of documents and the descriptors assigned in a mixed manual and semi-automatic indexing process. It then proposes a method for updating and adapting the controlled vocabulary. Specifically, it describes a semi-automatic model that makes it possible to enrich the controlled vocabulary used in a subject portal for academic journals through the keywords entered by users to access the articles.

The fact that the proposed model is semi-automatic is crucial in this context. First, in contrast to fully automatic models, under this model the proposed keywords are reviewed and evaluated by a human expert, thereby preventing the imprecision inherent to purely automatic approaches. At the same time, however, the computer-aided processing of large volumes of queries and the automatic identification of keywords significantly reduce the time a human expert needs to spend keeping the controlled vocabularies up to date (Moine et al., 2014; Spasić et al., 2008). Both features make this type of model optimal in the context of indexing.

The rest of this paper is organised as follows: first, the literature on controlled vocabularies is reviewed; second, the aims of the research are stated; third, the methodology used is described and, more specifically, the data sets, processing tools, and method for checking the suitability of the controlled vocabulary; fourth, the proposed model for updating controlled vocabularies is explained; and fifth, and finally, the conclusions are presented, along with possible future lines of work.

Literature review

The theoretical framework used in this study was the intersection of information seeking and retrieval with knowledge representation, from a library and information science perspective.

In this regard, the usefulness of controlled vocabularies in facilitating access to and retrieving information has been addressed extensively in the literature, particularly from the perspective of information searches at libraries. The two main approaches are: metadata-based indexing systems and full-text search systems.

In their exhaustive review of the literature generated by the debate over controlled vocabularies versus keyword searches, Gross and Taylor (2015) identified two main trends: (1) controlled vocabularies should be discontinued in favour of keywords; (2) the success of keyword searches depends on the role played by controlled vocabularies, particularly with regard to results. They also mentioned several proposals to do away with the *keyword search* vs *controlled vocabulary search* dichotomy. The main conclusions they reached were: first, that the two systems are complementary and not mutually exclusive; second, that keywords can be used to enlarge controlled vocabularies; and third, that the difficulty of applying controlled vocabularies can be lessened through the development of more user-friendly tools. They thus ratified the conclusion reached in their previous study (Gross and Taylor, 2005), namely, that controlled vocabularies continue to be essential tools for helping users search for information, even when the full text of abstracts and tables of contents are included, which is the specific contribution of their more recent work. This finding is also applicable to digital newspaper archives, the use of which is on the rise.

Numerous studies have also looked at the effectiveness of full-text keyword searches, particularly in relation to biomedical literature (Kostoff, 2010; Müller et al., 2008; Shah et al., 2003). In this regard, Beall (2008) offered a detailed overview of the main problems affecting keyword searches, including linguistic factors, search engine shortcomings, imprecision or lack of knowledge of the terms by users, and the opaque Web, among other challenges. To this end, proposals have emerged for hybrid systems that combine both metadata and document text fields to facilitate information retrieval (Kim et al., 2005).

The ANSI/NISO Z39.19-2005 standard, revised in 2010, establishes guidelines and conventions for controlled vocabularies. It groups them into the following levels, depending on their degree of complexity: lists of descriptors, synonym rings, taxonomies and thesauri (NISO, 2010, p. 16). The first are simply limited sets of descriptors provided in an alphabetical list, also known as a ‘selection list’. Synonym rings provide equivalent descriptors for a concept. Taxonomies, in contrast, organise concepts hierarchically. Finally, thesauri also include the semantic relationships between concepts.

The NISO standard also ascribes five purposes to controlled vocabularies. The first is translation, as they provide a means for converting natural language into a vocabulary that can be used to improve indexing and classification. The second is consistency, as they promote uniformity with regard to formats and the assignment of descriptors. The third is the indication of relationships among the descriptors. The fourth is labelling and browsing, as they provide consistent hierarchies in online navigation systems to help users locate content. The fifth and final is information retrieval, as they help users locate content.

This list should be expanded to include the concept of interoperability. ISO 25964-1:2011 (2011) defines it as the ‘ability of two or more systems or components to exchange information and to use the information that has been exchanged’ (definition 2.29). It moreover considers that ‘vocabularies can support interoperability by including relations to other vocabularies, by presenting data in standard formats and by using systems that support common computer protocols’. This feature facilitates end-user access to information from various platforms and even in different languages by making it possible to use a unified system.

Thus, the use of controlled vocabularies boosts efficiency in information retrieval. It helps avoid the problems of polysemy and homonymy, when a term has more than one meaning, and of synonymy, when a single concept can be designated by more than one word. With the former, it increases precision by returning only pertinent records. With the latter, it increases coverage, by including records that are designated with alternative terms.

Controlled vocabularies can be used in different stages of the information search process. They are thus compatible with keyword searches. For example, controlled vocabularies can be used to filter results by topic, facilitate recommendation systems, or even help to summarise content (Murphy et al., 2003). They are currently used at repositories (Haniewicz, 2012; White, 2013), and they have traditionally been widely used with specialised databases (Kharazmi et al., 2014; McKenzie, 2001), which require a high level of effectiveness with regard to searches.

Finally, it is necessary to review the various approaches used to identify the descriptors that make up a controlled vocabulary, i.e. the lexical units used to designate concepts within a subject-specific domain. The different terminology extraction methods can be classified into three approaches: linguistic, statistical, and hybrid (Estopà, 1999; Pazienza et al., 2005; Zhang et al., 2012).

The first method, the linguistic one, uses linguistic resources and natural language processing techniques to identify the terms. To this end, lexical-syntactical patterns are defined to identify strings of terms that match the defined structures (Golik et al., 2013; Vállez and Pedraza-Jiménez, 2007). Because of the type of knowledge and resources used, linguistic approaches tend to rely on a specific language and domain. Additionally, many records are required to build the lexical, morphological and syntactical patterns, making this method costly to implement.

The statistical approach identifies terminological units based on their frequency in a corpus. It also includes more complex, primarily probability-based measures, such as the Log-likelihood ratio, the Pearson's Chi-square test, the T-score measure, the Dice coefficient or Mutual Information (Lyse and Andersen, 2012; Nazar, 2011). This approach does not allow generalisations and its strategies are unrelated to the language. Furthermore, when only statistical knowledge is used, the results are conditioned by the size and type of the text corpus, as a large number of candidate terms remain unidentified due to their low frequency of use while terms with no terminological value may be identified as candidates.

Thus, the most common approach to term extraction is a hybrid one that uses both statistical and linguistic information (Meijer et al., 2014; Sclano and Velardi, 2007). In these systems, the order in which the various types of knowledge are applied matters. To optimise the terminology extraction process, the linguistic analysis must be performed first, followed by the use of statistical techniques. The statistical data provides information about the use of the terms and helps to clarify the exact context in which they are being used.

Aims of the research

The aims of this research were as follows:

1. To determine the extent to which a controlled vocabulary reflects how users actually access information through keyword searches in search engines.
2. To propose a model for updating and/or maintaining controlled vocabularies based on query mining, that is, on analysing the keywords entered by users to access Web content.

Based on the above aims, the following research questions were defined:

a) Do the keywords entered by users match the controlled-vocabulary descriptors? Do they match the descriptors used in manual and/or semi-automatic indexing processes?

b) Is the design of a model that uses the keywords entered by users to access Web content in order to maintain and update the controlled vocabularies used in a Web environment viable?

Methodology

The different data sets and processing tools used in this research are described below. This is followed by a final section explaining the method used to determine the extent to which the controlled vocabulary is suitable for the text corpus.

Data sets

The text corpus selected to conduct this evaluation consisted of a random selection of 100 articles, in HTML format and in Spanish, from *BiD Textos Universitaris de Biblioteconomia i Documentació*, a journal specialised in library and information science indexed in the portal *Temaria*. This portal indexes articles from Spanish journals on library and information science and can be accessed online. It currently includes articles published in 14 Spanish journals. Table I shows the corpus's main characteristics.

Table I Description of the text corpus

Text corpus	
Publication name	BiD Textos Universitaris de Biblioteconomia i Documentació
Publisher	Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona
Years	2005 - 2013
Number of documents	100
Number of terms	504,551

The articles were indexed with descriptors from the Library and Information Science Thesaurus *(Tesauro de Biblioteconomía y Documentación*), a controlled vocabulary developed by the Spanish *Instituto de Estudios Documentales sobre Ciencia y Tecnología* (IEDCYT) (Mochón and Sorli, 2002). Table II offers a summary of the elements and relationships established in the thesaurus.

Table II Description of the controlled vocabulary

Controlled vocabulary	
Name	Library and Information Science Thesaurus
Publisher	Instituto Español de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT)
Number of concepts	1,097
Number of non preferred terms	569
Number of broader terms	1,008
Number of narrower terms	1,072
Number of related terms	2,354
Total number of terms (descriptors)	1,481

Additionally, based on data obtained from Google Analytics, the search-query logs from a period of 7 years were analysed to determine how users accessed the text corpus from search engines (Jansen et al., 2000; White and Horvitz, 2014). Table III offers an overview of the query corpus, including: the period analysed, the number of documents from which data were extracted, the number of queries and of visits they generated, and the average number of keywords in each query. It also provides information on certain features of the queries, such as the number of queries formulated as questions, the number of queries using Boolean operators, and the number of literal queries.

Table III Description of user queries (based on data from Google Analytics)

User queries	
Period analysed	01-04-2006 / 31-03-2013
Number of documents	100
Number of queries *	4,297
Visits to documents	13,438
Average query length	4.8
Question-type queries **	357
Queries with Boolean operators	77
Literal queries ***	226

* only queries leading to visits lasting ≥ 120”

** queries with a question word or a question mark

*** queries with more than 15 words or quotes

On average, the queries were 4.8 words long. Queries containing 15 words or more were assumed to be literal queries, intended to locate the exact passage entered in a given document. They were thus considered not to be made up of representative keywords for the study being performed and were excluded.

Indexing and processing tools

The text corpus is indexed in two different ways: first, by human indexers for *Temaria*, a portal of electronic journals on library and information science; and, second, by the tool DigiDoc MetaEdit, a metadata editor that allows the description of the content of HTML pages (Pedraza-Jiménez et al., 2008). This tool can be configured to decide which aspects to assess when HTML documents are processed to assign keywords. DigiDoc MetaEdit included the same thesaurus used by human indexers in order to enable comparison of the results obtained with the indexing tool. In the manual indexing, each article was assigned between 2 and 8 descriptors, with an average of 4.14 descriptors and a standard deviation of 1.37. In the semi-automatic indexing, 5 or 10 descriptors were assigned in order to determine which cut-off yields the best results.

Separately, the *Natural Language Toolkit (NLTK)* (Bird, 2006) package was used, which enables text processing in a large number of languages (Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Porter, Portuguese, Romanian, Russian, Spanish and Swedish). The tool was used to process information and obtain statistical data on the corpus. It was also used to carry out the linguistic process of stemming. Thus, the experiments were performed with two corpora, the original and the stemmed version, thereby unifying singular and plural forms.

Finally, the *Ngram Statistics Package (NSP)* (Banerjee and Pedersen, 2003) was used. NSP is a suite of Perl programs that identifies significant multi-word units (n-grams) in written text using many different tests of association. N-grams are combinations of *n* consecutive words, for example, the keyword ‘information retrieval’. In this case, the software was used to obtain the representative keywords in each query for subsequent comparison (Gencosman et al., 2014). N-grams ranging from unigrams to four-grams meeting the following two conditions were extracted: (1) the first element had to be a noun; and (2) the final element could not be a stop word, a verb or an adjective. A total of 1,282 n-grams with a frequency greater than 3 and meeting both requirements were extracted. This list was then run through various filters to eliminate proper nouns, words in other languages, and n-grams forming part of a descriptor from the controlled vocabulary being used. In the end, a total of 340 n-grams were left that could be candidate terms for inclusion in the controlled vocabulary.

Method

The method used in this study tried to determine how well-adapted a controlled vocabulary was to the current manner of accessing information through keyword searches in search engines. To this end, first, the keywords entered by users were examined to determine how much overlap there was with the descriptors in the thesaurus. It was then determined whether the overlap was greater with the descriptors assigned by human indexers (i.e. through manual indexing) or those assigned through semi-automatic indexing.

The results revealed a need to better adapt the controlled vocabulary. Thus, a model for updating/maintaining controlled vocabularies using the keywords entered by users to access Web content was proposed. The model uses the tools described above to process user-entered keywords in order to produce a list of candidate terms for inclusion in the controlled vocabulary. The terms are then sorted according to a relevance formula that is used both to identify the most relevant terms and as an indicator of their terminological importance.

The defined relevance formula was as follows:

$$\frac{Term\ frequency}{Documents} \times Visits\ \times Avg.length\ visits\ \times\ n\_gram\ length \qquad (1)$$

The formula’s multiplication factors are described below.

The first part of the formula refers to the number of times a keyword appeared in the search-query corpus (term frequency) and the number of documents in which it appeared with a view to enabling identification of the most relevant keywords for the corpus. Specifically, a high keyword frequency in searches denotes a keyword’s importance; however, the formula modulates this importance by dividing it by the number of documents to which the keyword enables access. Therefore, less importance is given to the most common keywords, which have less power of discrimination. This is an adaptation of the TF-IDF measure to the document corpus used, which, in this case, consisted of user queries.

The second part of the formula comprises the number of visits to a document generated by a given keyword (visits) and the average length of the visit (measured in seconds). It is intended to determine the relevance of the user-entered keywords according to users’ ‘information needs’. A high number of visits indicates that a keyword is quite relevant. Likewise, longer visits are assumed to reflect greater user interest in the visited document. Therefore, the formula multiplies the number of visits generated by a keyword by the average visit length.

The last part of the formula considers the semantic relevance of the user-entered keywords by taking keyword length into account as an important variable. In the field of controlled vocabularies, the disambiguation of descriptors is a basic requirement. This disambiguation can only be achieved with precise keywords, which tend to be phrases. Therefore, the formula considers it to be a positive factor in terms of relevance when an identified keyword consists of more than one word.

The final output is thus a list of keywords prioritised to allow a human expert to assess whether or not they should ultimately be included in the controlled vocabulary.

Results

Table IV shows the degree of overlap between the controlled vocabulary and the text corpus and between the controlled vocabulary and the user queries.

Table IV Overlap between the controlled vocabulary and the the text corpus or user queries

	Text corpus	**User queries**
Stemming	52.4%	23.2%
Original	39.9%	18.6%

The data show that out of all descriptors comprising the controlled vocabulary, 52.4% appear at least once in the text corpus subjected to stemming. In other words, of the 1,481 descriptors that make up the controlled vocabulary, 776 appear in the documents. This percentage drops considerably, to 39.9%, when the list is compared with the original, non-stemmed corpus. When it is compared with user queries, the percentage falls even further, to 23.2%, since, of the 1,481 descriptors included in the thesaurus, only 344 were entered by users. In other words, 76.8% of the descriptors included in the controlled vocabulary did not reflect how users actually expressed themselves.

Table V shows the percentage of overlap between the keywords entered by users in their queries and the descriptors assigned by the indexer (Manual column) or the automatic system (Automatic column). The overlap was greater when the indexing process was performed automatically. Moreover, the overlap was even greater (53%) when only the first five descriptors offered by the automatic system were considered, indicating that the first descriptors automatically assigned by the DigiDoc MetaEdit tool are the most relevant (Vállez et al., 2015). On the other hand, although the overlap did increase by a few percentage points when the comparison was performed with the stemmed corpus, the gain was not substantial.

Table V Overlap between user queries and indexing

	Manual indexing	**Automatic indexing**	
		5 terms	**10 terms**
Stemming	34.5%	53%	40.1%
Original	30.2%	48.6%	35.1%

The results show that in the best-case scenario, the overlap was just 53%. This could be for three possible reasons: (1) the controlled vocabulary used in the indexing process was not well-adapted to the corpus; (2) the descriptors assigned during the indexing process were not the most suitable; and (3) users formulated their queries using keywords that were not the most representative ones for the documents in question. The last two possibilities were not considered in this study, as they cannot be objectively verified. Instead, this paper focuses on the first possibility and on proposing a model to improve and adapt controlled vocabularies to the contexts in which they are used.

The data in Tables IV and V show that controlled vocabularies should be adapted and updated. One effective way to do this is to take into account the keywords entered by users to access documents. This finding enabled the development of a model to update the descriptors in a controlled vocabulary that takes the perspective of users into account.

Model for updating controlled vocabularies

Based on the above results, a model was proposed built on the following points: query-log analysis, keyword processing and candidate terms.

This model, in which each point is also a step, consisted of: (1) processing and analysing the data obtained from the query logs in order to identify the user-entered keywords or phrases that allowed users to access the documents from search engines; (2) reprocessing the data using linguistic and statistical tools in order to exclusively isolate those keywords (n-grams) that could be included in the controlled vocabulary; and (3) applying the relevance formula to these keywords to obtain a prioritised list of candidate terms.

Query-log analysis

Search-query logs compiled with Google Analytics were processed to identify the keywords entered in search engines to access each document.

Table VI shows an example of the data collected and processed for each document. Each row corresponds to a query and includes: the keywords entered to access the document, the landing page (always the same document), the number of visits generated, and the average length of each visit.

Table VI Google Analytics data for a document from the text corpus

Keyword	Landing page	Visits	Av. length of visit
dspace manual	/bid/20rodri2.htm	102	121.59
applications php files	/bid/20rodri2.htm	65	126.71
rename language dspace	/bid/20rodri2.htm	51	343.90
d space	/bid/20rodri2.htm	39	120.05
install dspace in windows	/bid/20rodri2.htm	33	409.52
dspace settings	/bid/20rodri2.htm	27	223.56
how to install linux dspace	/bid/20rodri2.htm	13	147.69
dspace how to customize?	/bid/20rodri2.htm	11	332.18
how to create users dspace	/bid/20rodri2.htm	11	176.73
manual on dspace	/bid/20rodri2.htm	10	927.30
processes for managing documentation based	/bid/20rodri2.htm	9	473.67
dspace use	/bid/20rodri2.htm	9	232.67
cache: i-qabfgnflgj: www.ub.edu/bid/20rodri2.htm	/bid/20rodri2.htm	9	155.56
how to install dspace	/bid/20rodri2.htm	7	131.57
set statistics on dspace	/bid/20rodri2.htm	7	316.00
dspace configuration	/bid/20rodri2.htm	7	235.71
procedure to create dspace collection	/bid/20rodri2.htm	7	226.43
dspace ub	/bid/20rodri2.htm	6	215.17

In order to ensure that the keywords entered were relevant to accessing the document, only those queries that resulted in visits lasting more than 120” were included. This time was established as the minimum cut-off due to the type of documents under consideration, i.e. scholarly papers. A two-minute visit to a document can be considered a standard minimum time for users to get a complete picture of it, thereby indicating that the document was relevant to the query (Huntington et al., 2008). Thus, a total of 4,297 queries were processed for the 100 documents.

Keyword processing

Once the queries that had led users to relevant documents had been identified, they were processed to identify the significant keywords they contained. To this end, the software *Ngram Statistics Package (NSP)* was used to identify the n-grams, or strings of *n* consecutive words, that could be considered candidate terms for inclusion in the controlled vocabulary.

The Method subsection above describes the various stages of the process in more detail. Ultimately, the list of n-grams obtained consisted of 340 terms, along with a series of information for each one, including: its frequency in the query corpus, the number of component words, the number of documents containing it, the number of visits generated, and the average length of the visits. Table VII shows the top ten candidates by frequency.

Table VII List of n-grams obtained from the analysis of the queries

	Freq. n-gram	**n-gram length**	**Doc. with n-gram**	**Visits by n-gram**	**Avg. length of visits by n-gram**
competency evaluation	133	3	2	477	633
examples	89	1	41	202	486
social networks	86	2	5	972	586
uses	56	1	30	156	600
ideas	51	1	3	241	600
web 2.0	50	2	7	231	590
definition	45	1	16	167	597
proposals	35	1	8	107	576
analysis	34	1	15	95	475
management systems	33	3	6	78	752

The keywords obtained were then submitted to two further processes. The first consisted in excluding any unigrams, which are usually quite generic and not particularly pertinent to the subject area of the controlled vocabulary. The second consisted in excluding those candidate terms used to access only one document, which were deemed to be unrepresentative. In all, 260 candidate terms were eliminated through these processes, resulting in a final list of only 80 candidates for inclusion in the vocabulary.

Candidate terms

The final step of the model was to sort the candidate terms by processing the information on each one. Different relevance formulas were considered for this purpose. Ultimately the one described above was chosen, as it struck the best balance among the different inputs for each candidate. Table VIII shows the top ten candidates according to the applied relevance formula.

Table VIII Candidate terms for the controlled vocabulary

n-gram	Freq. n-gram	n-gram length	Doc. with n-gram	Visits by n-gram	Avg. length of visits by n-gram
competency evaluation	133	3	2	477	633
social networks	86	2	5	972	586
digital identity	30	2	2	159	642
Web 2.0	50	2	7	231	590
evaluation systems	19	3	2	42	852
evaluation of skills	21	3	2	64	488
management systems	33	3	6	78	752
assessment tools	24	3	5	67	673
institutional repositories	32	2	9	109	635
higher education	24	2	4	68	530

Finally, an expert was given the ordered list of candidate terms to assess them for potential inclusion in the controlled vocabulary.

Conclusion

To present the conclusions of this paper, it is necessary to revisit the two research questions underpinning the study:

1. Do the keywords entered by users match the controlled-vocabulary descriptors? Do they match the descriptors used in manual and/or semi-automatic indexing processes?

   This paper looked at whether user-entered keywords matched the descriptors of a controlled vocabulary. The degree of overlap was shown to be low, at just over 23% (Table IV). The paper also looked at the overlap between the indexing processes (both manual and semi-automatic) and the user-entered keywords. In this regard, it found that the best-case scenario offered an overlap of only 53% (Table V). The correlation was thus higher than in the first case. The answers to both questions show that the adaptation of controlled vocabularies is crucial to optimising the indexing process. To address this need, this paper has proposed a model that processes user-entered queries in order to produce a list of candidate terms for inclusion in the controlled vocabulary.

2. Is the design of a model that uses the keywords entered by users to access Web content in order to maintain and update the controlled vocabularies used in a Web environment viable?

   This paper proposed a new model to facilitate and expedite the process of updating controlled vocabularies. The list of candidate terms included current terms (e.g. Web 2.0, social networks, digital identity) that are often absent from traditional controlled vocabularies due to the difficulty of keeping them up to date. Because the proposed model is semi-automatic, the final output is reviewed by a human expert. It thus leverages the benefits of both automatic and manual systems.

The use of controlled vocabularies and the need to keep them up to date is justified by the shortcomings of search systems based exclusively on the use of keywords. As a result of the influence of Internet search systems, keyword searches are the most widely used information retrieval system today. However, numerous studies have shown that these systems are not sufficiently efficient in contexts in which the documentary objects are complex (e.g. academic publications). In these cases, the use of indexing languages is a viable option for optimising the information retrieval process, as they offer an efficient way of representing the documents' content. Therefore, controlled vocabularies remain an essential tool and they must be able to be suitably updated.

First, a new approach was used that made it possible to improve the controlled vocabularies used to describe documents. The proposed model thus directly impacts end users by offering additional access points to information, thereby enhancing information retrieval. Moreover, the system is simple and inexpensive to implement, as it uses open-source processing tools and the analytical data are available. Indeed, although the model works best in controlled Web environments, such as repositories, intranets or academic journals, the viability of its practical application may be one of its greatest strengths.

Finally, mention should be made of two lines of future research that are currently being considered: first, the development of an application to enable the use of the described model in various corpora in order to help test the proposal; and, second, the development of an environment to facilitate the described process in order to ensure that the controlled vocabulary is constantly updated.

Acknowledgements

This article is part of the projects 'Audiencias activas y periodismo' (Active audiences and journalism; CSO2012-39518-C04-02) and 'El turista en la Web: hábitos informacionales y selección de destino turístico' (Tourists online: information habits and choice of tourism destination; CSO2014-59896-P) funded through grants under the Spanish National Research, Development and Innovation Plan of the Spanish Ministry of Economics and Competitiveness.

References

Banerjee, S. and Pedersen, T. (2003), "The design, implementation, and use of the ngram statistics package", in Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science: Vol. 2588, Springer, Berlin, pp. 370–381.

Beall, J. (2008), "The weaknesses of full-text searching", *The Journal of Academic Librarianship*, Vol. 34 No. 5, pp. 438–444.

Bird, S. (2006), "NLTK: The natural language toolkit", *Proceedings of the 21st International Conference on Computational Linguistics*, COLING-ACL'06, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 69–72.

Bowen, P.L., O'Farrell, R.A. and Rohde, F.H. (2009), "An empirical investigation of end-user query development: the effects of improved model expressiveness vs. complexity", *Information Systems Research*, Vol. 20 No. 4, pp. 565–584.

Estopà, R. (1999), *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona, Spain, available at: http://www.tdx.cat/handle/10803/7489 (accessed 1 February 2015).

Gencosman, B.C., Ozmutlu, H.C. and Ozmutlu, S. (2014), "Character n-gram application for automatic new topic identification", *Information Processing & Management*, Vol. 50 No. 6, pp. 821–856.

Golik, W., Bossy, R., Ratkovic, Z. and Claire, N. (2013), "Improving term extraction with linguistic analysis in the biomedical domain", in Gelbukh, A. (Ed.), *Advances in Computational Linguistics*, Research in Computing Science, Centro de Investigación en Computación del IPN, México, Vol. 70, pp. 157–172.

Gross, T. and Taylor, A.G. (2005), "What have we got to lose? The effect of controlled vocabulary on keyword searching results", *College & Research Libraries*, Vol. 66 No. 3, pp. 212–230.

Gross, T., Taylor, A.G. and Joudrey, D.N. (2015), "Still a lot to lose: the role of controlled vocabulary in keyword searching", *Cataloging & Classification Quarterly*, Vol. 53 No. 1, pp. 1–39.

Haniewicz, K. (2012), "Local controlled vocabulary for modern web service description", in Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A. and Zurada, J.M. (Eds.), *Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science: Vol. 7267, Springer, Berlin, pp. 639–646.

Hjørland, B. (2012), "Is classification necessary after Google?", *Journal of Documentation*, Vol. 68 No. 3, pp. 299–317.

Huntington, P., Nicholas, D., Jamali, H.R. (2008), "Website usage metrics: A re-assessment of session data", *Information Processing & Management*, Vol. 44, No.1, pp. 358–372, doi:10.1016/j.ipm.2007.03.003.

IDC. (2014), *The digital universe of opportunities: rich data and the increasing value of the internet of things*, Massachusetts, USA: IDC Analyze the Future, available at: http://www.emc.com/leadership/digital-universe/2014iview/index.htm (accessed 20 April 2015).

ISO. (2011), *Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval. ISO 25964-1:2011*, Geneva, Switzerland: International Organization for Standardization, available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657 (accessed 20 April 2015).

Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life, real users, and real needs: a study and analysis of user queries on the web", *Information Processing & Management*, Vol. 36 No. 2, pp. 207–227.

Kajanan, S., Bao, Y., Datta, A., VanderMeer, D. and Dutta, K. (2014), "Efficient automatic search query formulation using phrase-level analysis", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 5, pp. 1058–1075.

Kharazmi, S., Karimi, S., Scholer, F. and Clark, A. (2014), "A study of querying behaviour of expert and non-expert users of biomedical search systems", *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, ACM, New York, NY, USA, doi:10.1145/2682862.2682871.

Kim, S.S., Myaeng, S.H. and Yoo, J.-M. (2005), "A hybrid information retrieval model using metadata and text", in Fox, E.A., Neuhold, E.J., Premsmit, P. and Wuwongse, V. (Eds.), *Digital Libraries: Implementing Strategies and Sharing Experiences*, Lecture Notes in Computer Science: Vol. 3815, Springer, Berlin, pp. 232–241.

Kostoff, R.N. (2010), "Expanded information retrieval using full-text searching", *Journal of Information Science*, Vol. 36 No. 1, pp. 104–113.

Lyse, G.I. and Andersen, G. (2012), "Collocations and statistical analysis of n-grams", in Andersen, G. (Ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, Studies in Corpus Linguistics, John Benjamins Publishing, Amsterdam, The Netherlands, pp. 79–109.

McKenzie, E.M. (2001), “Natural language searching: How win works in Westlaw”, *Legal Reference Services Quarterly,* Vol. 18 No. 4, pp. 39–47.

Meijer, K., Frasincar, F. and Hogenboom, F. (2014), “A semantic approach for extracting domain taxonomies from text”, *Decision Support Systems*, Vol. 62, pp. 78–93.

Mochón, G. and Sorli, A. (2002), *Tesauro de biblioteconomía y documentación*, CSIC, Madrid, Spain.

Moens, M.-F. (2002), “Automatic indexing: The assignment of controlled language index terms”, in Zhai, C. and de Rijke, M. (Eds.), *Automatic indexing and abstracting of document texts*, The Information Retrieval, Springer US, New York, pp. 103–132.

Moine, M.-P., Valcke, S., Lawrence, B.N., Pascoe, C., Ford, R.W., Alias, A., Balaji, V., et al. (2014), “Development and exploitation of a controlled vocabulary in support of climate modelling”, *Geoscientific Model Development*, Vol. 7 No. 2, pp. 479–493.

Müller, H., Rangarajan, A., Teal, T.K. and Sternberg, P.W. (2008), “Textpresso for neuroscience: Searching the full text of thousands of neuroscience research papers”, *Neuroinformatics*, Vol. 6 No. 3, pp. 195–204.

Murphy, L.S., Reinsch, S., Najm, W.I., Dickerson, V.M., Seffinger, M.A., Adams, A. and Mishra, S.I. (2003), “Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators”, *BMC Complementary and Alternative Medicine*, Vol. 3 No. 1, p. 3.

Nazar, R. (2011), “A statistical approach to term extraction”, *International Journal of English Studies*, Vol. 11 No. 2, pp. 159–182.

NISO. (2010), *Guidelines for the construction, format, and management of monolingual controlled vocabularies. ANSI/NISO Z39.19-2005 (R2010),* Baltimore, Maryland, USA: National Information Standards Organization, available at: http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf (accessed 20 April 2015).

Nowick, E.A. and Mering, M. (2003), “Comparisons between Internet users’ free-text queries and controlled vocabularies: a case study in water quality”, *Technical Services Quarterly*, Vol. 21 No. 2, pp. 15–32.

Olson, H.A. and Wolfram, D. (2008), “Syntagmatic relationships and indexing consistency on a larger scale”, *Journal of Documentation*, Vol. 64 No. 4, pp. 602–615.

Pazienza, M.T., Pennacchiotti, M. and Zanzotto, F.M. (2005), “Terminology extraction: an analysis of linguistic and statistical approaches”, *Knowledge Mining*, Studies in Fuzziness and Soft Computing, Springer, Berlin, pp. 255–279.

Pedraza-Jiménez, R., Codina, L. and Rovira, C. (2008), “Semantic web adoption: online tools for web evaluation and metadata extraction”, in Ruan, D. and Montero, J. (Eds.), *Computational Intelligence in Decision and Control*, Proceedings of the 8th International FLINS Conference, World Scientific Publishing Company, Madrid, Spain, pp. 121–126.

Pedraza-Jiménez, R., Codina, L. and Rovira, C. (2009), “Metadatos en la Web semántica: lenguajes de marcado para la organización de sistemas de información”, in Codina, L., Marcos, M.-C. and Pedraza-Jimenez (Eds.), *Web Semántica y Sistemas de Información Documental*, Trea, Gijón, Spain, pp. 13–42.

Rowley, J. (1994), “The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research”, *Journal of Information Science,* Vol. 20 No. 2, pp. 108–118.

Sclano, F. and Velardi, P. (2007), “Termextractor: A web application to learn the shared terminology of emergent web communities”, in Gonçalves, R.J., Müller, J.P., Mertins, K. and Zelm, M. (Eds.), *Enterprise Interoperability II*, Springer London, London, UK, pp. 287–290.

Shah, P.K., Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2003), “Information extraction from full text scientific articles: Where are the keywords?”, *BMC Bioinformatics*, Vol. 4 No. 20, doi:10.1186/1471-2105-4-20.

Spasić, I., Schober, D., Sansone, S.-A., Rebholz-Schuhmann, D., Kell, D.B. and Paton, N.W. (2008), “Facilitating the development of controlled vocabularies for metabolomics technologies with text mining”, *BMC Bioinformatics*, Vol. 9 No. Suppl 5, doi:10.1186/1471-2105-9-S5-S5.

Tejeda-Lorente, Á., Porcel, C., Peis, E., Sanz, R. and Herrera-Viedma, E. (2014), “A quality based recommender system to disseminate information in a university digital library”, *Information Sciences*, Vol. 261, pp. 52–69.

Vállez, M. and Pedraza-Jiménez, R. (2007), “Natural Language Processing in Textual Information Retrieval and Related Topics”, *Hipertext.net*, Vol. 5, available at: http://www.upf.edu/hipertextnet/en/numero-5/pln.html.

Vállez, M., Pedraza-Jiménez, R., Blanco, S., Codina, L. and Rovira, C. (2015), “A semi-automatic indexing system based on embedded information in HTML documents”, *Library Hi Tech*, Vol. 33 No. 2, doi:10.1108/LHT-12-2014-0114.

White, H. (2013), “Examining Scientific Vocabulary: Mapping Controlled Vocabularies with Free Text Keywords”, *Cataloging & Classification Quarterly*, Vol. 51 No. 6, pp. 655–674.

White, R.W. and Horvitz, E. (2014), “From health search to healthcare: explorations of intention and utilization via query logs and user surveys”, *Journal of the American Medical Informatics Association*, Vol. 21 No. 1, pp. 49–55.

Zhang, C., Niu, Z., Jiang, P. and Fu, H. (2012), “Domain-specific term extraction from free texts”, *Proceedings of the 9th International FSKD Conference*, Fuzzy Systems and Knowledge Discovery, IEEE, Sichuan, China, pp. 1290–1293.

[i] Taxonomy Warehouse (accessed 1 April 2015).