Vàllez, Mari; Pedraza-Jiménez, Rafael; Blanco, Saúl; Codina, Lluís; Rovira, Cristòfol (2015). A semi-automatic indexing system based on embedded information in HTML documents. *Library Hi Tech*, Vol. 33, No. 2, pp. 195-210.

A semi-automatic indexing system based on embedded information in HTML documents

Purpose – This paper describes and evaluates the tool DigiDoc MetaEdit which allows the semi-automatic indexing of HTML documents. The tool works by identifying and suggesting keywords from a thesaurus according to the embedded information in HTML documents. This enables the parameterization of keyword assignment based on how frequently the terms appear in the document, the relevance of their position, and the combination of both.

Design/methodology/approach – In order to evaluate the efficiency of the indexing tool, the descriptors/keywords suggested by the indexing tool are compared to the keywords which have been indexed manually by human experts. To make this comparison a corpus of HTML documents are randomly selected from a journal devoted to Library and Information Science.

Findings – The results of the evaluation show that there: (1) is close to a 50% match or overlap between the two indexing systems, however if you take into consideration the related terms and the narrow terms the matches can reach 73%; and (2) the first terms identified by the tool are the most relevant.

Originality/value – The tool presented identifies the most important keywords in an HTML document based on the embedded information in HTML documents. Nowadays, representing the contents of documents with keywords is an essential practice in areas such as information retrieval and e-commerce.

Keywords – Semi-automatic indexing; Keywords assignment; Metadata editor; Controlled language; Semantic web technologies.

Article Type – Research paper

Introduction

Representing the content of a document with keywords is a long-standing practice. Information retrieval systems have traditionally resorted to this method to facilitate the access to information, since it is a compact and efficient way of representing a document. This process is known as indexing. Thus, we will refer to indexing as the task of assigning a limited number of keywords to a document, keywords which indicate concepts that are sufficiently representative of the document.

Despite the advantages of using keywords, only a minority of documents have assigned keywords because it is expensive and time-consuming. Therefore, systems are needed to facilitate the generation of keywords. Our proposal tries to identify the most important terms of HTML documents with high frequency and semantic relevance from a controlled language.

In this paper we describe the tool DigiDoc MetaEdit that allows the semi-automatic indexing of HTML documents. The tool assigns keywords from a thesaurus with the objective of representing the semantic contents of the document efficiently. To do this, it follows some of the relevance criteria used by search engines. Furthermore, it can be customizable according to how frequently the terms appear in the document, the relevance of their position and the combination of both. In order to evaluate the efficiency of the indexing system, we compare the descriptors suggested by the tool to those used in a portal of electronic journals by human experts.

The article is organised into the following sections: first, a brief overview of the literature related to indexing and automatic indexing; second, the research objectives; third, the presentation of the tool DigiDoc MetaEdit to assign keywords to HTML documents; fourth, the methodology section with information about the experimental datasets, the configuration of the tool, and the evaluation process; fifth, the results obtained in the evaluation and the analysis of them; and finally, the conclusions and future lines of research.

Literature review

Indexing theory attempts to identify the most effective indexing process, for indexing to be executed as a science rather than as an art (Borko, 1977; Hjørland, 2011). In the academic literature, indexing process involves two main steps: one, identifying the subjects of the document, and two, representing them in a controlled language (Mai, 2001). This process is also known as subject indexing, in which the representation of the documents is conditioned by the controlled language structure. Some authors, Lancaster (2003) and Mai (1997) among them, analyze this procedure and the problems of identifying subjects. Others, such as Willis & Losee (2013) or Anderson (2001a, 2001b), review the most important aspects of manual and automatic subject indexing and also the differences between both systems.

Manual indexing involves an intellectual process using a controlled language, which results in this system being difficult, slow and expensive. It also entails a high number of inconsistencies, both external, when the task is conducted by multiple indexers, and internal, when a single indexer performs the work at different times (Olson and Wolfram, 2008; White et al., 2013; Zunde and Dexter, 1969) .

Moreover, automatic indexing can be approached from two main perspectives. The first one is keyword extraction, based on the keyword's appearance in the text and in the whole of a collection (Frank et al., 1999; Zhang, 2008; Beliga, 2014). The second technique is keyword assignment, based on the matching of terms between the text and a thesaurus (or some other controlled vocabulary) (Moens, 2002; Yang et al., 2014) .

The different approaches for the first technique —keyword extraction— can be grouped into three categories: systems based on machine learning; systems based on rules for patterns and systems supported by statistical criteria (Ercan and Cicekli, 2007; Giarlo, 2005; Kaur and Gupta, 2010). These different approaches can also be combined.

Firstly, machine learning systems rely heavily on probabilistic calculations from training collections (Abulaish and Anwar, 2012). They adapt well to different environments, but their drawbacks should also be mentioned: they require many examples, it is difficult to select appropriate sources for training, they consume considerable time before quality results appear, and their performance degrades when the heterogeneity of documents increases.

Secondly, systems based on rules for patterns depend on the experience of the person who develops them, therefore requiring specialists to define the extraction rules for each domain. This definition process might also include linguistic criteria in order to select the keywords (Hulth, 2003; Hu and Wu, 2006) and, as such, it involves morphological, syntactic and semantic analyses to perform the disambiguation process. These systems are complex and require devoting time to the configuration; also, it is difficult to introduce changes to them.

Finally, systems based on statistical criteria (Ganapathi Raju et al., 2011; Matsuo and Ishizuka, 2004) do not require a training phase, although in many cases they require big corpora in order to perform the calculations. Some statistical methods used are: word frequency, TF-IDF, mutual information, co-occurence, etc.

The approach to the second technique —keywords assigned from a thesaurus— has also been tackled from various perspectives (Gazendam et al., 2010). The following are examples of this kind of approach: Kamps' proposal (2004) resorts to a thesaurus and establishes a strategy for reordering keywords obtained through semantic relations. Likewise, Medelyan & Witten (2006a) resort to the semantic relations from a thesaurus to optimize the results obtained with machine learning techniques. Lastly, Evans et al. (1991) suggest combining natural language processing techniques with the information provided by a thesaurus. This approach is very common in areas with high scientific knowledge production and indexing is important, such as in biosciences, medicine or aeronautics (Glier et al., 2013; Névéol et al., 2009).

Thus, it can be observed that both the extraction and assignment of keywords are commonly present in hybrid systems combining the two methods (Hulth, 2004).

In any case, both models present disadvantages. Keyword extraction might present wrong results, particularly regarding words formed by several terms (that is to say, when the systems used have to identify n-grams). Regarding keyword assignment, the main problem is the difficulty of having controlled languages that cover the thematic diversity of the documents, as well as the constant need for updates, and both aspects are essential in contexts such as repositories and digital libraries (Tejeda-Lorente et al., 2014).

Research objectives

Automatic indexing systems have been available for several decades (Sharp and Sen, 2013; Spärck Jones, 1974). These allow you to process a lot of information quickly and cheaply, and also ensure the inter-indexer consistency. However automatic systems also present problems because of the complexity of natural language processing (Sinkkilä et al., 2011). Consequently the semi-automatic indexing approach is a good solution, because in addition to obviating the problems of the automatic indexing system it facilitates the the task of indexers by providing suitable term suggestions (Vasuki and Cohen, 2010).

In this context, the main goal of this research is evaluating the results obtained with DigiDoc MetaEdit, a web-accessible tool, that allows semi-automatic indexing based on the embedded information in HTML documents. The tool identifies the highlighted terms of HTML documents and assigns descriptors from a specialized thesaurus.

The specific objectives to reach this goal are: first, analyzing the results obtained with the different configurations of the tool to carry out the indexing; second, comparing the indexing proposed by the tool with the indexing carried out by professional indexers; third, identifying the descriptors incorrectly assigned by the tool; and finally, demonstrating the viability of the proposal with the results.

DigiDoc MetaEdit

DigiDoc MetaEdit is a metadata editor (Pedraza-Jiménez et al., 2008, Vállez et al., 2010) that allows the description of the content of HTML pages. The tool was created with the mission to help metadata assignment, focused in particular on identifying the keywords for the purpose of indexing. Describing contents with metadata aids development and optimization of internal search systems, such as search engines for digital repositories, intranets or corporate websites, where improved search tools are essential. It is worth noting that Semantic Web Case Studies from W3C show that improved search is, in terms of frequency, the second most popular application of semantic web technologies ("Improved search - Semantic Web Case Studies and Use Cases" n.d.). First place was taken by data integration.

The DigiDoc MetaEdit has an interface that lets users set the selection criteria for keywords assignment. Keywords are then proposed using a specialized controlled language, a thesaurus, to recommend synonyms, narrower terms, broader terms, and related terms to the words appearing in the document analysed. Once the keywords have been extracted, the tool produces an RDF file with the metadata and a report with the keywords scored.

DigiDoc MetaEdit has been developed as a free software application with a GLP licence. It is a dynamic application designed in Perl using MySQL for data storage. Its structure is modular, which makes it easier to add new features. The three main modules are:

- Customization module: its aim is to enable the customization of the tool in terms of the controlled language, the metaformats and the weight assignment to identify keywords.
- Extraction module: its aim is to extract the keywords and metaelements from the HTML documents.
- Output module: its aim is to present the extracted meta elements and to generate fragments of code with the metadata adapted to several standards, such as RDF or Dublin Core.

Figure 1 shows a summary of how DigiDoc MetaEdit is structured:

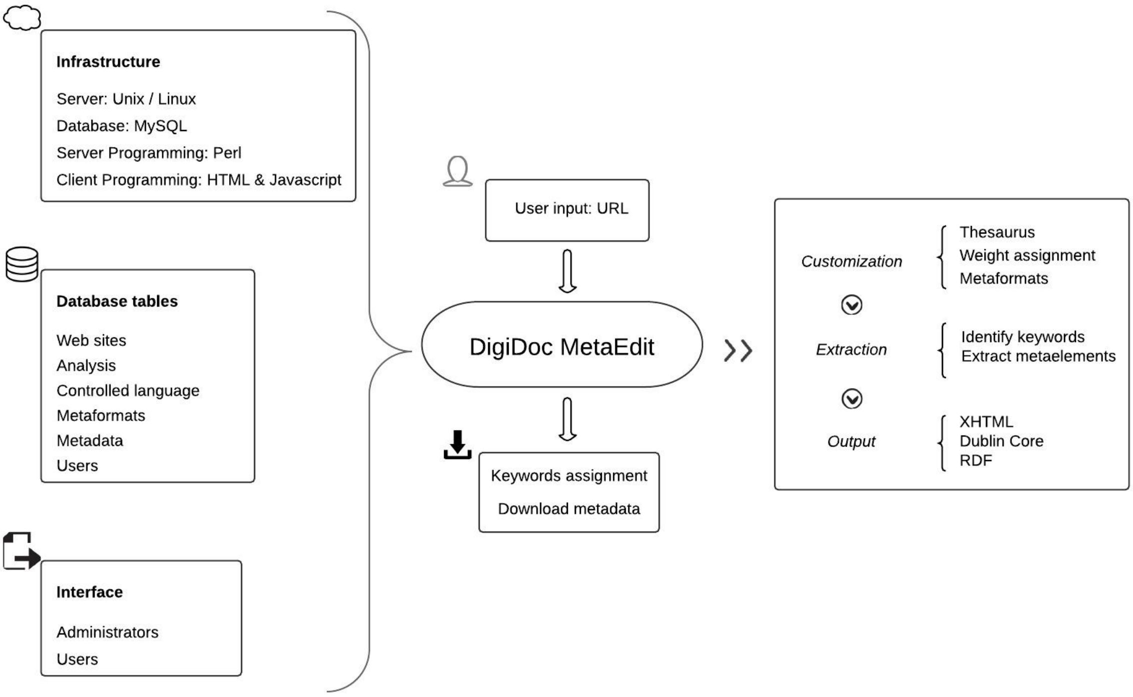


Figure 1. Components of the DigiDoc MetaEdit tool.

The tool contains the following components:

1. Data input interface: allows the user to indicate the URL of the HTML document or set of documents to be analysed.
2. Thesaurus: is the controlled language used to extract the keywords of the document.
3. Keyword weighting software: the tool presents mechanisms allowing the user the configuration of criteria and values for automatic keyword extraction, even though it already has a default configuration. The criteria which can be configured are based on some aspects considered in search engine optimization algorithms, such as:
   - term frequency: the number of times the term appears in the text,
   - location of the term (semantic markup): title, headers (h1, h2), URLs, anchors, emphasis, strong.
4. Text processing software: allows for the analysis of the textual contents of an HMTL document, and the extraction of its most significant keywords from the defined relevance criteria and the thesaurus.
5. Output interface: suggests formalized keywords as metadata of the document, in formats such as Dublin Core microformat, RDF and XHTML.

During the last years researchers and developpers from the Semantic Web and Linked Open Data community have made semantic tools for automatically editing and annotating web content. By example the applications developed by the Dbpedia community (http://wiki.dbpedia.org/Applications). Thereby, different platforms offer semantic annotation (Bukhari et al., 2013; Golbeck et al., 2002; Hu and Du, 2013), although in most cases they require complex infrastructure because they are part of a framework. In addition, there are a range of tools that offer similar solutions related to keyword research (Vállez, 2011), but most of them are based exclusively on statistical techniques to provide the proposed keywords, without taking into account the content structure and specific domain. Likewise, DigiDoc MetaEdit offers a range of different features from a single platform this is where our work breaks new ground.

Methodology

In order to evaluate the efficiency of the indexing proposal with the DigiDoc MetaEdit tool, this paper presents a comparison of the descriptors suggested by the system and those used by indexers in *Temaria* (http://temaria.net/), a portal of electronic journals on Library and Information Science. Regarding the present evaluation, we considered that the descriptors assigned by indexers were better to describe these documents. However sometimes the selection of a descriptor can be subjective (Coffman and Weaver, 2014; El-Haj et al., 2013).

Experimental datasets

The 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 Documentación* (http://bid.ub.edu/), a journal specialized in Library and Information Science indexed on the portal *Temaria*. This portal indexes articles from Spanish journals devoted to Library and Information Science and can be accessed online. It currently includes articles published in 14 Spanish journals.

The articles were indexed with descriptors from the *Tesauro de Biblioteconomía y Documentación* (Thesaurus on Library and Information Science), a controlled language developed by the Spanish *Instituto de Estudios Documentales sobre Ciencia y Tecnología* (IEDCYT) (Monchon and Sorli, 2002). Table 1 shows a summary of the elements and relations established in the thesaurus.

Table I.

Elements of the thesaurus

Number of concepts	1,097
Number of non preferred terms	569
Number of broader terms	1,088
Number of narrower terms	1,072
Number of related terms	2,354

The number of descriptors assigned to each article ranges between two and eight, with 4.14 descriptors on average and a standard deviation of 1.37. Taking this information as a starting point is contrasted to the descriptors assigned to each document with the 5, 10 and 15 keywords obtained with the DigiDoc MetaEdit. This checks that the first keywords ascribed are the most appropriate.

Configuration of the tool

The tool can be configured to decide which aspects to assess when HTML documents are processed to assign keywords. The *Keyword weighting software* lets you define settings to test different results. The configuration of the system has been conducted in different stages. In the beginning eleven parameterizations were defined that were subsequently grouped and delimited under three parameterizations:

- Frequency: based on the number of times a term appeared in the document.
- Semantics: based on the position of the term in the document according to embedded HTML information. This parameterization considers the location of the keywords in the HTML document, such as in the title of the page, in the metadata, in the headers, in the typographic emphasis (bold type or italic), in the alternative text of the images or links, and so on. This measure was determined by the semantic relevance of the word, hence the name.
- Mixed (Frequency and Semantics): the importance of the keywords was pondered by combining aspects of the two previously mentioned parameterizations, in so doing it attempts to find a balance between the frequency and the position occupied by the word in the document.

DigiDoc MetaEdit included the same thesaurus used by human indexers to compare the results obtained with the indexing tool.

Evaluation process

In order to test the robustness of the tool an evaluation system has been designed (shown in Figure 2). To do so, we first select a corpus of documents manually indexed using a thesaurus. Second, the metaeditor processes this corpus, and automatically suggests descriptors for each document. The output has three different indexing proposals for each document. Then, the descriptors assigned by indexers are compared to the descriptors suggested by each one of the parameterizations. An exact overloap is required. Next, the best parameterization is that which identifies a higher number of overlapping descriptors that proposed by human indexers. Finally, two indexers analyses whether the keywords assigned to each document are correct.

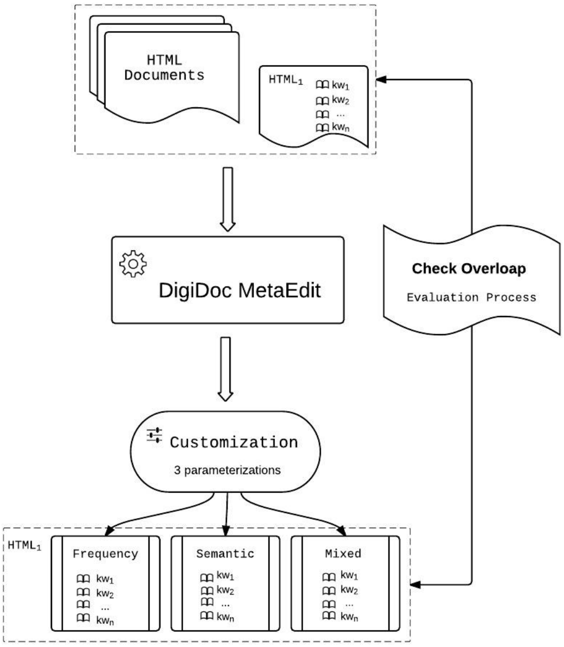


Figure 2. Evaluation process of the tool.

A number of routines have been written in Python to process and run comparisons of the settings. These routines process the files, identify the matching words and present the results. The coding in Python used object-oriented programming to aid replication of the experiment with other data sets and settings. The open source code can be found on the Git repository at: https://github.com/beauseant/MITAD.

The measures for evaluating the results offered by the tool have been those habitually used to assess automatic indexing (Medelyan and Witten, 2005; Verberne et al., 2014) as defined in Eq. (1-3):

$$Precision = \frac{\#\ \text{correct assigned keywords}}{\#\ \text{assigned keywords}} \tag{1}$$

$$Recall = \frac{\#\ \text{correct assigned keywords}}{\#\ \text{manually assigned keywords}} \tag{2}$$

$$F - measure = 2\,\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The *"# correct assigned keywords"* value corresponded to the number of correctly assigned keywords in an automatic way. They were considered to be correctly assigned when the automatically identified descriptors match the descriptors suggested by human experts. The *"# assigned keywords"* value was the total amount of keywords describing a document; that was to say, the amount of descriptors assigned by a human expert to the document. Lastly, *"# manually assigned keywords"* was the total amount of automatically suggested keywords.

Thus and as previously mentioned, it was possible to estimate both the percentage of automatically assigned keywords being relevant (precision) and the percentage of matches to the manually assigned keywords (recall). Since both measures act inversely (Cleverdon, 1972), the F-measure was used as a balanced combination of the two previous ones (Van Rijsbergen, 1977), allowing for the harmonious average of precision and recall.

Results and analysis

First, we compared the two indexing systems, manual and automatic. Then we evaluated the results in the automatic indexing system for the three parameterizations. The next step was to study how the assigned descriptors in both indexing processes were distributed and characterized. And lastly, we analysed the quality of the descriptors automatically assigned by DigiDoc MetaEdit.

Table 2 shows the keywords in common to both systems, and at the same time presents different semantic relations identified between words. The coincidental terms are identified in bold type in both indexing processes. The terms after the signs "<<" correspond to broader terms among those suggested in the manual indexing, and those headed by "**" to the related terms. This information is extracted from the semantic relations appearing in the thesaurus.

Table II.

Comparative process of keywords assigned to a document (translated to English for the reader's convenience)

Title: ***Cooperative repositories of the Digital Library of Catalonia***

Cut off: 5/10/15 terms

Manual indexing	**Automatic indexing**		
	Frequency	**Semantics**	**Mixed**
Digital libraries	Dissertations	**Open archives**	Dissertations
Libraries consortiums	<< Software	<< Software	<< Software
Copyrights	Academic journals	Dissertations	Academic journals
Open archives	Metadata	Authority control	Metadata
Free software	<< Libraries	** Computer programmes	** Computer programmes
	University centres	Academic libraries	<< Libraries
	** Computer programmes	**Copyrights**	University centres
	Visibility	** Electronic resources	**Open archives**
	Free software	e-Government	**Free software**
	Digitization	Library management	Visibility
	Authorship	Academic journals	Digitization
	Interoperability	Special collections	Academic libraries
	Academic libraries	Information Society	Academic community
	Academic community	Authors index	Electronic journals
	Electronic journals	Journal articles	Authorship

Note: Bold type = the coincidental terms identified in both indexing processes.
"<<" = broader terms of Manual indexing assigned.
"**" = related terms of Manual indexing assigned.

This example shows that the two indexing systems overlap, particularly when a cut off of ten terms is applied, as also shown in the Table 4. Thereby Mixed parameterization with ten terms cut off offers the best results and is used as a basis for following comparisons.

It is also important to emphasize that there is a high number of documents sharing descriptors in both indexing systems. Thus, 92 % of the documents match in both systems, meaning that 92 documents on average share almost two descriptors in both indexing systems. Therefore both systems offer similar results.

The second stage of the comparison process took into consideration the semantic relationship between the first ten keywords suggested by the metaeditor under the Mixed parameterization, with those assigned by the human indexers. In this case, the broader terms and related terms of the descriptors assigned by indexers were taken into account to calculate the exact overlap (Medelyan and Witten, 2006b). Table 3 shows how, by considering these semantic relations from the thesaurus, the automatic indexing increased recall. Thus, it shows that there is a semantic relation between the keywords used in both kinds of indexing. Although there is not an exact match, keywords assigned with the DigiDoc MetaEdit tool maintain a strong link with those used by indexers.

Table III.

Recall considering the relations of the thesaurus

	Mixed	Mixed + NT	Mixed + RT	Mixed + (NT+RT)
Recall	0.49	0.58	0.64	0.73

Note: NT = narrower terms
RT = related terms

Of the two semantic relations studied (narrower terms and related terms), recall showed the greatest increase –of 15 %– with the one including related terms. This kind of relationship allowed for the identification of keywords that did not present a “parent-child” relationship, a fact which contributed to obtaining terms with a higher semantic variety. This point is interesting because search engines take into account the variety of terms instead of just the frequency of a term.

Nevertheless, if the narrower terms were included, the increase amounted to only 9 %. Also significant in this case was the fact that most of the terms included mainly referred to the concept *Libraries*. This fact shows that narrower terms do not contribute semantic variety to the indexing system or help identify further meaningful keywords.

Table 4 below shows the averages of the evaluation measures used (recall, precision and F-measure) for the three parameterizations.

Table IV.

Average number of matches in each parameterization with manual indexing

	Cut off	Frequency	Semantics	Mixed
Recall	5	0.31	0.32	0.33
	10	0.47	0.48	0.49
	15	0.54	0.55	0.56
Precision	5	0.24	0.25	0.26
	10	0.19	0.19	0.20
	15	0.14	0.15	0.15
F-measure	5	0.26	0.25	0.27
	10	0.26	0.26	0.27
	15	0.22	0.23	0.23

In order to check the evolution of recall with regard to the number of keywords suggested by the metaeditor, additional intervals for analysis were created. Figure 3 shows how recall evolved under Mixed parameterization, with an increase in the number of keywords suggested by the automatic system: as more descriptors were considered, the increase in recall was gradually diminished. Thus, the ordering of the descriptors suggested by the metaeditor was working: the first terms identified by the tool were the most relevant, since they entailed a higher increase in recall.

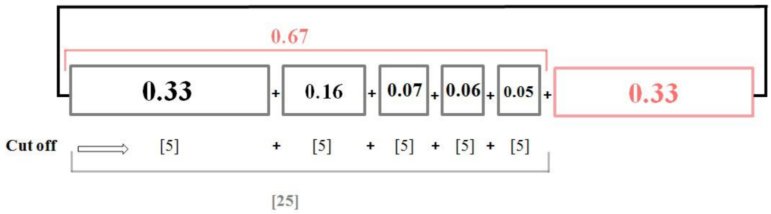
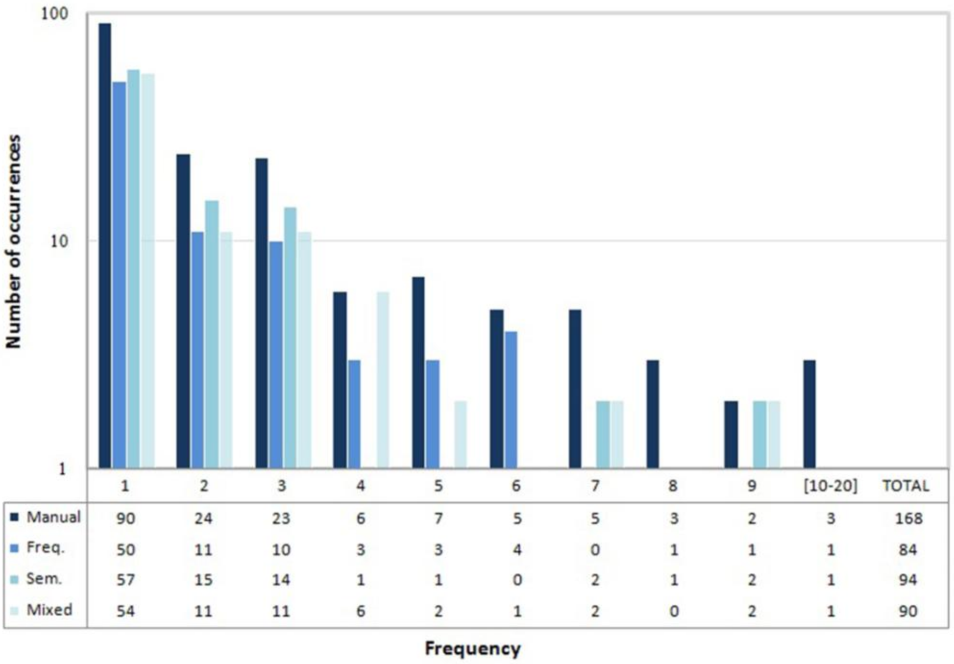


Figure 3. Recall increases for Mixed parameterization.

Regarding the distribution of keywords used in the corpus of documents, 168 of the 414 keywords used by the indexers were unique. Therefore the corpus of documents maintain a strong relationship among themselves. However, the Mixed parameterization with ten terms cut off provides 90 keywords unique from the 195 descriptors overlapped with those assigned by the human experts. On the basis of the calculations made with this data, automatic indexing offers more semantic variety.

Furthermore, similarly to what happens with natural language, and as stated by Zipf's law, in manual and automatic indexing many words presented a low frequency of use (long tail), whereas a few concentrated a high frequency. This information was significant since a good description must be characterized by its level of specificity. The high number of descriptors in the thesaurus that were seldom assigned is proof of that.

To study this aspect, it is interesting to see the distribution of keywords assigned according to their frequency (Figure 4). The logarithmic scale of the following bar chart presents the frequency of the unique terms assigned.



	1	2	3	4	5	6	7	8	9	[10-20]	TOTAL
■ Manual	90	24	23	6	7	5	5	3	2	3	168
■ Freq.	50	11	10	3	3	4	0	1	1	1	84
■ Sem.	57	15	14	1	1	0	2	1	2	1	94
■ Mixed	54	11	11	6	2	1	2	0	2	1	90

Figure 4. Frequency distribution of the keywords.

The first row of the table accompanying the chart shows the frequency distribution of the 168 unique descriptors assigned by the human indexers to describe the corpus of documents. Thus, it can be seen that 90 descriptors were used only once, which means that 54 % of the terms were not repeated. The subsequent rows show the frequency distribution of the assigned descriptors for each of the three parameterizations when these overlapped the manually assigned descriptors. Regarding the Mixed automatic assignment, 54 out of the 90 unique terms were used only once, which amounted to 60% of the terms. Thus, it was found that the metaeditor has practically the same or even slightly superior (6%) discriminating ability as the human expert when assigning keywords. This aspect is essential, since the specificity of the keywords contributes to improving the process of information retrieval because it provides more precise results. Additionally, the graph reveals that only three keywords were commonly used in the manual assignment, that is, they were employed more than fifteen times. Their use was nonetheless reduced to one case in automatic indexing.

Lastly, the quality of the terms provided by the metaeditor for Mixed parameterization was studied when ten descriptors were suggested through the analysis of human experts. Two indexers analysed whether the keywords assigned to each document with DigiDoc MetaEdit were relevant or not, as well as detecting the cases wherein it was a mistake to have assigned a specific keyword. Table 5 shows the percentages obtained in each case for the corpus of documents.

Table V.

Adaptation of the terms in Mixed parameterization with ten terms cut off

Mixed – 10 terms cut off			
Exact Terms (Precision)	**Relevant Terms**	**Not relevant Terms**	**Wrong Terms**
20%	69%	4%	7%

The exact terms match up in both indexing systems; the relevant ones are those that do not match up but clearly represent the contents of the document; the non relevant ones are those that, although appearing in the document, are not representative enough of its contents. On the other hand, those terms considered ‘wrong’ are those that cannot be used to describe the document because they are misleading. An example of this last case is the concept *Bibliography*, which appeared frequently because every article had a section with this epigraph, therefore being suggested by the metaeditor without being representative of the documents. From a practical point of view, the keywords presenting problems are not significant.

Conclusions

After conducting the different experiments, it is possible to conclude that the keyword assignment carried out by DigiDoc MetaEdit offers positive results and is, therefore, an efficient system. The indexing proposed by DigiDoc MetaEdit approaches a 50% match rate with manual indexing when taking into account the Mixed parameterization with ten terms cut off. Furthermore, it reaches 73% when related terms and narrower terms are considered. Besides, according to expert assessment, 89% of the words have been correctly assigned, with only 7% of misallocations and 4% of non-relevant assignments.

Nowadays in a situation of information overload, the identification of the most significant keywords in a document can serve various purposes. To a great extent, they can be focused on synthesizing the content of a document to facilitate access to it. The exponential increase of information brings about the need to automate this process to the utmost, and DigiDoc MetaEdit makes this task easier. Thus, it can be considered a useful tool for assisting and evaluating human indexing.

Besides, search optimization techniques have proved to be very effective in identifying keywords describing a collection of documents. The HTML documents described with these terms benefit from the fact that the terms are already optimized for search engine optimization.

Some short-term research lines are being considered to provide continuity to the research conducted. After studying the existing matches between manual and automatic indexing, it would be advisable to see if there is any relationship between the keywords chosen by the users of a search engine to access a document and those assigned in both manual and automatic indexing. Thus, the quality of the indexing process could be studied by testing whether the indexing terms present any links to the information needs of the users.

Acknowledgements

This article is part of the projects: “Audiencias activas y periodismo” (Active audiences and journalism). CSO2012-39518-C04-02 and "Comunicación online de los destinos turísticos" (Online communication of tourist destinations) CSO 2011-22691. Plan Nacional de I+D+i, Ministerio de Economía y Competitividad (Spain).

References

Abulaish, M. and Anwar, T. (2012), “A supervised learning approach for automatic keyphrase extraction”, *International Journal of Innovative Computing, Information and Control*, Vol. 8 No. 11, pp. 7579–7601.

Anderson, J.D. and Pérez-Carballo, J. (2001a), “The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort”, *Information Processing & Management*, Vol. 37 No. 2, pp. 255–277.

Anderson, J.D. and Pérez-Carballo, J. (2001b), “The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing”, *Information Processing & Management*, Vol. 37 No. 2, pp. 231–254.

Borko, H. (1977), “Toward a theory of indexing”, *Information Processing & Management*, Vol. 13 No. 6, pp. 355–365.

Bukhari, A.C., Klein, A. and Baker, C.J.O. (2013), “Towards interoperable bioNLP semantic web services using the SADI framework”, in Baker, C.J.O., Butler, G. and Jurisica, I. (Eds.), *Data integration in the life sciences*, Lecture Notes in Computer Science: Vol. 7970, Springer Berlin Heidelberg, pp. 69–80.

Beliga, S. (2014), Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, Rijeka.

Cleverdon, C.W. (1972), “On the inverse relationship of recall and precision”, *Journal of Documentation*, Vol. 28 No. 3, pp. 195–201.

Coffman, J. and Weaver, A.C. (2014), “An empirical performance evaluation of relational keyword search techniques”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26 No. 1, pp. 30–42.

El-Haj, M., Balkan, L., Barbalet, S., Bell, L. and Shepherdson, J. (2013), “An experiment in automatic indexing using the HASSET thesaurus”, in *Proceedings of the 5th Computer Science and Electronic Engineering Conference*, IEEE Xplore, Colchester, United Kingdom, pp. 13–18.

Ercan, G. and Cicekli, I. (2007), “Using lexical chains for keyword extraction”, *Information Processing & Management*, Vol. 43 No. 6, pp. 1705–1714.

Evans, D.A., Hersh, W.R., Monarch, I.A., Lefferts, R.G. and Handerson, S.K. (1991), “Automatic indexing of abstracts via natural-language processing using a simple thesaurus”, *Medical Decision Making*, Vol. 11 No. 4 Suppl, pp. 108–115.

Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. (1999), “Domain-specific keyphrase extraction”, in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 668–673.

Ganapathi Raju, N.V., Sukavasi, B., Rama Krishna Chava, S. and Rani Vadisala, V. (2011), “An application of statistical indexing for searching and ranking of documents - A case study on Telugu script”, *International Journal of Computer Applications*, Vol. 28 No. 3, pp. 22–27.

Gazendam, L., Wartena, C. and Brussee, R. (2010), “Thesaurus based term ranking for keyword extraction”, in *Workshop on Database and Expert Systems Applications, 21st DEXA Conference*, IEEE Xplore, Bilbao, Spain, pp. 49 –53.

Giarlo, M.J. (2005), *A comparative analysis of keyword extraction techniques*, Rutgers, The State University of New Jersey, available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.119.647 (accessed 16 August 2014).

Glier, M.W., McAdams, D.A. and Linsey, J.S. (2013), “An experimental investigation of analogy formation using the Engineering-to-Biology thesaurus”, in *Proceedings of the 25th International Conference on Design Theory and Methodology*, American Society of Mechanical Engineers, Portland, United States, Vol. 5, doi:10.1115/DETC2013-13160.

Golbeck, J., Grove, M., Parsia, B., Kalyanpur, A. and Hendler, J. (2002), “New tools for the semantic web”, in Gómez-Pérez, A. and Benjamins, V.R. (Eds.), *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, Lecture Notes in Computer Science : Vol. 2473, Springer Berlin Heidelberg, pp. 392–400.

Hjørland, B. (2011), “The importance of theories of knowledge: Indexing and information retrieval as an example”, *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 1, pp. 72–77.

Hu, H. and Du, X. (2013), “TAG: A Tag-as-You-Go online annotation tool for web browsing and navigation”, in Wang, M. (Ed.), *Knowledge Science, Engineering and Management*, Lecture Notes in Computer Science : Vol. 8041, Springer Berlin Heidelberg, pp. 298–309.

Hulth, A. (2003), “Improved automatic keyword extraction given more linguistic knowledge”, in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 216–223.

Hulth, A. (2004), *Automatic keyword extraction: combining machine learning and natural language processing*, Stockholm University, Edsbruk, Sweden, available at: http://people.dsv.su.se/~hulth/thesis_hulth.pdf (accessed 16 August 2014).

Hu, X. and Wu, B. (2006), “Automatic keyword extraction using linguistic features”, in *Data Mining Workshops, 6th IEEE International Conference on Data Mining*, IEEE Computer Society, Hong Kong, China, pp. 19–23.

Kamps, J. (2004), “Improving retrieval effectiveness by reranking documents based on controlled vocabulary”, in McDonald, S. and Tait, J. (Eds.), *Advances in Information Retrieval: Proceedings of the 26th European Conference on IR Research*, Springer, Sunderland, UK, Vol. 2997, pp. 283–295.

Kaur, J. and Gupta, V. (2010), “Effective approaches for extraction of keywords”, *International Journal of Computer Science*, Vol. 7 No. 6, pp. 144–148.

Lancaster, F.W. (2003), *Indexing and abstracting in theory and practice*, Facet Publishing, London, England, 3rd ed.

Mai, J.E. (1997), “The concept of subject: on problems in indexing”, in *Proceedings of the 6th International Study Conference on Classification Research*, International Federation for Information Documentation, The Hague, Netherlands, pp. 60–66.

Mai, J.E. (2001), “Semiotics and indexing: An analysis of the subject indexing process”, *Journal of Documentation*, Vol. 57 No. 5, p. 591.

Matsuo, Y. and Ishizuka, M. (2004), “Keyword extraction from a single document using word co-occurrence statistical information”, *International Journal on Artificial Intelligence Tools*, Vol. 13 No. 1, pp. 157–170.

Medelyan, O. and Witten, I.H. (2005), “Thesaurus-based index term extraction for agricultural documents”, in *Proceedings of the 6th Agricultural Ontology Service (AOS)*, Food and Agriculture Organization of the United Nations, Vila Real, Portugal, pp. 1122–1129.

Medelyan, O. and Witten, I.H. (2006a), “Measuring inter-indexer consistency using a thesaurus”, in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, Chapel Hill, NC, USA, pp. 274 –275.

Medelyan, O. and Witten, I.H. (2006b), “Thesaurus based automatic keyphrase indexing”, in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, New York, NY, USA, pp. 296–297.

Moens, M.-F. (2002), “Automatic indexing: The assignment of controlled language index terms”, *Automatic indexing and abstracting of document texts*, The Information Retrieval Series: Vol. 6, Springer US, pp. 103–132.

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

Névéol, A., Shooshan, S.E., Humphrey, S.M., Mork, J.G. and Aronson, A.R. (2009), “A recent advance in the automatic indexing of the biomedical literature”, *Journal of biomedical informatics*, Vol. 42 No. 5, pp. 814–823.

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.

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.

Van Rijsbergen, C.J. (1977), “A theoretical basis for the use of co-occurrence data in information retrieval”, *Journal of Documentation*, Vol. 33 No. 2, pp. 106–119.

Sharp, J. and Sen, B.A. (2013), “The viability of automatic indexing of biomedical literature”, *International Journal of Health Information Management Research*, Vol. 1 No. 1, pp. 55–66.

Sinkkilä, R., Suominen, O. and Hyvönen, E. (2011), “Automatic semantic subject indexing of web documents in highly inflected languages”, in Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P.D. and Pan, J. (Eds.), *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science: Vol. 6643, Springer Berlin Heidelberg, pp. 215–229.

Spärck Jones, K. (1974), “Automatic indexing”, *Journal of Documentation*, Vol. 30 No. 4, pp. 393–432.

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. (2011), “Keyword research: métodos y herramientas para identificar palabras clave”, *BiD: textos universitaris de biblioteconomia i documentació*, Vol. 27.

Vállez, M., Rovira, C., Codina, L. and Pedraza-Jiménez, R. (2010), “Procedures for extracting keywords from web pages, based on search engine optimization”, *Hipertext.net*, Vol. 8.

Vasuki, V. and Cohen, T. (2010), “Reflective random indexing for semi-automatic indexing of the biomedical literature”, *Journal of Biomedical Informatics*, Vol. 43 No. 5, pp. 694–700.

Verberne, S., D’hondt, E., van den Bosch, A. and Marx, M. (2014), “Automatic thematic classification of election manifestos”, *Information Processing & Management*, Vol. 50 No. 4, pp. 554–567.

White, H., Willis, C. and Greenberg, J. (2013), “HIVEing: The effect of a semantic Web technology on inter-indexer consistency”, *Journal of Documentation*, Vol. 70 No. 3, pp. 1–1.

Willis, C. and Losee, R.M. (2013), “A random walk on an ontology: Using thesaurus structure for automatic subject indexing”, *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 7, pp. 1330–1344.

Yang, S., Zhang, B., Li, S., Yu, C. and Hao, Q. (2014), “Keyword extraction using multiple novel features”, *Journal of Computational Information Systems*, Vol. 10 No. 7, pp. 2795–2802.

Zhang, C. (2008), “Automatic keyword extraction from documents using conditional random fields”, *Journal of Computational Information Systems*, pp. 1169–1180.

Zunde, P. and Dexter, M.E. (1969), “Indexing consistency and quality”, *American Documentation*, Vol. 20 No. 3, pp. 259–267.

"Improved search - Semantic Web Case Studies and Use Cases" (n.d.), available at http://www.w3.org/2001/sw/sweo/public/UseCases/ (accessed 26 November 2014).