NAVDEX, a helpful tool for the classification of environmental legislation

Geert De Meyer, VITO (Flemish Institute for Technological Research), Mol, Belgium, geert.demeyer@vito.be
G. Van Eetvelde, University Ghent, Ghent, Belgium, Greet.VanEetvelde@UGent.be

Abstract

Since its launch in 1998 the thematic indexation of the Flemish Environmental Navigator is carried out manually by legal experts of the University of Ghent, Belgium. However, due to the exponential growth of legal documents a physical indexation process eventually was no longer tenable, nor desirable.

Hence, a semi-automatic indexing tool for environmental legislation, called NAVDEX, was developed. A specific algorithm was determined, based on the presence of similar terms in law objects. A parameter was defined, reflecting the strength of the relation between law objects in order to computerise the return on a user's query. In view of managing the relations between law objects, a visualisation tool was created in order to provide the legal experts with a detailed overview of all associated law objects.

The testing corpus was decided to be VLAREA, a Flemish order concerning waste prevention and management. The evaluation of the test results was carried out by experts in environmental legislation, who computed the relative recall of several search terms. With an average score of 0.63 NAVDEX is able to retrieve nearly two third of the associated law objects. Consequently the evaluators' conclusions were unanimous so as to define NAVDEX as a useful tool to determine and visualise associated LawObjects.

1 Introduction

Through the years, the Internet has become a repository of human knowledge and culture. The rising success of the Internet was primarily based on the growing potential of information exchange. But each coin has its reverse site. Due to the lack of structure, it soon proved to be hard for the Internet user to find relevant information. A solution materialised with the introduction of search engines; thanks to commercial Internet firms, algorithms to index web pages, documents, images, video, etc. were developed. Simply by typing a keyword, the search engine returns a list of available relevant information on the Internet.

Dealing with an escalating amount of information, however, was new to the daily Internet user; yet it was not to lawyers. Since decades, the latter try to manage massive amounts of legal documents. Therefore it is not surprising that –to a large extent– research has been performed in the field of legal information retrieval seeking to improve the efficiency of law examination.

In Flanders, by the end of 1997 the Flemish government commissioned VITO, the Flemish Institute for Technological Research (www.vito.be), to build an environmental legal expert system for Flanders. In 1998 the Flemish Environmental Navigator, short the Navigator, was developed in collaboration with the University of Ghent. The purpose was as ambitious as simple, i.e. "*to ensure that all Flemish environmental legislation was electronically available for the benefit of jurisdiction, business and to whom it may concern*."

Since launch time, legal experts built manual indexes to make legal documents searchable by users. Due to the exponential growth of legal documents in the Belgian Official Journal, however, especially those concerning environmental issues, a physical indexation process was no longer tenable, nor desirable. Therefore, upon many years of online service the Navigator needed profound updating. Apart form a new interface and an on-the-fly generation of legal texts with version management and linked information, the renewal was mainly oriented towards enhancing the intelligence of the legal expert system.

This article describes the development of a semi-automatic index generator for environmental legislation. In section 1, a quick overview of the essential definitions is presented next to the general description of information retrieval techniques. Section 2 details the set-up of NAVDEX, the semi-automatic legal index generator used to search the Flemish environmental legislation. Based upon the well known index algorithms term frequency and inverse document frequency, a specific algorithm for legal texts is unfolded. In section 3 the results of the semi-automatic legal index generator are communicated and evaluated. To perform this scrutiny, a representative legal text was found in VLAREA, a Flemish order concerning waste prevention and management. Likewise, domain experts were chosen to evaluate NAVDEX. Finally, section 4 presents the conclusions and suggestions with respect to potential future work in the field of automatic generation of legal indexes.

2 Information retrieval

In literature, a diversity of definitions for information retrieval (IR) [1,2,3,4] is found. In short, IR "is the process that selects documents relevant to a user's query out of a well defined repository. Those documents can be texts, images, video, sounds, etc."

Accordingly, IR is an uncomplicated process as observed from the scheme in figure 1. Three components are defined: the input, the processor and the output.

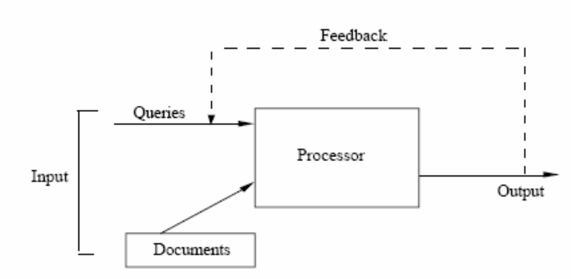


Fig. 1 : IR basic scheme [5]

The processor is the physical core of the IR system and is situated between the input and output module. It translates the natural language documents into computable parameters. To complete this task, it makes use of a predefined index algorithm, such as the term frequency (TF) and the inverse document frequency (IDF). Hence, the obtained index describes the information contents of the documents.

When a user launches a request to retrieve all the relevant documents, the output is generally produced as a list of significant documents with associated attributes. When the IR system is online, however, the user has the opportunity to refine his query during the search process. This is called the feedback process.

IR has proven to be a theme of constant interest for researchers, from the early 1950s until present. A bibliometric study of A. Pulgarin and I. Gil-Leiva [6] shows more than 800 research works between 1956 and 2000. As a pioneer in the area of automatic indexing, Luhn [7] indicated that frequency data could be used to extract words and sentences in an attempt to represent a document. He stated the idea that normally an author repeats certain words when writing on a subject. Words that are too common, thought, are defined as stop words, such as "and", "or", "because", etc.

Salton [8] contributed to the discussion by making a "*Blueprint for automatic indexing*". The most basic form of indexing exists of the following steps:

1. identify the individual text words occurring in the document;
2. use a stop list to delete common words;
3. compute the TF_{ij} of each term i for the content representation of document j;
4. compute the IDF_{ij} of each term i for the content representation of document j.

Additionally, suffix-stripping based on small but efficient algorithms or stemming based on morphological analysis can be incorporated in the indexing process. Stemming refers to the process of removing affixes (prefixes and suffixes) from words. In the information retrieval context, stemming is used to conflate word forms to avoid mismatches that may destabilise the recall. The most widely cited stemming algorithm was introduced by Porter (1980). The Porter stemmer applies a set of rules to iteratively remove suffixes from a word until none of the rules apply anymore [9].

3 Scope and methodology

The Navigator research served the understanding of the way to automatically index juridical information. Concisely, the project searched the question whether manual legal indexing could be replaced by automatic or semi-automatic indexing.

Prior to exploring the issue, it was believed that it would be difficult to gain a positive answer to the latter question. The main reason for this pessimism was the complexity of the environmental legislation. Caused by its semi-technical nature, it is recognised hard to produce congruent environmental legislation. On the other hand, the exceptional output of the concerned legislation over a two decade period of time turned automatic indexing into an interesting research tool.

Due to the massive amount of environmental legal texts, the scope of the research had to be narrowed. VLAREA, or the "*Order of the Flemish Government of December 5th 2003 for the Establishment of the Flemish regulations relating to Waste Prevention and Management*", was selected as a study object for automatic indexing. The text consists of 10 chapters and has, in brief, the purpose of protecting the health of persons and the environment against the harmful influence of waste. Likewise, it aims at indicting the wastage of raw materials and energy. VLAREA was merely chosen because of its electronic availability.

Indexing requires text partitioning in multiple Law Objects (LO). A LO is defined as an autonomous, self executing part of a legal text that consists of a defined number of terms $\{t_n\}$. In general, these are articles, but occasionally paragraphs are selected as well.
The VLAREA is composed of 689 Law Objects and 4718 different terms.

As mentioned above, some terms are of no use in representing informative content. Terms such as "and", "the", "of", "to", etc., known as stop words, and those lacking inquest relevance, such as numbers, physical units, auxiliaries and general terms, are referred to as bulk terms. They make up the majority of the terms without contributing to the core content of the considered reglementation and hence can be excluded without losing informative content. Exactly 3183 stop words were removed from the term list.

The importance of a term in representing informative content is known as its resolving power [10]. This weight factor indicates how well the issue can be resolved or whether a document is relevant or not to a user query. From a pure statistical point of view, the importance of one of the 1535 remaining terms can be calculated by using the TF and IDF table.

4 Results and discussion

Table 1 gives an overview of the 20 most frequently used terms in VLAREA. The highest frequency is recorded for the term "waste products" (435 hits). The acronym "OVAM", i.e. the Public Waste Agency of Flanders, is situated on the second place, followed by the terms "scrap" and "electronic".

Table 1: Overview of the 20 most frequent terms in VLAREA.

Term	Sum Of
waste products	435
OVAM	295
scrap	167
electronic	119
electric	119
demand	115
cars	107
producer	92
receipt	84
waste	84
acceptance obligation	72
importer	67
domestic	67
applicant	66
printed matter	57
accredited	56
producers	54
government	54
medical	53
collection	50

Relationships
Table 2 calculates the strength of a relation between a term and a specific LO. This strength is expressed as a quotient of the individual frequency of a term in a specific LO and the sum of the frequencies of all LO concerned.

Table 2: Inverse Document Frequency for the term "waste products".

Artid	Term	TF	Sum Of	IDF
156	waste products	7	435	0.016091954
378	waste products	4	435	0.009195402
689	waste products	3	435	0.006896552
687	waste products	1	435	0.002298851
445	waste products	6	435	0.013793103
146	waste products	1	435	0.002298851
148	waste products	1	435	0.002298851
150	waste products	1	435	0.002298851
675	waste products	1	435	0.002298851
151	waste products	1	435	0.002298851
152	waste products	1	435	0.002298851
154	waste products	3	435	0.006896552
155	waste products	1	435	0.002298851
439	waste products	1	435	0.002298851
669	waste products	1	435	0.002298851
438	waste products	4	435	0.009195402
668	waste products	1	435	0.002298851
667	waste products	2	435	0.004597701
666	waste products	4	435	0.009195402
281	waste products	1	435	0.002298851

For example Artid 156 corresponds to art 2.3.1 of VLAREA, containing 7 times the term "waste products":

Art.2.3.1. In accordance with article 3, § 5, of the Waste Decree, the following ***waste products*** *materials are additionally indicated as special* ***waste products****:*
...
2.the following ***waste products*** *that originate when maintaining, repairing or destroying motor vehicles, motor vessels, power planes and their appurtenances:*
...

*12. oil-bearing **waste products** such as oil filters, fuel filters, used absorbing material, **waste products** coming from oil/water separators, oil-bearing shock absorbers, packaging that has contained oil or has been soiled by oil and is no longer used;*

...

*3. paper and cardboard **waste products**;*

...

*16. PVC **waste products**;*

...

In total, the term "waste product" is detected 435 times in Vlarea. Hence, the IDF of this article is calculated as follows:

$$\mathrm{IDF}_{\text{art.2.3.1./waste products}} = 7/435 = 0.016$$

An IDF of less than 0.1 is considered low in terms of strength of the relationship between a term and a LO, i.e. an article.

In the example below, a strong relation is indicated between a term and a LO. Searching the Vlarea for the term "asphalt" yields only three hits, two of them appearing in art. 4.2.2.3.:

*§ 2. Tarry **asphalt** can only be used in listed work with a minimum scope of 1500 m3; used in a cold way in foundations consisting of **asphalt** granulate cement, provided that they satisfy the requirements of provisions of article 4.2.2.3, § 1, except for the maximum concentrations of polycyclic aromatic hydrocarbons and mineral oil.*

When calculating the $\mathrm{IDF}_{\text{art.4.2.2.3§2/asphalt}}$, a score of 0.667 or 2/3 is obtained for "asphalt", hence signifying a high specific term for art. 4.2.2.3, in contrast with "waste products".

The IDF is situated between 0 and 1. An IDF score of zero indicates a situation without any relation between a specific term and its LO. The opposite IDF-score points to a situation where a specific term exclusively appears in a LO. Hence, an IDF score of 1 denotes a very strong relationship between that term and the LO.

In this research, no linguistic tools like suffix stripping were used. The main reason for this decision was the absence of a performing suffix stripping algorithm for legal documents. The idea to develop a computable method with a minimum of human intervention strengthened this decision. From literature [11] too, on top of a random test, it is understood that the term use in legal documents is less diverse as observed in standard documents. Nonetheless it is recommended to perform further research on this issue.

Index algorithm

After indexing, the following step in the process is the development of a strategy to relate LO to each other, as observed in a legal index. The algorithm used is based on the IDF of each term in a LO. Starting with a LO_x and a LO_y the similarity r_{xy} between both law objects can be computed as follows:

$$r_{xy} = \sum_{i=1}^{n} IDF_i \;\Big|\; ti \in \{LO_x \cap LO_y\}$$

$$LO_x, LO_y \subset \{t_n\}$$

In this discussion, both previous LO are scrutinised. The relation between art.2.3.1. and art. 4.2.2.3.§2 can be calculated by focussing on the identical terms. In this case, the identical terms are "asphalt", "hydrocarbons" and "oil".

The IDF are respectively:

$\mathrm{IDF}_{\text{art.2.3.1./asphalt}} = 0.333333333$
$\mathrm{IDF}_{\text{art.2.3.1./hydrocarbons}} = 0.011764706$
$\mathrm{IDF}_{\text{art.2.3.1./oil}} = 0.219512195$
$\mathrm{IDF}_{\text{art.4.2.2.3.§2/asphalt}} = 0.666666667$
$\mathrm{IDF}_{\text{art. 4.2.2.3.§2/hydrocarbons}} = 0.011764706$
$\mathrm{IDF}_{\text{art. 4.2.2.3.§2/oil}} = 0.024390244$
yielding a sum of:
$r_{\text{art.2.3.1./art.4.2.2.3.§2}} = 1.267431850$

The r value reflects the strength of a relation between two specified LO. Based upon the frequency distribution shown in table 3, the above r value indicates a high relative strength, situated within the 10,4% range of the strongest relations between law two objects in this Flemish order.

Table 3 : Frequency distribution of the r value

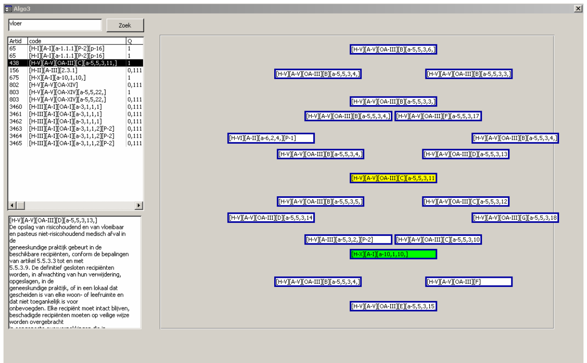
Class	Frequency (%)
0.001 < r < 0.01	0.1
0.01 < r < 0.1	16.9
0.1 < r < 1	72.6
1 < r < 10	10.3
10 < r < 100	0.1
	100

NAVDEX

To visualise the relation between law objects, a database was designed, called NAVDEX. Figure 2 shows a screenshot of the application, returning data on a full text search query. The search term input field is positioned in the upper left corner of the frame. The results of the query are presented underneath the input field, inviting the user to select a specific LO. The article corresponding to the selected LO is returned entirely below, thus completing downwards the left part of the window. The right part of the window is used as a content field concerning the selected LO.

The selected LO is situated in the centre of the content field (yellow box). Counter clockwise, the associated LOs are shown. The closer a suggested associated LO is positioned to the central LO, the stronger the relationship.

If an associated LO too includes the specific search query, the background of the box becomes colored (green box). This feature highlights the specificity of a search query.


Algo3
Zoek
[H-V][A-V][OA-III][B][a-5,5,3,6,]
[H-V][A-V][OA-III][B][a-5,5,3,4,]
[H-V][A-V][OA-III][B][a-5,5,3,3,]
[H-V][A-V][OA-III][B][a-5,5,3,3,]
[H-V][A-V][OA-III][B][a-5,5,3,4,]
[H-V][A-V][OA-III][F][a-5,5,3,17
[H-VI][A-II][a-6,2,4,][P-1]
[H-V][A-V][OA-III][B][a-5,5,3,4,]
[H-V][A-V][OA-III][B][a-5,5,3,4,]
[H-V][A-V][OA-III][D][a-5,5,3,13
[H-V][A-V][OA-III][C][a-5,5,3,11
[H-V][A-V][OA-III][B][a-5,5,3,5,]
[H-V][A-V][OA-III][C][a-5,5,3,12
[H-V][A-V][OA-III][D][a-5,5,3,14
[H-V][A-V][OA-III][G][a-5,5,3,18
[H-V][A-III][a-5,3,2,][P-2]
[H-V][A-V][OA-III][C][a-5,5,3,10
[H-X][A-I][a-10,1,10,]
[H-V][A-V][OA-III][B][a-5,5,3,4,]
[H-V][A-V][OA-III][F]
[H-V][A-V][OA-III][E][a-5,5,3,15

vloer

Artid	code	Q
65	[H-I][A-I][a-1.1.1][P-2][p-16]	1
65	[H-I][A-I][a-1.1.1][P-2][p-16]	1
438	[H-V][A-V][OA-III][C][a-5,5,3,11,]	1
156	[H-II][A-III][2.3.1]	0,111
675	[H-X][A-I][a-10,1,10,]	1
802	[H-V][A-V][OA-XIV]	0,111
803	[H-V][A-V][OA-XIV][a-5,5,22,]	1
803	[H-V][A-V][OA-XIV][a-5,5,22,]	0,111
3460	[H-III][A-I][OA-I][a-3,1,1,1]	0,111
3461	[H-III][A-I][OA-I][a-3,1,1,1]	0,111
3462	[H-III][A-I][OA-I][a-3,1,1,1]	0,111
3463	[H-III][A-I][OA-I][a-3,1,1,2][P-2]	0,111
3464	[H-III][A-I][OA-I][a-3,1,1,2][P-2]	0,111
3465	[H-III][A-I][OA-I][a-3,1,1,2][P-2]	0,111

[H-V][A-V][OA-III][D][a-5,5,3,13,]
De opslag van risicohoudend en van vloeibaar en pasteus niet-risicohoudend medisch afval in de
geneeskundige praktijk gebeurt in de beschikbare recipiënten, conform de bepalingen van artikel 5.5.3.3 tot en met
5.5.3.9. De definitief gesloten recipiënten worden, in afwachting van hun verwijdering, opgeslagen, in de
geneeskundige praktijk, of in een lokaal dat gescheiden is van elke woon- of leefruimte en dat niet toegankelijk is voor
onbevoegden. Elke recipiënt moet intact blijven, beschadigde recipiënten moeten op veilige wijze worden overgebracht

Fig. 2 : Screenshot NAVDEX

Evaluation

The most common way to evaluate the search results or the retrieval performance of a search robot is to measure the effectiveness by precision and recall (figure 3). Both parameters are calculated respectively as the portion of retrieved material that is actually relevant and the portion of relevant material that is actually retrieved in answer to a search request.

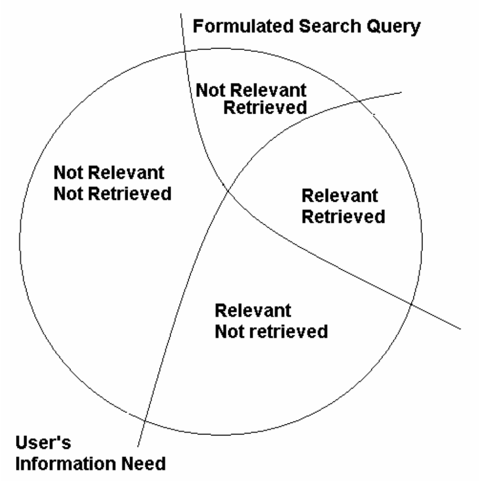


Fig. 3 : Precision and recall [12]

In this research a more subjective measurement strategy was followed, i.e. the relative recall. The relative recall is defined as the ratio between the number of relevant documents found by the system (= number of retrieved LO) and the number of relevant documents the user expects to find (= number of expected LO) [13]. This evaluation method is proven to be more factual, although it requires familiarity with the legal content.

As part of the research, two experts in the field of environmental legislation were asked to evaluate the NAVDEX database and comment on its performance. Both experts, masters in law with a minimum of four years of experience in environmental law, were asked to pick LO from the results list, whilst NAVDEX returned the associated LO.

For each query they computed the relative recall:

$$RECALL_{rel} = \frac{number\ of\ relevant\ documents\ retrieved\ by\ the\ system}{number\ of\ relevant\ documents\ expected\ by\ the\ domain\ expert}$$

In order to compute the average $\text{RECALL}_{\text{rel}}$ ratio the legal experts launched 10 different queries using the full text search tool.

Table 4 : Overview of the double test results

Search query	Selected LO	Nr LO retrieved	Nr LO expected	Recall$_{rel}$
list of waste	[H-I][A-II][a-1.2.1.][P-1]	13	13	1.00
acceptance obligation	[H-III][A-I][OA-I][a-1.1.2][P-1]	4	4	1.00
industrial waste	[H-II][A-II][a-2.2.1]	1	4	0.25
offset waste	[H-III][A-II][a-3.2.1][P-1]	5	10	0.50
wreck	[H-III][A-III][a-3.3.1][P-1]	13	15	0.87
Vlarem	[H-I][A-I][a-1.1.1.][P-2][p-4]	9	10	0.90
users certificate	[H-IV][A-III][a-4.3.1]	5	20	0.40
background values	[H-V][A-II][a-33]	4	6	0.67
register	[H-IV][A-I][a-22][P-1]	8	15	0.53
soil remediation standards	[H-V][A-I][a-31]	2	9	0.22

The first column of table 4 represents the search term that is launched in the full text search engine of NAVDEX. From the search results, each expert randomly picked out a specific LO as show in the second column. Finally the last column calculates the RECALL$_{rel}$ ratio.

As demonstrated in table 4, there is a high fluctuation of the RECALL$_{rel}$ ratio. In the test case, a range from 0.22 to the maximum score of 1.00 is observed. For example, when the first row of the table is examined, the search term "list of waste" returns 13 out of 13 expected LO. Hence in this case, NAVDEX is successful in finding the same associated LO as the human domain experts do. However, the search term "industrial waste" in the third row returns only 1 out of 4 associated LO, as expected by the domain experts.

In global, the average RECALL$_{rel}$ ratio of NAVDEX for the Vlarea order is 0.63. Hence, NAVDEX is able to retrieve nearly two third of the associated law objects. However, the question raises whether an approximating 2:3 ratio is considered as acceptable to return answers to search orders concerning environmental legislation.

Furthermore, it is recognised that selecting a right term to request a search is preconditional to getting the rigth answer. But with this statement the debate on law terminology is entered, without contributing to the project research.

In general, the experts emphasised their appreciation concerning the return and the visualisation of the search responses. In particular the highlighting of the suggested LO when exclusively containing the search term was recorded as a major step forwards in the e-consultation of environmental legislation.

5 Conclusions

When screening the (environmental) legislation for hits upon a query, NAVDEX is considered as a helpful tool. It computes the relation between law objects with an overall certainty of gaining nearly two out of three answers. Accordingly, when a critical search of the legislation is aimed at, it is prudent to double check the NAVDEX return with a law expert since complex strains only return one fifth of the expected law objects. On the other hand, in one out of five queries NAVDEX yields all law objects as compared to manual indexing.

The wide range of variety of the RECALL$_{rel}$ ratio evokes apprehensiveness so as to use NAVDEX as a full automatic indexation tool. For semi-automatic indexing, however, NAVDEX is proven to be successful. It suggests law objects that in due course may need confirmation by human experts, but in general are satisfactory. Moreover, the visualisation of the search results is regarded as helpful in yielding a quick view on the relation between law objects. The counter clockwise visualisation as well as the additional information based on background colors, was appreciated by a test panel.

Since promising for semi-automatic indexing, finetuning the NAVDEX tool is aimed at. A wide range of options is open to further research. Concerning the information retrieval part, the algorithm can be refined. Likewise, the study can focus on the accessibility of environmental parameters since it is considered unique in the domain of legal texts. Benchmarking with WEKA [14], an open source software machine learning and data mining toolkit, is considered. From the comments of the test panel, future work should also concentrate on the visualisation part of NAVDEX. In addition, classifying and versioning (technical) legal texts should be examined since legislation, in particular environmental legislation is proven to have a high turnover. Hence versioning is crucial to correlate the right law objects.

In general, the main objective of further research is to improve the average $\text{RECALL}_{\text{rel}}$ ratio.

6 References

[1] Lancaster, F.W. Towards Paperless Information Systems Academia Press New York, 1978

[2] Bing, J. *Designing Text retrieval systems for conceptual searching*, International Conference on Artificial Intelligence and Law archive Proceedings of the 1st international conference on Artificial intelligence and law table of contents Boston, Massachusetts, United States, 1987

[3] N. Goharian, D. Grossman, N. Raju, O. Frieder, "Migrating Information Retrieval from the Graduate to the Undergraduate Curriculum", Journal of Information Systems Education (15:1), April 2004.

[4] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989

[5] van RIJSBERGEN, CJ, 'File organization in library automation and information retrieval', Journal of Documentation, (32), 1976, pp.294-317

[6] Pulgarin A., Gil-Leiva I. *Bibliometric analysis of the automatic indexing literature 1956-2000*, Information Processing and Management (40), 2004, pp. 365-377

[7] Luhn, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, (2), 1958, pp. 159-165

[8] G. Salton, *A blueprint for automatic indexing*, ACM SIGIR Forum (1981)

[9] M.F. Porter, *An algorithm for suffix stripping, Program* (14: 3), 1980 pp. 130-137.

[10] Maarek, Y., Berry, D. & Kaiser, G., *An information retrieval approach for automatically constructing software libraries*. IEEE transactions on software engineering, 1991, 800–813

[11] E. Schweighofer, A. Rauber, M. Dittenbach *Automatic text representation, classification and labeling in European law*, International Conference on Artificial Intelligence and Law archive, 2001

[12] C.W. Cleverdon, J. Mills, and E.M. Keen. Factors determining the performance of indexing systems, vol. 2: Test results. Technical report, *Aslib Cranfield Research Project, Cranfield, England, 1966*

[13] R. Baeza-Yates, Ribeiro-Neto, B Modern Information Retrieval Addison-Wesley Longman, *Reading MA, 1999*