A Greek Retrieval System in the Legal Domain

Apostolis Apostolou, Lilian Mitrou, Theodore Kalamboukis
Information Processing Laboratory,
Department of Informatics,
Athens University of Economics and Business,
76 Patission Street,
Athens 10434, Greece

Abstract

In this paper, a new pilot Greek legal information retrieval system is described. The automatic generation of hypertext links on top of the original printed PDF document is the major contribution of this work. Navigation through hyperlinks give the user the possibility to instantly access a referenced document without submitting a new query to the system. The retrieval model is based on page-size passages and a scoring relevance function is proposed that takes into account information from the structure of documents. The evaluation of the system was made by experts with very encouraging results.

1. Introduction

Information retrieval has a long history in the research community since the early sixties, but it is just the last thirty years that it gained a huge popularity due the explosion of information in the web. In the law domain, information retrieval revealed its significance at a quite early stage. The first text retrieval in the legal domain dates back to 1963, with the LITE ('Legal Information Thru Electronics') system of the USA Air Forces. In 1970, Professor Spiros Simitis [1] in his book "Informationskrise des Rechts und Datenverarbeitung" stated that there is a strong need to improve the performance of legal research in order to meet the increased requirements of the modern welfare state.
Since then, several computerized legal information services were developed, which today are used in any jurisdiction, such as Reed Elsevier's LEXIS-NEXIS service, or Westlaw that is one of the primary online legal research services for lawyers.
The increased demand for retrieving complex information needs fast and accurate from one side and the peculiarities of the language appearing in legal texts from the other side resulted in the development of a distinct research area in the legal domain. This explains the development of an independent research field in Artificial intelligence and Law, mainly concerned with research and applications on deriving computational models of legal knowledge reasoning and decision making.
Recently legal retrieval tracks have been introduced in the context of international conferences, such as Text Retrieval Conference (TREC) legal track[1], with the goal to develop search technology that satisfies the needs of lawyers in effective discovery in digital document collections.

In Greece there are mainly two, operational Legal Information Retrieval (LIR) systems: "Nomos"[2], a private system and the Legal Information Bank, "Isokratis"[3] of Athens Bar Association. Access to "Nomos" is subject to subscription (based on time charges) while "Isokratis" is freely accessible to lawyers. Documents from a variety of sources are included, such as Greek and European legislation as well as Greek and European jurisprudence. Both systems are constantly updated and provide extensible thematic indexes of the legal material. This is an important feature that offers to users easy navigation towards their information needs. Structured and free search tools are offered, accompanied by a legal term-based hierarchical index to support search. When users search for a specific provision or law, the system returns its latest modification with links at the top of the result pages that lead to the initial legislation.
Users can search for provisions by selecting the appropriate topic from the thematic index or by selecting the year, the issue and the number of the "Official Government Gazette of the Hellenic Republic (OGG)

[1] http://trec-legal.umiacs.umd.edu/
[2] http://lawdb.intrasoftnet.com/nomos/nomos_frame.html
[3] http://www.dsanet.gr/1024x768.htm

they are interested in. Both platforms offer also the possibility of free text search, giving users the chance to search within the documents for specific terms or phrases.

Furthermore, the official National Printing House website offers some basic search functions as well. The National Printing House website[4] is the State's official on-line system and operates since 1996. Every legal document, including Laws, Presidential Decrees, Ministerial Decisions etc. since 1833, is available to download -free of charge- in PDF format. Each document is digitally signed ensuring the integrity of its contents and additionally it is an exact digital reproduction of the printed original copy. National Printing House website is not a retrieval system, so its free search capabilities are very limited and the user interface is not particularly friendly.

In this paper, we present the development of a new Greek Legal Information Retrieval System. The main objective of the system, developed initially in the context of a postgraduate thesis, is the implementation of modern practices of information retrieval, such as the passage based retrieval, the navigation from document to document converting "hard wired" references to hyperlinks, the simultaneous usage of structured and unstructured search and the presentation of the documents using the exact layout of their original printed counterparts. In order to achieve the best possible retrieval performance, each document was divided into unique passages which were independently retrievable. The basic "passage unit" we used was the content of each -physical- page. Extensive effort was made to create a set of tools that allow the fully automated creation and maintenance of the system's document collection.
The automatic generation of hypertext links on top of the original printed PDF document is the major contribution of this work. Hyperlink navigation means that users can instantly read a referenced document without submitting a new query to the system. Another point which is worth mentioning here is the use of a scoring relevance function that takes into account information from the structure of a text. This function significantly improved the quality of the results compared to other traditionally used functions. The system is available to the users through web address 195.251.252.22:8084/laws/.
The rest of this paper is organized as follows: In the second section a requirements analysis defines the fundamental options that should be followed by a modern LIR system, regarding the document collection, the browsing of documents, indexing and searching strategies. The third section deals with the architecture of the system, alongside with the presentation of the necessary utility modules which were developed. In the next section, the evaluation method and results are presented. Finally, in the fifth section we draw our conclusions and discuss on future extensions of the system.

2. Requirements Analysis

Legal texts are a special category of documents: they include both structured and unstructured information and they are written in a very particular language. Titles, paragraphs, provisions, references and dates may be considered as structured information. Legislators and judges tend to express themselves in exceptionally long and quite complicate sentences. In addition, many terms have a specific semantic meaning within the legal context. A detailed analysis of legal texts concerns the vocabulary, syntax and semantics of individual sentences, clauses and phrases can be found in [2, 4]. This analysis has shown that language tools, like thesauri, and classification of the collection documents into a hierarchical scheme are useful to enhance results in traditional legal information retrieval.
The size of a document (law) may vary from a single page to a few hundred pages, including multiple sections, possibly dealing with a large variety of subject matters.

Legal sources are mainly divided into three categories: legislation, judicial decisions and literature. Some legal texts are related to the judicial proceedings, such as judgments and court decisions while other refer to legal relationships between private (legal and/or natural) persons, for example contracts. However, most commonly used documents in retrieval systems are statutes (legislation) and court decisions (cases).
The retrieval procedure of legal documents, exhibits many interesting features and the implementation of such systems requires a special approach. In order to define a minimum set of requirements for the design of a legal information system, we should analyze the behavior of the users of legal content, the type of the

[4] www.et.gr

content itself, and the work they typically perform. In general, a LIR system should reflect the look and-feel of the traditional printed document version as far as possible, in order to make it easier for users to orientate themselves. All screens produced by the system should have an identical look with the printouts, to ensure that they are easy to read and acceptable in court. Emphasis should be given, on user-friendliness through a clear and simple interface. The search system should be as simple as possible for the non-experienced user.

Another important aspect that distinguishes the legal domain from other domains is related to the material itself and the way it is used. Regarding the structure of legal material, a legal information system should leverage the structure of the legal material as far as possible, by providing search tools that are related to this structure (e.g. year filtering). Traditional systems do not provide mechanisms to move along the network of cross-references and citations inside the same document or across different documents, thus requiring complex search operations on the part of the user. Besides being highly structured, legal documents tend to be longer than those found in other domains. Legal material is also extensively cross-referenced and interrelated. Cross-references exist within documents, pointing to other documents, other sections or footnotes. The best way to navigate in such a collection of legal documents appropriately is via hyperlinks. However the mere number of cross-references in large collections makes the manual insertion of hyperlinks a highly resource-intensive, if not impossible task.

The collection of documents in our framework contains only legislation. This decision was based on the fact that searching for legislation is the main task of any professional related to the Legal Domain and secondly it was taken due to the lack of adequate recourses. The content of these documents is freely available from the National Printing House Website. The statutes vary in size from one page of few kilobytes to hundredths of pages of several megabytes. A collection including so large documents requires special approach in terms of processing, storing, indexing and evaluating so that the system is both fast and effective. Our framework stores and processes documents in an appropriate repository, by splitting documents into pages, so that single pages can be retrieved efficiently, while browsing whole documents is still possible.
It is known from research on the query-logs from web search that queries are quite short. These findings led to the idea of a simple Google-style search interface supporting the retrieval of documents, by exploiting knowledge about their structure and adopting ranked retrieval techniques. For simplicity each statute was divided into two fields: the title and the body (content). These two fields should receive different weights. A query term that matches the title of a statute should, result in higher score than a hit in the body. Finally, another requirement was that searching by structured criteria had to be possible. For instance, entering the number (FEK) of a statute, returns directly the desired document, bypassing the search results page.

Legal material is undergoing continuous change and new material is constantly accumulated. The IR system should thus facilitate the update and addition of content, or, to be more precise, it should allow for new documents to be added at runtime, and for a number of documents to be indexed without having to rebuild the whole index.

3. Systems' Architecture

The implementation procedure of the system is divided into three stages. In the following we give a high level description of the major modules of our prototype architecture that is, crawling, indexing, and searching. In the first stage, the statute collection is created, by automatically downloading the statute-documents from the National Printing House website. Apart from downloading the files, we also acquired the appropriate structured data on which most features of the system rely. The second stage includes the preprocessing of the documents. References were automatically detected and transformed into hyperlinks and the documents were indexed for being retrievable by the free search engine. Finally, in the third stage, the system's web site was developed and the evaluation sub-system was incorporated. The first two stages

were completed with two separate modules that were developed as stand-alone applications with graphical user interface.

3.1 Crawling and Preprocessing

The most challenging task we faced was the automatic creation of the collection. Due to the large number of documents, this task had to be performed without human intervention. The documents were downloaded from the National Printing House website, which offers a search form that allows users to search and download any statute- document in PDF format. Our crawler automatically “fills” that form, creating a request with the appropriate data and submits it to the server. Server responds with an HTML page which contains a link to the PDF file that satisfies our search. The crawler fetches server's response, detects the link and starts downloading the document. When download is completed, the file is stored locally using a fully descriptive file name, according to the document's issue number, issue type and year. That step was necessary for creating a browsable collection, because the files downloaded from the NT website are originally named in a meaningless way (all documents are named “document.pdf”).

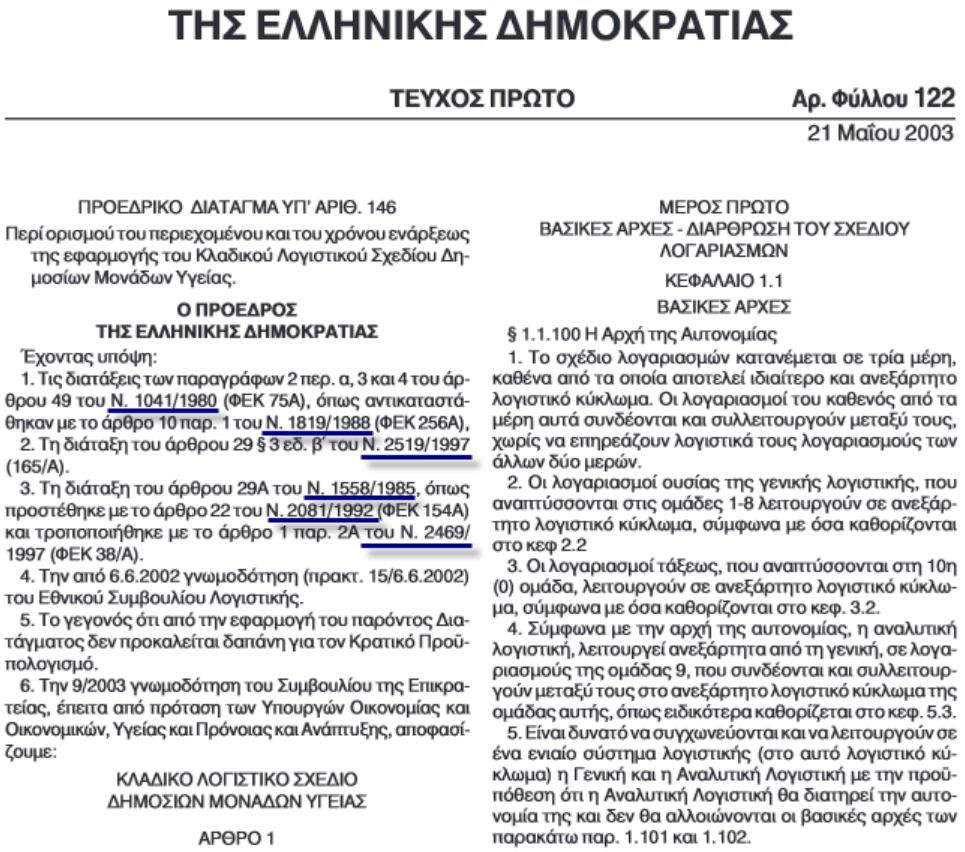
The major feature of the system concerns the automatic conversion of law-references to hyperlinks in a way that integrates structured-data search with unstructured text search. For this purpose an index was built, where each Greek Law ever published since 1883 and each Presidential Decree since 1973 was linked to the document in which was first published. For example, Law no. 4000/1929 was published in the year's 1929 72A issue. Our system needs to know this information when a structured search query is processed. In addition to this, when the link detection module meets a law reference, sets the links' destination to the correct document using the Structured Search Index. Building this index was technically similar to the creation of the collection. Except from all the necessary details for each document-law, a brief description of each law is also included in the structured search index.

In order to build the index for text based retrieval, we first had to extract the contents of the PDF files and save them as plain text. For this purpose, we used the open source library IcesoftICEpdf, which provides methods for processing PDF files. To support our passage retrieval approach (described later), we had to extract the text of each document page by page and save each page into separate TXT files. Extracting text from our collection was not a trivial task. There had been several problems during the conversion of the PDF legislation-files into plain text format. This occurred due to the unstructured and not rigorous editing format of the original files. The original forms contain many scanned images, signatures, different text formats, tables, charts etc. In general, there was no steady structure and formatting in each document. All documents before 2000 are scanned images saved in PDF files and text extraction was impossible without a sophisticated OCR algorithm whose implementation was not part of our research. A lot of documents, mostly the pre-2006 ones, where encoded in a way that plain text extraction was impossible. In this case, the originally Greek characters of the documents were extracted as unreadable ASCII characters. Solving this issue required the creation of encoding tables where all the unreadable characters were mapped to the appropriate Greek characters. When the PDF Text Extraction tool detected that a document produced unreadable results, a simple algorithm exploiting this dictionary, was invoked to ensure the success of the procedure.

After we have obtained the plain ASCII text, we proceeded with a morphological analysis, which included:

- Removal of end-of-line hyphenation. Although the use of hyphenation of words across lines makes printed documents to have a professional look they destroy the words in their electronic representation. This problem is more frequent in multicolumn documents, like the OGG Greek statute documents. In order to reduce those errors we parse the documents and remove the hyphens at the end of lines.

- Tokenization. The text is parsed and individual words are recognized.
- Removal of stopwords. A stopword list contains a set of the most frequently occurred terms within the document corpus such as articles, prepositions were removed.
- Stemming: The reduction of the remaining words to their stem form by removing word suffixes [9].
- Weighting of indexed terms: The computation of the importance indicator or term weight of each stemmed word, based on the TF*IDF weighting scheme [Manning, 2008]. It is generally assumed that terms that occur frequently in a text and infrequently in the complete document corpus are good index terms.



ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ

ΤΕΥΧΟΣ ΠΡΩΤΟ Αρ. Φύλλου 122

21 Μαΐου 2003

ΠΡΟΕΔΡΙΚΟ ΔΙΑΤΑΓΜΑ ΥΠ' ΑΡΙΘ. 146

Περί ορισμού του περιεχομένου και του χρόνου ενάρξεως της εφαρμογής του Κλαδικού Λογιστικού Σχεδίου Δημοσίων Μονάδων Υγείας.

Ο ΠΡΟΕΔΡΟΣ
ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΔΗΜΟΚΡΑΤΙΑΣ

Έχοντας υπόψη:

1. Τις διατάξεις των παραγράφων 2 περ. α, 3 και 4 του άρθρου 49 του Ν. 1041/1980 (ΦΕΚ 75Α), όπως αντικαταστάθηκαν με το άρθρο 10 παρ. 1 του Ν. 1819/1988 (ΦΕΚ 256Α),
2. Τη διάταξη του άρθρου 29 § 3 εδ. β' του Ν. 2519/1997 (165/Α).
3. Τη διάταξη του άρθρου 29Α του Ν. 1558/1985, όπως προστέθηκε με το άρθρο 22 του Ν. 2081/1992 (ΦΕΚ 154Α) και τροποποιήθηκε με το άρθρο 1 παρ. 2Α του Ν. 2469/1997 (ΦΕΚ 38/Α).
4. Την από 6.6.2002 γνωμοδότηση (πρακτ. 15/6.6.2002) του Εθνικού Συμβουλίου Λογιστικής.
5. Το γεγονός ότι από την εφαρμογή του παρόντος Διατάγματος δεν προκαλείται δαπάνη για τον Κρατικό Προϋπολογισμό.
6. Την 9/2003 γνωμοδότηση του Συμβουλίου της Επικρατείας, έπειτα από πρόταση των Υπουργών Οικονομίας και Οικονομικών, Υγείας και Πρόνοιας και Ανάπτυξης, αποφασίζουμε:

ΚΛΑΔΙΚΟ ΛΟΓΙΣΤΙΚΟ ΣΧΕΔΙΟ
ΔΗΜΟΣΙΩΝ ΜΟΝΑΔΩΝ ΥΓΕΙΑΣ

ΑΡΘΡΟ 1

ΜΕΡΟΣ ΠΡΩΤΟ
ΒΑΣΙΚΕΣ ΑΡΧΕΣ - ΔΙΑΡΘΡΩΣΗ ΤΟΥ ΣΧΕΔΙΟΥ ΛΟΓΑΡΙΑΣΜΩΝ

ΚΕΦΑΛΑΙΟ 1.1

ΒΑΣΙΚΕΣ ΑΡΧΕΣ

§ 1.1.100 Η Αρχή της Αυτονομίας

1. Το σχέδιο λογαριασμών κατανέμεται σε τρία μέρη, καθένα από τα οποία αποτελεί ιδιαίτερο και ανεξάρτητο λογιστικό κύκλωμα. Οι λογαριασμοί του καθενός από τα μέρη αυτά συνδέονται και συλλειτουργούν μεταξύ τους, χωρίς να επηρεάζουν λογιστικά τους λογαριασμούς των άλλων δύο μερών.
2. Οι λογαριασμοί ουσίας της γενικής λογιστικής, που αναπτύσσονται στις ομάδες 1-8 λειτουργούν σε ανεξάρτητο λογιστικό κύκλωμα, σύμφωνα με όσα καθορίζονται στο κεφ 2.2
3. Οι λογαριασμοί τάξεως, που αναπτύσσονται στη 10η (0) ομάδα, λειτουργούν σε ανεξάρτητο λογιστικό κύκλωμα, σύμφωνα με όσα καθορίζονται στο κεφ. 3.2.
4. Σύμφωνα με την αρχή της αυτονομίας, η αναλυτική λογιστική, λειτουργεί ανεξάρτητα από τη γενική, σε λογαριασμούς της ομάδας 9, που συνδέονται και συλλειτουργούν μεταξύ τους στο ανεξάρτητο λογιστικό κύκλωμα της ομάδας αυτής, όπως ειδικότερα καθορίζεται στο κεφ. 5.3.
5. Είναι δυνατό να συγχωνεύονται και να λειτουργούν σε ένα ενιαίο σύστημα λογιστικής (στο αυτό λογιστικό κύκλωμα) η Γενική και η Αναλυτική Λογιστική με την προϋπόθεση ότι η Αναλυτική Λογιστική θα διατηρεί την αυτονομία της και δεν θα αλλοιώνονται οι βασικές αρχές των παρακάτω παρ. 1.101 και 1.102.

Figure1. The user can navigate to related legislation through the links underlined with blue color.

3.2 Link Detection

Achieving hypertext-style navigation through the collection demanded that each PDF file was scanned for references to other documents. The references had to be efficiently detected and converted into clearly visible links. Using the extracted plain text, with the help of a finite state automaton, the documents are parsed searching for certain patterns of characters (e.g. "ν. 4000/1929", "π.δ. 63/2005") that appear when references are made. Unfortunately, documents use no strictly predefined syntax when referencing to Laws or Presidential Decrees. As an effect, the list of possible pattern variations was quite long and the detection expensive in processing power terms. The destination of the links is determined from the Structured Search Index exploiting the fact that each reference itself provides adequate structured information (year and Law or P.D. number) to make a successful structured query. When the detection stage is completed, the list containing the parts of the document that must be converted into links and their target file is fed to a PDF processing algorithm that marks the corresponding words in the PDF file by underlining them with blue color (figure 1).

3.3 Indexing and ranking system

Another crucial aspect of the new system was the availability of multiple and accurate search tools. The prototype system offers three basic categories for searching within the collection: structured search using clearly defined criteria (citations), traditional free search based on user submitted queries and a combination of the former two methods. Users may search using a simple free text query and then refine the results by applying structured filters such as publishing year. In this way the time spent for searching and browsing results is reduced significantly.

4. Passage Retrieval

Passage retrieval, [6, 3] is the task of retrieving only the portions of a document that are relevant to a particular information need. The technique could be useful for reducing the amount of non-relevant text presented to a user. Many definitions of passages have been proposed in the literature. Some of these definitions rely on semantic properties such as sentence boundaries, and others on fixed length sequences of words. In [7], it was observed that techniques that made less use of semantics were in general most effective. Results showed that the greatest retrieval effectiveness is achieved with passages of fixed-length sequences of words that can start at any point in a document.

As we mentioned before statute documents are lengthy and usually contain sections that deal with a variety of completely unrelated subjects. Thus, in our case, passage retrieval was used to improve performance due to document length normalization. The documents were divided into smaller sub-documents or passages and the retrieval score of these passages is combined to give the final retrieval score of the whole document.

For simplicity, documents were segmented into page-size passages. It has been shown [8] that combining similarities from the best passages in a document can be more effective than taking a document as a whole. In this approach, documents are still returned in response to queries, providing the passages satisfying the information need. A retrieval score was calculated for each passage extracted and the sum of the scores of all retrieved passages was taken as the document's final score.
We mention here that the number of passages in a collection is much larger than the number of documents so ranking cost at query time is higher.

5. User Interface

The user interface can be described in three words: simple, friendly and, fast. Presenting the documents without altering their original layout was a very important milestone for the new system. In addition to that, our system had to fulfill this goal even if our documents had to be preprocessed so that references to other documents to be transformed into hyperlinks using a clearly visible annotation.

6. Implementation details

The freetext search index was built using the open source search engine library Apache Lucene[1]. The TXT files produced from the PDF Text Extraction tool representing the document of our collection were preprocessed so that they could be indexed. The procedure included the typical steps of document preprocessing: case normalization, tokenization, stop word removal and stemming. Each page was indexed as an independently retrievable document, in accordance with our passage retrieval approach. Except from the contents of the page, index file stored fields containing structured information (document's issue number, publication year, page number, etc.) and a field with the document's summary.

The system returns all legislation from 1894 if they are accessed through hyperlinks. All documents before 2000 are accessed with their citations only (FEK number, year, legislation number) through the NT web site. At present our collection contains 7090 documents (legislations). From those only 3094 from years 2000 till today have been indexed and the rest are accessed through hyperlinks. The collection of documents is split into 32094.

The core of the system's website is a simple JSP Application. When users submit the search form, our server receives a request containing the query. Depending on the search type, the application calls either the free search or the structured search algorithms that retrieves the relevant documents using the appropriate index. Another module takes on and dynamically produces the HTML results page, which is sent to the users. This page contains the list of results divided into pages of 10 documents each, a brief summary and the most relevant passages of each document. The whole process, from receiving the request to sending the HTML response, lasts but a few milliseconds and makes very limited use of the server's resources.

7. Systems Evaluation

Traditional retrieval performance measurements are based on ad-hoc collections. Specific queries with predefined results are submitted to the system. The system can be eventually evaluated by comparing the produced results to the expected –correct- ones (“golden standard”). The measures employed to evaluate an information retrieval system are precision and recall [5]. Precision is the fraction of the documents retrieved that are relevant to the user's information need. Recall is defined by the fraction of the relevant documents that are successfully retrieved. Unfortunately, no such testing collection of Greek legal documents exists.

This led us to use Web Retrieval measurements for evaluating our system. What really matters is rather how many relevant documents are included on the first page of the results. This leads to measuring precision at fixed low levels of retrieved results, such as 10 or 30 documents. This is referred to as “Precision at k”. This measure has the advantage of not requiring to estimate the exact number of relevant documents and the disadvantage that it does not average well, since the total number of relevant documents for a query has a strong influence on precision at k.

For the evaluation we relied on volunteers who agreed to participate in the evaluation of the system. For this purpose, an evaluation module was implemented and the volunteers received the evaluator login credentials to enter the evaluation section of the web site. This section was a carefully designed sub system that had to record every single evaluation session. Volunteers could submit their evaluation just by using the system's search tools and ticking on a checkbox next to each result to mark it as relevant, if it was satisfying their information needs. The submitted results accompanied by some supplementary important data, such as the documents marked as relevant and the evaluator's profession, were automatically appended to a comma separated value file (CSV), which was used later in a Spreadsheet application to calculate the evaluation results.

As we already mentioned, during indexing, each page was independently indexed. This means that our search algorithm is able to rate each page of the collection separately. Thus the ranking algorithm performs in two stages: In the first stage we select to top, retrieved pages returned to the submitted query. This parameter was set to 200 in our experiments. In the second stage we parse the returned results and calculate the total relevance of each document, by summing up the scores of the corresponding retrieved pages. Practically, this means, the more pages of a document have been retrieved and the more relevant these pages are to the query, the more relevant the entire document is.
Furthermore each document was divided two fields: title and body. If we suppose that the body, b, contains k pages (passages), b= (p1,....,pk) the final score of a document d in respect to a query, q, is estimated by the following relation

$$SCORE(d,q) = w_1 * score(q,title) + w_2 * score(q,b)$$ where

$$score(q,b) = \sum_{j=1}^{k} score(q, p_j)$$

The weights w1, w2 were estimated experimentally.

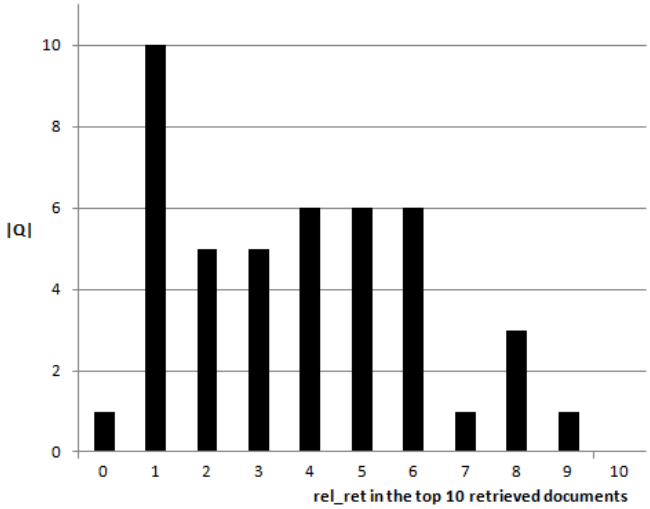


Figure 2. Distribution of relevant retrieved in the top 10 retrieved documents vs the number of queries, |Q|.

For the evaluation of our system we used professional volunteers in the legal domain. For this purpose a web-based module was developed where the participants could submit their queries, ticking those from the returned documents which were satisfying their information needs. All submitted queries were automatically stored in an spreadsheet file for further processing. The participants determined a set of relevant documents for each query which we used as a golden standard in our evaluation of system's parameters. In total 44 queries were selected for the evaluation. The averaged length of the submitted queries was 2.42 words and the results showed an averaged precision in the top 10 retrieved documents precision@10=39.54% with equal weights in the parameters w_1=0.5, w_2=0.5. The best results precision@10=41,13% were achieved for values of w_1=0.6, w_2=0.4. From figure 2 follows that in average 4.2 documents per query were found relevant and only one query failed to retrieve a relevant document in the top tenretrieved documents

8. Conclusions and Future developments

We have described the implementation of a legal information retrieval system for the Greek legislation. All the tools we developed are cross-platform applications with a friendly graphical user interface. They are easy to use, without demanding any complex technical knowledge. The system's search engine allows both free and structured search. The combination of free and structured search acts as a filter to increase the precision and reduces the time needed for browsing through the results.

To increase the precision of retrieval, a combination of field-based, passage retrieval was applied. Each legislation document is divided in two fields: title and body. The title is automatically extracted and the documents are split into pages which form the basic passage unit. The system initially evaluates and retrieves pages independently, and calculates the total document relevance-score by summing up the weighted scores of the title and body. This scoring technique proved to enhance retrieval performance in the evaluation section.

Two features are the main contributions of this prototype: Documents are presented with their exact physical (printed) layout, and navigation from document to document is possible using automatically

assigned hyperlinks in the PDF files. The interface is simple and friendly and offers useful extra information to users when browsing through the results. Finally, the system continues being usable even in low resolution screens of mobile devices or older computers.

Our system provides all the necessary administration tools to create and manage a document collection from scratch. Every procedure, from the initial download of the documents to the reference detection and conversion to links is automated and demands no human inspection. These tools include the Crawler, a free search index builder, a structured search Index builder and a converter of references to hyperlinks.

At the moment, the link module, detects references to Laws and Presidential Decrees. Developing a standalone client side PDF reader, specially focused on the presentation of legal documents, would allow the system to include a lot more extra information than hyperlinks. A list of laws that have altered the articles of the displayed document or an indication of whether the law being displayed has been superseded, are some of the important features that are currently under investigation.

As we have already mentioned the official Greek legislation published by OGG does not follow a rigorous editing style and before year 2000 all legislation is saved in image form. Even after the year 2000,several different encodings are used in the PDF files, a fact that toughens up their conversion to txt and their preprocessing. It is important to note that metadata is incredibly useful for codifying, categorizing and annotating documents. Adding metadata to primary legal information is an exciting area of legal technology that will help delivering better information both to lawyers and public. Consequently, there is a need for entering metadata at the release time of legislation, if we want to have more intelligent Law Information Retrieval systems in the Greek language in the future.

9. References

1. Spiros Simitis. Informationskrise des Rechts und Datenverarbeitung, Karlsruhe, Verlang C.F. Müller, 1970.
2. Marie-FrancineMoens, Innovative techniques for legal text retrieval, Artificial Intelligence and Law 9: 29–57, 2001.
3. Ismael Hasan, Javier Parapar, Roi Blanco, "Segmentation of Legislative Documents Using a Domain-Specific Lexicon," 19th International Conference on Database and Expert Systems Application, dexa, pp.665-669, 2008.
4. Kenneth A. Yates, Charles E. Shapiro, Establishing a Sustainable Legal Information System in a developing country: A Practical Guide, The Electronic Journal on Information Systems in Developing Countries, 42, 8, 1-20, 2010.
5. Christopher D. Manning, PrabhakarRaghavan and HinrichSchütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
6. G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In SIGIR ’93, pages 49–58.
7. Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited, Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '97), Nicholas J. Belkin, et al. (Eds.). ACM, New York, NY, USA, 178-185.
8. Hearst, M. A. &C. Plaunt (1993). Subtopic structuring for full-length document access, 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, 27-30 June, 1993, pages 59-68, New York.
9. T.Z. Kalamboukis, (1995) "Suffix stripping with modern Greek", Program: electronic library and information systems, Vol. 29, pp.313 – 321