TEXT EXTRACTION AND WEB SEARCHING IN A NON-LATIN LANGUAGE

FOTIS LAZARINIS

A thesis submitted in partial fulfilment of the requirements
of the University of Sunderland for the degree of
Doctor of Philosophy
by existing published works

May 2008

School of Computing and Technology
University of Sunderland
Sunderland, United Kingdom

Abstract

Recent studies of queries submitted to Internet Search Engines have shown that non-English queries and unclassifiable queries have nearly tripled during the last decade. Most search engines were originally engineered for English. They do not take full account of inflectional semantics nor, for example, diacritics or the use of capitals which is a common feature in languages other than English. The literature concludes that searching using non-English and non-Latin based queries results in lower success and requires additional user effort to achieve acceptable precision.

The primary aim of this research study is to develop an evaluation methodology for identifying the shortcomings and measuring the effectiveness of search engines with non-English queries. It also proposes a number of solutions for the existing situation. A Greek query log is analyzed considering the morphological features of the Greek language. Also a text extraction experiment revealed some problems related to the encoding and the morphological and grammatical differences among semantically equivalent Greek terms. A first stopword list for Greek based on a domain independent collection has been produced and its application in Web searching has been studied. The effect of lemmatization of query terms and the factors influencing text based image retrieval in Greek are also studied. Finally, an instructional strategy is presented for teaching non-English students how to effectively utilize search engines.

The evaluation of the capabilities of the search engines showed that international and nationwide search engines ignore most of the linguistic idiosyncrasies of Greek and other complex European languages. There is a lack of freely available non-English resources to work with (test corpus, linguistic resources, etc). The research showed that the application of standard IR techniques, such as stopword removal, stemming, lemmatization and query expansion, in Greek Web searching increases precision.

Acknowledgements

I would like to thank Prof. John I. Tait, my supervisor, and the rest of the staff of the University of Sunderland for accepting and supporting my candidacy for the Doctor of Philosophy. Also I would like to express my gratitude to Dr. Mark Sanderson of Sheffield's University for accepting to act as my external examiner.

I am deeply grateful to my son and my new born baby girl for making my life happier and for their numerous distractions away from the computer. This thesis could not be completed without their love and the continuing support of my wife. I am also indebted to my parents for their moral support and their invaluable help in some of my everyday duties.

Finally, I would also like to thank all my Greek students and colleagues who willingly participated in the Greek Web searching evaluation experiments performed during the last 3 years and reported in my papers.

To my children Eleftherios and Vasiliki

Contents

List of papers on which the PhD is based

1. Lazarinis, F. (1998). Combining information retrieval with information extraction. *20th Annual BCS-IRSG Colloquium on IR*, Autrans, France, pp. 162-174 http://www.bcs.org/server.php?show=ConWebDoc.4410.

2. Lazarinis, F. (2005a). A rule based tool for mining textual data from Greek calls for papers. *2nd Balkan Conference in Informatics*, Ohrid, Skopje, pp.126-133.

3. Lazarinis, F. (2005b). Do search engines understand Greek or user requests "sound Greek" to them? *Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology*, Compiegne France, pp. 43-46.

4. Lazarinis, F. (2005c). Evaluating user effort in Greek Web searching. *10th Pan-Hellenic Conference in Informatics*, University of Thessaly, Volos, Greece, pp. 99-109, ISBN: 960-8029-39-2.

5. Lazarinis, F. (2006). Automatic extraction of knowledge from Greek Web documents. *6th Dutch-Belgian Workshop in Information Retrieval (DIR 06)*, Delft, the Netherlands, pp. 33-37.

6. Lazarinis, F. (2007a). How do Greek searchers form their Web queries? *3rd International Conference on Web Information Systems and Technologies, (WEBIST)*, Barcelona, Spain, pp. 404-407.

7. Lazarinis, F. (2007b). Web retrieval systems and the Greek language: Do they have an understanding? *Journal of Information Science*, 33(5), 622-636, doi:10.1177/016555150607639

8. Lazarinis, F. (2007c). An initial exploration of the factors influencing retrieval of Web images in Greek queries*, Euro American Conference on Telematics and Information Systems (EATIS 07), ACM Digital Library*, doi:

doi.acm.org/10.1145/1352694.1352765

9. Lazarinis, F. (2007d). Lemmatization and stopword elimination in Greek Web searching. *Euro American Conference on Telematics and Information Systems (EATIS 07), ACM Digital Library*, doi: doi.acm.org/ 10.1145/1352694.1352757

10. Lazarinis, F. (2007e). Engineering and utilizing a stopword list in Greek Web retrieval. *Journal of the American Society for Information Science and Technology*, 58(11), pp. 1645-1652, doi: 10.1002/asi.20648

11. Lazarinis, F. (2007f). Forming an instructional approach for teaching web searching to non-English users. *Program: Electronic Library and Information Systems*, 41(2), pp. 170-179, doi: 10.1108/00330330710742935

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation for the Research

Since its conception in 1992 (Berners Lee et al., 1992) the World Wide Web (WWW or Web) has rapidly become one of the most widely used services of the Internet along with email. Its friendly interface and its hypermedia features attract a significant number of users around the globe. As a result, the Web has become a pool of various types of data, dispensed in a measureless number of locations. Finding information that satisfies specific criteria is a regular daily activity of almost every Web user. Web search engines provide searching services through their uncomplicated interfaces.

According to recent statistics 64.2% of the online population are non-English users (Global Reach, 2004). As the Web population continues to grow more non-English users will be amassed online. Recent studies showed that non-English queries and unclassifiable queries have nearly tripled since 1997 (Spink et al., 2002). Even though several Web search engines exist, most of their features and virtues are catered for the English language only. For example, the query "Bookshop New York" in Google retrieves Web pages mentioning the semantically related terms "book", "books" and "bookstore" as well. This is easily understood as the matching terms are emboldened. In contrast, the queries "Librairie Paris" in French, "Libreria Roma" in Italian, "Librería Madrid" in Spanish and "Βιβλιοπωλείο Αθήνα" in Greek, retrieve only pages which include exactly the query terms as they are typed in the query. This query could be more problematic in more complex European, Asian and Afri-

can natural languages. Other information retrieval (IR) techniques such as stemming are employed by international search engines. For example, the query "stemming site:www.dcs.gla.ac.uk/Keith" in Google returns pages containing the words "stem" and "stemming". These terms are emboldened as they are considered matching terms.

English is considered a compact language (www.english-test.net, www.adlcommunity.net). The average English word for example is shorter than in other languages (e.g. German). That is because English verbs, nouns and adjectives do not usually have endings, unless in past tense. There is almost no declension and no conjugation in the English language which makes it much easier to form simple sentences that are grammatically correct. There is only one definite article in the English language whereas there are many variations of the definite article in languages like Greek.

Greek is a linguistically complex language based on a non-Latin alphabet. The Greek alphabet consists of 24 lower and 24 upper case letters. The vowels may get an accent mark when they are in lower case, which is usually absent in upper case, e.g. "υπολογιστής" (computer) and "ΥΠΟΛΟΓΙΣΤΗΣ". Considering the nouns, there are in total 39 different suffixes in all their forms (Triantafyllidis, 1941). Adding the adjectives in all their inflections, there are 17 more different suffixes. Counting also all the possible verb inflections, there are 110 more different suffixes. So, for the general forms of the main inflectional types of the Greek language there are 166 different suffixes. Other non-English natural languages, like Spanish for example, are also quite complicated. More than 20 variation groups for gender inflection and more than 10 variation groups for number inflection have been identified in Spanish nouns and adjectives (Vilares Ferro, 1997). These irregularities influence Web retrieval in Spanish.

International search engines like Google and Yahoo are preferred over the local ones in non-English text searching (Lazarinis, 2005b; Lazarinis, 2005c),

as they employ better interfaces and searching mechanisms. However, both international search engines and domestic Web retrieval systems do not really utilise all the characteristics of other spoken languages than English. For instance, the Greek queries "υπολογιστής" (computer) and "ΥΠΟΛΟΓΙΣΤΗΣ" retrieve different Web pages in most search engines. Correlation between top ranked results is low. These observations are true for text based image retrieval as well (Lazarinis, 2007c). Existing activities like CLEF [http://www.clef-campaign.org] and NTCIR [http://research.nii.ac.jp/ntcir/] are not sufficiently focussed on the requirement to build better search engines for all forms of non-English queries and documents in practice.

Based on the previous observations the motivation behind this thesis is to evaluate the effectiveness of search engines in non-English queries and to propose techniques and tools for improving their effectiveness. The emphasis of the research is on Greek Web searching but the outcomes and the inferences made are applicable to other non-English and non-Latin natural languages.

1.2 Aims and Objectives

The fundamental aims of this research study are to methodologically identify the shortcomings and measuring the effectiveness of search engines in non-English queries and to propose a number of solutions for improving the existing situation. The main focus of the research is on the Greek language. Therefore, the main objectives of this research derived from the aim are as follows:

- Review existing search engine evaluation studies related to non-English Web searching.
- Experiment with extracting textual information from Greek documents.
- Analyze Greek query logs.
- Define a methodology for evaluating the effectiveness of search engines in non-English queries.

- Evaluate search engines using Greek queries and measure the additional user effort using the structured evaluation methodology.
- Propose extensions based on standard IR techniques to search engines in order to improve Greek Web retrieval.
- Develop tools for improving Greek Web searching.
- Propose teaching strategies for helping users improve their searching skills.
- Discuss adaptations of the methodology to other non-English languages.

1.3 Research Hypothesis

The main hypothesis of this work is that the international search engines do not take account of all the grammatical and morphological idiosyncrasies of non-English and non-Latin languages and a disciplined evaluation methodology is needed to provide information about the shortcomings of the existing search engines with respect to a specific natural language.

This work claims that the application of basic IR techniques, such as stop-word removal, in non-English and non-Latin searching could improve the effectiveness of search engines.

1.4 Originality of the Work

The primary originality of this research is the identification of the shortcomings of search engines in Greek queries. Concept based image searching using Greek textual queries was also reviewed. The problems identified span from the inability of some search engines to support Greek queries to the lack of localized interfaces to a differentiation in the precision in semantically identical queries which differ in morphology.

The second novelty of this work is the analysis of a large Greek query log based on the morphological features of the Greek language. A text extraction

experiment using Greek text is another contribution towards the understanding of the extra difficulties confronted in non-Latin text processing. Another point of originality is the creation of a publicly available stopword list for the Greek language and the study of its application in Web searching. A document collection of 5,124 texts was assembled for this purpose. The effect of lemmatization of query terms is also studied in Greek Web queries. The last element of originality is the formulation of a strategy for teaching students and adult learners how to effectively utilize search engines. This instructional approach considers the explanation of the internal search engine intelligence and inefficiencies with respect to non-English natural language as its basic structural element.

1.5 Research Methodology

Initially, previous studies on information retrieval evaluation and search engine evaluation were reviewed. Studying of the papers focusing on non-English queries was of crucial importance so as to identify the criteria and record the suggestions of the researchers. The features of major search engine such as Google and Yahoo were reviewed with English queries so as to record the conveniences offered to English speaking Web searchers. Additionally, studying of research articles analyzing Web query logs was required so as to realize what users search for and how long or short their queries are.

Having acquired the necessary knowledge a Greek query log of 5,698 queries was analyzed so as to record the additional issues emerging in the case of a non-Latin language with complex accentuation and grammar such as Greek (Lazarinis, 2007a). A system using heuristic rules was also created in order to extract specific information from Greek calls for papers (Lazarinis, 2005a, 2006). Using this system the existence of multiple variations of semantically similar information into Greek texts was identified and the importance of creating tools which take account of these differences made apparent. The text ex-

traction system was based on an older extraction tool which was used in experiments with English texts (Lazarinis, 1998).

Following these experiments, the evaluation methodology was formed and it was applied into Greek Web searching (Lazarinis, 2005b, 2005c, 2007b). A number of authentic user-provided queries were run and evaluated into international and local search engines. The results of these runs were the combined estimates of multiple users. Additionally, the abilities of text based image retrieval in Google, Yahoo and MSN were reviewed (Lazarinis, 2007c).

Based on the findings of the importance of stopword elimination, a stopword list was developed for the Greek language, using a collection of 5,124 domain independent texts which was assembled for this purpose (Lazarinis, 2007e). A number of 32 queries were run with and without stopwords and the relevance of the results was recorded. Also the effect of lemmatization into Greek Web searching was studied (Lazarinis, 2007d).

Finally, a teaching strategy was formulated which takes into account the shortcomings of search engines related to Greek queries (Lazarinis, 2007f). This strategy aims at strengthening the searching skills of Greek users. Its effectiveness was tested with the aid of 4 student groups.

In all the experiments presented in this thesis, evaluation was completed with the aid of real users and authentic queries provided by them. In this way the research presented in this study reflects the real user needs and the results are validated by the persons who express the information need.

1.6 Commentary Outline

The remainder of this thesis is organised as follows. The focus of Chapter 2 is on text extraction from English and Greek texts. The differences between extracting data from English and non-Latin texts are discussed. Chapter 3 presents the analysis of a Greek query log. The analysis is performed with a num-

ber of criteria related to the morphology and the grammar of the queries. Chapter 4 presents and comments an evaluation methodology for identifying the capabilities and shortcoming of search engines in non-English searching. Also a study focusing in text based image retrieval is commented. Chapter 5 discusses stopword elimination and lemmatization for improving Greek Web searching. Next, in Chapter 6, an instructional methodology for teaching users how to efficiently utilize search engines is analyzed. The learning activities and aims of this teaching strategy are adapted to the shortcomings of search engines. Chapter 7 provides conclusions for my research including the summary of achievements and contributions, review of hypothesis and future research directions.

Chapter 2
Text extraction

The following sections comment and present the main findings of the work presented in (Lazarinis, 1998, 2005a, 2006). The full papers are presented in Appendices A1, A2, and A5 respectively.

2.1 Extracting data from English texts

In (Lazarinis, 1998) I present a tool which extracts specific information from English conference announcements. The tool utilized a number of heuristic rules triggered off when specific dictionary terms are identified in the textual description of the Call for Papers (CfPs). The lexicons contained the 12 month names and country names. The surrounding text was then scanned to identify specific patterns so as to eventually identify the location and date of the conference. These structured data were combined with the unstructured textual data in a retrieval system. The goal of this work was to improve the response of information retrieval systems by identifying and utilising in queries the key attributes of documents.

The effectiveness of the extraction procedure was tested with the entire collection consisting of 1927 calls for papers. A basic IR system based on the cosine measure was developed for the evaluation experiments. The final evaluation experiments showed that combining structured database entries with unstructured textual data improved the precision of the combined system against SMART (Buckley, 1985). Although SMART outperformed my IR system in

general queries, it performed significantly worse in queries related to location and dates.

2.2 Extracting data from Greek texts

Utilizing the techniques and rules presented in my previous work (Lazarinis, 1998) I developed and experimented with a tool which extracts specific information from Greek texts (Lazarinis, 2005a, 2006). This tool extracts keywords, titles, dates and locations from Greek conference announcements. Data are mined with the aid of lexicons containing rule activation terms. The success of the extraction procedure is evaluated on a document collection consisting of 145 meeting announcements.

To improve the success of the extraction process, texts were first normalized and the lexicons were augmented with all the morphological variations of month names and locations. Rule activation terms are significantly increased compared to the extraction of data from English texts. Rules are more complicated as well, since the morphology of the surrounding text varies from case to case even within the same document.

2.3 Discussion

The Greek extraction experiments showed that extraction of information in a complex non-Latin language like Greek is a more difficult task than in English because:

- Several morphological variations of data which convey the same information can be found within a document or across the document collection.

- Rules and rule activation terms need to be more complex than in English text extraction in order to improve the success of the extraction process.
- Grammatical errors and coding problems influence the extraction of data from Greek texts.
- There are no free resources available (e.g. text collections, spell checkers, dictionaries) to use and to experiment on.

The work reported in the papers which discuss text mining from Greek documents is novel because most of the text extraction experiments and techniques are developed for the English or other Latin script based natural languages. These papers try to identify some of the additional inconveniences caused in the extraction of text in a complex non-Latin script natural language.

However, this work could have been generalized by providing specific patterns of the data and by constructing a number of extensible rules which could be used as a mining engine for future experiments. Also the success of the extraction process should be evaluated in other domains as well to estimate their effectiveness.

Chapter 3
Non-latin Web queries

The following sections comment and present the main findings of the work presented in (Lazarinis, 2007a). The full paper is presented in Appendix A6.

3.1 Analysis of a Greek query log

Previous studies of query logs rely primarily on English queries and their main objective is to identify the topics that users search for and to produce various statistics about how these change or handled over time. My study (Lazarinis, 2007a) focus on the morphology and the grammar of the query terms in addition.

The query log analyzed in my study contains 5,698 query strings in Greek. The user search strings of a number of academic departments were accessible via the Web and they were statistically analyzed. The assembled data expand in a period of 12 months (November 2005-October 2006).

Queries were analyzed mainly in terms of the following six factors:

(i) Query length.

(ii) Capitalization.

(iii) Accentuation.

(iv) Lemmatized form.

(vi) Existence of stopwords.

These factors were selected because the length and the form of the query terms may influence Web searching. The main aim of this paper was to understand how Greek users form their queries and to realize whether queries of equal information weight are morphologically or grammatically differentiated.

The statistical analysis showed that the majority of queries contain 2 or 3 terms and that, although, queries appear mostly in lower case a significant number of queries are typed in upper case or in title case. Queries are usually in non lemmatized form and 26.61% of the queries contain words of low discriminatory value. Diacritics are often omitted and a number of typographic errors were identified. Further, a number of Greek queries were Latinized.

3.2 Discussion

The general conclusion of the analysis of the Greek query log is that users express the same information need in various forms, e.g. in upper or lower case or in different declensions. In other words, it was shown that queries of identical meaning differ either in morphology or in grammar. As it will be discussed in the next Chapter these subtle differences produce different results in search engines.

The paper commented in this section analyzed the Web query logs from the perspective of the morphological and grammatical characteristics of a non-Latin language. Another contribution of this work is the compilation of a Greek query log which could be used for further experiments.

This work could be extended in a number of ways. Initially, some of the issues were not extensively studied and therefore they must formally be measured. The analysis of the query log could lead to a methodology which could be applicable to other non-Latin and non-English languages.

Greek is a complex natural language with several exceptions of grammatical rules. For instance, the position of the accent mark in a word may change

its meaning. So omission of accent marks may lead to retrieval of erroneous results. A more extended linguistic analysis of the query logs is needed to realize which grammatical rules may cause vagueness in a query and eventually lead to a class of exceptions which could fine tune Web searching.

Several instances of the query log were either Latinized or a mixture of Greek and English terms. These mixed queries should be further studied to realize the need underlying this mixture and also to comprehend whether they follow specific patterns which could be exploited in Web searching. For instance, transliteration of Greek characters to Latin characters is not standardized and this leads to different Latinized versions of the same Greek terms.

Chapter 4
Evaluating non-English Web searching

The following sections comment and present the main findings of the work presented in (Lazarinis, 2005b, 2005c, 2007b, 2007c). The full papers are presented in Appendices A3, A4, A7 and A8 respectively.

4.1 An evaluation methodology

Previous studies on Web searching provide frameworks and guidelines on how to evaluate search engines. However most of the evaluation efforts focus on precision and recall neglecting other factors, such as user effort for instance or the language of the queries.

Over 60% of the online population are non-English speakers (Global Reach, 2004; Internet World Statistics, 2007) and it is probable the number of non-English speakers is growing faster than English speakers. Recent studies showed that non-English queries and unclassifiable queries have nearly tripled since 1997 (Spink et al., 2002). Most search engines were originally engineered for English. They do not take full account of inflectional semantics nor, for example, diacritics or the use of capitals. Further, it has been argued that existing search engines may not serve the needs of many non-English speaking Internet users (Chung et al., 2004).

Based on the previous observations and on my pilot studies on Greek Web searching (Lazarinis, 2005b, 2005c) in (Lazarinis, 2007b) I propose an evaluation methodology for non-English queries fitting into the multilingual and mul-

ticultural Web environment. As seen in Figure 4.1 the methodology consists of two classes of attributes: (i) interface, (ii) searching effectiveness. The criteria of the proposed assessment procedure are collected from the previous research studies and their aggregation aims at constructing a compact yet efficient model for measuring the "understanding" of international and local search engines with respect to a specific language.

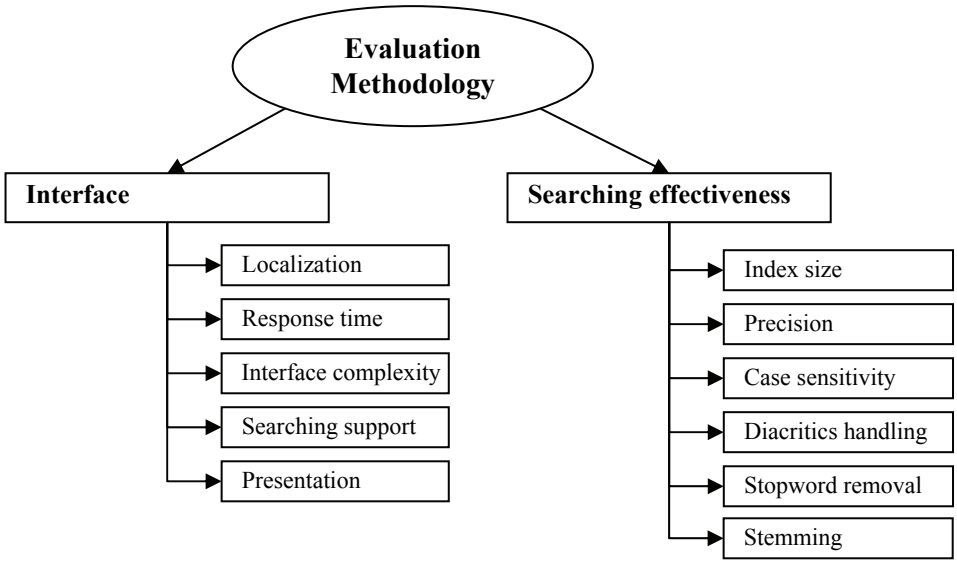


Figure 4.1 Criteria of the Web searching evaluation methodology

4.2 Evaluating Greek supporting search engines

The criteria of the proposed methodology were applied in the evaluation of Greek supporting search engines. A number of authentic Greek queries were run in Greek and in worldwide search engines. 31 users assessed these search engines with the criteria of the evaluation methodology.

The outcomes of this evaluation are:

- Some international search engines are not able to handle Greek queries at all.

- International search engines maintain richer index files than local Greek Web retrieval systems.
- Precision of international search engines is higher than the precision of local Greek search engines.
- In some case summaries were in English which deters users from visiting these Web sites.
- The localization of a search engine to other natural languages is an important factor concerning its international acceptability and usability. None of the international search engines localizes all of its services to Greek. Yahoo, MSN, AltaVista, AlltheWeb, AOL and ASK do not even localize their basic searching interface.
- International search engine do not take account of most of the grammatical and morphological idiosyncrasies of the Greek language. This leads to a differentiation of the retrieved Web pages among semantically identical queries. Queries in upper case retrieve different results than queries in lower case and queries where diacritics are omitted cannot retrieve Web pages where the query terms appear with the diacritics.
- Native Greek search engines do take account of the morphology of user queries and serve some user requests more effectively.
- Greek nouns, verbs, adjectives and even first and last names have conjugations. None of the reviewed search engines exploits this feature so as to produce better results.
- Common words affect negatively the retrieval. Since there are is no stopword list for the Greek language search engines cannot discriminate among query terms and they treat all terms as equivalent. Thus common words such as articles and prepositions influence retrieval as much as the important terms of the queries.

4.3 Text based image searching in Greek

In (Lazarinis, 2007c) I present an initial exploration of the factors which influence the retrieval of image in Greek textual queries. A number of queries are submitted to Google, Yahoo and MSN and their results are analyzed. The results of this analysis are on par with the research presented in the previous section. That is queries which differ only on their morphology but not on their content recalled different images. My study showed that features such as the filenames of images cannot be exploited since they are in Greeklish form (Greek Latinized words).

The ideas of a flexible searching tool run on top of Google were also presented. This tool is aware of some of the linguistic features of the Greek language and combines the images recalled from queries with similar content but with different morphology. The initial evaluation of the system in single word queries showed a significant increase in the number of relevant images in the top 20 ranked images.

4.4 Discussion

In this section my work on Web searching evaluation using Greek queries is presented. The presented methodology includes a number of criteria which measure the adaptability of search engines to other natural languages than English. My contribution is original because this is the first work which evaluates search engines in non-English queries using a specific methodology which expands and quantifies previous criteria on IR evaluation.

Additionally these studies are the first to discuss Greek Web searching and to consider issues related to the morphology of the queries. The queries and the evaluators were authentic and therefore the experiments assess real user needs.

My paper, regarding concept based image retrieval using Greek queries, was an initial attempt to realize the shortcomings of the image retrieval service of search engines in a non-Latin language. This study showed that the morphology of the queries, the omission of diacritics, the form (plural or singular), and the case (nominative, accusative, etc) of the queries influence retrieval of the queries. Filenames and alternative text of images cannot be utilized by search engines because they are either in English or in Greeklish.

Furthermore, the papers commented in this chapter provide examples of problematic Web queries in other non-English natural languages like German, French, Italian, Spanish, Russian and Serbian in an attempt to motivate other IR researchers.

My work can be expanded in several ways. First, more tests are needed with more Greek queries to discover new potential problems. The presented methodology needs to be applied to other natural languages to realize its applicability, its universality and its limitations. A further expansion will be possible that way as well.

The evaluation presented involved user assessments of specific search engines. Users had initially to assess the interface of the search engines and then to provide combined estimates about the precision after the application of certain information retrieval techniques, such as stopword removal or removal of diacritics. During the assessment of the interface, users were divided into groups and were sequentially shown and used the interfaces of search engines. The order of the search engines was identical in all groups. Users were initially shown the most complex interfaces and the minimalist interface of Google was the last one. During the evaluation users had to run some queries and their problems were recorded by direct observation. At the end of the experiments they had to assess and to report the problems they faced with the aid of a questionnaire. Performance on a series of tasks or a composite assessment, like one the described here, often depends on the order in which the task or subtasks are

assessed. In our experiments, the order of the tasks was the same in all evaluation tasks. Therefore it might be that in some cases the treatment and the assessment of search engines influenced the final evaluation of their interfaces. For example, as users gradually became more competent they may have considered the search engines assessed later as friendlier and easier to use than the initially assessed ones. Therefore, in the future some experiments need to be rerun with a random distribution in the order of the evaluation tasks across the evaluators, ensuring a high validity of the results.

Text based image searching in Greek queries has been studied with a limited number of queries and users. Most importantly it has not been studied with the aid of the structured methodology as in the case of text Web searching. Solutions such as extended image metadata which have been proposed in other studies (Begelman et al., 2006) need to be adapted to non-English text based image searching.

Chapter 5
Improving text based Web searching in Greek

The following sections comment and present the main findings of the work presented in (Lazarinis, 2007d, 2007e). The full papers are presented in Appendices A9 and A10 respectively.

5.1 Constructing a stopword list for Greek

As discussed in the previous Chapters, stopwords influence negatively the precision in Greek Web searching (Lazarinis, 2007b). Also the analysis of the Greek query log showed that users do include stopwords in their queries (Lazarinis, 2007a). Therefore in (Lazarinis, 2007d) I discuss the construction procedure of a stopword list for the Greek language.

For constructing the stopword list a domain independent document collection consisting of 5,124 text documents was assembled. Texts were tokenized and 77,913 unique lemmas were produced. The stopword list was constructed based on the frequency of the terms (tf) as in previous studies (Fox 1990; Savoy, 1999). The stopword list consist of the first top 99 words as after these words the frequency drops considerably and is not capable to classify a word as common.

The main conclusions of the stopword construction procedure are:

- There are no resources available for experimentation and therefore a significant effort was required to assemble the document collection.

- The tokenization procedure revealed similar problems to the text extraction experiments. That is several Greek words were either completely or partially encoded in Latin characters.
- In several occasions Latin delimiters were used to separate sentences.
- Several spelling errors were identified even in words of 3 or 4 letters.
- Morphological variations of the same terms need to harmonize to a single variation before calculation of the term frequency.
- The final stopword list has to contain all the morphological variations for each entry to reduce the required computation in future Greek IR experiments which involve usage of stopwords.

5.2 Eliminating stopwords from Greek Web queries

After the construction of the stopword list, 13 users provided 32 queries. 20 of these queries contained stopwords. These queries were run by the users with and without the stopwords in Google. The average number of relevant pages in these 20 queries increased from 4.85 to 6.30 in the first 10 highest ranked Web pages when stopwords were eliminated.

To test the significance of the previous observation we run a significance test (O'Mahony, 1986; http://www.socialresearchmethods.net/). The t-test checks whether the mean precision between the query group with stopwords and the query group without stopwords were indeed different and whether the mean precision was higher after the elimination of the stopwords. The t-test assesses whether the means of two groups are statistically different from each other.

The data used for the t-test are found in Table 2 of Appendix A10. These data sets are dependent since they regard the precision of the same queries before and after the elimination of the stopwords and therefore a paired t-test was used. The null hypothesis of the t-test was that the means of the query runs

(Group A: queries with stopwords, Group B: queries without stopwords) are equal. A risk level value (alpha value) of 0.05 was used. As shown in Table 1 the value of *t* after the test is -4.781. The highest the absolute value of t, the less similar the means of the two samples are. The probability P(T<=t) shows that by rejecting the hypothesis there is a probability of less that 0.01% of being wrong, which clearly shows that it can be concluded that the populations have different means. Also since the confidence interval for the mean does not include 0, we can be 95% confident that there is indeed a difference between the two means. The negative signs confirm that the mean of Group B is higher than the mean of Group A.

Table 4.1 Results of the t-test of the importance of stopword elimination

	Group A	Group B
Mean	4.85	6.3
Variance	8.134	9.905
Sample size	20	20
Hypothesized Mean Difference	0	
Degrees of freedom	19	
t stat	-4.781	
P(T<=t)	0.0000649	
95% confidence interval for Mean of Group A-B	-2.085 thru -0.8152	

The most important findings of these experiments which are reported in (Lazarinis, 2007d, 2007e) are:

- Users express their information needs in a natural way including articles, prepositions and other connecting words of low discriminatory value.
- Stopword elimination from Greek queries increases the number of relevant Web pages.

- All the linking terms (articles, prepositions, etc) existing in the queries are contained in the Greek stopword list. This is a positive indication about the completeness of the stopword list.

5.3 Lemmatization in Greek Web queries

In (Lazarinis, 2007e) I introduce a Greek lemmatizer which operates on nouns only and through a set of rules identifies their inflectional suffixes. Then it attempts to create the lemma of the noun. Through a set of nested if-then-else rules (see figure 5.2 for example) the longest possible inflection is identified and the word is matched to its equivalent singular nominative form. An initial estimation of 300 nouns in various forms and declensions resulted in 95.67% (287/300) success in the lemmatization procedure.

```
if term has suffix "ΑΔΕΣ" or "ΑΔΩΝ"
{
            replace suffix with "Α"
}
```

Figure 5.2 An example of a suffix replacement rule

The lemmatizer was integrated into my basic IR system (Lazarinis, 1998) and the new expanded version was tested against Google with 10 queries. These queries were run in the document collection which was constructed for the stopword list experiments.

Table 4.2 reports the results of a significance test run on the data presented in Table 1 of Appendix A9. Group A refers to the precision of non-lemmatized versions of 10 queries and group B to precision of the lemmatized versions of the query strings. Again the null hypothesis is that the two groups have equal means. The value of *a* (alpha) is set to 0.05. After the completion of the t-test the value of *t* is -3.503 (see Table 4.2). This difference is statistically significant, since the probability is 0.00669, which is much lower than the alpha value

(i.e. 0.05). The negative signs confirm that the mean of Group B is higher than the mean of Group A.

Table 4.2 Results of the t-test of the importance of lemmatization

	Group A	Group B
Mean	3.8	5,3
Variance	12.178	8,9
Sample size	10	10
Hypothesized Mean Difference	0	
Degrees of freedom	9	
t stat	-3.503	
P(T<=t) one tail	0.00669	
95% confidence interval for Mean of Group A-B	-2.469 thru -0.5314	

Summarizing the experiments it can be argued that:

- There was indeed an increase in the retrieval of relevant Web pages among the first 10 Web pages in lemmatized queries.
- In all queries my simplistic IR tool could retrieve more relevant documents than Google.
- In three queries where Google could not retrieve any documents my system retrieved 2 or 3 text files.

5.4 Synopsis

This Chapter presents a number of techniques which could be embedded in search engines so as to produce more precise results among the top ranked Web pages. An important contribution of my work is the publicly available stopword list, which was constructed based on a statistical analysis of a domain independent corpus. There is limited research on the effect of stopword elimination from non-Latin Web queries and as shown in my studies stopword

elimination and lemmatization increase the number of relevant documents. The lemmatizer, although is still under development, is a useful add on for search engines for a highly inflectional language like Greek.

Although the experiments reported in my papers showed a positive difference in relevance, these experiments should be considered as initial experimentation only. First, the evaluation experiments could be expanded with more queries and users. Secondly, the tools should be tested in various forms of searching (text web searching, image web searching, searching in e-commerce sites) to put them in an overall context and to measure their effectiveness. The presented tools should be further developed and they could be distributed as open source tools. In the stopword elimination experiments, stopwords were manually eliminated from the queries. A filter is needed which could take as input some text and produce the same text without the stopwords. These further developments could produce a number of constructive conclusions and a suite of tools for further promoting Greek IR research.

Chapter 6
Teaching the user

The following sections comment and present the main findings of the work presented in (Lazarinis, 2007f). The full paper is presented in Appendix A11.

6.1 Introduction

Locating information on the internet is an important skill in the Information Society. The previous chapters showed that searching using non-English terms is a more demanding task than searching in English. Based on these observations, my work presented in (Lazarinis, 2007f) applied the Instructional System Design (ISD) methodology to analyse, design and implement a training course for Greek users. This course aims at teaching users how to effectively use search engines and utilizes the knowledge acquired in the previous experiments to make users aware of the limitations of search engines in non-English queries.

6.2 The instructional approach

The instructional approach analyzed in (Lazarinis, 2007f) considers the explanation of the internal search engine intelligence and inefficiencies related to Greek Web searching as its basic structural element. The instructional approach was developed following the steps of the Instruction Methodology Design (ISD, 2006).

Figure 6.1 shows the learning steps, identified during the design phase, required to teach learners how to use search engines. These steps were further analyzed to specific activities. For each activity I provided a suitable example which demonstrates its main idea.

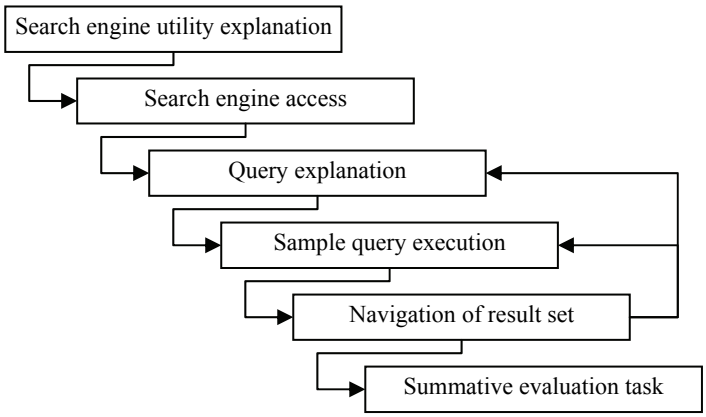


Figure 6.1 Learning steps

The efficiency of the instructional approach was tested in two groups of learners. Students who followed my disciplined teaching strategy which explains the inefficiencies of search engines in non-English queries were more successful in their searches.

During the application of the teaching methodology it was observed that:

- Users are not aware of the limitations of search engines which relate to non-English queries.
- Searching in a non-English language like Greek is a more difficult task and users need to be more creative to increase their possibility to retrieve relevant Web pages.
- Most users are not aware of the advanced searching capabilities of search engines and do not utilize them.

6.3 Discussion

This Chapter comments a methodological teaching approach for searching in-

formation on the Web. This methodology is applicable to both English and other spoken languages with inflections and intonation. Although Web searching is a common everyday activity of Web surfers the bibliography on how to teach the user to be an effective searcher is limited. My paper is the first attempt to put the instruction of basic Web searching skills to a specific context. This framework is adapted to the limitations of search engines to non-English queries. Its modular structure allows for easy upgrade.

The methodology needs to be applied to other non-English and non-Latin languages to test its applicability and to possibly expand it. Some limitations of this work is that it has not been applied to an augmented set of student groups and that it does not consider at all the advanced options of the search engines or the other options (e.g. image search, video search, etc). If these limitations are overcome in future versions of the methodology then a robust and holistic method will be produced.

Chapter 7
Conclusions

7.1 Introduction

This thesis addressed the problem of Web searching in Greek queries. This Chapter concludes the whole work that has been discussed in the thesis. The findings and the hypothesis are reviewed and the contributions of this thesis are discussed. Finally, future research ideas are presented.

7.2 Review of findings

The overall finding of this project is centred on the capabilities of search engines in non-English queries and more specifically in Greek queries. The information extraction studies presented in Chapter 2 showed that text processing in a non-Latin language poses additional difficulties originating from character encoding and from the differences in semantically related terms (Lazarinis, 2005a, 2006). The morphological analysis of the Greek query log presented in Chapter 3 showed that users usually type their queries in lower case mode (Lazarinis, 2007a). However, a significant number of queries were in upper case mode or without diacritics.

The application of the evaluation methodology, which was presented in Chapter 4, showed that most of the international search engines do not take account of the morphology of non-English queries and thus they retrieve different Web pages in semantically identical queries (Lazarinis, 2005b, 2005c, 2007b). Also none of the international search engines, e.g. Google and Yahoo, adapts

all of its services to Greek, which obstructs users from taking full advantage of the existing services. The evaluation showed that local search engines are weaker in terms of efficiency, speed and index composition than the worldwide search engines. However, they distinguish their results according to morphology of the query terms. Native Greek search engines do take account of the morphology of user queries and serve some user requests more effectively. The same conclusions apply in image Web retrieval (Lazarinis, 2007c).

In Chapter 5 the effect of stopword elimination and lemmatization of query terms was examined (Lazarinis, 2007d, 2007e). An increase in precision is reported in all these techniques. The process of constructing a stopword list is also presented.

Finally, in Chapter 6 a structured teaching approach is presented and evaluated. This teaching strategy aims at methodologically instruct students on how to use search engines and how to revise their queries based on the limitations of search engines. The student groups who were instructed using this approach were more successful in their Web searches (Lazarinis, 2007f).

7.3 Review of hypothesis

The hypothesis of this work was that the international search engines do not effectively support Greek queries and therefore a disciplined evaluation methodology was needed so as to provide information about the shortcomings of the existing search engines with respect to a specific natural language. Additionally, the work claimed that the application of basic IR techniques, such as stopword removal, in non-English and non-Latin searching could improve the effectiveness of search engines.

The hypothesis has been proven through a series of experiments. In Chapters 2 and 3 it was shown that the extraction of specific information from Greek texts and the formation of Greek queries depend on the morphology of Greek

words which is also important when searching the Web (Lazarinis, 2005a, 2006, 2007a). The evaluation of the features of Web search engines through the structured sequence of criteria and the evaluation of image searching using textual queries presented in Chapter 4 demonstrated the limitations of search engines in Greek queries (Lazarinis, 2005b, 2005c, 2007b).

Also in Chapter 4 it was briefly discussed that mixing the results of queries which differ in morphology increases the possibility of retrieving more relevant images in textual queries (Lazarinis, 2007c). Stopword elimination and lemmatization experiments presented in Chapter 5 confirmed that these two techniques increase precision in Greek Web searching (Lazarinis, 2007d, 2007e). These experiments are solid indications that standard information retrieval techniques could improve the capabilities of search engines in non-English queries. The instructional approach presented in Chapter 6 illustrated that indeed users are not aware of the limitations of search engines in Greek queries and that searching in Greek requires additional effort so as to retrieve more relevant results (Lazarinis, 2007f).

7.4 Thesis contributions

This thesis contributes to the literature on evaluation of search engines and on the analysis of query logs. The presented evaluation methodology quantifies and assembles previous evaluation criteria used in Web searching evaluation studies. The evaluation performed with the aid of real users and authentic queries. The evaluation methodology proposed in this work takes into account non-English users and deciphers a number of abstract evaluation criteria proposed in previous research.

Previous studies acknowledged the importance of studying query logs. However, these studies focused on the topics users search for and the length of

the queries and on how these topics changed over time. In this thesis user queries are additionally examined morphologically.

Finally, the construction of a stopword list for Greek and the formation of an instructional methodology for teaching users how to effectively search the Web have not been discussed previously.

7.5 Discussion and Future work

This thesis focuses primarily on understanding the problems in Greek Web searching and on proposing and testing specific techniques for improving the effectiveness of search engines. As discussed above, the aims of the current research, as set in Chapter 1, have been achieved. Nevertheless, a number of issues arise from the information presented in the previous chapters though which could be utilized in further extending the reported research.

The thesis includes a number of experiments with human subjects. The persons who willingly participated to the experiments had to assess the interface of search engines and the improvement in the accuracy of the returned results after the application of specific techniques. As discussed in section 4.4 of chapter 4, the order of the evaluation tasks was identical. Similar approaches applied in the stopword removal experiments and in the teaching experiments. Order effects can confound experiment results when the order is the same or the different orders are systematically associated with particular conditions. This practice may cause what is known as experimenter's bias (http://en.wikipedia.org/wiki/Experimenter's_bias). That is the outcome of the experiment may be biased towards a result expected by the human experimenter. This issue should be addressed to future experimentation with more queries and users and a random order of the evaluation tasks.

Further, at the end of each of Chapters 2 to 6, a number of suggestions have been proposed which could expand the current research. Summarizing these suggestions, the following issues can be considered for further work:

- Apply the proposed methodologies and techniques to other non-English languages and especially in non-Latin natural languages to identify the searching problems in these languages and also to inspect the applicability of the presented techniques and to propose a set of possible extensions.
- Build a model for evaluating query logs as the studies which analyze Web query logs use a number of criteria which are not uniform across them.
- More thorough examination of the grammatical exceptions of the Greek language is needed to see if such exceptions apply in Web searching and if they are detected in queries. These exceptions may cause searches to fail by producing entirely erroneous results.
- The searching behaviour and the user needs of non-English users should be further examined as they tend to either run Latinized queries or mixtures of Greek and English terms, for example. The reformulation approaches applied by the users in these cases should be further studied in order to understand the users and their needs.
- The effectiveness of summaries provided by search engines should be evaluated and effective summarization techniques should be developed and adapted to the characteristics of non-English languages as it was observed that some summaries were in English or in Greeklish (i.e. Greek words written in Latin scripts).

References

Begelman, G., Keller, P., Smadja, F. (2006). Automated tag clustering: improving searching and exploration in the tag space. *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*. Available at http://www.pui.ch/phred/automated_tag_clustering/ (accessed 1 June 2007).

Berners Lee, T., Caillau, R., Groff, J., Pollermann B. (1992). World Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, 2 (1), 52-58.

Buckley, C. (1985). Implementation of the SMART information retrieval system. *Technical Report TR85-686*, Cornell University. Available at ftp://ftp.cs.cornell.edu/pub/smart/ (accessed 1 October 2006).

Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T., Chen, H., (2004). Internet searching and browsing in a multilingual world: an experiment on the Chinese business intelligence portal (CBizPort). *Journal of the American Society for Information Science and Technology*, 55(9), 818–831.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

Global Reach (2004). *Global internet statistics (by language)*, Available at: www.global-reach.biz/globstats (accessed 31 July 2006).

Internet World Statistics, (2007). *Internet world users by language* (19/Mar/07). Retrieved May 15, 2007, available at http://www.internet worldstats. com/stats7.htm (accessed 19 March 2007).

ISD (2006). *Instructional System Design*. Available at: http://www.nwlink.com/~donclark/hrd/sat.html (accessed 31 July 2006).

Lazarinis, F. (1998). Combining information retrieval with information extrac-

tion. *20th Annual BCS-IRSG Colloquium on IR*, Autrans, France, pp. 162-174 http://www.bcs.org/server.php?show=ConWebDoc.4410.

Lazarinis, F. (2005a). A rule based tool for mining textual data from Greek calls for papers. *2nd Balkan Conference in Informatics*, Ohrid, Skopje, pp.126-133.

Lazarinis, F. (2005b). Do search engines understand Greek or user requests "sound Greek" to them? *Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology*, Compiegne France, pp. 43-46.

Lazarinis, F. (2005c). Evaluating user effort in Greek Web searching. *10th Pan-Hellenic Conference in Informatics*, University of Thessaly, Volos, Greece, pp. 99-109, ISBN: 960-8029-39-2.

Lazarinis, F. (2006). Automatic extraction of knowledge from Greek Web documents. *6th Dutch-Belgian Workshop in Information Retrieval (DIR 06)*, Delft, the Netherlands, pp. 33-37.

Lazarinis, F. (2007a). How do Greek searchers form their Web queries? *3rd International Conference on Web Information Systems and Technologies, (WEBIST)*, Barcelona, Spain, pp. 404-407.

Lazarinis, F. (2007b). Web retrieval systems and the Greek language: Do they have an understanding? *Journal of Information Science*, 33(5), 622-636, doi:10.1177/016555150607639

Lazarinis, F. (2007c). An initial exploration of the factors influencing retrieval of Web images in Greek queries*, Euro American Conference on Telematics and Information Systems (EATIS 07), ACM Digital Library*, doi: doi.acm.org/10.1145/1352694.1352765

Lazarinis, F. (2007d). Lemmatization and stopword elimination in Greek Web

searching. *Euro American Conference on Telematics and Information Systems (EATIS 07), ACM Digital Library*, doi: doi.acm.org/ 10.1145/1352694.1352757

Lazarinis, F. (2007e). Engineering and utilizing a stopword list in Greek Web retrieval. *Journal of the American Society for Information Science and Technology*, 58(11), pp. 1645-1652, doi: 10.1002/asi.20648

Lazarinis, F. (2007f). Forming an instructional approach for teaching Web searching to non-English users. *Program: Electronic Library and Information Systems*, 41(2), pp. 170-179, doi: 10.1108/00330330710742935

O'Mahony, M. (1986). *Sensory Evaluation of Food: Statistical Methods and Procedures*. CRC Press, 487. ISBN 0-824-77337-3.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944 – 952.

Spink, A., Jansen, B. J., Wolfram, D., Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. IEEE Computer, 35(3), 107–111.

Triantafyllidis, M. (1941). *Modern Greek Grammar*. Institute M Triantalyllidis (in Greek).

Vilares Ferro, M., Graña Gil J., Alvariño Alvariño, P. (1997). Finite-state morphology and formal verification. *Journal of Natural Language Engineering*, special issue on Extended Finite State Models of Language, 3(4), 303-304.

Appendices

Papers A1-A11 for the PhD by published works

- Appendix A1 -

Combining Information Retrieval with Information Extraction for Efficient Retrieval of Calls for Papers

Fotis Lazarinis

Department of Computing Science, University of Glasgow
Glasgow, Scotland

Abstract

In many domains there are specific attributes in documents that carry more weight than the general words in the document. This paper proposes the use of information extraction techniques in order to identify these attributes for the domain of calls for papers. The utilisation of attributes into queries imposes new requirements on the retrieval method of conventional information retrieval systems. A new model for estimating the relevance of documents to user requests is also presented. The effectiveness of this model and the benefits of integrating information extraction with information retrieval are shown by comparing our system with a typical information retrieval system. The results show a precision increase of between 45% and 60% of all recall points.

1 Introduction

Information retrieval (IR) systems, also called text retrieval systems, facilitate users to retrieve information which is relevant or close to their information needs.

Even though specific words may be key attributes of a domain, conventional IR systems process them as ordinary terms using general statistical methods [18, 26]. Usually such terms appear several times in a document collection and so they lose their power to discriminate among documents. However, a term may appear several times in a document collection but with different significance each time. In calls for papers (CFPs), for example, there exist some past dates along with the conference's date. When users pose queries about conferences held in a specific month all the calls for papers where the specific month name appears are retrieved even though most of them are irrelevant. Another problem of the conventional approach is caused by the fact that in collections about specific subjects, synonyms and/or abbreviations are often encountered. Traditional IR systems treat the variations of a term as different terms. Stemming algorithms [10] attempt to partly solve the problem with variations but they cannot effectively cope with synonyms and abbreviations. This affects the retrieval and requires either the integration of a thesaurus [22] or users to specify all the alternative forms in their query if they wish to retrieve all the relevant documents. As we will see in section 3 this is not a problem in our system because we implicitly use a thesaurus for the important terms of our domain. The last problem of typical text retrieval systems is that they consider two terms to be equally important if they exist the same times in a document or in a document set. For example, imagine a document collection of medical case records. Certain disease names will be treated equally in the retrieval with other words that are not important simply because their frequencies of occurrence are equal.

These problems seriously affect the retrieval in collections about specific subjects such as medical cases or financial news where the important terms are encountered several times. The solution to the above problems proposed in this paper is to employ information extraction (IE) [6] techniques in order to identify the useful information that would lose its significance if it was processed by a standard text retrieval system. Since IE is a highly domain-dependent task we concentrate on calls for conference papers and our aim is to automatically identify conferences' date and location. In calls for papers there exist many other locations and dates in addition to the conference's date and location. With a typical IR system when a user wishes to retrieve meeting announcements held in a specific place or in a specific date all the CFPs that contain the user specified data will be retrieved since the IR system cannot distinguish among the appearing dates and locations. As soon as the attributes, i.e. location and date, have been identified the rest of a CFP's content is processed using standard IR techniques.

After this processing, documents are represented by a set of keywords (index terms) describing the subject of a document and a set of solid attributes, i.e. date and location in our system. The most important issue arising in this case concerns the computation of the similarity between documents and queries. In typical information retrieval systems the similarity between documents and queries is based either on the probabilistic model [26] or on the vector space model [19]. With the incorporation of attributes the direct use of these models is not suitable, at least for queries based partly on attributes. A general model for estimating the relevance between queries based on

attributes and documents and a model for mixing the results of queries based on both content (index terms) and attributes are presented and analysed later on the paper.

The rest of the paper explains the algorithms employed in CIERS (Combined Information Extraction and Retrieval System). CIERS is an information retrieval system that combines the strengths of IR and IE. Figure 1 shows a prototype user interface in Java for CIERS (http://www.dcs.gla.ac.uk/~lazarinf/project.html).

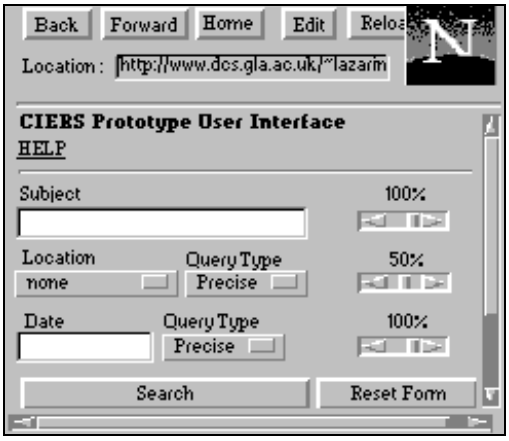


Figure 1: CIERS Prototype User Interface

The standard techniques employed in IR engines and thus in the standard IR module of CIERS are described briefly in the next section. Section 3 presents information extraction and the rules that achieve the automatic identification of location and date from meeting announcements. Section 4 describes the model on which the similarity computation between documents and queries is based. Finally, we present the results of experiments that show that information extraction techniques can benefit information retrieval.

2 Information Retrieval Engines

The basic operations of typical information retrieval systems can be grouped into two main categories: indexing and matching (or retrieval). The purpose of the indexing process is to identify the most descriptive words existing in a text. After the elimination of the stopwords and the identification of the unique stems of the remaining words, the term frequency (tf) and the inverse document frequency (idf) of each unique stem are calculated. Each document is then described by a set of keywords along with their tf and idf [9].

The aim of the query matching process is to derive a list of documents ranked in decreasing order of relevance to a given query. When a query based on content (expressed in natural language) is submitted to the system, it undergoes a process similar to the indexing process. Now both documents and queries can be represented as weighted vectors where each term's weight is usually a combination of tf and idf. In this case the similarity between documents and queries is based on the vector space model [19]. In this model documents and queries are viewed as n-dimensional vectors, where n corresponds to the number of unique index terms. The similarity between query q and document d is computed by measuring the cosine of the angle between their vectors (figure 2).

$$\text{Similarity}(q,d) = \frac{\sum_{i=1}^{n} Wiq * Wid}{\sqrt{\sum_{i=1}^{n} Wiq^2 * \sum_{i=1}^{n} Wid^2}}$$

where

w_{iq}, the weight of query term i

w_{id}, the weight of document term i

n, the total number of terms

Figure 2: Cosine similarity measure

3 Information Extraction

A relatively new and increasingly important area in text processing is information extraction [6]. Information extraction aims at identifying special kind of data from domain-specific document collections. IE systems process documents trying to identify pre-defined entities and the relationships between them, filling a structured template with the extracted information. Hence, an IE system can be considered as converting some elements of unstructured text documents into structured database entries.

3.1 Related Work

Information extraction systems have been employed in the summarisation of medical case records by extracting diagnoses, test results, and treatments [17]. Postma et al. [14], and Chowdhury and Lynch [2] have used chemistry papers to extract data such as names, and scientific terms. In general technical reports are perfect candidates for information extraction because they contain data that have standard forms, e.g. references. Business IE systems extract details about companies, products, and services and other details of interest to businesspersons, e.g. in the message understanding domain (MUC) [23].

These examples are standalone IE systems and cover only special cases of text processing. Although information retrieval and information extraction are complementary there has been little work aimed at integrating the two areas. The most notable work is that of Gaizauskas and Robertson [11]. In their work they used the output of Excite [7] as input to an IE system, called VIE (Vanilla IE System). Their domain was management succession events and their scenario was designed to track changes in company management. The results of Excite searches were passed to VIE which produced a template filled with the company's name, the old manager's name, the new manager's name, etc. A natural language summary was also produced for the retrieved documents by populating the empty fields of a fixed-structure summary.

Since the purpose of Gaizauskas and Robertson was to create a system that would construct a structured data resource from free text they evaluated only the success of the information extraction procedure. Whereas we also evaluate the extraction procedure in order to measure its effectiveness we are more interested in the performance of the combined system. Our goal is to improve the efficiency of conventional IR systems, at least in some special cases. Therefore, we evaluate CIERS using the standard IR method and we compare it with a typical text retrieval system. Before we report the results of the evaluation of the combined system we need to explain the location and date extraction procedure and the model on which CIERS is based.

3.2 Extraction of Attributes

As many other applications of natural language processing information extraction systems rely on domain-specific dictionaries to extract specific kind of information from free text [3]. Such dictionaries contain lexical items that enable the recognition of the desired entities. For instance, if we are interested in English full names the dictionary must consist of all the first English names.

These domain-specific dictionaries contain only the minimal necessary information for the extraction procedure. When an item in the text is matched with an item in the dictionary a rule is activated which enables the extraction of the desired information. To continue the last example, when a word is found to be a proper first name then a simple rule like "the next word is a potential surname" may be activated.

Unfortunately, simple rules rarely have high success rates and complex rules (or heuristics) are often needed. This need arises from the fact that the same kind of data exhibit considerable variation in both the information they carry and in the way they are presented. For instance, the name of a person may appear in several different forms, e.g. "Peter Smith", or "Peter M. Smith", or "Smith Peter". In addition, the information may be scattered across several sentences. Finally, several instances of the same type of information may appear in the same text, e.g. many names may exist in a text in addition to the one of interest.

In order to construct rules that will enable the successful extraction of the desired facts, one has to examine thoroughly a representative sample of documents of her/his domain. This will also allow the accurate construction of the dictionary. As already mentioned, our aim was to automatically extract conferences' location and date from meeting announcements. Therefore, we examined 250 CFPs, a small part of our document collection consisting of 1927 meeting announcements[1]. This analysis allowed us to realise the different patterns of date and location and

[1] The CFPs collection can be found at http://www.dcs.gla.ac.uk/~lazarinf/CFPcoll.html

construct the extraction rules. The following two sections analyse briefly the location and date extraction procedure[2] respectively and section 3.2.3 presents the results of the evaluation of the IE module.

3.2.1 Location Extraction Procedure

In almost every of the 250 CFPs we examined, the conference's city and country was named while the continent was cited only in few announcements. Time limitations prevented the identification of the city because the required set of rules would be rather complicated and hence it was decided to detect only the country of each conference. Nevertheless, when a country of a conference is extracted it can be easily connected to its continent as we will see below. The second conclusion reached was that more than 50% of the conferences were held in USA. Almost all of these CFPs mentioned the state of the conference. Therefore, in order to offer users a wider choice of queries it was decided to extract the state name as well for conferences held in USA. Hence, CIERS identifies US states and countries for the rest of the world. In other words US states are treated as ordinary countries and USA as a continent.

In order to recognise country and state names the dictionary should contain all the formal country names, e.g. "Greece", and state names such as "California". Additionally, all the variations of a country's name, e.g. "Hellas" for "Greece", and all the state codes, e.g. "CA" for "California", should be incorporated into the dictionary because they are used very frequently to indicate a conference's location.

As previously explained when a word of a document is matched with a dictionary entry a rule is activated. However, this cannot work with country names consisting of more than one words because it is impossible to automatically decide which words probably constitute a country and search them in the dictionary as one text element. As a result of this, apart from the proper country names and their variations the dictionary contains the rarer of the words making a country name (the rarer word is used to minimise the activation of rules), e.g. "Kingdom" for "United Kingdom".

In order to associate the location dictionary entries we add an attribute, named country type, to the dictionary. If an entry is a full country name then the country type's value is the country's continent. If it is a variation or a state code then the country type points to the proper country or state name. Finally, if an entry is a part of a country's name then the country type shows how many words before or after this part are needed in order to constitute a proper name.

The last conclusion reached from the examination of the 250 CFPs was that although several countries or US states may be mentioned in a CFP, in nearly all the cases the state or country that appears first is the conference's location.

So, if the matching term is a formal country name then the identification procedure ends successfully. If it is a variation or a state code is mapped to the formal country name and again the conference's country has been extracted. When the processed term is part of the name of a country the necessary previous or next words are taken and the new potential country name is searched in the dictionary. If it is found in the dictionary then the country name has been identified; otherwise the extraction procedure ends. Finally, if a proper country name has been detected it is connected to its continent via its country type value.

The above set of rules is only a subset of the actual rules employed in the implementation of CIERS. Space limitations prohibit us from explaining the rest of the rules that cover special cases such as collisions of state codes with stopwords, e.g. "IN" is both a stopword and the state code for "INDIANA". Even in that case the above description verifies that an IE system cannot be used in any document collection but only in some specific domain because the rules depend on the characteristics of the collection.

3.2.2 Date Extraction Procedure

Again the first step in the identification of a conference's date is the construction of a suitable dictionary containing the necessary terms that will activate the rules for the extraction procedure. The date dictionary is less populated than the location dictionary as it must contain only the 12 full month names and their 11 abbreviations ("May" is both the month's full name and the contraction).

The second observation made after the analysis of the 250 CFPs is that the latest date existing in a call for papers is usually the conference's date. Unfortunately in a very small percentage of calls for papers, some future dates announcing future meetings appear. But this problem does not significantly affect the identification procedure as it typically leads to a minor error (table 1).

[2] For a full description of the extraction procedure please consult [12].

Whenever a month is found in the text, CIERS first checks the succeeding words until the end of the sentence and then the preceding words until the beginning of the sentence. This search aims at identifying the day and the year of the conference. If a number from 1 to 31 or a word that starts with a number from 1 to 31 and ends in “st”, “nd”, “rd”, “th”, e.g. 1st, 2nd, 3rd, 4th, is found then this is the day of the conference. If a four-digit number is found which starts either with 19 or 20 then it is the desired year. As soon as a date is identified it is compared with the previous extracted one, if any, and the latest one is kept.

The description of the rules employed in the extraction of dates (again we omitted the description of the specialised rules that handle month names such as “may” which is both a month and an English modal verb) and locations leads us to some important conclusions. First, the context surrounding the activation terms is really important and is this that actually allows a rule to succeed. Second, the rules are complicated even if the desired information is simple and its alternative forms are limited. Moreover, the extraction of knowledge can only be based on heuristics because they depend entirely on the characteristics of the extracted entities and the context in which it appears. Finally, the dictionaries that contain the activation terms can be used as thesauri and can be utilised in queries. For example, by consulting the dictionaries user queries can be expanded to include all the variations of a term, thus retrieving more relevant documents.

3.2.3 Evaluation of the IE Module

Although regularly used in the evaluation of IR systems, the performance of an IE system can be measured using *Precision (P)* and *Recall (R)* [15, 16]. Precision measures the ratio of the correctly extracted information against all the extracted information. Recall measures the ratio of the correct information extracted from the texts against all the available information.

The effectiveness of the extraction procedure was tested with the entire collection consisting of 1927 calls for papers. These CFPs were gathered from the Internet from various archives and contain announcements for conferences mainly covering various fields of computer science. Also a significant portion of this collection is made up of psychology, engineering, and physics conference announcements.

Despite the diversity of the collection the system works extremely well and the employed rules achieve high rates of precision and recall. The results are summarised in the next table.

	Correctly extracted	**Erroneously extracted**	**Not extracted**	**Precision**	**Recall**
Country	1796	112	19	94.12%	93.20%
Date	1881	41	5	97.86%	97.61%

Table 1: Precision and Recall for the attribute extraction procedure

This high accuracy will eventually result in improved performance of the combined system over a typical IR system where queries based on attributes are expressed as ordinary queries based on content.

Before moving on to the next section two remarks should be made about the occasional failure of CIERS to detect date and location. Sometimes the system fails because the date and location do not appear in a call for papers. Whereas the system is not responsible for this failure, in the evaluation we accounted it to CIERS and thus we got slightly worse precision and recall values. Furthermore, a failure analysis should have followed the testing of the IE module. This analysis would have helped us to realise the possible sources of error and modify the set of the employed rules. However, time limitations prevented the analysis of the erroneous instances in both cases.

4 Combined System

CIERS supports two types of queries: attribute and content queries. Users are able to ask either individual content or attribute queries or any combination of them. The first issue arising is how a single estimate of relevance is derived in any combination. A possible solution is to use data fusion techniques [8, 24, 27] and to merge the ranked lists for the different types of queries. In the next section we define a model for merging the results of the different types of queries. The subsequent issue is how a list of relevant documents in attribute queries is obtained. A model for computing the relevance or closeness of documents to attribute queries is proposed below.

4.1 Merging of Results

Data fusion is a technique used for combining the results of different retrieval strategies into one unified output. In traditional IR systems data fusion is used for improving performance by allowing each strategy's relevance estimate to contribute to the final result. The combined list is typically more accurate than any of the individual results.

Fusion of results takes two different forms. Data fusion aims at merging the results of different strategies for a given query on a single document set. In its second form, known as collection fusion [27], the goal is to combine the retrieval results from multiple, independent collections into a single result such that the performance of the combined system will exceed the performance of searching the entire collection as a single collection.

Our work is a combination of the two different forms. CIERS uses three different data sets, i.e. index terms, date, and location attributes. Also different strategies are used for each query type. Before we proceed in the explanation of how the combination of the output for the different kinds of queries is achieved, it is worth mentioning some work done in the area of data fusion.

A number of different methods for combining the results of different strategy implementations of the same query have been proposed. Fox et al. [8] determine documents' overall relevance by adopting the maximum score for each document of all the strategy outputs. Thompson [24] combines the results of the different methods based on the performance level of each method. That is, each method is evaluated independently and its performance level is measured. Then the methods are combined in proportion to their performance level. Both of these approaches are acceptable in a single query and a single set of data because the aim is to enable the best strategy to affect most the retrieval. Nevertheless, in our case this is not desirable as it will result in retrieval biased against one query type.

Similarly to the approaches described above we use a linear combination of the output lists. That is, the overall relevance estimate is the sum of scaled estimates of the individual queries (figure 3).

$$S(q,d) = \Theta_1 S_1(q_1,d) + \Theta_2 S_2(q_2,d) + \Theta_3 S_3(q_3,d)$$

Figure 3: Parameterised mixture of the individual relevance estimates

$S(q,d)$ is the combined estimate for the combined query q and document d. $S_i(q_i,d)$ [3] is the similarity between the subpart q_i of the original query and document d[4]. Θ_i are free parameters in the model set by users, granting them with full flexibility over the retrieval (as shown in figure 1). Θ_i is the scale of the query subpart q_i; Θ_1 is the scale of the query about subject, Θ_2 of the query about the location[5], and Θ_3 of the query about the date. So, if only the subject and the date are of interest then Θ_1 and Θ_3 will be 1 (100%) and Θ_2 will be 0 (0%). That way, searchers get the combined estimate of only these two subqueries. Furthermore, the last equation allows users to specify their rate of interest in each subquery. For instance, if the subject and the date of conferences are essential and location is less important then users can define Θ_1 and Θ_3 to be 1 (100%) and Θ_2 to be 0.5 (50%) or less.

At this point it is necessary to underline that the results of the individual queries must be consistent so before the merging of the individual estimates each list is divided by its maximum score. That way the partial scores lie between 0 and 1 and in the merging of the results each counts as much as its scale, i.e. Θ_i, defines.

4.2 Semantics of Attribute Query Matching

As explained sometimes CIERS fails to identify a conference's date or location so the attribute values will be missing. This fact originates partially from the occasional failure of the system to identify the attributes and partially from the fact that in some conference announcements (usually preliminary) the date and/or location are not cited.

The missing values of attributes could be denoted with the special value "null". Null values have been used extensively in databases to denote that a value is missing [4, 5, 25]. However, indicating a missing value with null is not adequate because no distinction is made between missing and non-applicable attribute values.

[3] $S_1(q,d)$ is the estimate for the content query and is based on the Vector Space model (section 2).

[4] The equation of figure 3 can be extended if more attributes are incorporated by adding the scaled estimates of the strategies based on the new attributes.

[5] The location of a conference may be its country or its continent.

A more suitable solution for differentiating the two cases is to use two special values, Not-Known (NK) and Not-Applicable (NA) [13]. Not-known represents attribute values that the system fails to identify and not-applicable denotes attribute values that are not defined in a CFP. In addition to the problem of missing values another issue stems from the fact that usually conferences last more than one day. This means that the date attribute cannot actually be represented by a single value but by a range of values denoting the conference's duration. Morrissey [13] uses a special value named p-range. A p-range is denoted with a lower and upper limit, indicating the limits of the range of values. An example of p-range would be [19/7/97-25/7/97] meaning that a conference starts on 19 July 1997 and ends on 25 July 1997. The current implementation supports only the NK special value. The other special values are included in the model for future expandability, e.g. calls for journals where location is not applicable could be processed by the system. Below the system, the queries, and the similarity computation are formally described.

4.2.1 System

D: the finite set of all stored documents. An individual document is denoted by d_i.

A: the finite set of all attributes. a_i represents a specific attribute.

V: the set of all possible different value sets. A specific attribute value set is denoted by V_{a_i} and the value for a specific attribute is denoted by u_i.

F: the set of all functions that map documents to attribute values. For reasons of convenience a function is denoted as the attribute $a_i(d) = u_i$, e.g. country(d) = UK.

4.2.2 Queries

Since information retrieval aims at identifying those objects that are relevant or close to a user's needs our model supports two different types of attribute queries, namely *Precise* and *Close*.

In precise attribute queries users are interested in documents that definitely satisfy their needs. In close queries searchers are additionally interested in conferences that are close to their information need, so in their requests they specify the desired value for an attribute and the tolerance for the values of that attribute. For example, one may be quite interested in conferences held on 15 July 1997 but she/he may be also interested in conferences held a few days before or after the specified date. This type of attribute queries would be also useful in queries about countries. If the location dictionary is extended to include the concept of neighbouring countries the system will be able to retrieve conferences held close to the original place. Users simply have to specify how far from their original goal they are willing to deviate. The current implementation supports only the concept of continents for grouping countries in a geographic area. Consequently, all the CFPs of conferences held in countries of the same continent will be retrieved but ranked lower than conferences held in the user specified country.

Formally an attribute query is denoted as $Q_{u_i}^{a_i}$, which means that stored documents must have value u_i for the attribute a_i in order to satisfy the query. There are three sets of documents that match attribute queries; those that exactly (are known to the system to) match a query denoted as $K_{u_i}^{a_i}$, those that possibly satisfy a query denoted as $P_{u_i}^{a_i}$, and those that are close to the initial request represented as $C_{u_i,t}^{a_i}$. Below each set of relevant documents is defined formally.

$$K_{u_i}^{a_i} = \bigcup \{d : a_i(d) = u_i\}$$

$$P_{u_i}^{a_i} = \bigcup \{d : a_i(d) = \text{NK or } a_i(d) = \text{NA}\}$$

$$C_{u_i,t}^{a_i} = \bigcup \{d : a_i(d) = u'_i, \text{where } dist(u_i, u'_i) \leq t\}$$

where

$d \in D$, $a_i \in A_i$, u_i and $u'_i \in V_{a_i}$

t, specifies the limit of the range of values
a user considers close to her/his query

Figure 4: Formal definition of the sets satisfying attribute queries

u'_i is a value for the attribute a_i for which the distance (dist) between it and the query specified value is less or equal than the specified tolerance (t).

Precise attribute queries are satisfied by $K_{u_i}^{a_i} \bigcup P_{u_i}^{a_i}$ which means that those documents that have value u_i for the attribute a_i and those for which the value of the specified attribute is missing, (possibly) satisfy the query. For example the $K_{\text{July 1997}}^{\text{date}}$ contains documents like date(d1) = 15/7/1997 or date(d2) = [12/17/1997 – 18/7/1997] and the $P_{\text{July 1997}}^{\text{date}}$ is consisted of documents such as date(d3) = NK or date(d4) = NA.

The set of documents that satisfy close attribute queries is the union of all the three discrete sets described above. For instance, a user may be particularly interested in conferences held on 15 July 1998 but she/he would be interested in meetings that happen 15 days before or after that date. Examples of documents belonging to $C_{\text{15 July 1998, 15}}^{\text{date}}$ are: date(d1) = 12/7/1998 or date(d2) = 30/7/98 or date(d3) = [18/7/1998-19/7/1998].

In any of the previous cases, documents that definitely satisfy the query should be ranked first, those that are close (in close queries) should be ranked second, and last should be ranked those that possibly satisfy a query. The ranking of documents close to the initial request should depend on how close a document is to the original query. In the last example, CFPs for conferences held 5 days after the specified date should be ranked before conference announcements held 10 days later. Additionally, the ranking of documents close to a request should depend on the number of different possible close values. For example, if a user poses a query about country then close CFPs should be ranked higher if the close countries are 5 than when the close countries are 10 because in the second case the deviation from the initial request is greater.

4.2.3 Query Matching

Similarity between attribute queries and documents is computed by the equation of figure 5. This equation is based on entropy [20, 21]. Entropy calculates the uncertainty associated with the decision to retrieve an object that satisfies a query. The uncertainty is inversely proportional to the information the system has for an object. Since we are interested in finding how close a document is to a given query, entropy has been modified and our score is maximum when a document matches the attribute query and minimum when a document possibly matches it.

$$\text{Similarity}(q,d) = 1 - \frac{-\sum_{k=1}^{m} P_k * \log_2 P_k}{\log_2 n}$$

where

n, the number of all different values of the query specified attribute

m, the number of attribute values that match the query

P_k, the probability of a document to match the query

Figure 5: Similarity score based on entropy

The last equation depends on the set of documents that match a request and the number of attribute values that match it. In $K_{u_i}^{a_i}$ m is 1, since only one value for the specified attribute satisfies the query. In $C_{u_i,t}^{a_i}$ m is equal to the number of values that match a query or in other words to the distance between a close attribute value and the query specified one. For example, in a query about date, m is 5 for conferences held five days later than the desired date and 10 for conferences held ten days later. In the last set of documents, $P_{u_i}^{a_i}$, since a_i may have any of the n different values m equals n.

For all the three document sets P_k is the probability of a document to correspond to a query, i.e. to have such an attribute value that matches the query. While P_k varies for different attribute values and should depend on the error in the extraction of attributes, for simplifying its calculation it is assumed that every document of a particular set has the same chance to match a query and it is defined to be 1/m. Let us now illustrate the use of the last equation with an example from our document collection[6].

[6] The number of different countries in our location dictionary is 254.

Query: "List all calls for papers of conferences held in California".

For documents of $K^{country}_{California}$: $Similarity(q,d) = 1 - \frac{-\sum_{k=1}^{1} \frac{1}{1} * \log_2 \frac{1}{1}}{\log_2 254} = 1$

For close documents (51 close states): $Similarity(q,d) = 1 - \frac{-\sum_{k=1}^{51} \frac{1}{51} * \log_2 \frac{1}{51}}{\log_2 254} = 0.28$

For $P^{country}_{California}$: $Similarity(q,d) = 1 - \frac{-\sum_{k=1}^{254} \frac{1}{254} * \log_2 \frac{1}{254}}{\log_2 254} = 0$

The last example shows that the equation of figure 5 meets the requirements set in the previous section where the two different attribute query types were defined. Nevertheless, the assumption for the probability leads to equal treatment of the different attribute values. For example, conferences for which the system has no definite information, i.e. NK or NA, have the same chance to take place in "Greece" and in "Italy" . Clearly, this assumption is incorrect but time restrictions prevented us from analysing our collection and calculating the probability in each case.

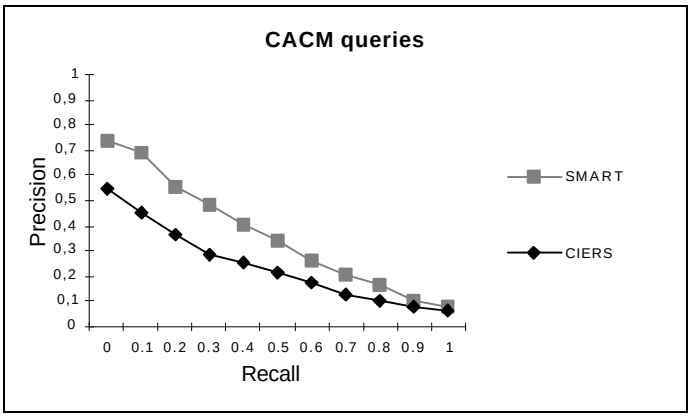
5 Evaluation

Our experiments were divided into three phases. First the system was tested with only content queries, then with only attribute queries, and finally with combined queries. That way it was possible to determine the performance of each query type and thus to realise the effect of employing information extraction techniques in IR systems. In all three cases, CIERS was tested against SMART [1], one of the most widely used experimental IR systems.

For our experiments we constructed our own document collection comprising of calls for papers because none of the existing experimental collections was suitable for CIERS as they did not fit the IE module. A set of realistic requests was provided by some members of the Computing Science Department of the University of Glasgow. Considering that it is a tedious and time consuming task we made the relevance judgements for these queries ourselves. However, this may lead to questionable estimates as it is difficult to decide if the subject of a conference matches a query. On the contrary, in queries about country, continent, and date the relevant documents can be easily determined as these data are easily accessible. Therefore, before the tests with our document collection, CIERS was tested with the CACM standard IR test collection in order to establish the baseline performance of the CIERS standard IR engine. In other words with this test we estimated the efficiency of CIERS against SMART in content queries.

5.1 CACM tests

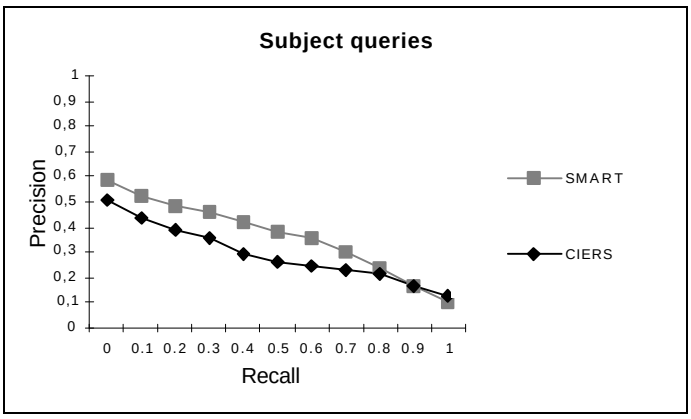
For the CACM tests 50 (content) queries were used and 5 different tests were run, each using a different combination of term weights for documents and queries. Due to space restrictions we present only the first test where the difference between the two systems was maximum.

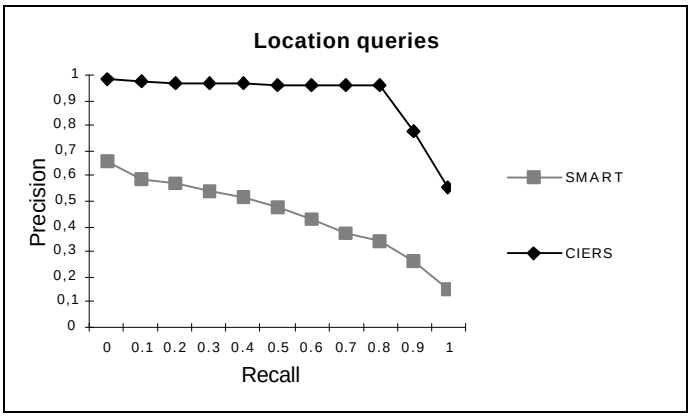


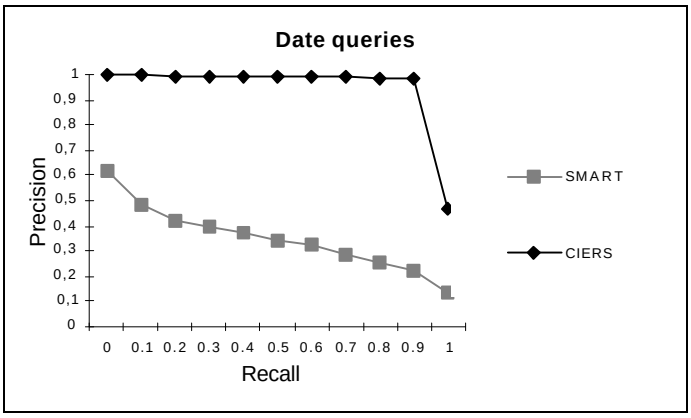
As expected in all the tests CIERS performed worse than SMART due to more advanced indexing, weighting, and matching schemes used in SMART. Their difference in performance lies between 8% and 12%.

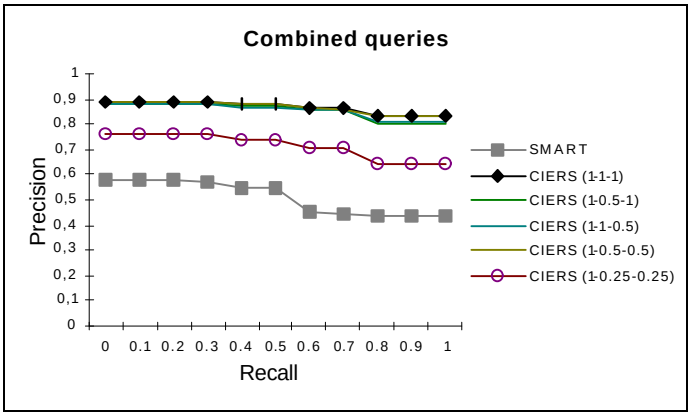
5.2 CFPs Collection tests

A set of 25 queries specifying the desired conference characteristics, i.e. subject, date, and location, were used for these experiments. The experiment with our document collection was divided into four stages. First we tested the system with queries concerning conferences' subject, then with queries about location and date, and finally with combined queries. In order to be fair to SMART in these experiments we used the weighting function of the first CACM test where SMART performed best.









Before explaining the significance of the last tests it is important to make some remarks. First, as it was seen in section 3 CIERS acts as a thesaurus. For example, users can retrieve CFPs of conferences held in "United Kingdom" even if their query is about "Great Britain". Again, to be fair to SMART all variations of a country's name were specified in queries about countries with more than one name. Moreover, location queries were concerned only with countries and not with continents since SMART would be unable to retrieve any of the matching documents. Additionally, only 12 distinct date queries were submitted each specifying one of the 12 months. This was necessary because SMART discards numbers in the indexing phase and it would not be able to retrieve the matching documents. Finally, all the queries were precise since CIER's notion of close queries is not supported by SMART.

Even though we tried to be fair to SMART, the last graphs show that the performance of CIERS is significantly improved over SMART in attribute queries. As anticipated, location queries performed slightly worse than queries about date since the error in the detection of location is greater than the error in the date extraction procedure. In both tests, the sudden drop in precision at the high recall points is because of the error in the extraction procedure.

The last test aimed at measuring the performance of CIERS against SMART with queries based on all the three different data sets, i.e. index terms, location, and date attributes. For the 25 combined queries posed to the system only those CFPs that met all the three conditions set in every query were considered relevant. While SMART was executing each request as one content query, CIERS was dividing them into three subparts. Each subpart was executed separately and the individual results were combined into one unified output. Five different combinations of the partial query subpart results were tested. In all these tests the objective was to examine the influence of the attribute queries in the combined retrieval. As it can be seen in the last graph the performance of the system is extremely high compared to the performance of SMART even when the attribute queries count only 25% each (CIERS 1-0.25-0.25). This means that the increase in precision is vital even when the retrieval is based primarily on the content and not on the attributes.

6 Conclusions and Further Work

The goal of our work was to improve the response of information retrieval systems by identifying and utilising in queries the key attributes of documents. This was achieved by employing information extraction techniques. Even when the extracted information is simple the extraction rules are complicated and depend on the context surrounding these entities. Nevertheless, the benefits gained are important. First the increase in precision is substantial and lies between 45% and 60% in attribute queries. At this point it must be underlined that the average precision of SMART would be lower if only one variation of a country's name and full dates were used. The increase in precision in combined queries is significant as well, even when the attribute subqueries count only 25% each. The significance of this increase is further realised if we take into account the performance level of the standard IR module of CIERS. The second advantage of CIERS is that it acts as a thesaurus. As we have seen the same information may be expressed with several alternative ways but users have to define only one variation in their requests. The last benefit of CIERS is that it supports a wider range of query expression. For example, it is possible to retrieve conference announcements for conferences held in a specific continent or in a specific date. But, as explained, for a fair comparison these features of CIERS where not used when comparing performance with SMART.

Although CIERS embodies a number of novel features there are several ways to improve its functionality and investigate its effectiveness. First, the standard IR module of CIERS should be improved by employing advanced IR techniques, such as the relevance feedback technique [18,26]. Furthermore, the erroneous instances in the extraction procedure should be analysed and the extraction procedure should be modified accordingly. Also our work should be applied in other domains to explore the difficulty of porting an IE system and its performance thereafter.

In general, our work can be considered as initial experimentation towards the integration of IR and IE. There are still several open research issues in creating an integrated IR-IE system. For example, the usability of information extraction into text retrieval should be investigated with more complex pieces of data. That way it would be possible to realise the effects of IE in IR when the success in the extraction procedure would not be as high as in our work. Moreover, as mentioned in section 3, IE systems convert unstructured text elements to structured database entries. It would be rather interesting to investigate the impact of integrating a database system to a combined IR-IE system. The database system would provide more efficient storing mechanisms and enhanced modelling capabilities. The model for estimating the relevance of documents to attribute queries could be embedded in the database system. This would also allow the automatic computation of the prior probability P_k because the database system could compute it for each attribute value by simply analysing the stored objects.

To sum up we can say that information extraction can benefit information retrieval, especially when the success in the identification process is high. However, improvements are required and expected in both fields before their integration provides a powerful tool for text retrieval.

7 Acknowledgements

I would like to thank my supervisor Dr. Mark Dunlop for his valuable guidance and support throughout this project. Also I would like to acknowledge the help of the IR group of the Computing Science Department of the University of Glasgow.

8 References

1. Buckley C. Implementation of the SMART Information Retrieval System. Technical Report 85-686, Department of Computer Science, Cornell University, Ithaca, 1985

2. Chowdhury G G, Lynch M F. Automatic Interpretation of the Texts of Chemical Patent Abstracts. Journal of Chemical Information and Computer Science 1992; 32: 463-467

3. Church K. W., Rau L. Commercial Applications of Natural Language Processing. Communications of the ACM 1995; 38: 71-79

4. Codd E F. Extending the Database Relational Model to Capture more Meaning. ACM TODS 1979; 4: 397-434

5. Codd E F. Understanding Relations. ACM SIGMOD 1975; 7: 23-28

6. Cowie J, Lehnert W. Information Extraction. Communications of the ACM 1996; 39: 80-91

7. Excite. Keep It Simple, Searchers. WebWeek Magazine 1997, 7

8. Fox A E, Koushik M P, Shaw J, Modlin R, Rao D. Combining Evidence from Multiple Searches. TREC-1 Conference Proceedings, National Institute for Standards and Technology Special Publication 500-207, 1993, pp 319-328

9. Fox C. Lexical Analysis and Stoplists. In: Frakes W B, Baeza-Yates R (eds) Information Retrieval: Data Structures and Algorithms. Prentice-Hall, New Jersey, 1992, pp 102-130

10. Frakes W B. Stemming Algorithms. In: Frakes W B, Baeza-Yates R (eds) Information Retrieval: Data Structures and Algorithms. Prentice-Hall, New Jersey, 1992, pp 131-160

11. Gaizauskas R, Robertson A. Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web. RIAO'97 Conference Proceedings, Canada, 1997, pp 356-370

12. Lazarinis F. Combining Information Extraction with Information Retrieval. MSc Thesis, Computing Science Department, University of Glasgow, Glasgow, Scotland, 1997

13. Morrissey M J. A Treatment of Imprecise Data and Uncertainty in Information Systems. PhD Thesis, Department of Computer Science, University College Dublin, Dublin, Ireland, 1987

14. Postma G J, Van der Linden J R, Smits J R M, Kateman G. TICA: A System for the Extraction of Analytical Chemical Information from Texts. In: Karjalainen E J (ed) Scientific Computing and Automation. Elsevier, Amsterdam, 1990, pp 176-181

15. Robertson S E. The Parameter Description of Retrieval Systems: Overall Measures. Journal of Documentation 1969; 25: 93-107

16. Robertson S E. The Parameter Description of Retrieval Systems: The Basic Parameters. Journal of Documentation 1969; 25: 1-27

17. Sager N. Natural Language Information Processing: A Computer Grammar of English and its Applications. Addison-Wesley, Reading, Massachusetts, 1981

18. Salton G, McGill M. Introduction to Modern Information Retrieval. McGraw-Hill Book Company, Computer Science Series, New York, 1983

19. Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. Communications of the ACM 1975; 8: 613-620

20. Salton G. Automatic Text Processing. Addison-Wesley, Reading, Massachusetts, 1989

21. Shannon C E. A Mathematical Theory of Communications. Bell System Technical Journal 1948; 27: 379-423 & 623-656

22. Srinivasan P. Thesaurus Construction. In: Frakes W B, Baeza-Yates R (eds) Information Retrieval: Data Structures and Algorithms. Prentice-Hall, New Jersey, 1992, pp 161-218

23. Sundheim B M. (ed). Proceedings of the Fifth Message Understanding Conference. Morgan Kaufmann, San Francisco, 1993

24. Thompson P. Description of the PRC CEO Algorithm for TREC. TREC-1 Conference Proceedings, National Institute for Standards and Technology Special Publication 500-207, 1993, pp 337-342

25. Ullman J. Universal Relation Interfaces for Database Systems. Information Processing 1983; 83

26. Van Rijsbergen C J. Information Retrieval. 2nd ed, Butterworths, London, 1979

27. Voorhees M E, Gupta K N, Johnson-Laird B. The Collection Fusion Problem. TREC-3 Conference Proceedings, National Institute for Standards and Technology Special Publication 500-236, 1995, pp 95-104

- Appendix A2 -

A Rule Based Tool for Mining Textual Data from Greek Calls for Papers

Fotis Lazarinis

Technological Educational Institute
30200 Mesolonghi, Greece
lazarinf@teimes.gr

Abstract. This paper presents a system which automatically extracts data from Greek Calls for Papers. Five categories of data are mined utilizing vocabularies of rule activation terms, term co-occurrence and domain dependent rules. Prior to the extraction procedure textual input is normalized, a necessary step in languages with inflections and intonation, like Greek. Normalization leads to fewer rules and vocabulary entries, thus to less execution time and greater success in the mining process. The success of the extraction procedure is evaluated and finally conclusions and future work are discussed.

1 Introduction

Calls for Papers (CFPs) announce upcoming conferences and workshops and are important to the academic community and to field practitioners. Early notification of an event's date and theme is quite important as it allows prospective participants to promptly organize their work and schedule. With the rapid expansion of the Web most meeting announcements circulate via the web. We believe that an automated system for identifying the most significant data, i.e. title, submission date, conference date, etc, of CFPs would be quite useful. From this system XML descriptions of the events could be produced which in turn could be utilized in automatically constructing conference announcement indices. These Web pages will be thematically sorted and automatically and regularly updated, with advanced searching capabilities thus enabling users to find everything in one place.

This paper reports the work done and the conclusions drawn on the first phase of an ongoing project. In this first phase a tool for extracting the title, thematic area(s), event date, submission deadline and location is being developed. This tool is based on the identification of patterns and on knowledge lexicons (dictionaries) for mining the previously mentioned data from Greek CFPs. We focus on Greek meetings because as Greek is a language with inclinations and intonation it yields more linguistic opportunities and challenges than English. As it was shown common search engines supporting Greek do not actually understand specific characteristics of the language [1, 2], so utilizing a general purpose search engine to discover events would demand more effort by the user resulting also in lower success. Our main aim is not simply to build up a practical system with extraction capabilities but to explore additional in-

conveniences and present solutions applicable in mining data from Greek corpuses which show considerable diversity.

2 Related Work

Text mining systems analyze unrestricted text in order to extract specific kind of information. Contrary to Information Retrieval (IR) systems [3] and Web search engines, they do not try to process all of the text in the documents but they are restricted on those portions of each document that contain the specific information of interest. They process documents trying to identify pre-defined entities and the relationships between them, filling a structured template with the mined information. In the previous decade text mining systems were usually called Information Extraction systems [4]. Such systems have been implemented to extract data such as names and scientific terms from chemistry papers [5, 6]. Gaizauskas and Robertson [7] used the output of a search engine as input to a text extraction system. Their domain was management succession events and their scenario was designed to track changes in company management.

More contemporary work uses co-occurrence measurement in order to identify relationships and to extract specific data from Web pages [8]. Han et al [9] extract personal information from affiliation, such as emails and addresses, based on document structure. Efforts on Greek information extraction are recorded as well. In [10] a rule based approach to classify words from Greek texts was adapted. Rydberg-Cox [11] describes a prototype multilingual keyword extraction and information browsing system for texts written in Classical Greek. This system automatically extracts keywords from Greek texts using term frequency.

Our work differs from these attempts in that it tries to identify specific information based on rules and on vocabularies of rule activation terms. Also a technique for recognizing term relationships is explored. Additionally classic IR techniques such as suffix and stopword removal [3] are utilized and evaluated in Greek texts.

3 Extracting textual data from Greek CFPs

Many natural language processing and text extraction applications rely on domain specific dictionaries to extract certain kind of information from free text. Such dictionaries contain lexical items that enable the recognition of the desired entities. For instance, if we are interested in US city names the vocabulary must consist of all the city names and their variations. These domain-specific dictionaries contain only the minimal necessary information for the extraction procedure. When an item in the text is matched with an item in the word list a rule activates initiating the extraction process of the desired information. To continue the last example, when a word is found to be a proper city name then a simple rule like “the next word is a potential state name” may be activated.

The relevant work done so far, focus mainly on English text neglecting other languages, which are more demanding and challenging in terms of recognition of patterns. In languages like Greek the same information may appear in many different forms, e.g. 11 Μαΐου 2005 or 11 MAIOY 2005 or Μάιο 11 or 11 Μάη 2005 (11 May 2005), and still convey exactly the same message.

In our system, information mining relies on rule formalisms for each identified entity. Each extraction sub-procedure ends up with one of four alternative results: *(i)* identified (IDN), *(ii)* possibly identified (PDN), *(iii)* not identified (NDN), *(iv)* not applicable (NA). Strong rule paths produce IDN results while weak rule paths end up in PDN. Strong rules are those which definitely identify the information that accurately falls into one of the known and well defined patterns. Weak rules are those who rely on probability and heuristic methods to infer the data. Failing to identify some entity may be due to one of two reasons:

1. A rule activates but it fails to complete, so the data is not identified because of our system's inability. These cases, denoted as NDN, could be used for retraining the system and eventually improve mining of data.
2. The detection of an entity is not possible because it does not exist in the CFP. For example in preliminary announcements the exact conference's date is not yet decided. So NA, adopted by Morrisey's work [11], denotes nonappearance of the hunted piece of information.

The extracted data form an XML file based on a short DTD. That way CFP's data can be presented in many different ways and utilized by other applications. In order to construct rules that will enable the successful extraction of the desired facts, one has to examine thoroughly a representative sample of documents of her/his domain. This will also allow the accurate construction of a vocabulary. For constructing our system, we examined 25 CFPs, a small part of our collection consisting of 145 meeting announcements. This analysis allowed us to realize the different patterns the desired data follow and construct the rules. The remaining 120 CFPs used in the evaluation.

3.1 Normalization procedure

From the analysis of the textual data it was considered necessary to normalize them first. Words are capitalized and accents or other marks are removed. In addition, simple suffix removal techniques (i.e. a primitive Greek stemmer) were applied. It has been proved that these factors influence searching of the Greek Web space as well [1, 2]. Finally, abbreviations were automatically replaced by their full form. For example, month names appear abbreviated quite often, e.g Jan (Ιαν) stands for January (Ιανουάριος). Normalization leads to fewer rules and vocabulary entries, thus to less execution time and greater success in the mining process. In English text normalization procedure is simpler. As a final normalization point, multiple spaces, html tags and other elements, which are not useful at this first version of the system, are removed. We should indicate though that html tags could prove significant especially in correctly identifying the title and the thematic area, as they provide structure to the information.

3.2 Attribute extraction

Title
Extraction of the title of a conference is based on heuristic rules. The basic idea is that titles appear on the top part of a CFP and they follow a "title" format, i.e. words are in capital letters or start with a capital letter, etc. Obviously normalization should be done after the identification of title as the form of words plays an important role here. Another rule employed is based on the surrounding text and in keywords, like conference, symposium, congress and meeting. As we will see in the evaluation section title identification is quite successful, though some extracted titles are truncated.

Thematic area
Correct identification of the title is also important for classifying the meeting. Classification means the detection of some keywords which describe the meeting. At the moment we base the classification on the co-occurrence of pairs of terms derived from the title and from short lists of terms [8]. We first remove stopwords and then we construct a list of pairs of neighboring terms. Then we try to measure the co-occurrence of these pairs. We define co-occurrence of two terms as terms appearing in the same Web page. If two terms co-occur in many pages, we can say that those two have a strong relation and the one term is relevant to the other. This co-occurrence information is acquired by the number of retrieved results of a search engine using the following coefficient measure.

$$r(a,b) = \frac{\left|a \cup b\right|}{\left|a\right| + \left|b\right| - \left|a \cup b\right|}$$

With |a| we symbolize the number of documents retrieved when we search using term a. Similarly |b| is the number of documents relevant to term b and |a ∪ b| is the number of pages containing both terms. The co-occurrence is measured for every pair of terms and the top results are kept, based on a fixed cut off value. So if a conference is about New Technologies in Adult Education "in" is removed and the pairs "New Technologies", "Technologies Adult", "Adult Education" are formed. Then these pairs along with the terms "New", "Technologies", "Adult", "Education" are searched in the Web and the coefficient measure of the term pairs is decided.

Date
The first step in the identification of dates is the construction of a suitable vocabulary containing the normalized month terms that will activate the rules for the extraction of the conference's date. The identification of the date is based on a simple observation. The latest dates, appearing in a CFP, are most probably the event's start and end dates. Our purpose is to recognize both start and end dates. For example from a date 11-13 June 2005 we extract 11 June 2005 as the start date and 13 June 2005 as the end date.

The date detection procedure initiates when a month or a full date (e.g. 12/05/2006) is found in the text. In that case we first check the succeeding words until the end of the sentence and then the preceding words until the beginning of the sentence. This search aims at identifying the day and the year of the conference and keywords which verify that it is actually the meeting's date. Thus the system needs to be able to keep information preceding and succeeding the rule activation keyword. If more than one date or date range is discovered then the system searches for appropriate keywords. Rules are a set of If then else and sub ifs. Document is processed line by line and term by term. At the end of the rule formalism the result is stored in the CFPs XML repository.

Submission date
Submission date is trickier than the event's date as is absent in many cases, especially in short announcements. This procedure is complimentary to the previous one as dates which are denoted as meeting's start and end dates should not be checked again. After the extraction of a proper date the surrounding text is scanned for words like deadline (υποβολή), or other synonyms. Clearly these rules are domain dependant and have a high error probability. This procedure ends up mostly with one of the codes PDN, NDN, NA.

Location
For extracting an event's location we used a knowledge base with the major Greek cities and the prefecture in which they belong. This listing also models bordering city and county relations. A city's name will trigger off the rules for the identification of the desired information. It was proved that normalization of locations names is absolutely essential as they appear in many different forms, e.g. Αθήνα, Αθηνών, Αθήνας (Athens). One problem in the identification of the location arises when a conference is co-organized by more than one institution. In this case many locations co-exist. Mining is then based on the surrounding context or on the location's tf (term frequency) measured in the whole CFP. If a strong decision is made then the procedure ends up, whereas when a weak decision is made the procedure initiates again when new activation terms appear up.

3.3 System architecture

System is developed in Java using techniques such as Servlets and Java Server Pages. As seen in figure 1 a CFP is submitted via a Web interface, as simple unformatted text, and is processed by independent subsystems. The results are stored in a central XML Repository accessible via the Web. System's client server Web architecture enables easy access and integration with other components. At the moment the CFP index is projected back to the client as static HTML, however we plan to create dynamically produced pages, based on client's criteria and on the certain and uncertain extracted values and to create and explore more sophisticated search and ranking mechanisms.

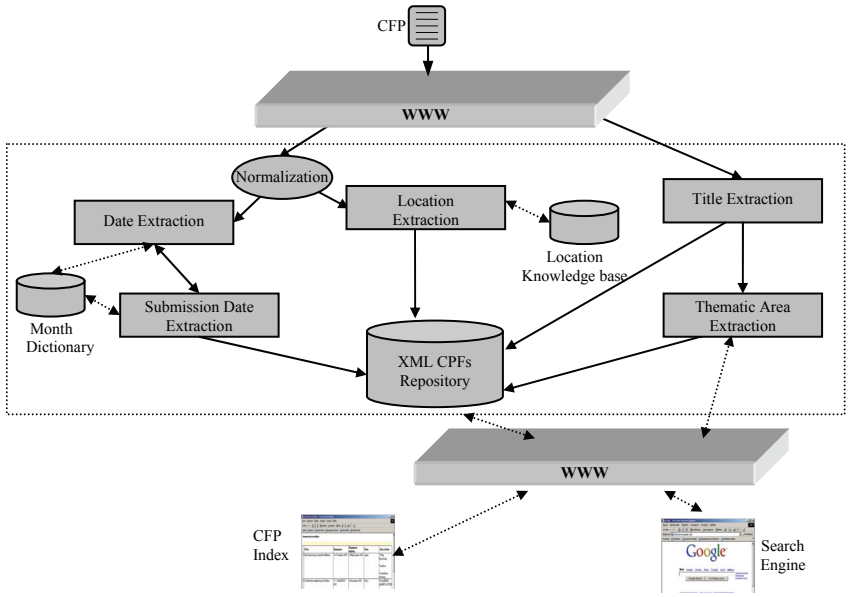


Fig. 1. Overview of the system's architecture.

4 Evaluation

The performance of an information extraction system can be measured using Precision (P) and Recall (R), as in Information Retrieval systems [3, 13]. Precision measures the ratio of the correctly extracted information against all the extracted information. Recall measures the ratio of the correct information extracted from the texts against all the available information. Despite the diversity of the collection the system works adequately well and the employed rules achieve high rates of precision and recall, especially in the attributes where a dictionary is used.

Table 1. Precision and Recall for the attribute extraction procedure.

Attributes	Correctly Extracted	Erroneously Extracted	Not Extracted	Precision	Recall
Title	77	29	14	72,64%	64,17%
Thematic area	39	65	16	37,50%	32,50%
Date	107	8	5	93,04%	89,17%
Submission date	89	19	12	82,41%	74,17%
Location	110	7	3	94,02%	91,67%

The results of the evaluation are summarized in table 1. As expected, title and thematic area show a higher error percentage. Clearly more sophisticated rules are

needed. A possible solution would be the exploitation of tagging information and the usage of lexicons which model domain relationships as well. It should be noted that partially extracted titles, even those with only one not identified word, were accounted as erroneously extracted. So with slight improvements we can achieve higher precision and recall. Date and location rules achieve high precision and recall scores. Their extraction is relying on specific word lists and they follow better structured patterns.

In order to realize the effects of normalization and to get an indication of the additional difficulties posed in Greek we evaluated the system's performance, on date, submission date and location extraction, without extensive normalization. That is words were only capitalized and short forms replaced by their full forms. The evaluation showed that system's precision reduced by more than 30%. One could argue that in this case more rules should be employed in order to achieve higher precision. While this could be partially true, we need to take into account that more rules means increased execution time as more searches are needed and a higher error probability as more heuristics and weak rules will be employed.

A final evaluation task was performed utilizing Google. A set of five queries concerning specific locations and a second set concerning dates consisting of months and years were run in our collection using Google. Then we evaluated the precision of each query (table 2). Clearly Google retrieves many irrelevant files which diminish precision and recall. This is because every file containing the query terms or one of them is retrieved. Furthermore, announcements where terms appear in different forms than the requested ones are not retrieved. In our tool vocabularies act as thesauri as well allowing retrieval of meetings where locations or month names appear in another form or inclination. Of course table 2 shows an initial estimation. A more thoroughly designed evaluation is needed with more queries to safely reach useful conclusions.

Table 2. Precision and Recall of location and date queries run in Google.

Location	Precision	Recall	Date	Precision	Recall
Query 1	57,50%	76,00%	Query 1	42,31%	60,00%
Query 2	42,86%	83,33%	Query 2	32,14%	52,38%
Query 3	77,78%	83,33%	Query 3	43,75%	75,00%
Query 4	55,88%	64,29%	Query 4	40,63%	50,00%
Query 5	50,00%	65,71%	Query 5	37,50%	54,29%

5 Summary

This paper presents a system which automatically extracts data from Greek Calls for Papers. Five categories of data are mined utilizing various techniques and approaches. For the first two categories rules are based on text's position, on context surrounding the information and on a coefficient measure. The last three types of data are mined with the utilization of lexicons which contain rule initiation terms. Then the surrounding text is again exploited. It was shown that simple removal of endings and accents and other adjustments, specific to Greek language, improve the mining procedure and

lead to increased Precision and Recall and to less elaborated rules. Vocabularies act as thesauri permitting retrieval of CFPs where terms appear in different forms than the requested ones.

However more work needs to be done in order to achieve high rates of precision. Tagging and formatting information should be utilized in the identification of complex textual information. Metadata and link tracking, in the case of html or xml files, could be utilized. Links usually point to more detailed announcements in which all the data are applicable. Domain vocabularies are necessary in order to identify classification terms.

References

1. Lazarinis, F.: Evaluating User Effort in Greek Web Searching. 10th PanHellenic Conference in Informatics, University of Thessaly, Volos, Greece, (2005), (to appear)
2. Lazarinis, F.: Do Search Engines Do search engines understand Greek or user requests "sound Greek" to them?, International Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology, (2005) 43-46
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, ACM Press, New York, (1999).
4. Cowie, J, Lehnert, W.: Information Extraction. Communications of the ACM, 39 (1996) 80-91.
5. Postma, G. J., Van der Linden, J. R., Smits, J. R. M., Kateman, G.: TICA: A System for the Extraction of Analytical Chemical Information from Texts. In: Karjalainen E J (ed) Scientific Computing and Automation. Elsevier, Amsterdam, 1990, 176-181.
6. Chowdhury, G. G., Lynch, M. F.: Automatic Interpretation of the Texts of Chemical Patent Abstracts, Part 1: Lexical Analysis and Categorisation. Journal of Chemical Information and Computer Science, 32 (1992) 463-467.
7. Gaizauskas, R., Robertson, A.: Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web. RIAO'97 Conference Proceedings, Canada, (1997) 356-370.
8. Mori, J., Matsuo, Y., Ishizuka, M., Faltings, B.: Keyword Extraction from the Web for Creation of Person Metadata, ISWC2004, (2004) (in press)
9. Han, H., Giles, L. C., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.: Automatic Document Metadata Extraction using Support Vector Machines. Proceedings of the ACM IEEE Joint Conference on Digital Libraries, (2003) 37-48.
10. Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.: Resolving part-of-speech ambiguity in the Greek language using learning techniques, Proceedings of the ECCAI Advanced Course on Artificial Intelligence, Chania, Greece, (1999).
11. Rydberg-Cox, A. J: A prototype multilingual document browser for ancient Greek texts, The New Review of Hypermedia and Multimedia, 7(1) (2002) 103 – 113.
12. Morrissey, M. J.: A Treatment of Imprecise Data and Uncertainty in Information Systems, PhD Thesis, Department of Computer Science, University College, Dublin, Ireland, (1987).
13. Robertson, S. E.: The Parameter Description of Retrieval Systems: Overall Measures. Journal of Documentation, 25 (1969) 93-107.

- Appendix A3 -

Do search engines understand Greek or user requests "sound Greek" to them?

Fotis Lazarinis
Technological Educational Institute of Mesolonghi
30200 Mesolonghi, Greece
lazarinf@teimes.gr

Abstract

This paper presents the outcomes of initial Greek Web searching experimentation. The effects of localization support and standard Information Retrieval techniques such as term normalization, stopword removal and simple stemming are studied in international and local search engines. Finally, evaluation points and conclusions are discussed.

1. Introduction

The Web has rapidly gained popularity and has become one of the most widely used services of the Internet. Its friendly interface and its hypermedia features attract a significant number of users. Finding information that satisfies a particular user need is one of the most common and important operations in the WWW. Data are dispensed in a measureless number of locations and so utilization of a search engine is necessary.

Although international search engines like Google and Yahoo are preferred over the local ones, as they employ better searching mechanisms and interfaces, they do not really value other spoken languages than English. Especially in languages like Greek which has inclinations and intonation, it seems that the majority of the international search engines have no internal (indexing) or external (interface) localization support. Thus the user has to devise alternative ways so as to discover the desired information and to adapt themselves to the search engine's interface.

This paper reports the results of initial experimentation in Greek Web searching. The effect of localization support, upper or lower case queries, stopword removal and simple stemming is studied and evaluation points are presented. The conclusions could be readily adapted to other spoken languages with similar characteristics to the Greek language.

2. Experimentation and evaluation

Interface simplicity and adaptation is maybe the most important issue which influences user satisfaction and acceptance of Web sites and thus search engines [1, 2]. User acceptance factor is obviously increased when a search engine changes the language and maybe its appearance to satisfy its diversified user basis. This is significant especially to novice users.

Stopword removal, stemming and capitalization or more generally normalization of index and query terms are amongst the oldest and most widely used IR techniques [3]. All academic systems support them. Commercial search engines, like Google, explicitly state that they remove stopwords, while capitalization support is easily inferred. Stemming seems to not be supported though. This may be due to the fact that WWW document collection is so huge and diverse that stemming would significantly increase recall and possibly reduce precision. However simple stemming, like final sigma removal which will be presented later in the paper, may play an important role when seeking information in the Web using Greek query terms.

These four issues were examined with respect to the Greek language. For conducting our assessment we used most of the predominately known worldwide .com search engines: Google, Yahoo, MSN, AOL, Ask, Altavista. The .com search engines were selected based on their popularity [4]. Also, for comparison reasons, we considered using some native Greek search engines: In (www.in.gr), Pathfinder (www.pathfinder.gr) and Phantis (www.phantis.gr).

2.1. Interface issues

Ten users participated in the interface related experiment and they also constructed some sample queries for the subsequent experiments. Users had varying degrees of computer usage expertise. We needed end users with technical expertise and obviously increased demands over the utilization of web searchers. On the other hand we should measure the difficulties and listen to the people who have just been introduced to search engines. This combination of needs reflect real everyday needs of web "surfers".

The following sub-issues extracted from a more complete evaluation study of user effort when searching the Greek Web space utilizing international search engines [5]. Here we extend (with more users and search engines) and present only the issues connected with whether search engines really value other spoken

languages than English, like Greek, or not.

2.1.1. Localization support. The first issue in our study was the importance of a localized interface. All the participants (100%) rated this feature as highly important as many users have basic or no knowledge of English. Although search engines have uncomplicated and minimalist interfaces their adaptation to the local language is essential as users could easily comprehend the available options.

From the .com ones only Google automatically detects local settings and adapts to Greek. Altavista allows manual selection of the presentation language with a limited number of language choices and setup instructions in English. Also if you select another language, search is automatically confined to this country's websites (this must be altered manually again).

2.1.2. Searching capability. In this task users were asked to search using queries with all terms in Greek. All search engines but AOL and Ask were capable of running the queries and retrieving possibly relevant documents. AOL pops-up a new Window when a user requests some information but it cannot correctly pass the Greek terms from the one window to the other. So no results are returned. However, when requests typed directly to the popped-up window then queries are run but presentation of the rank is problematic again.

Ask does not retrieve any results, meaning that indexing of Greek documents is not supported. For example zero documents retrieved in all five queries of section 2.2. For these reasons AOL and Ask left out of the subsequent tests.

2.1.3. Output presentation. An important point made by the participants is that some of the search engines rank English web pages first, although search requests were in Greek. For example in the query "Ολυμπιακοί αγώνες στην Αθήνα" (Olympic Games in Athens) Yahoo, MSN and Altavista ranked some English pages first. This depends on the internal indexing and ranking algorithm but it is one of the points that increase user effort because one has to scroll down to the list of pages to find the Greek ones.

2.2. Term normalization, Stemming, Stopwords

Trying to realize how term normalization, stemming and stopwords affect retrieval we run some sample queries. We used 5 queries (table 1) suggested by the participants of the previous test. They were typed in lower case sentence form with accent marks leaving the default options of each search engine. A modified version of Recall and Precision [6] are used for comparing the results of the sample queries. Recall refers to the number of retrieved pages, as indicated by search engines, while precision (relevance) was measured in the first 10 results.

Table 1. Sample queries.

No	Queries in Greek	Queries in English
Q1	Μορφές ρύπανσης περιβάλλοντος	Environmental pollution forms
Q2	Εθνική πινακοθήκη Αθηνών	National Art Gallery of Athens
Q3	Προβλήματα υγείας από τα κινητά τηλέφωνα	Health problems caused by mobile phones
Q4	Συνέδριο πληροφορικής 2005	Informatics conference 2005
Q5	Τεστ για την πιστοποίηση των εκπαιδευτικών	Tests for educators' certification

Table 2 presents the number of recalled pages for each query. From table 2 we realize that In and Pathfinder share the same index and employ exactly the same ranking procedure. The result set was identical both in quantity and order. Their only difference was in output presentation. Altavista and Yahoo had almost the same number of results, ranked slightly differently though.

Table 2. Recall in lower case queries.

	Q1	Q2	Q 3	Q4	Q5
Google	867	3400	805	15500	252
Yahoo	820	933	527	11200	186
MSN	1357	1537	542	6486	272
Altavista	821	939	515	11400	191
In	251	343	67	689	49
Pathfinder	251	343	67	689	49
Phantis	33	63	22	88	6

In all cases the international search engines returned more results than the native Greek local engines. However, as seen in table 3, relevance of the first 10 results is almost identical in all cases, except Phantis, which maintains either a small index or employs a crude ranking algorithm. Query 4 retrieves so many results because it contains the number (year) 2005. So, documents which contain one of the terms and the number 2005 are retrieved, increasing recall significantly.

Table 3. Precision of the top 10 results.

	Q1	Q2	Q 3	Q4	Q5
Google	5	7	9	8	8
Yahoo	5	7	8	7	8
MSN	4	7	8	6	7
Altavista	5	7	8	7	8
In	5	7	8	6	8
Pathfinder	5	7	8	6	8
Phantis	2	2	2	1	0

We confined the relevance judgment to only the first ten results so to limit the required time and because the

first ten results are those with the highest probability to be visited. Relevance was judged upon having visited and inspected each page. The web locations visited had to be from a different domain. So if two consecutive pages were on the same server only one of them was visited.

An interesting point to make is that although recall differs substantially among search engines precision is almost the same in all cases. Another point of attention is that the third query shows the maximum precision. This is because in this case terms are more normalized, compared to the other queries. This means that they are in the first singular or plural form which is the usual case in words appearing in headings or sub-headings. Consequently a better retrieval performance is exhibited. But, as we will see in section 2.2.3, it contains stopwords which when removed precision is positively affected and reaches 10/10.

2.2.1. Term normalization. We then re-run the same queries but this time in capital letters with no accent marks. Recall (table 4) was dramatically diminished in most of the worldwide search enabling sites while it was left unaffected in two of the three domestic ones (In and Pathfinder). Precision was negatively affected as well (table 5), compared to results presented in table 3.

Table 4. Recall in upper case queries.

	Q1	Q2	Q 3	Q4	Q5
Google	22	3400	41	673	252
Yahoo	18	229	2	116	8
MSN	10	233	2	379	10
Altavista	18	239	2	117	9
In	251	343	67	689	49
Pathfinder	251	343	67	689	49
Phantis	4	63	3	14	6

These observations are true for Yahoo, MSN and Altavista. Google and Phantis exhibit a somehow unusual behavior. In queries 2 and 5 Google and Phantis retrieve the same number of documents in the same order and have the same precision therefore. Upper case queries 1, 3 and 4 recall only a few documents compared to the equivalent lower case queries. Correlation between results is low and precision differs.

Trying to understand what triggers this inconsistency we concluded that it relates to the final sigma existing in some terms of queries 1, 3 and 4. The Greek capital sigma is *Σ* but lower case sigma is *σ* when it appears inside a word and *ς* at the end of the word. Phantis presents the normalized form of the query along with the result set. Indeed it turns out that words ending in capital Σ are transformed to words with the wrong form of sigma, e.g. "ΜΟΡΦΕΣ" (forms) should change to "μορφες" but it changes to "μορφεσ".

These observations are at least worrying. What would happen if a searcher were to choose to search only in capital letters or without accent marks? Their quest would simply fail in most of the cases leading novice users to stop their search. In English search there is no differentiation between capital and lower letters. The result sets are identical in both cases so user effort and required "user Web intelligence" is unquestionably less.

Table 5. Precision of the top 10 results.

	Q1	Q2	Q 3	Q4	Q5
Google	4	7	3	10	8
Yahoo	3	8	0	5	7
MSN	3	6	0	7	7
Altavista	3	8	0	5	7
In	5	7	8	6	8
Pathfinder	5	7	8	6	8
Phantis	0	2	0	0	0

Wrapping up this experiment one can argue that in Greek Web searching the same query should be run both in lower and in capital letters, so as to improve the performance of the search. Sites where there are no accent marks or contain intonation errors will not be retrieved unless variations of the query terms are used. Greek search engines are superior at this point and make information hunting easier and more effective. From the international search engines only Google has recognized these differences and try to improve its searching mechanism.

2.2.2. Stemming. Another factor that influences searching relates to the suffixes of the user request words. For example the phrases "Εθνική πινακοθήκη Αθηνών" or "Εθνική πινακοθήκη Αθήνας" or "Εθνική πινακοθήκη Αθήνα" all mean "National Art Gallery of Athens". So while they are different they describe exactly the same information need. Each variation retrieves quite different number of pages. For example Google returned 3400, 722 and 5420 web pages respectively. Precision is different in these three cases as well, and correlation between results is less than 50% in the first ten results.

One could argue that such a difference is rational and acceptable as the queries differ. If we consider these queries solely from a technical point of view then this argument is right. However if the information needed is in the center of the discussion then these subtle differences in queries which merely differ in one ending should have recalled the same web pages. Stemming is an important feature of retrieval systems [3] (p. 167) and its application should be at least studied in spoken languages which have conjugations of nouns and verbs, like in Greek. Google partially supports conjugation of English verbs.

2.2.3. Stopwords. Google and other international search engines remove English stopwords so as to not influence

retrieval. For instance users are informed that the word *of* is an ordinary term and is not used in the query "National Art Gallery of Athens". Removal of stopwords [3] (p. 167) is an essential part of typical retrieval systems.

We re-run, in Google, queries 3 and 5 removing the ordinary words. Queries were in lower case and with accent marks so results should be compared with tables 2 and 3. Query 3 recalled 839 pages and precision equals 10 in the first 10 ranked documents. Similarly for the fifth query Google retrieved 275 documents and precision raised from 8 (table 2) to 10. As realized, recall was left unaffected but precision increased by 10% and by 20% respectively. This means that ranking is affected when stopwords are removed. However more intense tests are required to construct a stopword list and to see how retrieval is affected by Greek stopwords

4. Conclusions

This paper presents a study regarding utilization of search engines using Greek terms. The issues inspected were the localization support of international search engines and the effect of stopword removal, capitalization and stemming of query terms. Our analysis participants identified as highly important the adaptation of search engines to local settings. Most of the international search engines do not automatically adapt their interface to other spoken languages than English and some of them do not even support other spoken languages. At least these are true for Greek.

In order to get an estimate of the internal features of search engines that support Greek, we run some sample queries. International search engines recalled more pages than the local ones and they had a small positive difference in precision as well. However they are case sensitive, apart from Google, hindering retrieval of web pages which contain the query terms in a slightly different form to the requested one. Even if the first letter of a word is a capital letter the results will be different than when the word is typed entirely in lower case.

Endings and stopwords are not removed automatically, thus affecting negatively recall of relevant pages. Stopwords are removed from English queries making information hunting easier, looking at it from a user's perspective. Terms are not stemmed though even in English. However in a language with inclinations, like Greek, simple stemming seems to play an important role in retrieval assisting end users. In any case more intensive tests are needed to realize how endings, stopwords and capitalization affect retrieval.

Trying to answer the question posed in the article's title it can be definitely argued that international search enabling sites do not value the Greek language and possibly other languages with unusual alphabets. Google is the only one which differs than the others and seems to be in a process of adapting to and assimilating the additional characteristics.

5. References

[1] J. Nielsen, R. Molich, C. Snyder, S. Farrel, *Search: 29 Design Guidelines for Usable Search* http://www.nngroup.com/reports/ecommerce/search.html,2000.

[2] Carpineto, C. et al., "Evaluating search features in public administration websites", *Euroweb2001 Conference*, 2001, 167-184.

[3] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison Wesley, ACM Press, New York, 1999.

[4] D. Sullivan, *Nielsen NetRatings: Search Engine Ratings* http://searchenginewatch.com/reports/article.php/2156451, 2005.

[5] Lazarinis, F., "Evaluating user effort in Greek web searching", *10th PanHellenic Conference in Informatics*, University of Thessaly, Volos, Greece, 2005 (to appear)

[6] S. E. Robertson, "The Parameter Description of Retrieval Systems: Overall Measures", *Journal of Documentation*, 1969, 25, 93-107.

- Appendix A4 -

Evaluating User Effort in Greek Web Searching

Fotis Lazarinis

Technological Educational Institute of Mesolonghi
30200 Mesolonghi, Greece
lazarinf@teimes.gr

Abstract. Searching is one of the most important operations in the Web because it helps users to satisfy their information needs. Although several local and international search engines exist, offering a rich set of capabilities and options, most of which expect users to fine tune their queries and become skilled so as to be successful in their quest. This is even more demanding in the case of multilingual information seek out. Thus far evaluation attempts of search engines focus only on their retrieval performance. In the current study we concentrate on the user effort and we aim at identifying difficulties and knowledge prerequisites when using a Greek supporting search engine. We then suggest a series of improvements which could diminish the required user effort and knowledge and increase user satisfaction in Greek web searching. Our conclusions could be readily adapted to other spoken languages as well.

1 Introduction

The World Wide Web (WWW or the Web) has rapidly gained popularity and has become one of the most widely used services of the Internet along with email. WWW has gained such great publicity that many people erroneously equate it with the Internet. The friendly interface and the hypermedia features of the Web attract a significant number of users. As a result, the Web has become a pool of various types of data, dispensed in a measureless number of locations.

Finding information that satisfies a particular user need is one of the most common and important operations when using the Web. Although there are a significant number of automated search engines that facilitate Web searching, little or no attention has been given to the user effort required in utilizing one. Especially when searching is in a language other than English, efficient search engine utilization requires an increased level of knowledge. This is because most search engines have no internal (indexing) or external (interface) localization support and thus the user has to devise alternative ways so as to discover the desired information. Baeza-Yates and Ribeiro-Neto [1] (p. 391) suggest teaching users methods for effective searching. Clearly this is not a feasible solution as the potential student group would be enormous.

The purpose of this paper is to identify some of the problems of searching the Web using Greek terms. Based on the findings we suggest ways to decrease the required user effort so that even beginners can exploit the full power of search engines. The

conclusions of our survey could be utilized in the enhancement of the Greek supporting web search engines.

2 Evaluation of Web Search engines

Evaluation is an important aspect in an Information Retrieval system [2, 3]. Cleverdon [2] listed six criteria that could be used to evaluate information retrieval systems: 1. coverage, 2. time lag, 3. recall, 4. precision, 5. presentation and 6. user effort. Of these criteria, recall and precision have most frequently been applied in measuring information retrieval.

With the deployment of information retrieval in WWW, these evaluation criteria re-shaped to fit in this environment [4, 5]. Chu and Rosenthal [4] evaluated the capabilities of three search engines, Alta Vista, Excite, and Lycos and proposed a methodology for evaluating WWW search engines in terms of five aspects:

1. Composition of Web indexes (coverage) – collection update frequencies and size can have an effect on retrieval performance.
2. Search capability – they suggest that search engines should include "fundamental" search facilities such as Boolean logic and scope limiting abilities.
3. Retrieval performance (precision, recall, time lag) – such as precision, recall, and response time.
4. Output option (presentation) – this aspect can be assessed in terms of the number of output options that are available and the actual content of those options.
5. User effort – how difficult and effortful it is to use the search engine by typical users.

Most search engine evaluation attempts focus on the third criterion. For example Dunlop [6] used the expected search length to construct graphical evaluation methods to measure retrieval performance from AltaVisa. These graphs were introduced as supplementary to precision-recall graphs. Altavista, Infoseek, Lycos, and Open Text used in another evaluation study [7]. The authors employed the measured precision and partial precision for the first twenty hits returned by the search engines. They also defined an evaluative measure that compared ratings of relevance on a 5-point scale. Many more comparisons and assessments of Web search engines were performed by academics and trade magazines [8, 9].

In all these studies the basic aim was to measure precision and recall, as was done traditionally in IR systems [10, 11]. However if we consider that search engines are widely spread and accessible by non expert and occasional Web users then it seems that "user" as an evaluation parameter is quite important. The last three criteria could be considered from the user point of view and could lead to important conclusions about the skills required to successfully use a search engine. In this study we focus on the last three factors pointed above and we try to extract some conclusions which could be initial research points for alleviating multilingual web search engines.

3 Evaluation of Greek Supporting Search Engines

3.1 Search engines

For conducting our assessment we used most of the predominately known worldwide .com search engines along with purely .gr Greek robots. Table 1 summarizes the names and the URLs of the search engines studied. The .com search engines were selected based on their popularity [12] or because they were used in other evaluations as well [4]. The native Greek search engines are empirically selected.

Table 1. Names and urls of the search engines used in our study.

Search engine	URL
Google	www.google.com
Yahoo	www.yahoo.com
MSN	www.msn.com
Altavista	www.altavista.com
In	www.in.gr
Pathfinder	www.pathfinder.gr
Robby	www.robby.gr
Anazitisis	www.anazitisis.gr

3.2 Users

Four computer science (CS) graduates and four non-computer related graduates participated in the experiments described in the next sections. We needed end users with technical expertise and obviously increased demands over the utilization of web searchers. On the other hand we should measure the difficulty and listen to the people who have just been introduced to search engines. This combination of needs reflect real everyday needs of web "surfers". Although CS graduates were aware on the purpose and the functions of automated web searchers, a short introduction on how to use a search engine was given to all the participants. The second category of participants had basic e-skills, i.e. knowledge of MS-Office tools.

3.3 Evaluation methodology and topics

User assessments were recorded using interviews, self reporting and field observation [13] (p. 228). Users had to report back problems they encountered or they were asked to rate some issues on a 5-point scale or we were observing and recording their behavior. Evaluation topics included localization support, interface complexity and sample searches. Sample searches were in Greek and variations of the queries were run so as to measure the effort and required user knowledge when utilizing a Greek supporting search engine.

Localization

The first issue in our study was the importance of a localized interface. All the participants (100%) rated this feature as *highly important* (5 on the 5-point scale) as many users have basic or no knowledge of English. Although search engines have uncomplicated and minimalist interfaces their adaptation to the local language is essential as users could easily comprehend the available options (e.g. advanced preferences).

From the .com ones only Google automatically detects local settings and adapts to Greek. Altavista allows manual selection of the presentation language with a limited number of choices though and setup instructions in English. Also if you select another language, search is automatically confined to this country's websites (this must be altered manually again).

Interface complexity

With the exception of Google, Altavista and Anazitisis all the other sites also act as Web portals containing categorized links, news, photos and animated Gifs. These features led to increased downloading time, which can be irritating when connection speed is low. Also it can cause confusion and disorientation to users as the textbox where the query is typed and the procedure's initiation button are not easily viewable.

To formally measure these assumptions, users were instructed to visit each search engine and look for sites containing information about *Ολυμπιακοί αγώνες στην Αθήνα* (Olympic Games in Athens). Users were not allowed to consult each other but they were only permitted to ask assistance from us on encountering any problems. Queries and difficulties raised during this procedure appear on table 2.

Table 2. Problems/questions posed during first usage of search engines.

User problems/questions	Occurrences	URL
Slow downloading	2	www.in.gr www.robby.gr
In which textbox to type the query	5	www.in.gr www.yahoo.com
Which button to click on	3	www.anazitisis.gr www.altavista.gr
Is login required	1	www.pathfinder.gr
Activation of additional features	3	www.anazitisis.gr www.in.gr

The two most important problems met are the second and third of table 2. These difficulties obstructed users in completing their task. In the case of the In search engine even a CS graduate was confused because one can search solely in the in.gr site or the whole web or the Greek web space. Based on the desired case, the searcher must additionally select one of the two available textboxes to type their request.

The third issue recorded in table 2 was raised because the button which the user had to click on was not the standard "search" button but something different like

"go". Two non CS graduates were unsure of what to do after typing the query and this posed a question to us.

The other problems do not impede completion of the task but they provoke confusion. For instance a login textbox exists close to the query textbox in Pathfinder. One participant wondered if login is necessary before utilizing the service. Odd as it may sound we must keep in mind that Web services are utilized by many non technologically skilled people and should balance between simplicity and options therefore.

Output presentation

In this task users were required to search again for Olympic Games in Athens, to visit the first result and then go back and visit the second web page retrieved.

All search engines present the retrieved urls in the same way (fig 1). That is a list of clickable phrases, i.e. the titles of the site they point on, a brief summary of the site's contents and the actual url at the end of the brief summary. In and Robby differ a little from the other ones. The first one presents the results in a more condensed form without leaving much space between results and the second one presents the findings with smaller letters with a quite short or no summary and in a condensed mode as well.

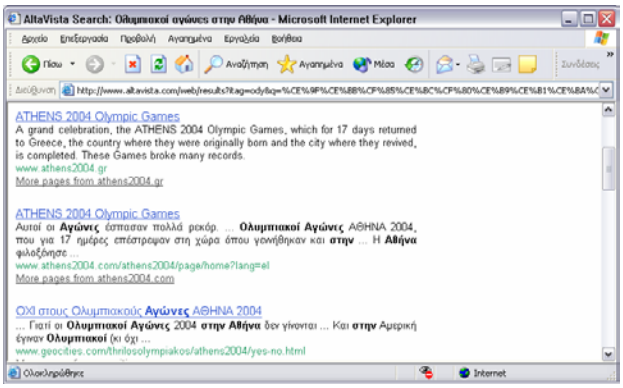


Fig. 1. Output of a standard search engine.

All participants (100%) showed dissatisfaction with the condensed presentation output because it was more difficult to distinguish between the resulting URLs. Also short summaries increase human effort as they have to first visit the web page and then decide if it is relevant. Summarization is a quite difficult task in information retrieval and most systems provide inadequate summaries [14]. This task is even harder when the document collection is enormous and of varying human languages as in the Web.

After this assessment, users had to visit the first result, then return back to the list of results and visit the second retrieved page. Every participant was able to perform the requested tasks in the seven out of the eight engines under survey. The only problem appeared in the utilization of the www.in.gr engine. This particular one opens a

new window when a result is selected. Three out of the four non CS graduates (3/4-75%) had to consult us on how to return to the list of results. Evidently this design decision instead of helping end users causes problems especially to novices.

Another important point made by the participants is that some of the search engines rank English web pages first1, although search was in Greek. This depends on the internal indexing and ranking algorithm but it is one of the points that increase user effort because one has to scroll down to the list of pages to find the Greek ones.

Sample searches

Trying to get an estimate of the index size, the ranking procedure and the relevance of the results of the engines under survey the participants experimented using the following sample queries. They were typed in Greek leaving the default search options. Table 3 shows the sample queries in Greek and in English. These questions suggested by the participants themselves so to reflect actual user needs.

Table 3. Sample queries.

No	Query terms in Greek	Query terms in English
1	Μορφές ρύπανσης περιβάλλοντος	Environmental pollution forms
2	Εθνική πινακοθήκη Αθηνών	National Art Gallery of Athens
3	Προβλήματα υγείας από τα κινητά τηλέφωνα	Health problems caused by mobile phones

Result set

Table 4 presents the number of retrieved pages for each query. It seems that the last two search engines do not operate properly because in all runs one returned no results whereas the other returned an unexpectedly large number of pages which indicates malfunctioning.

Table 4. Number of results retrieved in sample queries.

Search engine	Query 1	Query 2	Query 3
Google	527	896	521
Yahoo	588	713	301
MSN	1228	2099	1080
Altavista	582	715	303
In	237	320	45
Pathfinder	237	320	45
Robby	0	0	0
Anazitisis	646859	20352	78457

Another conclusion drawn by table 4 is that In and Pathfinder share the same index and employ exactly the same ranking procedure. The result set was identical both in quantity and order. Their only difference was in output presentation. Altavista and Yahoo had almost the same number of results, ranking differently though.

The international search engines returned approximately an equivalent number of relevant web pages, with the exception of MSN which provided more results. It is obvious that the .com search engines employ better indexing and retrieval algorithms than the local ones, although national engines have a better understanding of the local language properties.

Relevance

Participants were asked to judge the relevance of the first 5 pages in the rank. Although this task could be performed by us we preferred to let users decide on the relevance of the pages so as to see if the pages really satisfy their requests. We confined the relevance judgment to only the first five results so to limit the required time and strength. Relevance was judged upon having visited and inspected each page. The web locations visited had to be from a different domain. So if two consecutive pages were on the same server only one of them was visited.

Table 5. User relevance estimate of the top 5 ranked results.

Search engine	Query 1	Query 2	Query 3
Google	2	4	4
Yahoo	4	2	4
MSN	1	2	3
Altavista	2	2	4
In	2	2	4
Pathfinder	2	2	4
Robby	0	0	0
Anazitisis	0	0	0

Table 5 presents the lower user relevance estimate. That is if a user considers that 4 results were relevant and another that 3 results were relevant for the same search engine then we adopt the opinion of the second one.

As we can see, the first 6 search engines had almost the same performance. An interesting point to make is that although MSN seems to have a recall lead over the others its precision is lower in every case. The last two engines should be left out of the next tests as they do not recall any relevant documents.

Capitalization

We then asked our participants to run the same query terms but this time in capital letters with no accent marks. 100% of them initially believed that they would get exactly the same results. However tables 6 and 7 failed this assumption and made users think that searching is more tricky and complicated than it should be.

Recall was dramatically diminished in the worldwide search enabling sites while it was left unaffected in the domestic ones. Precision was negatively affected as well. It must be underlined that the accurately pulled out sites were different in this test. As stated, 100% of our survey subjects, although half of them technologically proficient, expected no difference in the extraction outcome. This observation and the results of

our survey should alert us. What would happen if a searcher were to choose to search only in capital letters or without accent marks? Their quest would simply fail.

In English search there is no differentiation between capital and lower letters. The result sets are identical in both cases (see www.google.com/intl/en/help/ for example) so user effort is unquestionably less.

Table 6. Number of results retrieved when sample queries typed in capital letters.

Search engine	Query 1	Query 2	Query 3
Google	21	115	3
Yahoo	13	63	1
MSN	10	171	4
Altavista	13	63	1
In	237	320	45
Pathfinder	237	320	45

Wrapping up this experiment one can argue that in Greek Web searching the same query should be run several times, in lower and in capital letters, so as to improve the performance of the search. Sites where there are no accent marks or contain intonation errors will not be retrieved unless variations of the query terms are used. Greek search engines are superior at this point and make information hunting easier and more effective.

Table 7. User relevance estimate of the top 5 ranked results.

Search engine	Query 1	Query 2	Query 3
Google	1	3	0
Yahoo	1	3	0
MSN	2	3	0
Altavista	2	3	0
In	2	2	4
Pathfinder	2	2	4

Catalexis

Another factor that influences searching is the catalexis of the query terms. For example the second sample query could be altered from *Εθνική πινακοθήκη Αθηνών* to either *Εθνική πινακοθήκη Αθήνας* or *Εθνική πινακοθήκη Αθήνα* and still describe exactly the same information need. When someone searches using the second or the third query forms they will get quite different results. For example Google returns 1520 and 623 web pages respectively while it produced 896 in the first query variation. Precision is different in these three cases as well.

One could argue that such a difference is rational and acceptable as the queries differ. If we consider these queries solely from a technical point of view then this argument is right. However if the information needed is in the center of the discussion then these subtle differences in queries which merely differ in one ending should have recalled the same web pages. Stemming is an important feature of retrieval sys-

tems [1] (p. 167) and its application should be at least studied in spoken languages which have conjugations of nouns and verbs, like in Greek. Once again the majority of our users, 7/8-87.5%, was not aware of this fact and apparently could not benefit from it.

Stopwords

Google and other international search engines remove English stopwords so as to not influence retrieval. For instance users are informed that the word *of* is an ordinary term and is not used in the query *National Art Gallery of Athens*. The third Greek query of table 3 contains two quite ordinary words: *από* (preposition) and *τα* (article). These terms are not automatically removed and therefore affect retrieval. Removal of stopwords [1] (p. 167) is an essential part of a typical retrieval system.

Table 8. Quantity of results and relevance estimates in third query after stopwords removed.

Search engine	Result set	Relevant pages
Google	543	5
Yahoo	304	5
MSN	1075	3
Altavista	305	5
In	96	5
Pathfinder	96	5

Users were instructed to rerun the third query of table 2 but with the stopwords removed. Table 8 summarizes the number of results and the relevant pages. Comparing these results with tables 4 and 5 the reader can see a small recall increase and an important precision boost. The precision is measured only in the first 5 results and in almost every case the relevant document increased from 4 to 5, a 25% increase. This observation is important though more intense tests are required to construct a stopword list and to see how retrieval is affected by Greek stopwords.

4 Conclusions and Future Work

This paper presents a survey regarding utilization of search engines using Greek terms. The issues raised were evaluated and assessed by users themselves. The main purpose of our survey was the identification of inconveniences which affect user satisfaction and increase the required knowledge for successfully utilizing a Web searcher.

Our participants identified as highly important the adaptation of search engines to local settings. Additionally they considered that search engines with a dual role, that is search engines which are also WWW portals, is confusing and leads to increased downloading time, which in turn is frustrating and time consuming. Also results should be clearly presented with enough space among them and if possible with an

explanatory summary. These conclusions relate to the external engines's environment and affect a search engine's acceptance factor and the needed user effort.

In order to get an estimate of the internal features of search engines that support Greek, users had to run some sample queries. International search engines recalled more pages than the local ones and they had a small positive difference in precision as well. However they are case sensitive hindering retrieval of web pages which contain the query terms in a slightly different form to the requested one. Even if the first letter of a word is a capital letter the results will be different than when the word is typed entirely in small letters.

Endings and stopwords are not removed automatically, thus affecting negatively recall of relevant pages. Stopwords are removed from English queries making information hunting easier, looking at it from a user's perspective. Terms are not stemmed though even in English. However in a language with inclinations, like Greek, stemming may play an important role in retrieval assisting end users. In any case more intensive tests are needed to realize how endings, stopwords and capitalization affect retrieval.

As a general conclusion it can be argued that Greek users need to be more creative and knowledgeable than English searchers when utilizing search engines. This conclusion may apply to other spoken languages with similar characteristics. International search engines need to be more adaptable internally and externally to be truly multilingual, domestic should follow design principles applied in the worldwide ones and all of them should avoid performing multiple roles.

References

1. Baeza-Yates, R. Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, ACM Press, New York (1999)
2. Cleverdon, C.W., Mills, J., Keen, E.M.: An Inquiry in Testing of Information Retrieval Systems. Aslib Cranfield Research Project, College of Aeronautics, Cranfileld, U.K. (1966)
3. Robertson, S. E.: The Parameter Description of Retrieval Systems: Overall Measures. Journal of Documentation, (1969), 25, 93-107
4. Chu, H., Rosenthal, M.: Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. Proceedings of the Annual Conference for the American Society for Information Science, (1996), 127-135
5. Oppenheim, C., Morris, A., McKnight, C.: The Evaluation of WWW Search Engines. Journal of Documentation, (2000), 56 (1), 71-90
6. Dunlop, M.D.: Time, Relevance and Interaction Modelling for Information Retrieval. Proceedings of ACM/SIGIR, (1997), 206-213
7. Su, L. T., Chen, H. L., Dong, X. Y.: Evaluation of Web-based Search Engines from an End-User's Perspective: A pilot study. Proceedings of the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA., (1998), 348-361
8. Overton, R.: Search Engines get Faster and Faster, but not Always Better. PC World, (1996), 14. http: //www.pcworld.com/workstyles/online/articles/sep96/1409_engine.html
9. Lake, M.: 2nd Annual Search Engine Shoot-out: AltaVista, Excite, HotBot, and Infoseek Square off, (1997) http://www.zdnet.com/pccomp/features/excl0997/sear/sear.html
10. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London, England, (1979)

11. Salton, G, McGill M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company, Compute Science Series, New York, (1983)
12. Sullivan, D.: Nielsen NetRatings: Search Engine Ratings (2005) http://searchenginewatch.com/reports/article.php/2156451
13. Avouris, N.: Introduction in Human Computer Interaction (in Greek). Diavlos Editions, Athens, Greece (2000)
14. Ruthven, I., Tombros, A., Jose, J.: A Study on the Use of Summaries and Summary-based Query Expansion for a Question-Answering Task. ECIR 2001, (2001), 41-53

- Appendix A5 -

Automatic Extraction of Knowledge from Greek Web Documents

Fotis Lazarinis
Technological Educational Institute of Mesolonghi
30200 Mesolonghi, Greece
0030-26310-58148
lazarinf@teimes.gr

ABSTRACT
Extracting textual data from Greek corpuses poses additional difficulties than in English texts as inclinations and intonation differentiate terms of equal information weight. Pre-processing and normalization of text is an important step before the extraction procedure as it leads to fewer rules and lexicon entries, thus to less execution time and greater success of the mining process. This paper presents a system accessible via the Web which automatically extracts data from Greek texts. The domain of conference announcements is utilized for experimentation purposes. The success of the extraction procedure is discussed on the basis of an evaluative study. The conclusions and the techniques discussed are applicable to other domains as well.

Categories and Subject Descriptors
H.2.8 [**Database Application**]: Data mining

H.3 [**Information Systems**]: Information Search and Retrieval

Keywords
Web mining, information extraction, XML storage, multilingual retrieval

1. INTRODUCTION
Some recent studies showed that common search engines supporting Greek do not actually understand specific characteristics of the language [7, 8] so utilizing a general purpose search engine to discover specific information such as dates, keywords or even general purpose terms demand more effort by the user resulting also to lower success. This is mainly due to differences in Greek terms caused by inclinations, intonation and lower and upper case forms.

Proceedings of the sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)

In this paper we present a tool for extracting the title, keywords, event date, submission deadline and location of conference announcements. This tool is based on the identification of patterns and on knowledge lexicons (dictionaries) for extracting the previously mentioned data. Pre-processing and normalization of text is an important step before the extraction procedure as it leads to fewer rules and lexicon entries and to greater success of the mining process. Our main aim is not simply to build a system with extraction capabilities but to explore additional inconveniences and present solutions applicable in mining data from Greek corpuses which show considerable grammatical diversity although they carry the same information weight. The conclusions of this work could be applied to other spoken languages with similar characteristics to the Greek language.

2. EXTRACTING TEXTUAL DATA
Information extraction systems analyze unrestricted text in order to extract specific kind of information. They process documents trying to identify pre-defined entities and the relationships between them, filling a structured template with the mined information. Such systems have been implemented to extract data such as names and scientific terms from chemistry papers [2, 12]. Gaizauskas and Robertson [4] used the output of a search engine as input to a text extraction system. Their domain was management succession events and their scenario was designed to track changes in company management.

More contemporary work uses co-occurrence measurement in order to identify relationships and to extract specific data from Web pages [9]. Han et al [5] extract personal information from affiliation, such as emails and addresses, based on document structure. Efforts on Greek information extraction are recorded as well. In [11] a rule based approach to classify words from Greek texts was adapted. Rydberg-Cox [14] describes a prototype multilingual keyword extraction and information browsing system for texts written in Classical Greek. This system automatically extracts keywords from Greek texts using term frequency.

Our approach differs from the ones described in the previous paragraphs in that it tries to identify specific information based on rules and on vocabularies of rule activation terms. Also a technique for recognizing term relationships is explored. Additionally classic IR techniques such as suffix and stopword removal [1] are utilized and evaluated in Greek texts.

3. SYSTEM OVERVIEW

The relevant work done so far, focus mainly on English text neglecting other languages, which are more demanding and challenging in terms of recognition of patterns. In languages like Greek the same information may appear in many different forms, e.g. 11 Μαΐου 2005 or 11 MAIOY 2005 or Μάιο 11 or 11 Μάη 2005 (11 May 2005), and still convey exactly the same meaning.

In our system, information extracting relies on rule formalisms for each identified entity. Each extraction sub-procedure ends up with one of four alternative results:

(i) identified (IDN)
(ii) possibly identified (PDN)
(iii) not identified (NDN)
(iv) not applicable (NA)

Strong rule paths produce IDN results while weak rule paths end up in PDN. Strong rules are those which definitely identify the information that accurately falls into one of the known and well defined patterns. Weak rules are those who rely on probability and heuristic methods to infer the data.

Failing to identify some entity may be due to one of two reasons: *i.* A rule activates but it fails to complete, so the data is not identified because of our system's inability. These cases, denoted as NDN, could be used for retraining the system and eventually improve mining of data.

ii. The detection of an entity is not possible because it does not exist in the announcement. For example in preliminary announcements the exact conference's date is not yet decided. So NA, adopted by Morrisey's work [10], denotes nonappearance of the hunted piece of information. NDN and NA are preferred over null as they provide the system with different semantics which could be utilized for improving the system's functionality and the searching capabilities.

The extracted data form an XML file based on a short DTD. That way data can be presented in many different forms and utilized by other applications. In order to construct rules that will enable the successful extraction of the desired facts, we examined 25 text files, a small part of our collection consisting of 145 meeting announcements. This analysis allowed us to realize the different patterns the desired data follow and construct the rules. The remaining 120 call for papers were used in the evaluation.

3.1 Text Normalization

From the analysis of the textual data it was considered necessary to normalize the data first. Words are capitalized and accents or other marks are removed. In addition, simple suffix removal techniques were applied. The primitive Greek stemmer, which is analytically described in [8] removes final Greek sigma and transforms some endings such as "ει" and "ηκε" to "ω" among other mild transformations. It has been proved that the factors described in the previous paragraph influence searching of the Greek Web space as well [6, 7].

Abbreviations were automatically replaced by their full form. For example, month names appear abbreviated quite often, e.g Jun (Ιουν) stands for June (Ιούνιος). As a final normalization point, multiple spaces, html tags and other elements, which are not useful at this first version of the system, are removed. We should indicate though that html tags could prove significant especially in correctly identifying the title and the thematic area, as they provide structure to the information.

The normalization procedure leads to fewer rules and vocabulary entries, thus to less execution time and greater success in the mining process. In English text normalization procedure is simpler as there are no differences between upper and lower case forms, there are no inclinations of verbs and nouns (apart from minor differences between singular and plural forms) and accent marks are absent unlike in Greek.

3.2 Title extraction

Extraction of the title of a conference is based on heuristic rules. The basic idea is that titles appear on the top part of an announcement and they follow a "title" format, i.e. words are in capital letters or start with a capital letter, etc. Obviously normalization should be done after the identification of title as the form of words plays an important role here. Another rule employed is based on the surrounding text and in keywords, like conference, symposium, congress and meeting. As we will see in the evaluation section title identification is quite successful, though some extracted titles are truncated.

3.3 Keyword extraction

Correct identification of the title is also important for classifying the meeting. Classification means the detection of some keywords which describe the meeting. At the moment we base the classification on two techniques. We try to identify sort list of terms by discovering terms such as "conference topics".

Furthermore we explored a technique for constructing pairs of terms describing the conference. This technique is based on co-occurring terms [9]. We define co-occurrence of two terms as terms appearing in the same Web page. If two terms co-occur in many pages, we can say that those two have a strong relation and the one term is relevant to the other. Using words from the top part of an announcement we construct a list of pairs of neighboring terms. Then we try to measure the co-occurrence of these pairs. This co-occurrence information is acquired by the number of retrieved results of a search engine using the coefficient measure $r(a, b) = |a \cup b| / (|a| + |b| - |a \cup b|)$. With $|a|$ we symbolize the number of documents retrieved when we search using term a. Similarly $|b|$ is the number of documents relevant to term b and $|a \cup b|$ is the number of pages containing both terms. The co-occurrence is measured for every pair of terms and the top results are kept, based on a fixed cut off value. So if a conference is about New Technologies in Adult Education "in" is removed and the pairs "New Technologies", "Technologies Adult", "Adult Education" are formed. Then these pairs along with the terms "New", "Technologies", "Adult", "Education" are searched in the Web and the coefficient measure of the term pairs is decided.

Although our first heuristic approach performed well the second technique produced several "bad" instances among some useful two-term keywords. For example in a conference about "Educational Software" the keywords "Educational Games" were produced, which is acceptable and was not stated explicitly in the announcement, but the bizarre keyword "Adult Software" was also produced. Clearly this technique, although promising, needs certain refinements so as to be useful.

3.4 Extraction of dates

3.4.1 Conference's date

The first step in the identification of dates is the construction of a suitable vocabulary containing the normalized month terms that will activate the rules for the extraction of the conference's date. The identification of the date is based on a simple observation. The latest dates, appearing in a call for papers, are most probably the event's start and end dates. Our purpose is to recognize both start and end dates. For example from a date 11-13 June 2005 we extract 11 June 2005 as the start date and 13 June 2005 as the end date.

The date detection procedure initiates when a month or a full date (e.g. 12/05/2006) is found in the text. In that case we first check the succeeding words until the end of the sentence and then the preceding words until the beginning of the sentence. This search aims at identifying the day and the year of the conference and keywords which verify that it is actually the meeting's date. Thus the system needs to be able to keep information preceding and succeeding the rule activation keyword. If more than one date or date range is discovered then the system searches for appropriate keywords.

Rules are a set of *If then else* and *sub ifs*. Document is processed line by line and term by term. At the end of the rule formalism the result is stored in the XML repository. A simplified part of the date extraction procedure in pseudo code is shown below.

```
While not eof and date not identified do
    Separate current line to terms
    While not eof term set do
        Look up Vocabulary
        If month name is found then
            Scan Previous Terms
            Scan Next Terms
            If ... then
                ...
            Else if ... then
                ...
            End
        End
    End
End
Update conference XML Repository accordingly
```

3.4.2 Submission date

Submission date is trickier than the event's date as is absent in many cases, especially in short announcements. This procedure is complimentary to the previous one as dates which are denoted as meeting's start and end dates should not be checked again. After the extraction of a proper date the surrounding text is scanned for words like deadline (υποβολή), or other synonyms. Clearly these rules are domain dependant and have a high error probability. This procedure ends up mostly with one of the codes PDN, NDN, NA.

3.5 Location extraction

For extracting the location we constructed and utilized an ontology with the major Greek cities and the prefecture in which they belong. This listing also models bordering city and county relations. A city's name will trigger off the rules for the identification of the desired information. It was proved that normalization of locations names is absolutely essential as they appear in many different forms, e.g. Αθήνα, Αθηνών, Αθήνας (Athens). One problem in the identification of the location arises when a conference is co-organized by more than one institutions. In this case many locations co-exist. Mining is then based on the surrounding context or on the location's tf (term frequency) measured in the whole announcement. If a strong decision is made then the procedure ends up, whereas when a weak decision is made the procedure initiates again when new activation terms appear up.

4. SYSTEM ARCHITECTURE

The system is implemented in Java using JSP and Servlets. For processing the textual information a version of the jflex utility (http://jflex.de) is used. A flowchart of the system is shown in figure 1. The conference announcement is submitted either as a url pointing to an html file or it pasted in a text box on the system's web page.

The extracted information is stored in an XML file which is then accessible by the retrieval component of the system. This component, which is currently under development, dynamically forms an index of the processed conferences based on the information found in the XML repository. When projected to the client's browser conferences are classified as open or past and they are categorized based on their date. This tool will also allow multiriteria retrieval of conferences, such as "show me conferences in Athens or near Athens which are about Web mining and will take place this summer". Supporting these queries will be based on the location knowledge base and on the month dictionary.

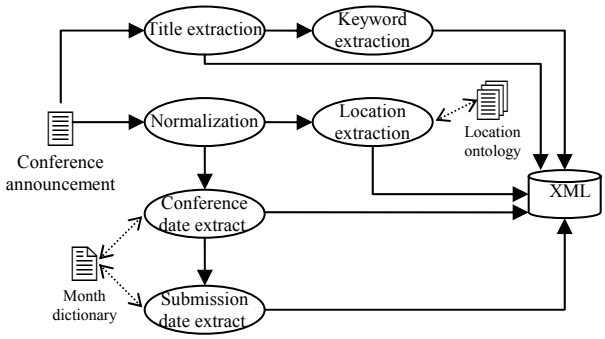


Figure 1. Flowchart of the extraction procedure.

5. EVALUATION

The performance of an information extraction system can be measured using Precision (P) and Recall (R) [13], as in Information Retrieval systems. Precision measures the ratio of the correctly extracted information against all the extracted information. Recall measures the ratio of the correct information extracted from the texts against all the available information. Despite the diversity of the collection the system works adequately well and the employed rules achieve high rates of precision and recall, especially in the attributes where a dictionary is used.

Table 1. Precision and Recall of the extraction procedure

	Title	Keyword	Conf date	Subm date	Location
Correct	77	39	107	89	110
Wrong	29	65	8	19	7
Not extra	14	16	5	12	3
Precision	72,64%	37,50%	93,04%	82,41%	94,02%
Recall	64,17%	32,50%	89,17%	74,17%	91,67%

The results of the evaluation are summarized in table 1. As expected, title and keywords show a higher error percentage. Clearly more sophisticated rules are needed. A possible solution would be the exploitation of tagging information and the usage of lexicons which model domain relationships as well. It should be noted that partially extracted titles, even those with only one not identified word, were accounted as erroneously extracted. So with slight improvements we can achieve higher precision and recall. Date and location rules achieve high precision and recall scores. Their extraction is relying on specific word lists and they follow better structured patterns.

In order to realize the effects of normalization and to get an indication of the additional difficulties posed in Greek we evaluated the system's performance, on date, submission date and location extraction, without extensive normalization. That is words were only capitalized and short forms replaced by their full forms. The evaluation showed that system's precision reduced by more than 30%. It could be argued that in this case more rules should be employed in order to achieve higher precision. While this could be partially true, we need to take into account that more rules means increased execution time as more searches are needed and a higher error probability as more heuristics and weak rules will be employed.

A final evaluation task was performed utilizing Google. A set of five queries concerning specific locations and a second set concerning dates consisting of months and years were run in our collection using Google. Then we evaluated the precision of each query (tables 2 and 3). Clearly Google retrieves many irrelevant files which diminish precision and recall. This is because every file containing the query terms or one of them is retrieved. Furthermore, announcements where terms appear in different forms than the requested ones are not retrieved. In our tool vocabularies act as thesauri as well allowing retrieval of meetings where locations or month names appear in another form or inclination. Of course tables 2 and 3 show an initial estimation. A more thoroughly designed evaluation is needed with more queries to safely reach useful conclusions.

Table 2. Precision and Recall of location queries in Google

Location	Precision	Recall
Query 1	57,50%	76,00%
Query 2	42,86%	83,33%
Query 3	77,78%	83,33%
Query 4	55,88%	64,29%
Query 5	50,00%	65,71%

Table 3. Precision and Recall of date queries in Google

Date	Precision	Recall
Query 1	42,31%	60,00%
Query 2	32,14%	52,38%
Query 3	43,75%	75,00%
Query 4	40,63%	50,00%
Query 5	37,50%	54,29%

6. SYNOPSIS AND FUTURE WORK

This paper presents an under development system which automatically extracts data from Greek conference announcements. Five categories of data are mined utilizing various techniques and approaches. For the first two categories rules are based on text's position, on context surrounding the information and on a coefficient measure. The last three types of data are mined with the utilization of lexicons which contain rule initiation terms. Then the surrounding text is again exploited. It was shown that simple removal of endings and accents and other adjustments, specific to Greek language, improve the extraction procedure and lead to increased Precision and Recall and to less elaborate rules. Vocabularies act as thesauri permitting retrieval of text where terms appear in different forms than the requested ones.

However more work needs to be done in order to achieve high rates of precision. Tagging and formatting information should be utilized in the identification of complex textual information. Metadata and link tracking, in the case of html or xml files, could be utilized. Links usually point to more detailed announcements in which all the data are applicable. Domain vocabularies are necessary in order to identify classification terms. Also, when fully developed, the system should be evaluated against the existing manual or semi automatic conference engines so as to realize all the advantages of our automated system.

Ultimately we aim at building a more complicate system which continually scans the Web to find future conferences, symposiums and congresses. From this combined system XML descriptions of the events could be produced which in turn could be utilized in automatically constructing conference announcement indices. These Web pages will be thematically sorted and automatically and regularly updated, with advanced searching capabilities thus enabling users to find everything in one place. Many issues related to information retrieval are open in the intended system, from categorization of events to summarization and to multicriteria and multilingual retrieval.

7. REFERENCES

[1] Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, ACM Press, New York, 1999.

[2] Chowdhury, G. G., Lynch, M. F. Automatic interpretation of the texts of chemical patent abstracts, part 1: lexical analysis and categorisation. *Journal of Chemical Information and Computer Science*, 32, (1992), 463-467.

[3] Cowie, J, Lehnert, W. Information extraction. Communications of the ACM, 39, (1996), 80-91.

[4] Gaizauskas, R., Robertson, A. Coupling information retrieval and information extraction: a new text technology for gathering information from the web. In *Proceedings of the RIAO'97 Conference*, (Canada), 1997, 356-370.

[5] Han, H., Giles, L. C., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.. Automatic document metadata extraction using support vector machines. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*, 2003, 37-48.

[6] Lazarinis, F. Do search engines understand Greek or user requests "sound Greek" to them? In *Open Source Web Information Retrieval Workshop* (in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology, France), 2005, 43-46.

[7] Lazarinis, F. Evaluating user effort in Greek web searching. In *Proceedings of the 10th PanHellenic Conference in Informatics* (University of Thessaly, Greece), 2005, 99-109.

[8] Lazarinis, F. Old information retrieval techniques meet modern Greek Web searching. In *Data Mining and Information Engineering Proceedings, 2006* (accepted)

[9] Mori, J., Matsuo, Y., Ishizuka, M., Faltings, B. Keyword extraction from the web for foaf metadata. In *1st Workshop on Friend of a Friend, Social Networking and the Semantic Web* (1-2 September 2004, Galway, Ireland), 2004.

[10] Morrissey, M. J. *A treatment of imprecise data and uncertainty in information systems*. PhD Thesis, Department of Computer Science, University College, Dublin, Ireland, 1987.

[11] Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C. Resolving part-of-speech ambiguity in the Greek language using learning techniques, In *Proceedings of the ECCAI Advanced Course on Artificial Intelligence* (ACAI, Chania, Greece), 1999.

[12] Postma, G. J., Van der Linden, J. R., Smits, J. R. M., Kateman, G. TICA: a system for the extraction of analytical chemical information from texts. In Karjalainen E J (ed) *Scientific Computing and Automation*. Elsevier, Amsterdam, 1990, 176-181.

[13] Robertson, S. E. The parameter description of retrieval systems: overall measures. Journal of Documentation, 25, 1969, 93-107.

[14] Rydberg-Cox, A. J. A prototype multilingual document browser for ancient Greek texts. *The New Review of Hypermedia and Multimedia*, 7(1), 2002, 103-113.

- Appendix A6 -

HOW DO GREEK SEARCHERS FORM THEIR WEB QUERIES?

Fotis Lazarinis
Department of Applied Informatics in Management & Finance
Technological Educational Institute, Mesolonghi 30200, Greece
lazarinf@teimes.gr

Keywords: Web searching, search engine evaluation, web queries, Greece

Abstract: This paper presents an initial analysis of a large log of Greek Web queries. The main aim of the study is to understand how users form their queries. The analysis showed that users include terms of low discriminatory value and form their queries in various non lemmatised forms. Lower case queries are the most common case, although several query instances are in upper case. Accent marks are usually left out by query terms. These conclusions could be utilized by local and worldwide search engines so as to improve the services offered to the Greek Web community and to users of other morphologically complex natural languages.

1 INTRODUCTION

Searching the Web is a daily activity of almost all Internet users. Users form their queries in various manners and it has been argued that this may depend on the nationality and cultural background of the user (Jansen and Spink, 2005). There is a growing body of research examining the search patterns of users of predominantly US search engines (Silverstein et al., 1999; Jansen & Pooch, 2001; Spink et al., 2002). All these studies focus on understanding about what topics people search for and how short or long are their queries. Clearly this is important, as search engines could be refined based on their findings. However one of the limitations of these studies is that they focus mainly on English Web queries or more general in queries based on the Latin alphabet. In languages with different alphabets, like Greek or Russian or Arabic, additional difficulties could be raised by the way users form their queries. In these languages capitalization or diacritics in query terms plays an important role in relevance of documents (Moukdad, 2004; Bar-Ilan & Gutman, 2005, Lazarinis, 2005; Lazarinis, in press).

In this study we focus on the Greek language and try to understand how users form their Web queries. By identifying the query patterns we will eventually be able to suggest improvements to search engines so as to better adapt to and handle Greek queries. The findings of our statistical analysis may be directly applicable to other languages with non Latin alphabets, and noun, adjective and verb declensions.

2 THE STUDY

2.1 Data Collection

The query data were obtained from four Greek academic institutions. The user search strings of specific departments are accessible via the Web and they were analyzed statistically in our study. Data of the last 12 months (November 2005-October 2006) were assembled to form our user query data collection. These queries were redirected to Google or Yahoo through the local search engines of the academic departments. Queries were submitted by members of the Academic staff and by students.

In total, 48 html files were examined containing 211,172 unique queries. 205,474 of these search strings were in English and the remainder 5,698 queries were in Greek. In some cases the Greek queries contained English terms as well. In the following sections we focus on and analyze the Greek search strings.

2.2 Data Analysis

The html files contained the query strings and some statistics. We did not analyze or utilize the existing

statistics which focus mainly on the number of times and on the time and the date a query has been submitted. Motivated by some of our previous work on the theme of Greek Web retrieval (Lazarinis, accepted) and the work of Jansen and Spink (2005), we analyzed the data from a number of different angles. The data analysis and the conclusions of each test are presented below.

2.2.1 Query length

As seen in Table 1, the majority of queries (66.95%) contain 2 or 3 words which is an indication that users are aware that 1-word queries are usually too broad to retrieve reliable results. On average, each of the 5,698 queries is consisted of approximately 2.47 terms, i.e. 14096 in the 5,698 queries.

Table 1: Lengths of Greek queries.

Number of words	Number of queries	
	n	%
1	1,005	17.64
2	2,275	39.93
3	1,540	27.03
4	619	10.86
5	178	3.12
6+	81	1.42

2.2.2 Lower and Upper case

Capitalization of query terms is an important factor in retrieval of Web documents. Lazarinis (submitted) showed that international search engines like Yahoo, MSN and even Google, retrieve different numbers of pages with different precision in lower and upper case queries. In our sample, 1,028 (18.04%) queries were in upper case and 4,670 (81.96%) were in lower case or in title case (i.e. first letter of each word was capitalized). There was no difference in the distribution of query lengths in upper and lower case so as to make any valid inference. However it seems that upper case queries are finer grained as they are usually abbreviations or titles or person and organization names. In these cases retrieval is probably more effective.

In any case, the percentages of lower and upper case queries show that although users search using lower case terms mostly, a considerable number of queries are in upper case. In English Web searching there is no differentiation between results in upper and lower case queries. In Google and Yahoo, for example, the queries "Ancient Athens" and "ANCIENT ATHENS" retrieve the same number of Web documents ranked identically. However, in Greek the queries "Αρχαία Αθήνα" and "ΑΡΧΑΙΑ ΑΘΗΝΑ" retrieve different numbers of Web pages and therefore it is up to the Greek users to run the queries in both forms to get the maximum number of relevant documents.

2.2.3 Accent marks

The Greek language is a morphologically complex language compared to English and to some of the European languages which are based on the Latin alphabet. Modern Greek words use accent marks and umlaut in vowels in lower case letters. In capital letters accent marks are not regularly used.

It has been reported that when diacritics are absent, precision drops significantly in Web searching (Lazarinis, accepted). Table 2 illustrates that 46.21% of the lower case queries contain at least one word without accent marks and that more than 1/4 of the query sample are typed entirely without accent marks. 5,251 out of the total 11,700 (44.88%) words of the lower case queries had no diacritics.

Table 2: Number of user queries without diacritics.

Queries with all words without diacritics	Queries with at least one word without diacritics
1,542 – 27.06%	2,633 – 46.21%

The problem is more serious in the case of umlaut. By searching the query sample we found 6 variations of the word "Ευρωπαϊκή" (European). 5 of these variations were typed without umlaut. This is maybe to user lack of knowledge of how to input umlaut in vowels. In any case it influences negatively the recall and relevance of pages. For instance, in Yahoo the word "Ευρωπαϊκή" retrieves 1,250,000 pages, the term "Ευρωπαική" 33,400 and the term "Ευρωπαικη" 32,300 pages. In the latter two queries relevance in the first 10 results is significantly lower than the normal form. Google has identified this difference and retrieves the same pages in all three variations.

2.2.4 Lemmatised form

The query "Bookshop New York" retrieves pages having as matching terms the words "Bookshops", "Book", "Books" and "Bookstore" in Google. In other words synonyms and lemmas of a word are matched to the query terms to help the searcher locate more relevant pages.

Nouns, adjectives, verbs and even first names have conjugations in Greek (nominative, genitive, etc). Lemmatization involves the reduction of words to their respective headwords (i.e. lemmas). For

example, the terms "speaks" and "speaking", resulting from a combination of a sole root with two different suffixes ("s" and "ing"), are brought back to the same lemma "speak".

With the aid of a dictionary we calculated that 4,135 lower case queries were not in lemmatised form (Table 3). The percentage is lower in upper case queries (31.03%) as most of these terms are abbreviations or person and organization names (see Table 3). Subtle differences in queries (e.g. "Πανεπιστήμιο Αθήνας", "Πανεπιστήμιο Αθηνών" – University of Athens) are capable of differentiating the retrieved pages in Google, Yahoo and in the other international and even national search engines, which supposedly have a better understanding of the Greek language.

Table 3: Number of non lemmatized queries.

Lower case queries in non lemmatised form	Upper case queries in non lemmatised form
4,135 – 88.54%	319 – 31.03%

Lemmatization would be quite helpful in Greek Web searching since most of the queries and obviously Web pages are not in lemmatised form and their matching is apparently not possible.

2.2.5 Stopwords

Stopwords are the terms which appear too frequently in documents and thus their discriminatory value is low (van Rijsbergen, 1979). Elimination of stopwords is one of the first stages in typical information retrieval systems. In English Web searching stopwords are removed or they do not influence the retrieval process significantly. Stopword lists have been constructed for most of the major European languages (see http://snowball. tartarus.org for example) and they could be utilized by search engines. Such a listing does not exist for the Greek language. Usual candidates of the stopword list are articles, prepositions and conjunctions (Baeza-Yates & Ribeiro-Neto, 1999).

Using all 5,698 lower and upper case queries we identified the articles, prepositions and conjunctions existing in our query collection. Such common words exist in 1,516 queries. That is 26.61% of the queries contain common words. These words occurred 2,032 times within these 1,516 queries. Thus they account for the 14.42% of the total words of the Greek queries.

These statistics indicate that users do utilize common words in their queries and therefore the construction of a Greek stopword list and its application to Web retrieval should be further studied.

2.2.6 Other Issues

Although the analysis of the data is still in progress, the most important issues were discussed above. A number of other issues were also identified by observing the user queries but they have not been thoroughly examined as yet.

A number of queries in the English part contained the string "www" or were in a semi url form. For instance, a user typed the query "travel to Greece.gr". This is an indication that some users are not competent in search engine usage. Proper training or presentation of proper examples on the search engine's page could help users work out their misconceptions.

By inspecting the first 100 queries of the sample we located 3 spelling errors. We run these queries in Google and we got either no results or pages with the same spelling errors as in the query. International search engines aid English users even in spelling errors with "Did you mean" tips. For instance, Yahoo presents the message "Did you mean: confidentiality" if a user types the word "confidentiallity" in its searching box.

In 12 Greek queries the "*" wildcard was used at the end of the query. As known, users get no additional results if they use wildcards. Additionally, the wildcard was not properly used as a space was included between the wildcard and the last word. This observation, along with the inclusion of "www" in the queries, is an indication that a few search engine users are confused and therefore training is needed.

"GreekEnglish" is a term shared among Greek Internet users. It refers to the typing of Greek words using English characters. For example, the word "Athina" in GreekEnglish, is the word "Αθήνα" in Greek and "Athens" in English. GreekEnglish originates from the time Greek were not supported in some operating systems or in e-mail clients and it was invented as a communication means so as to assure readability. Several users still follow this logic. We observed several instances of GreekEnglish queries in our sample. However, it cannot be decided whether it was a conscious action or this behavior results, again, from user misconceptions about the ability to use or not Greek characters in searching.

Advanced options such as site or file specification were sporadically detected. However, we cannot derive valid conclusions from this finding

since queries are submitted to Google and Yahoo through the local search engine. So advanced options are not immediately visible and available to these users.

3 CONCLUSIONS

This paper presents the initial analysis of a large query log. Although the analysis is not complete as yet some important findings resulted from this study. It is easily understood that Greek users include common words and form their Web queries in various declensions. Lower case queries are the most common case, although several query instances were in upper case. Accent marks are usually left out. By observing the queries we realized that, as anticipated, users do some spelling errors and they erroneously use wildcards and other not proper characters or strings.

These facts affect negatively Web searching using Greek terms. Some of these problems have been effectively dealt by Google. However the techniques which could substantially reduce user effort and have already been applied in English searching are not adapted to the Greek language. Probably similar problems are faced by other non Latin users. Search engines should try to value these natural languages. One way to achieve this is through the user queries.

REFERENCES

Baeza-Yates, R., & Ribeiro-Neto, B., 1999. *Modern Information Retrieval*, Addison Wesley, ACM Press. New York.

Bar-Ilan, J., Gutman, T., 2005. How do search engines respond to some non English queries? *Journal of Information Science, 31*(1), 13–28.

Jansen, B. J., Pooch, U., 2001. Web user studies: a review and framework for future work. *Journal of the American Society of Information Science and Technology*, *52*(3), 235–246.

Jansen, B., Spink, A., 2005. An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management, 41*, 361–381.

Lazarinis, F., 2005. Do search engines understand Greek or user requests "sound Greek" to them? In *Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology,* pp. 43-46.

Lazarinis, F., in press. Evaluating the searching capabilities of Greek e-commerce Web sites. *Online Information Review Journal*.

Lazarinis, F., accepted. Web retrieval systems and the Greek language: Do they have an understanding? *Journal of Information Science*.

Moukdad, H., 2004. Lost in Cyberspace: How do search engines handle Arabic queries? In *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, Winnipeg.

Silverstein, C., Henzinger, M., Marais, H., Moricz, M., 1999. Analysis of a very large Web search engine query log. *SIGIR Forum*, *33*(1), 6–12.

Spink, A., Jansen, B. J., Wolfram, D., Saracevic, T., 2002. From e-sex to e-commerce: Web search changes. *IEEE Computer*, *35*(3), 107–111.

van Rijsbergen, C.J., 1979. *Information Retrieval*, Butterworths. London, 2nd edition.

JIS

Web retrieval systems and the Greek language: do they have an understanding?

Fotis Lazarinis

Technological Educational Institute, 30200 Mesolonghi, Greece

Abstract.

Searching the web is a common activity of web users. English and non-English speakers utilize international or local search engines so as to satisfy their information needs. Most of the attempts at evaluation of search engines focus on English queries and on English document collections. In this paper an evaluation methodology is presented and the capabilities of international and local web retrieval systems using Greek queries are evaluated based on this method. We aim at identifying difficulties and knowledge requirements when using a Greek supporting search engine. The importance of interface localization and the effects of standard information retrieval techniques such as case insensitivity, stopword removal and simple stemming are studied in international and local search engines. The evaluation methodology is applicable to other non-English natural languages as well.

Keywords: Search engine evaluation; non-English retrieval; stemming; stopwords; Greek

1. Introduction

The world wide web has gained great popularity and has become one of the most widely used services of the internet along with email. The web has gained such publicity that many people erroneously equate it with the internet. The friendly interface and the hypermedia features of the web attract a significant number of users around the globe. As a result, the web has become a pool of various types of data, dispensed in a measureless number of locations. Finding information that satisfies specific criteria is a regular daily activity of almost every web user. Web search engines provide searching services through their uncomplicated interfaces.

According to Global Reach [1], 64.2% of the online population are non-English speakers. This makes the web a multicultural and multilingual information space. The preferences and requests of non-English speaking users should undoubtedly be taken into account in the design of any web information system and especially in web retrieval systems since these are utilized on a daily basis by virtually every web surfer.

Correspondence to: Fotis Lazarinis Technological Education Institute, 30200 Mesolonghi, Greece. E-mail: lazarinf@teimes.gr

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

Even though several web search engines exist to facilitate searching, not enough attention has been given to other spoken languages than English. Efficient search engine utilization requires an increased level of knowledge on the part of users. This is because most search engines have no internal (indexing) or external (interface) localization support and thus the user has to devise alternative ways to discover the desired information. Baeza-Yates and Ribeiro-Neto [2, p. 391] suggest teaching users methods for effective utilization of retrieval systems. Clearly this is not a feasible solution in our case, as the potential student target group would be enormous. Therefore the shortcomings of search engines should be identified and efforts should be made in order to amend them.

The purpose of the present paper is to create a methodology for identifying some of the deficiencies of searching the web using non-English queries. The criteria of the methodology are applied in Greek web searching as an initial evaluation experiment. The paper is structured as follows. Section 2 provides an overview of the criteria used in the evaluation of web retrieval systems and reviews the literature related to non-English web searching. Section 3 presents and analyses the evaluation methodology and Section 4 presents the results of applying its criteria using Greek queries. Finally, the last section synopsizes the results of the evaluation experiment.

2. Literature review

2.1. Evaluation of search engines

A number of criteria have been proposed for the evaluation of information retrieval systems (coverage, time lag, recall, precision, presentation, user effort) [3, 4]. Of these criteria, recall and precision have most frequently been applied in measuring information retrieval. Information retrieval on the web is fairly different from retrieval in traditional indexed databases. This difference arises from the high degree of dynamism of the web, its hyper-linked character, the absence of a controlled indexing vocabulary, the heterogeneity of document types and authoring styles, and the easy access that different types of users may have to it [5].

Therefore the criteria have been reshaped to fit in the dynamic web environment. The capabilities of three search engines, AltaVista, Excite, and Lycos have been evaluated in terms of five aspects [6]:

(1) Composition of web indexes (coverage) – collection update frequencies and size can have an effect on retrieval performance.

(2) Search capability – they suggest that search engines should include 'fundamental' search facilities such as Boolean logic and scope limiting abilities.

(3) Retrieval performance (precision, recall, time lag) – such as precision, recall, and response time.

(4) Output option (presentation) – this aspect can be assessed in terms of the number of output options that are available and the actual content of those options.

(5) User effort – how difficult and effortful it is for typical users to use the search engine.

Most search engine evaluation attempts focus on the third criterion. For example in [7] eight search engines were reviewed and their effectiveness was calculated based on the traditional information retrieval measures of recall and precision at varying numbers of retrieved documents. Dunlop [8] used the expected search length to construct graphical evaluation methods to measure retrieval performance from AltaVista. These graphs were introduced as supplementary to precision-recall graphs. AltaVista, Infoseek, Lycos, and Open Text were used in another evaluation study [9]. The authors employed measured precision and partial precision for the first 20 hits returned by the search engines. They also defined an evaluative measure that compared ratings of relevance on a five-point scale. Similar approaches have been used in more recent studies [10, 11]. Other research papers focus additionally on other issues such as the search interface and the response pace of search engines [12].

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

2.2. Non-English web retrieval

Although the studies reviewed in the previous section provide frameworks and models for evaluating the capabilities of search engines they usually focus on precision and recall, neglecting other factors such as user effort, for instance, and more importantly they focus only on English queries. It has been argued that existing search engines may not serve the needs of many non-English-speaking internet users [13]. The latter observation proves that the multicultural and multilingual dimensions of the web have been overlooked, especially in search engines. That is why a few recent studies have assessed web retrieval systems taking into consideration the language of the users and focused on non-English and non-Latin queries.

Polish supporting search engines were examined in [14]. Polish versions of English language search engines and homegrown Polish search engines were assessed. The searching capability and retrieval performance were considered. Major emphasis was given to the precision criterion, which was based on relevance judgments for the first 10 matches from each search engine. Of the five search engines evaluated, Polski Infoseek and Onet.pl had the best precision scores, and Polski Infoseek turned out to be the fastest web search engine.

The performances of general and Arabic search engines were compared based on their ability to retrieve morphologically related Arabic terms. The findings highlight the importance of making users aware of what they miss by using the general engines, underscoring the need to modify these engines to better handle Arabic queries [15].

Experimentation with Russian, French, Hungarian and Hebrew queries revealed some of the inefficiencies of worldwide search engines related to issues such as capitalization and singular and plural forms of query terms [16]. Their results indicate that in the examined cases the general search engines ignore the special characteristics of non-English languages, and sometimes they do not even handle diacritics well.

Another research article explored the characteristics of the Chinese language and how queries in this language are handled by different search engines [17]. Queries were entered in two major search engines (Google and AlltheWeb) and two search engines developed for Chinese (Sohu and Baidu). Criteria such as handling word segmentation, number of retrieved documents, and correct display and identification of Chinese characters were used to examine how the search engines handled the queries. The results showed that the performance of the two major search engines was not on a par with that of the search engines developed for Chinese.

The capabilities of the local Greek search engines of e-commerce sites were reviewed in [18]. This study focused mostly on the existence of search engines and on interface issues. Yet a few inefficiencies of the local e-shops' search engines related to the attributes of the Greek languages were revealed. For instance most of the search engines are case sensitive and let stopwords negatively influence the retrieval of products. In [19] an initial evaluation of the capabilities of web search engines revealed some of the deficiencies of international and domestic search engines in Greek queries.

All these studies try to understand and identify the inefficiencies of search engines with respect to non-English and non-Latin languages. They also try to understand the regional differences and trends in web searching [20]. Additionally, CLEF experiments aim to test, tune and evaluate information retrieval systems operating in European languages in both monolingual and cross-language contexts [21].

The previous research papers and experiments reveal a lot of the qualities and inefficiencies of stand alone information retrieval systems and search engines in non-English queries and try to engineer algorithms for increasing the effectiveness of the retrieval systems. However, each study assesses web searching information systems from a different perspective, although some criteria are common. In this paper we focus on creating and applying a generalized evaluation methodology restricted to search engines only. This methodology combines interface issues, e.g. adaptation to the local language, with searching effectiveness, e.g. case insensitivity or effect of removal of stopwords. The methodology is presented and explained and then it is applied to evaluate the capabilities of Greek supporting web search engines. This framework can serve as the basis for evaluating the effectiveness of web retrieval systems in non-English text retrieval. Another difference with the previous

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

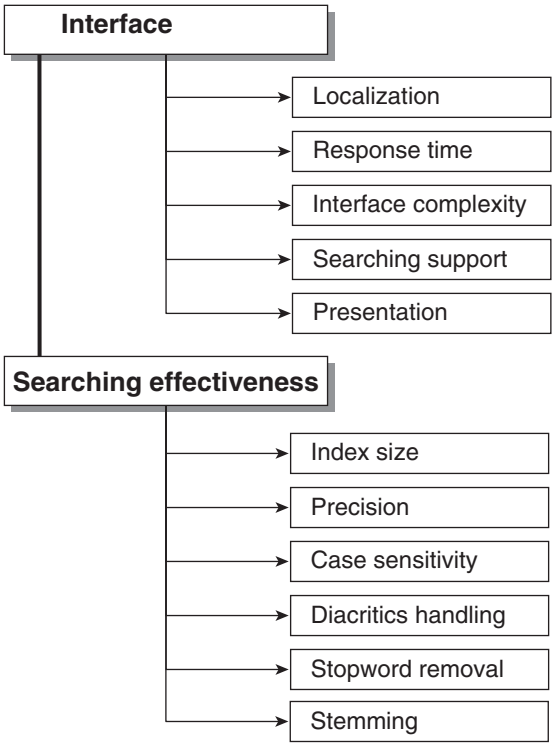


Fig. 1. Criteria of the evaluation methodology.

studies is that formation of sample queries and assessment of specific characteristics of the search engines are performed with the aid of users in authentic environments.

3. Evaluation methodology

The evaluation method suggested in this paper consists of two classes of attributes:

(1) interface; and

(2) searching effectiveness.

These sets of attributes are analysed further as shown in Figure 1. The criteria of the proposed assessment procedure are collected from the previously presented relevant studies and their aggregation aims at constructing a compact yet efficient model for measuring the 'understanding' of international and local search engines with respect to a specific language. The criteria of the evaluation procedure are quantitative or qualitative. Some of them are measured by experts and some are measured by real users searching for specific information.

The aim of the presented model is twofold. On the one hand we aim at identifying the qualities and shortcomings of search engines in non-English queries and on the other hand we intent to investigate the effects of the standard information retrieval techniques in web retrieval. Our basic objective is to be able to suggest improvements, based on the findings of the evaluation, in search engines so as to increase their searching effectiveness and reduce the required user effort in monolingual non-English queries.

The criteria assembled under the 'Interface' heading aim at measuring the intuitiveness, the simplicity and the speed of response of international and domestic search engines. They all relate to language issues as some search engines may present their results in a readable and clear form in English but this presentation may be problematic in another natural language with more accent

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

marks or with a non-Latin alphabet. Or, for example, response time is an important issue for users when choosing between a local and an international search engine. Support for searching in a particular language is clearly an issue related to language and perhaps to complexity as well, as specific domestic search engines tend to act as web portals offering searching facilities as well, so their interface is more complicated. The 'effectiveness' class groups a number of attributes which are important when searching. These attributes relate to the language used in searching and the aim of their grouping is to identify how they influence the retrieval process.

3.1. Analysis of the evaluation methodology

A brief explanation of each attribute and its assessing method is given below.

Localization is an indicator that refers to the ability of a search engine to adapt its interface to the local language. A search engine may be denoted as: *not localized, partially localized, fully localized*. The first value means that a search engine does not adapt at all to the language of the interest, the second type refers to the adaptation of certain interface parts and services of the search engine and the last indicates that all the provided services and interface components are localized.

Response time is a quantitative measure and can be analysed into two sub-categories: the time to load the initial search engine's web page, and the time required to retrieve the relevant set of documents. This attribute can be mechanically measured using the same internet connection speed and a number of queries that can be used to measure the average retrieval time.

Interface complexity refers to the information presented in the initial web page. A number of search engines act as web portals. This approach may lead to increased downloading time, which can be irritating when the speed of the internet connection is low. Additionally they may cause confusion and disorientation to users as the textbox where the query is typed and the procedure's initiation button are not easily viewable. Interface complexity can be assessed by users themselves as our opinion would be subjective due to our expertise in utilizing search engines.

Searching support for other languages than English is obviously an essential attribute. Some search engines do not handle non-Latin queries and some may not handle effectively terms with diacritics in natural languages which are based on the Latin alphabet. This attribute is noted as *supported* or *not supported*.

Presentation is an attribute used in assessing standard retrieval systems and web search engines. In our work this feature is related to the presentation of the potentially relevant documents. This attribute is qualitative and not quantitative and is used to assemble the observations and problems raised by users.

Index size is an element which cannot be conclusively measured unless the search engine has revealed the actual index size. But even then this number would have to be divided according to the language codes of the pages contained in the index. Since this is not possible, the only way to get a rough idea is by running some sample queries in different search engines. The recalled set of documents will then provide an estimate of the index size of each search engine.

Precision (relevance) is a standard measure used in information retrieval systems [2, 4]. Here precision can be measured at specific recall points. In other words, as in previous studies, precision can be measured in the top ranked documents [6]. For example, it can be calculated in the first 10 or 20 results which hold the highest possibility to be viewed by users [22].

Case sensitivity is a feature that does not affect English web searching. For uexample the queries 'olympic games' and 'OLYMPIC GAMES' produced exactly the same results in Google. However the results differ between the queries 'ολυμπιακοί αγώνες' and 'ΟΛΥΜΠΙΑΚΟΙ ΑΓΩΝΕΣ' (both queries mean 'Olympic games' in Greek). Assessment of this attribute is objective as it can be noted as *supported* or *not supported*.

Diacritics handling concerns the intonation marks and other accent marks, such as umlaut, which many spoken languages support. For example, the term 'European union' is written in Greek as 'Ευρωπαϊκή ένωση'. Both intonation and umlaut are used. Other languages, like French or Serbian, contain more accent marks. Search engines should be able to handle diacritics to efficiently support user requests. Efficient handling of diacritics is important as diacritics may change the meaning of two morphologically equal terms. For instance, in Greek the word 'μόνο' means only and alone while the word 'μονό' means single. These Greek terms differ only in the position of the accent mark.

Stopword removal is supported by Google and other international search engines in English queries. For instance users are informed that the word 'of' is an ordinary term and is not used in the query 'National Art Gallery of Athens'. Removal of stopwords [2] (p. 167) is an essential part of typical information retrieval systems. Although significant relevant work has been performed in English information retrieval and suitable stopword lists have been constructed, such stopwords lists have not been constructed for most of the other major European, Asian and African languages. Thus the effect of stopwords in retrieval has not been thoroughly studied in these languages. A possible way to study the influence of stopwords in web retrieval is by running composite queries containing both significant terms and stopwords and then running the same query without the stopwords. This way one could get an initial estimate of the positive or negative influence of non-significant words in web retrieval and realize if an international search engine values all the attributes of a language.

Stemming is the process of reducing a word to its stem or root form. This procedure equalizes the morphological variants of words that have similar semantic interpretations. This feature is partially supported in Google. For example the query 'evaluating web sites' retrieves documents which contain the terms 'evaluate web sites' or the terms 'evaluation websites' as can be concluded from the highlighted matching terms of the relevant documents. In web retrieval, stemming may lead to recall of countless web documents and thus may be an inapplicable technique. However, Greek, and other languages, exhibit notable morphological variance in terms while the content remains the same. This is due to tense, noun and adjective inflections, plural and singular forms and composite words. For example, all three queries 'Εθνική πινακοθήκη Αθηνών', 'Εθνική πινακοθήκη Αθήνας' and 'Εθνική πινακοθήκη Αθήνα' mean 'National art gallery of Athens' but they are expressed with different inflections. Nevertheless they express exactly the same information need. Light stemming, like suffix removal (e.g. removal of final sigma in Greek), could possibly improve recall and precision of search engines, at least in the highly ranked results.

4. Applying the evaluation methodology

The methodology described in the previous section was applied in the evaluation of Greek supporting search engines. For conducting our assessment we used most of the predominately known worldwide .com search engines: Google (www.google.com), Yahoo (www.yahoo.com), AlltheWeb (www.alltheweb.com), MSN (www.msn.com), AOL (search.aol.com), Ask (www.ask.com), and AltaVista (www.altavista.com). The .com search engines were selected based on their popularity [23]. Also, for comparison reasons, we considered using some native Greek search engines: In (www.in.gr), Pathfinder (www.pathfinder.gr), Robby (www.robby.gr) and Anazitisis (www.anazitisis.gr).

4.1. Interface

To assess the interface issues and some of the issues related to searching effectiveness we asked 31 users to participate in a 'retrieval experiment'. Participants were also asked to construct a number of sample queries for the subsequent experiments. Users had varying degrees of computer usage expertise. We needed end users who knew how to use search engines effectively and therefore had increased demands on the utilization of web searching systems. On the other hand we also needed to listen to people who had just been introduced to search engines and measure their difficulties.

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

Table 1
Time needed to load search engines and to respond to user queries

Search engine	Load up time (s)	Average response time
www.google.com	1	1
www.yahoo.com	10	3.67
www.alltheweb.com	3	3.33
www.msn.com	11	2.67
search.aol.com	11	4.67
www.ask.com	4	3.33
www.altavista.com	2	3.67
www.in.gr	9	5.67
www.pathfinder.gr	13	3.67
www.robby.gr	9	5.33
www.anazitisis.gr	3	8.33

This combination of needs reflects the real everyday needs of web 'surfers'. The trial searches were conducted at the end of June 2006 and lasted two days. They were carried out in a computer lab sharing the same internet connection. Each session lasted two didactic hours. The .gr engines were assessed first because if users were to use an uncomplicated interface first, like Google's, their judgments would be influenced in favour of Google later.

4.1.1. Localization The first issue in our study was the importance of a localized interface. All the participants (100% – 31/31) rated this feature as 'highly important' as many users have basic or no knowledge of English. Although search engines have uncomplicated and minimalist interfaces their adaptation to the local language is essential so users can easily comprehend the available options.

From the .com ones only Google automatically detects local settings and adapts to Greek. AltaVista allows manual selection of the presentation language with a limited number of language choices though and setup instructions in English. Also if you select another language, search is automatically confined to this country's websites (this must be altered manually again).

Nevertheless none of the reviewed web retrieval systems qualifies for the *fully localized* label. Google merely adapts to Greek its basic searching services. For instance, Froogle, Book search, Scholar and Video search are services Google offers in English only. Non-English web searchers may not even be aware of these services. Indeed, 80.64% (25/31) of our participants were not aware of these features and clearly could not benefit from them.

4.1.2. Response time The time to load the initial page is important, especially when the internet connection speeds are slow. Table 1 presents the time needed to load the homepage of the search engines of our study. Time was measured using a fast internet connection and the Opera browser's built in utilities. Search engines which needed several seconds to load up are actually web portals.

Additionally, we ran three queries consisting of one, two and three words respectively. Table 1 also presents the average time required for each engine to return the list of relevant web documents in the three queries. The objective of these two calculations was to determine which search engine offers the fastest searching mechanism. As anticipated Google was the winner again. An important observation resulting from this distribution is that the local Greek search engines are slower than their international competitors. This is true for both parameters of the response time attribute of the evaluation methodology. At this point it should be noted that the Ask and AOL engines experience problems in Greek searching as will be discussed in the following sections.

Afterwards, participants were asked to run the two word query using every search engine. They were then requested to comment on their experience and to try to identify problems and advantages of particular search engines. Their replies cannot be quantitatively evaluated but the main conclusion is that the preferred retrieval systems were Google, AlltheWeb and AltaVista because they are faster and they have an uncomplicated interface. Anazitisis has a straightforward interface but the average searching time is significantly longer than the rest of the engines. Especially in some of the sample queries used later in Section

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

4.2, the retrieval time was approximately 60 s, which is clearly prohibiting in environments where searching is a frequent operation.

4.1.3. Interface complexity Yahoo, MSN, AOL, In, Pathfinder and Robby act as web portals containing categorized links, news, photos and animated Gifs. These features led to increased downloading time as seen in Table 1, which can be irritating when the connection speed is low. Also it can cause confusion and disorientation to users.

The most important problems brought up by the users were 'Slow downloading', 'In which textbox to type the query' and 'Which button to click on'. These difficulties obstructed a few users from completing their tasks and they had to consult us. Even two computer science graduates were confused when utilizing www.in.gr because one can search solely in the www.in.gr site or the whole web or the Greek web space. Based on the case, the searcher must additionally select one of the two available textboxes to type their request.

4.1.4. Searching support This task relies on the previous sample runs, using queries with all terms in Greek. All search engines but AOL and Ask were capable of running the queries and retrieving possibly relevant documents. When a Greek query is run in www.aol.com the information cannot correctly pass from the one window to the other, at least in some browsers. So no results are returned. However, when requests are typed directly using the search.aol.com window, then queries are executed but presentation of the rank is problematic again. Ask did not retrieve any results at all, meaning that indexing of Greek documents is not supported. For example, zero documents were retrieved in all three queries run in the previous section. Ask and search.aol.com were included in the subsequent tests only for comparison purposes, even though none of the users would actually end up using these tools since the first retrieves no results and the second is malfunctioning or Greek searching is not supported through its home page www.aol.com.

4.1.5. Presentation An important point made by the participants is that some of the search engines ranked English web pages first, although search requests were in Greek. For example, in the query 'Ολυμπιακοί αγώνες Αθήνας' (Olympic Games in Athens) Yahoo, MSN and AltaVista ranked some English pages first. This depends on the internal indexing and ranking algorithm but it is one of the points that increase user effort, because one has to scroll down to the list of pages to find the Greek ones.

AlltheWeb and two of the Greek search engines present the rank in a condensed form, without leaving adequate space between results and present the findings with smaller letters with a brief or no summary. All participants (100%) showed dissatisfaction with the condensed presentation output, because it was more difficult to distinguish between the resulting URLs. Also, short summaries increase human effort as users first have to visit the web page and then decide if it is relevant. Summarization is a quite difficult task in information retrieval and most systems provide inadequate summaries. This task is even harder when the document collection is enormous and of varying natural languages as in the web.

4.2. Searching effectiveness

Trying to realize whether user requests 'sound Greek' to the web retrieval systems or not, or in other words if they value the Greek language, we executed six authentic queries (Table 2) suggested by the participants of the previous test. They were typed in lower case sentence form with accent marks, leaving the default options of each search engine.

4.2.1. Index size Table 3 presents the number of retrieved pages for each query as they are indicated by the search engines. Before we explain the results we have to note that AOL is 'enhanced by Google' as it states and since it shows the number of pages which contain potentially relevant links and not the actual number of retrieved web documents we multiply this number by 10 (the number of results presented per page) to get an estimation of the number of retrieved documents.

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

Table 2
Sample queries

No	Queries in Greek	Queries in English
Q1	οδυσσέας ελύτης	Odysseus Elytis (Greek Nobel prize- winning poet)
Q2	μορφές ρύπανσης περιβαλλοντος	Environmental pollution forms
Q3	εθνική πινακοθήκη αθήνας	National Art Gallery of Athens
Q4	προβλήματα υγείας από τα κινητά τηλεφωνα	Health problems caused by mobile phones
Q5	ευρωπαϊκό δικαστήριο	European Court [of Justice]
Q6	τεστ για την πιστοποίηση των εκπαιδευτικών	Tests for certification of educators

It is clear that Google, Yahoo and AltaVista maintain larger indexes than the other search engines and definitely larger compared to the local retrieval tools. AlltheWeb and MSN follow and AOL, Ask, In, Pathfinder, Robby and Anazitisis extract the smallest number of documents compared to the major international search engines.

Although this experiment could be perceived as an estimation of the recall only, it is evident that search engines that maintain larger indexes (and better ranking algorithms) retrieve more documents. In any case the intention of this experiment is merely to get an estimation of the index size. The index size is important, as it is an indication that 'richer' search engines could retrieve more results, which would probably be more precise. Search engines like Anazitisis retrieve only a few documents compared to Google (see Table 3). Clearly the likelihood of satisfying user needs with Anazitisis is smaller than with Google.

4.2.2. Precision To measure the precision of the ranked set of documents we divided the users that participated in our survey randomly, into six groups. Each group had to assess the relevance of the top 20 results of each search engine in a specific query from Table 2. Every member of the group had to visit and explore each of the first 20 results. Then, altogether, they had to decide whether the information presented in the page could be considered relevant to the given query. In this way the relevance judgment was the result of unbiased team work.

Table 4 illustrates the number of pages judged relevant by each group. Although the international search engines returned more results than the native Greek local engines (see Table 3), the relevance

Table 3
Number of retrieved pages in lower case queries

	Q1	Q2	Q 3	Q4	Q5	Q6
Google	36,000	23,700	58,100	20,900	276,000	467
Yahoo	5150	1190	2670	774	142,000	151
AlltheWeb	2570	882	1140	743	69,900	123
MSN	3399	817	1043	423	9046	207
AOL	1970	1390	3030	900	1570	240
Ask	0	0	0	0	0	0
AltaVista	4520	1180	1666	1240	114,000	143
In	3251	561	1366	525	12,114	96
Pathfinder	3262	571	1368	791	14,890	97
Robby	9	35	131	1831	17	4144
Anazitisis	149	120	78	51	429	28

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

of the first 20 results is almost identical in all cases, except in Robby and Anazitisis. These two retrieval systems either maintain a shorter index or employ a crude ranking algorithm. Especially the Anazitisis search tool requires a prolonged time to retrieve the potentially relevant files.

Again some participants mentioned that some international engines rank pages with English content first, although they contain some Greek text as well. These pages could be characterized as non-relevant without the need to visit them as they would probably not be visited by Greek users in real search cases. This tactic would in turn reduce the precision of the search engines in some cases. However it was avoided for reasons of uniformity. Thus every page was visited even if its summary was in English. Then if no useful information was contained in the visited page it was judged as non-relevant. Naturally the potentially relevant information had to be in Greek.

Another observation is that precision is diminished when the number of query terms increases. One would expect the opposite to happen. However most of the words in queries 4 and 6 are common words (stopwords) and possibly trigger this behaviour on the part of the searching tools. In query #2 the last word is in a less used conjugation and this may cause the drop in precision.

4.2.3. Case sensitivity The next part of the experiments was the re-run of the same queries but this time in capital letters with no accent marks. The number of retrieved documents (Table 5) was dramatically diminished in the worldwide search enabling sites while it was left unaffected in three of the domestic ones (In, Pathfinder and Robby). We also measured precision as in the previous experiment. Precision was affected as well (Table 6), compared to results presented in Table 4. In half of the cases precision was increased and in the other half precision dropped. Figures 2 and 3 elegantly depict these results. Figure 2 shows Google's number of retrieved documents, in logarithmic scale, for the same lower and upper case queries and the average recall in the lower and upper case queries. Similarly, Figure 3 portrays the precision ups and downs in lower and upper case queries.

Trying to understand what triggers these inconsistencies in recall and precision we created a short list of potential reasons:

- Final sigma: the Greek capital sigma is 'Σ' but lower case sigma is 'σ' when it appears inside a word and 'ς' at the end of the word. Probably words ending in sigma are transformed internally to words with the wrong form of sigma when they are capitalized or vice versa, e.g. 'ΜΟΡΦΕΣ' (forms) should change to 'μορφες' but it may change to 'μορφεσ', as was concluded in [19]. This leads recall to be reduced. Indeed, the variants 'ελύτης', 'ελυτης' and 'ΕΛΥΤΗΣ' produce, in Google, 64,700, 65,600 and 973 web documents respectively, while all the variants 'ελυτη', 'ελύτη and 'ΕΛΥΤΗ' produce exactly 58,400 results. The first group of variants represents the surname of the poet Elytis in the nominative case and the second group in the genitive case with and without accent marks. In Yahoo the variants 'ελύτης', 'ελυτης' and 'ΕΛΥΤΗΣ' produce 18,500, 99 and 639 web documents respectively.

Table 4
Precision in the first 20 results

	Q1	Q2	Q 3	Q4	Q5	Q6
Google	18	9	14	11	18	10
Yahoo	18	8	14	10	18	10
AlltheWeb	18	7	12	8	17	10
MSN	17	8	13	10	17	10
AOL	17	7	12	10	17	10
Ask	0	0	0	0	0	0
AltaVista	18	8	14	10	18	10
In	16	6	12	10	17	9
Pathfinder	16	6	12	10	17	9
Robby	2[a]	4	3	0	0[a]	0
Anazitisis	13	6	5	4	15	6

[a] The query returned less than 20 results. See Table 3.

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

Table 5
Number of retrieved pages in upper case queries

	Q1	Q2	Q 3	Q4	Q5	Q6
Google	657	37	52,900	14,100	284,000	472
Yahoo	435	15	38	3	432	16
AlltheWeb	346	13	23	1	320	15
MSN	691	10	48	3	232	8
AOL	300	150	3070	370	15,810	1780
Ask	0	0	0	0	0	0
AltaVista	436	15	40	3	433	16
In	3251	561	1366	525	12,114	96
Pathfinder	3262	571	1368	791	14,890	97
Robby	9	35	131	1831	17	4144
Anazitisis	41	5	12	1	31	3

Table 6
Precision in the first 20 results

	Q1	Q2	Q3	Q4	Q5	Q6
Google	19	2	16	9	19	7
Yahoo	18	1[a]	16	0[a]	19	5[a]
AlltheWeb	18	1[a]	13	0[a]	19	4[a]
MSN	19	1[a]	14	0[a]	19	3[a]
AOL	19	1	13	0	19	5
Ask	0	0	0	0	0	0
AltaVista	18	1[a]	16	0[a]	19	6[a]
In	16	6	12	10	17	9
Pathfinder	16	6	12	10	17	9
Robby	2[a]	4	3	0	0[a]	0
Anazitisis	11	0[a]	4[a]	0[a]	16	0[a]

[a] The query returned less than 20 results. See Table 5.

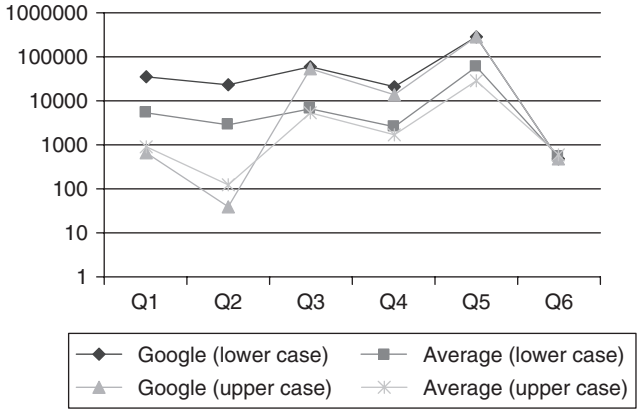


Fig. 2. Number of retrieved pages of the six lower and upper case queries in Google. The average number of retrieved documents of all the search engines is also portrayed.

- Accent marks: accent marks are not used with capital letters and this may cause the inconsistencies in retrieval of pages. Experimentation with the Elytis surname, presented in the previous paragraph, is an indication that intonation is smoothly handled, at least in Google. In Yahoo the variants 'ελυτη', 'ελύτη' and 'ΕΛΥΤΗ' produce 406, 17,800 and 434 documents respectively. In this case absence of accent marks causes recall to drop.

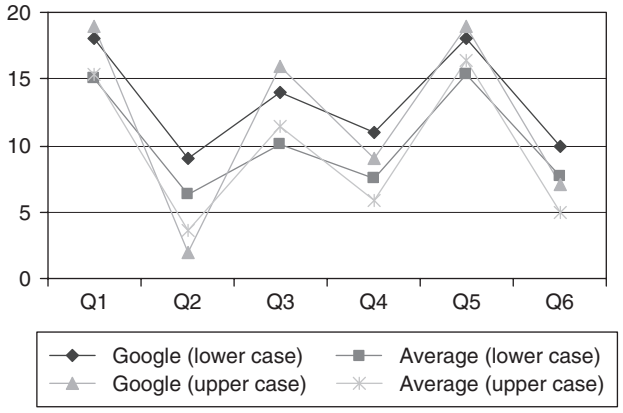


Fig. 3. Precision of the six lower and upper case queries in Google. The average precision of all the search engines is also portrayed.

- Query content and form: queries 1, 3 and 5 which result in better precision when run in upper case form are more normalized than the other queries. They contain only two or three significant words and no stopwords and they are names of persons or organizations. Thus they usually appear in titles when they are in capital letters and therefore precision is better.

These observations are at least alarming. What would happen if a searcher were to choose to search only in capital letters or without accent marks? Novice users are not aware of these differences and they are usually confused [24]. In English search there is no differentiation between capital and lower letters. The result sets are identical in both cases so user effort and required 'user web intelligence' is unquestionably less.

Wrapping up this experiment one can argue that in Greek web searching the same query should be run both in lower and in capital letters, so as to improve the performance of the search. Information from sites where there are no accent marks or which contain intonation errors will not be retrieved unless variations of the query terms are used. Greek search engines are superior at this point and make information hunting easier and more effective. From the international search engines only Google has recognized some of the differences and tried to improve its searching mechanism.

4.2.4. Diacritics handling Handling of diacritics refers to efficient handling of intonation and accent marks such as grave and acute accents. To form an idea of how search engines handle queries when diacritics are used and how they respond when they are not, we executed the queries 'δικαστήριο'and 'δικαστήριο' (court) and the queries 'ευρωπαϊκό' and 'ευρωπαϊκό' (european). The first two variations differ in intonation and the second group of queries differ in umlaut.

Table 7 presents the results of these runs. Google, In, Pathfinder and Robby made no differentiation between the queries. All the other search engines act as simple grep utilities and do not base the retrieval process on the content. AOL does not distinguish the results in the case of intonation but it produces a different number of results when the umlaut is omitted. We further examined this result and it proved to be the normal behaviour of the AOL search engine.

This behaviour on the part of the international search engines and Anazitisis indicates that search engines do not have full understanding of the special characteristics of the Greek language. We assume that this mode of operation would make searching in languages like French, German, Serbian and other more morphologically complicated languages even more demanding.

4.2.5. Stopword removal Google and other international search engines remove English stopwords so as not to influence retrieval. Queries #4 and #6 were re-run in Google, Yahoo and In removing the

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

Table 7
Number of retrieved pages in queries which differ in accent marks

	δικαστήριο	δικαστηριο	ευρωπαϊκό	ευρωπαικό
Google	893,000	893,000	2,920,000	2,920,000
Yahoo	498,000	11,500	1,560,000	18,700
AlltheWeb	207,000	3200	793,000	13,400
MSN	28,930	943	133,686	3959
AOL	51,070	51,070	194,010	178,670
Ask	0	0	0	0
AltaVista	349,000	10,900	1,080,000	18,000
In	7336	7336	24,231	24,231
Pathfinder	92	92	25,792	25,792
Robby	1	1	16	16
Anazitisis	1090	121	2432	240

ordinary words (από – from, τα – the, για – for, την – of, των – of). Queries were in lower case and with accent marks so results should be compared with Tables 3 and 4.

Evidently stopwords affect web retrieval of Greek documents. Table 8 shows that both the number of retrieved documents and precision have been increased. Although more intensive tests are required to construct a stopword list and to see how retrieval is affected by Greek stopwords, this short experiment proves that retrieval performance is increased when stopwords are removed.

4.2.6. Stemming Another factor that influences searching relates to the suffixes of the user request words. The phrases 'Εθνική πινακοθήκη Αθηνών' or 'Εθνική πινακοθήκη Αθήνας' or 'Εθνική πινακοθήκη Αθήνα' mean 'National Art Gallery of Athens'. While they are morphologically different they describe exactly the same information need. Each variation retrieves a different number of pages. For example, Google returned 49,400, 58,000 and 56,500 web pages respectively. Precision is different in these three cases as well and the correlation among the first 20 results is less than 50%.

One could argue that such a difference is rational and acceptable as the queries differ. If we consider these queries solely from a technical point of view then this argument is right. However, if the need for information is the focal point of the discussion then these subtle differences in queries, which merely differ in one ending, should have recalled similar web pages with the same precision. Stemming is an important feature of retrieval systems and its application should be at least studied in spoken languages which have conjugations of verbs and declension of nouns, like in Greek. Google partially supports conjugation of English verbs. Although some Greek stemmers have been created, they have been tested only on their stemming accuracy [25, 26]. The effect of stemming in retrieving Greek web documents is still an issue for research.

5. Discussion

This paper presents a study regarding utilization of search engines using Greek terms. Initially a methodology was described on which the evaluation of Greek web retrieval was based. Regarding interface

Table 8
Number of retrieved pages and precision in queries without stopwords

	Q4		Q6	
	Number of pages	Precision	Number of pages	Precision
Google	20,900	14	1060	12
Yahoo	772	14	165	11
In	616	14	138	12

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

issues, adaptation to local language, interface simplicity, ranking of Greek documents first, quick response and unambiguous presentation of the results are the main demands of users. Google is the unquestionable winner in all these categories which proves that it tries to adapt itself to the demands of other languages than English. Unfortunately most of the international search engines do not offer localized interfaces and some of them do not even support other spoken languages. At least these findings are true for Greek. Additionally, Google does not offer localized versions of all its services.

To estimate the searching effectiveness of search engines that support Greek, we executed a number of sample queries suggested by the participants. International search engines recalled more pages than the local ones and they had a small positive difference in precision as well. However, they are case sensitive, hindering retrieval of web pages which contain the query terms in a slightly different form to the requested one. Terms with accent marks produce different ranks than queries without accent marks. This search engine behaviour requires that users be alerted when they enter a query. On the contrary English users are additionally supported by 'did you mean' tips when they mistype a word in Google.

Endings and stopwords are not removed automatically, thus affecting negatively the retrieval of relevant pages. Stopwords are removed from English queries making information hunting easier, looking at it from a user's perspective. Terms are not stemmed though, even in English. However, in a language with conjugations, like Greek, simple stemming may play an important role in retrieval assisting end users. In any case more intensive tests are needed to see how endings, stopwords and case sensitivity affect retrieval.

The evaluation methodology analysed in this paper tries to identify the deficiencies and the extra user effort required so as to utilize a search engine effectively. The methodology can also be applied in the evaluation of the capabilities of web search engines in other natural languages. For instance, Cyrillic based languages exhibit notable morphological variance in terms while the content remains the same, as in Greek. Our methodology could be applied in assessing search engines with respect to these languages. Some work has already been done towards identifying some of the deficiencies of web search engines in particular languages, e.g. [14–17], and some work has been carried out in the area of the classical IR topics, such as construction of stopword lists and stemming [27–29]. The individual issues negotiated in these studies could be combined in our methodology to measure web search engines' 'understanding' of a particular natural language.

Trying to answer the question posed in the article's title, it can be argued that international search enabling sites do not value the Greek language and possibly other languages with unusual alphabets. Google is the only exception as it seems to be in a process of adapting to and assimilating the additional characteristics. Although domestic search engines 'understand' more features of the Greek language, they are slower, with worse recall and precision and their interfaces are more complicated.

References

[1] Global Reach, *Global Internet Statistics (by Language)* (2004). Available at: www.global-reach.biz/globstats (accessed 31 July 2006).

[2] R. Baeza-Yates and B. Ribeiro-Neto (eds), *Modern Information Retrieval* (Addison Wesley/ACM Press, New York, 1999).

[3] C.W. Cleverdon, J. Mills and E.M. Keen, *An Inquiry in Testing of Information Retrieval Systems* (Aslib Cranfield Research Project, College of Aeronautics, Cranfield, 1966).

[4] S.E. Robertson, The parameter description of retrieval systems: overall measures, *Journal of Documentation* 25 (1969) 93–107.

[5] J. Gwizdka and M. Chignell, *Towards Information Retrieval Measures for Evaluation of Web Search Engines* (1999). Available at: www.imedia.mie.utoronto.ca/~jacekg/pubs.html (accessed 31 July 2006). [Unpublished manuscript, 1999]

[6] H. Chu and M. Rosenthal, Search engines for the World Wide Web: a comparative study and evaluation methodology. In: S. Hardin (ed.), *Proceedings of the Annual Conference for the American Society for Information Science, Baltimore, MD, 1996* (Information Today, Medford, NJ, 1996) 127–35.

[7] M. Gordon and P. Pathak, Finding information on the World Wide Web: the retrieval effectiveness of search engines, *Information Processing and Management* 35(2) (1999) 141–80.

Journal of Information Science, 33 (5) 2007, pp. 622–636 © CILIP, DOI: 10.1177/0165551506076394

[8] M.D. Dunlop, Time, relevance and interaction modelling for information retrieval. In: N.J. Belkin et al. (eds), *Proceedings of ACM/SIGIR, Philadelphia, PA, 1997*, (ACM Press, 1997) 206–13.
[9] L. Su, H. Chen and X. Dong, Evaluation of Web-based search engines from an end-user's perspective: a pilot study. In: *Proceedings of the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA., 1998*, 348–61.
[10] I. Hsieh-Yee, The retrieval power of selected search engines: how well do they address general reference questions and subject questions? *The Reference Librarian* 28/60 (1998) 27–47.
[11] C. Oppenheim, A. Morris and C. McKnight, The evaluation of WWW search engines, *Journal of Documentation* 56(1) (2000) 71–90.
[12] M. Courtois, W. Baer and M. Stark, Cool tools for searching the Web: a performance evaluation, *Online* 19(6) (1995) 14–32.
[13] W. Chung, Y. Zhang, Z. Huang, G. Wang, T. Ong and H. Chen, Internet searching and browsing in a multilingual world: an experiment on the Chinese Business Intelligence portal (CBizPort), *Journal of the American Society for Information Science and Technology* 55(9) (2004) 818–31.
[14] M. Sroka, Web search engines for Polish information retrieval: questions of search capabilities and retrieval performance, *International Information & Library Research* 32(2) (2000) 87–98.
[15] H. Moukdad, Lost in cyberspace: how do search engines handle Arabic queries? In: *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, *Winnipeg, 2004*. Available at: www.cais-acsi.ca/proceedings/2004/moukdad_2004.pdf (accessed 31 July 2006).
[16] J. Bar-Ilan and T. Gutman, How do search engines respond to some non-English queries? *Journal of Information Science* 31(1) (2005) 13–28.
[17] H. Moukdad and H. Cui, How do search engines handle Chinese queries? *Webology* 2(3) (2005) Article 17. Available at: www.Webology.ir/2005/v2n3/a17.html (accessed 31 July 2006).
[18] F. Lazarinis, Evaluating the searching capabilities of Greek e-commerce Web sites, *Online Information Review Journal* (forthcoming).
[19] F. Lazarinis, Do search engines understand Greek or do user requests "sound Greek" to them? In: M. Beigbeder and W.G. Yee (eds), *Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology*, *Compiegne, France, 19 September 2005* (IEEE, 2005) 43–6.
[20] B. Jansen and A. Spink, An analysis of Web searching by European AlltheWeb.com users, *Information Processing and Management* 41(2) (2005) 361–81.
[21] *Cross-Language Evaluation Forum (CLEF).* Available at: www.clef-campaign.org/ (accessed 20 February 2007).
[22] C. Silverstein, M. Henzinger, J. Marais and M. Moricz, *Analysis of a Very Large Alta Vista Query Log, Technical Report 1998–014* (COMPAQ Systems Research Center, Palo Alto, CA, 1998).
[23] D. Sullivan, *Nielsen NetRatings: Search Engine Ratings* (2006). Available at: http://searchenginewatch.com/ reports/article.php/2156451 (accessed 31 July 2006).
[24] F. Lazarinis, Searching the Greek Web space: an instructional approach for effective utilization of search engines, *Journal of Science Education* 7 (2006) 28–30. [Special Issue II]
[25] T.Z. Kalamboukis, Suffix stripping with modern Greek, *Program* 29(3) (1995) 313–21.
[26] G. Ntais, *Development of a stemmer for the Greek language* (MSc Thesis, Stockholm University, 2006). Available at: www.dsv.su.se/~hercules/papers/Ntais_ greek_stemmer_thesis_final.pdf (accessed 31 July 2006).
[27] J. Savoy, A stemming procedure and stopword list for general French corpora, *Journal of the American Society for Information Science* 50(10) (1999) 944–52.
[28] S. Tomlinson, Finnish, Portuguese and Russian retrieval with Hummingbird SearchServerTM, *CLEF 2004*, 221–32. Available at: http://clef.iei.pi.cnr.it/2004/working_notes/WorkingNotes2004/21.pdf (accessed 20 January 2007).
[29] F. Zou, F.L. Wang, X. Deng and S. Han, Automatic identification of Chinese stop words, *Research on Computing Science* 18 (2006) 151–62. [Special issue on Advances in Natural Language Processing]

- Appendix A8 -

An initial exploration of the factors influencing retrieval of Web images in Greek queries

Fotis Lazarinis
Technological Educational Institute
30200 Mesolonghi, Greece
lazarinf@teimes.gr

Abstract

A number of queries are submitted to the image searching mechanisms of Google, Yahoo and MSN and their results are analyzed. The analysis shows that queries which are in different forms (e.g. upper case and lower case) but have exactly the same content retrieve different images. The paper also presents a tool which takes as input a user query and based on knowledge of the linguistic characteristics of Greek produces different forms of this query. It then submits the new queries to Google and merges the recalled images.

Keywords: Search engines, text based image retrieval, context based image retrieval, Greek

1. Introduction

The number of images stored on the Internet increases daily and will continue to expand as the storage media become cheaper. To retrieve images from the Web, related to a specific subject, one has to type one or more keywords in a popular search engine, e.g. Google, Yahoo! and MSN. These search engines retrieve images based on the context and not on the content of an image. More specifically they scan the surrounding text or the captions and the filenames of the images in order to retrieve images relevant to the user queries. Although it sounds quite simplistic to discover images relevant to a topic the reality is different, especially in non English and non Latin queries.

The motivation of the present study originates from the inability of Google to retrieve photos relevant to a very specific question knowing that relevant images do exist. The query submitted to Google, Yahoo and MSN was a Greek surname of a book writer. The personal pages of this person are indexed by the three search engines and the name appears twice inside the page on the right of the existing Gif or JPEG images and once in the title of the page. The name appears in all the three cases as in the query. By running the query, Google retrieved 1 image from an external web site, Yahoo retrieved 14 images from the writer's homepage and MSN presented 3 images from the same Web location as Yahoo.

This simple example demonstrates that Google, which is the most popular search engine [1], fails to retrieve images related to a quite narrow Greek query. This observation, along with the fact that image filenames are in English and not in Greek, decreases the easiness of finding relevant images and increases the required user effort and knowledge.

2. Greek Web retrieval

The Greek language is grammatically more complex than the English language. It has conjugations and morphologically complex words. Articles, verbs, nouns, first names and surnames may be in various cases (nominative, genitive, etc), in singular or plural form and they are differentiated according to their gender (masculine, feminine, neuter). Additionally, diacritics are used, which are shifted according to the case of the word. So user queries may appear in various modes. For example, all the queries "εκπαίδευση σκύλος", "εκπαίδευση σκύλου", "εκπαίδευση σκύλων" mean "dog training" and appear in different cases in singular and plural words. In these three queries only the second term is altered. If the case of the first word is altered as well and if the diacritics are omitted then more queries describing the same user information need will be formulated.

In a previous study on query formulation methods of Greek users, it was identified that users omit diacritics when type the queries and that they type queries in capital or lower case forms [2]. It was also shown that only Google from the international search engines handles effectively some variations by not differentiating the results. But even Google, handles efficiently only the upper or lower case differences and the exclusion of diacritics. Queries which differ simply in one ending produce different results [3].

However it seems that image retrieval is even trickier as queries which differ in diacritics produce different results as well. Additionally, since most of the image filenames

are in a "Greeklish" mode (i.e. Greek words typed in Latin letters), they are neither proper English words nor Greek words and thus search engines cannot exploit them so as to offer more accurate results.

3. Retrieving Web images using Greek queries

To recognize some of the problems in retrieving images using Greek queries we run a number of sample queries and their results were evaluated in terms of relevance of the results.

3.1. Single word queries

As reported in [4] the majority of the user queries submitted in search engines contain one or two terms. Therefore in this first experiment two single term queries, which are the simplest type of Web queries, were run in Google, Yahoo and MSN. The first query "σκύλος" (dog) is a general purpose word and the second "Σάμος" (Samos) is a Greek island.

Query	# of images retrieved		
	Google	Yahoo	MSN
σκύλος	469	321	100
σκυλος	7	45	7
ΣΚΥΛΟΣ	120	57	28
σάμος	862	475	155
σαμος	35	43	3
ΣΑΜΟΣ	273	147	45

Table 1. Number of images retrieved in queries which differ in their form.

Lazarinis studied the form of the Greek queries in [2] and reports that of the total 5,698 queries, the 1,028 (18.04%) queries were in upper case form and the rest 4,670 (81.96%) were in lower case or in title case (i.e. first letter of each word was capitalized). Also it was found that 46.21% of the lower case queries contain at least one word without accent marks. Accent marks are not used in upper case queries. Based on these findings we run both the sample queries in various forms as seen in Table 1. Queries were run in lower case with and without accents and in upper case form. Table 1 presents the number of retrieved images for each query as they are indicated by the search engines. The experiment was conducted on the same day during December 2006.

The most important inference made from the number of images retrieved is that the form of the query severely affects the number of retrieved images. Instead of focusing on the user information need, all the search engines used in the experiment focus on the form of the query. Baeza-Yates and Ribeiro-Neto [5] (p. 2) describe this behavior as data retrieval and not as information retrieval. In English image searching, results are identical in upper and lower case queries. For example, the queries "dog" and "DOG" retrieve identical results. This problem affects other natural languages as well. For instance, the German queries "Bücher Berlin" and "Bucher Berlin" (Books Berlin) retrieve 9.820 and 330 images respectively in Google. In the second case umlaut were omitted.

Another conclusion is that although the omission of diacritics does not influence text Web retrieval [3], at least in Google, the recall of images drops in all the search engines when accent marks are not used. This tactic reduces user satisfaction and increases the required user knowledge and effort on behalf of Greek users.

3.2. Queries in "Greeklish" form

The next step of the experiment was to run the first query of the previous experiment in "Greeklish" form; that is to type the Greek words using Latin characters. The query became "skylos". The motivation behind this experiment arises from the fact that although the filenames of images are in Latin letters they are usually transformations of existing Greek words.

Google retrieves 463, Yahoo 140 and MSN 32 images respectively. The number of images differs from the previous experiments and in all cases the Greeklish version of the query retrieved more images than the second and third query forms of the word "σκύλος" (dog) (see Table 1). This is due to the fact that the current form of the query exploits the file naming conventions used by many users.

The second query "σάμος" is typed as "samos" in Greeklish and in English. Therefore this query was not used in this trial run as it would have definitely produced many relevant images.

3.2.1. Relevance. These differences in behavior were further investigated in Google. The queries "σκύλος" and "skylos" were evaluated in terms of relevance. The first query retrieves 13 relevant results in the 20 initial images while the second retrieves 12 different results in the 20 top ranked images. Behaving as real users, relevance estimation was based on the image's content and not on the surrounding text. Furthermore, only 3 images are the same between the first 20 images of the two query forms.

Using the advanced query Google's searching options we run the combined query "σκύλος OR skylos". This query retrieves 828 results. 16 of the initial 20 images are relevant in this case. Thus relevance increases when queries are combined.

3.3. Queries with stopwords

Stopwords are the terms which appear very frequently in documents and thus their discriminatory value is low for them to be useful index terms [6, 7]. Usual candidates of the stopword list are articles, prepositions and conjunctions, although specific nouns, verbs or other grammatical types could be of low importance in terms of information retrieval in specific domains.
Greek users use articles, prepositions and conjunctions in their queries [2]. This is true in Web image retrieval as well. We asked two of our students to search for images of beaches in the island of Samos. The first query was "παραλίες στη σάμο" (beaches in samos) while the second formed the query "παραλίες της σάμου" (beaches of samos). Both contain a different common word and they are in a different declension.
Table 2 presents the results of the queries run in lower case with and without the stopwords. We observe that the elimination of stopwords increases the number of images.

Query	# of images retrieved		
	Google	Yahoo	MSN
παραλίες στη σάμο	0	18	3
παραλίες σάμο	13	86	3
παραλίες της σάμου	118	68	4
παραλίες σάμου	121	138	6

Table 2. Number of images retrieved in queries with and without stopwords.

3.4. Queries in different declensions

Another important inference can be made on Table 2. Users express the same information need in different forms. The lemmatized version of the queries presented in the previous section is "παραλίες σάμος" (Beaches Samos). Both words are in the nominative case. The lemmatized version retrieves 32 images in Google, 86 in Yahoo and 5 in MSN.
Although the differences between the three forms ("παραλίες σάμο", "παραλίες σάμου", "παραλίες σάμος") are subtle, the number of retrieved images (13, 21, 32) differs among them. In queries with more words this could lead to greater differences in the rank and in the number of the recalled images.

4. An enhanced Web image searching tool

English is a morphologically simple natural language compared to most of the European languages (e.g. German, Greek, Scandinavian languages). Non Latin languages are even more complex even from a technical point of view. The previous examples showed that retrieving images in Greek is more demanding than in English because of the variations in query forms.
To help Greek users retrieve more relevant images an enhanced searching mechanism was created. This tool pipelines the queries through a series of successive subroutines (see figure 1). Initially queries are normalized. The normalization module eliminates the stopwords based on the Greek stopword list presented in [8] and removes any unnecessary punctuation marks. The alternative queries module creates two versions of each query; a query in title case and a query in capital letters. Finally, the produced queries are submitted to Google and their results are merged into a single result set. More specifically, the recalled images from each query are divided in groups of tens. The first two groups of 10 images from each query create a group of 20 images. Then the next 10-image groups create another group of 20 images and so on.

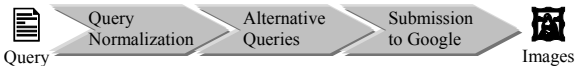


Figure 1. Flowchart of the searching mechanism

At the moment the new queries are submitted to Google only. But as it was mentioned in the introductory section and as it can be concluded from table 2, in some cases Google's performance is worse than Yahoo and MSN. Taking this fact into account, in the future the system will be expanded to submit the queries to Yahoo and to MSN as well. Then the results of all the three search engines will be merged and projected back to the users. Additionally, the system will be further developed to create more versions of the initial queries. For example, creating lemmatized versions of the queries submitted (i.e. in nominative case) and utilizing them with the original query would be a rather interesting research path in a natural language like Greek.

4.1. Evaluation

Although the system is still under development, its alpha version was tested so as to realize its potentials and its shortcomings. The two students who helped us in the previous experiments created a set of ten 1-word Greek queries. All the queries were general purpose terms, such as "Κιθάρα" (guitar) or "Γάτα" (cat). The initial queries were already in title case with accent marks. Since this is a first experiment and is performed by users who are aware of the limitations identified in section 3, we wanted to be fair to Google. So the initial queries contained accent marks so as to allow Google to retrieve a reasonable number of images.

Figure 2 shows the first 20 images retrieved by Google in the query "Κιθάρα" (guitar) and figure 3 depicts the first 20 images for "ΚΙΘΑΡΑ". Our tool merges these results and presents the first 10 images of figure 2 and the first 10 images of figure 3. Evidently, the results of merging Google's outputs are better than the results of each of the individual queries. In all the ten queries there was an increase in the number of relevant documents produced by merging the results. This increase varied from 3 to 5 images in the first 20 results and from 1 to 4 in the second 20 images.

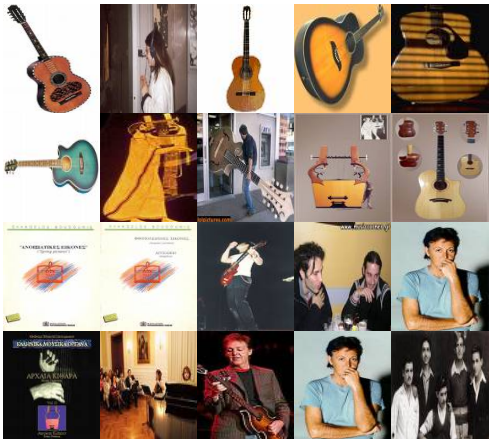


Figure 2. The first 20 images for "Κιθάρα" (guitar)

Figure 3. The first 20 images for "ΚΙΘΑΡΑ" (guitar)

5. Conclusions

This short paper explored some of the factors influencing the retrieval of Web images in Greek queries. A number of queries were submitted in some major search engines. Queries which differ only on their form and not on their content recalled different images. This behavior affects natural languages with complex grammatical rules and multiple diacritics (e.g. Greek, French, German, Serbian, etc).

The ideas of a flexible searching tool run on top of Google were also presented. This tool is aware of some of the linguistic features of the Greek language and combines the images recalled from queries with similar content but with different form. The initial evaluation of the system in single word queries showed a significant increase in the number of relevant images in the top 20 ranked images.

Our work should be expanded based on the query patterns of Greek users. Lemmatization and other information retrieval techniques, such as spelling detection and correction techniques [9], should be applied to reformulating the queries. Merging of the results should be then evaluated against the original queries. The user's behavior during image searching should be further studied so as to propose improvements on search engines adapted to the query patterns and the linguistic characteristics of the query's natural language.

6. References

[1] D. Sullivan, *Nielsen NetRatings: Search Engine Ratings* (2006) http://searchenginewatch.com/

[2] F. Lazarinis, "How do Greek searchers form their Web queries?" *3rd WEBIST* (Barcelona 2007) to appear

[3] F. Lazarinis, "Web retrieval systems and the Greek language: Do they have an understanding?", *Journal of Information Science*, SAGE Publications (in press).

[4] B. Jansen, and A. Spink, "An analysis of Web searching by European AlltheWeb.com users", *Information Processing and Management,* 2005, 41, 361–381.

[5] Baeza-Yates R. and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley and ACM Press, New York, 1999.

[6] van Rijsbergen, C.J., *Information Retrieval*, Butterworths, London, England, 1979.

[7] Salton, G, and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, Compute Science Series, New York, 1983.

[8] F. Lazarinis, "Engineering and utilizing a stopword list in Greek web retrieval", *Journal of the American Society for Information Science* (in press)

[9] K. Kukich, "Techniques for automatically correcting words in text", *ACM Computing Surveys*, 1992 24(4), 377–439.

- Appendix A9 -

EATIS 2007

Lemmatization and stopword elimination in Greek Web searching

Fotis Lazarinis
Technological Educational Institute
30200 Mesolonghi, Greece
lazarinf@teimes.gr

Abstract

This paper explores the effect of noun lemmatization and stopword removal in Greek Web searching. A light lemmatizer is presented and applied in a retrieval experiment. Stopwords are removed from user queries. In both experiments an increase in precision is reported. The main purpose of our work is to adapt and apply some "ancient" information retrieval techniques in non Latin queries and measure their effect in the retrieval of relevant documents.

Keywords: Lemmatization, stemming, stopword, information retrieval, search engines, Greek

1. Introduction

Stemming and stopword removal are amongst the oldest and most widely used information retrieval techniques [1, 2]. Classical information retrieval (IR) systems support them as it has been proved that retrieval effectiveness is positively influenced. Commercial search engines, like Google, do support these techniques, at least partially. For example, the queries "I won a nobel" and "won a nobel" retrieve the same documents in Google. The query "stemming site:www.dcs.gla.ac.uk/Keith" retrieves three pages where both the words "stem" and "stemming" are emboldened as they are considered as matching terms. Here we restricted our search in Keith's van Rijsbergen site. In other natural languages, though, and especially in non Latin languages like Greek, these two techniques are not supported. In [3, 4] it has been shown that some common words could influence Greek retrieval positively if they are removed and that queries which have the exact same meaning but differ solely in one ending retrieve different pages. Experimentation with other languages revealed some of the inefficiencies of worldwide search engines related to singular and plural forms of query terms [5, 6].

In general it has been argued that existing search engines may not serve the needs of many non-English-speaking Internet users [7]. The purpose of the current study is to report the initial findings on the effect of lemmatization and stopword elimination in Web searching using Greek terms.

The Greek language is grammatically more complex than the English language. It has conjugations and morphologically complex words. Articles, verbs, nouns, first names and surnames may be in various cases (nominative, genitive, etc), in singular or plural form and they are differentiated according to their gender (masculine, feminine, neuter). The application of standard IR techniques, such as stemming, lemmatization and stopword elimination would possibly have positive effects in Web retrieval in such a complex natural language.

2. Lemmatization/Stemming

Stemming is the process of reducing a word to its stem or root form. The most well-known stemmer is the rule based algorithmic stemmer of Porter [8]. Stemmers have been implemented for other languages as well, including Greek [9, 10, 11]. The Greek stemmers try to be exhaustive, meaning that they try to find the minimum stem for a word. This could retrieve lots of non relevant documents if applied in Web searching. For example the words "ΔΕΝΩ" (tie) and "ΔΕΝΟΜΟΥΝ" (tied) are reduced to the stem "ΔΕΝ" which is the same as the word "ΔΕΝ" (not). Furthermore, they have been tested only on their stemming accuracy and not on a search engine or an IR system.

Lemmatization involves the reduction of words to their respective headwords (i.e. lemmas). In the linguistic dictionaries every entry corresponds to a lemma that defines a set of words with the same lexical root. Lemmatization is closely related to stemming. The difference is that a stemmer finds the stem of a word while a lemmatizer tries to find the lemma for a given word. For example, the lemma for the words "ΔΕΝΩ" (tie) and "ΔΕΝΟΜΟΥΝ" (tied) is the word "ΔΕΝΩ".

The basic idea behind our work is to create a tool which is actually a semi-stemmer-semi-lemmatizer and utilize it on Web searching. We could neither call it stemmer nor lemmatizer. It is not a stemmer since it does not find the stem of a term and it is not a full lemmatizer as it actually

operates on nouns only and identifies their inflectional suffixes.
Since the system is still under development, our lemmatizer operates on noun and identifies the inflectional suffix of a noun and then, based on a set of rules it may reduce a suffix so as to create the lemma for a word. For example, if the word is "ΓΙΑΓΙΑΔΕΣ" or "ΓΙΑΓΙΑΔΩΝ" (grandmothers), then the word "ΓΙΑΓΙΑ" (grandmother) is produced. If the word is already a lemma it does not change it. In Greek, nouns have singular and plural forms and in each form there are four inflections, namely nominative, genitive, accusative and vocative. Identification of the inflections is based on a set of 39 different suffixes for all the forms of the nouns, as in [11]. These suffixes were taken from a well-known Greek grammar book [12]. Through a set of nested if then else rules (see figure 1 for example) the longest possible inflection is identified and the word is matched to its equivalent singular nominative form.

```
if term has suffix "ΑΔΕΣ" or "ΑΔΩΝ"
{
     replace suffix with "Α"
}
```

Figure 1. An example of a suffix replacement rule

An initial estimation of 300 nouns in various forms and declensions resulted in 95.67% (287/300) success in the lemmatization procedure. The 13 erroneous instances will be used so as to refine the lemmatizer, which is still under development, though. A dictionary based lemmatizer could improve the effectiveness of the lemmatization process and the handling of diacritics but could slow down the procedure.

2.1. Lemmatized versus Non Lemmatized queries

The first version of the lemmatizer was used in some sample queries, run in Google and in a custom made simplistic IR system, so as to estimate the importance of lemmatization and the precision improvement in Web searching. Our objective is to eventually propose ways to support what Google partially supports in English, i.e. questions like "evaluating web sites" which retrieve also documents containing the words "evaluation web sites". But we first need to evaluate and measure the significance of the lemmatization procedure.
Therefore with the aid of two students we constructed a set of 10 queries. These queries contained 1, 2 or 3 words. In total there were 17 different words. 9 of these terms needed to be lematized. With the aid of the lemmatizer we created 10 new queries equivalent to the first ones but with all terms as lemmas. For example, the query "Μορφές Ρύπανσης Περιβάλλοντος" (Environmental Pollution Forms) was transformed to the query "Μορφή Ρύπανση Περιβάλλον" (Environmental Pollution Form). Queries were carefully selected so as to prohibit the lemmatizer from producing errors. This tactic is biased towards the lemmatizer but since the focus of the experiment is the influence of the lemmatization in Web searching it does not affect the validity of our experiments.
Using Google and a basic IR system, based on cosine similarity [13], we run these 10 queries in both forms using a 5,124 text collection of various sources. The text collection was indexed using our IR system which was expanded with the lemmatizer. Table 1 shows that the lemmatized versions of the queries improve relevance in the first top 10 retrieved documents. Queries 2, 7 and 10 were one-word queries in plural form. Google could not retrieve any relevant result in these cases. On the contrary, our system could retrieve a few relevant pages which contained the query term in singular form.

Query No	Non lemmatized	Lemmatized
	# of relevant docs. in top 10	# of relevant docs. in top 10
1	6	7
2	0	3
3	10	10
3	2	2
5	5	5
6	7	9
7	0	4
8	6	8
9	2	3
10	0	2

Table 1. Number of relevant documents in the top 10 results in lemmatized and non lemmatized queries.

We could not base this experiment entirely on Google since the document collection had to be indexed using our lemmatizer in order to measure the potentials or shortcomings of our technique. Running the user queries in Google, first in non lemmatized form, i.e. as specified by users, and then in lemmatized form would have retrieved different sets of documents and thus no safe conclusions could be made. Therefore we utilized our simplistic IR system against the powerful Google.
Google supports lemmatization and even retrieval of synonyms in English retrieval. For example, the query "Bookshop New York" retrieves documents having as matching terms the words "Bookstore" or "Bookshops" or "Book" or "Books". In Greek Web retrieval only documents containing the exact query terms are retrieved. This is true for most of the European languages. For instance, the query "Libreria Roma" (Bookshop Rome) in

Italian, retrieves Web documents based on the exact form of the query terms. Therefore lemmatization is a feature which should be applied to other natural languages as well so as to reduce the required user effort and increase the possibilities of retrieving more relevant documents.

3. Stopword Elimination

Stopwords are the terms which appear too frequently in documents and thus their discriminatory value is low [1, 2]. It has been claimed that a word which appears in 80% of the documents in a document collection is useless for purposes of retrieval [1]. Therefore they are eliminated during the indexing and querying phases. Usual candidates of the stopword list are articles, prepositions and conjunctions, although specific nouns, verbs or other grammatical types could be of low importance in terms of information retrieval in specific domains.

Stopword lists have been constructed and utilized in English information retrieval [1, 2]. Stopword lists have been engineered for some European languages as well (see http://snowball.tartarus.org). In [14] the construction process and the stopword list for the French language are presented. A common word list for Chinese is presented in [15]. The Greek stopword list used in our experiments is presented in [16].

Word	Freq.	Word	Freq.
για (for)	7	του (the)	2
και (and)	4	να (to)	1
στην (in)	4	που (where/that)	1
των (the)	4	πως (how)	1
της (the)	3	σε (in)	1
τις (the)	3	στα (into)	1
από (from)	2	τα (the)	1
ο (the)	2	την (the)	1
στο (into)	2	το (the)	1

Table 2. Stopwords and their frequencies in the 32 query sample.

In the present paper we study the effect of stopword removal in Web searching, with the aid of 32 user constructed queries. These queries were supplied by 13 users who frequented an introductory seminar related to WWW. 20 of the total 32 queries contained 18 terms of the stopword list presented in [16]. For example in the query "ζώα που ζουν στο νερό και στην ξηρά" (animals that live in water and in mainland) the underlined words are stopwords. The 18 stopwords appeared 41 times in these 20 queries. All common words contained in the queries and instinctively considered as stopword candidates, were indeed part of the stopword list presented in [13]. This is an indication that the stopword list is thoroughly constructed, at least with respect to the sample queries. Table 2 presents the unique stopwords and the number of times they occurred within the query sample.

As explained the Greek language is grammatically more complex than the English language. It has conjugations and more morphologically complex words. That is why in Table 2 the terms "των", "της", "ο", "του", "τα", "την" are translated to "The" in English. These terms are articles in various cases (nominative, genitive, etc), in singular or plural form and concern the three genders.

Query No	With Stopwords	No stopwords
	# of relevant docs. in top 10	# of relevant docs. in top 10
2	9	10
3	5	6
6	6	6
7	3	6
9	4	4
13	7	10
14	10	10
16	10	10
17	4	8
18	0	0
22	2	3
23	5	7
24	7	10
25	0	0
26	4	4
27	4	7
28	7	9
29	3	5
30	4	7
31	3	4
Average	4.85	6.30

Table 3. Number of relevant documents in the top 10 results in queries with and without stopwords.

Users were then asked to run the 20 queries in Google. Each user was given two forms (with and without stopwords) of one or two of the 20 queries. Then they had to estimate the relevance on the top 10 results. From their estimations it was made clear that stopwords affect negatively Web searching in Greek. As seen in Table 3, in the 13/20 queries that contained stopwords, an increase in the relevance in the top 10 documents was reported. In the other 7/20 queries the relevance remained unaltered. So, the precision was either increased when queries were run without stopwords or remained unaffected. There was no single query instance where precision dropped when a query was run with stopwords. In general, the average

precision increased from 4.85 to 6.3 per 10 relevant documents.

4. Synopsis

This short paper reviews the effect of lemmatization and stopword removal in Greek Web searching. The principles and ideas of an under development lemmatizer for Greek nouns were also discussed. With the aid of authentic queries it was made obvious that after the application of lemmatization and stopword elimination more relevant documents are retrieved. In a natural language with conjugations and intonation, like Greek or in another morphologically complex language, Web retrieval is trickier than in English. Therefore more experiments are needed so as to realize the effect of information retrieval techniques such as lemmatization, stemming and stopword removal from user queries and index files. The combination of all these techniques should be also further studied. Efficient light stemmers or light lemmatizers could be applied in e-shop catalog searching as well. Alternative queries could be constructed and suggested to users in "no result" queries. Refined stopword lists could be constructed based on large user query samples.

The initial conclusions presented in this study are applicable to other natural languages as well. For example if similar techniques are applied in the Italian language then the query "Librerie di Roma" could retrieve Web pages containing the words "Libreria" and "Libro". Also the word "di" would not influence the retrieval of documents significantly. In German Web retrieval, queries containing the term "Bücher" (books) would also retrieve Web documents containing the word "Buch" (book). These examples show that there are still a lot to do before non English users enjoy the full facilities offered to English users by international search engines.

5. References

[1] van Rijsbergen, C.J., *Information Retrieval*, Butterworths, London, England, 1979.

[2] Salton, G, and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, Compute Science Series, New York, 1983.

[3] F. Lazarinis, "Evaluating the Searching Capabilities of Greek e-commerce Web sites", *Online Information Review Journal* (in press)

[4] F. Lazarinis, "Web retrieval systems and the Greek language: Do they have an understanding?", *Journal of Information Science*, SAGE Publications (in press).

[5] J. Bar-Ilan, J., and T. Gutman, "How do Search Engines respond to some non-English queries?", *Journal of Information Science,* 2005, 31(1), pp. 13–28.

[6] H. Moukdad, "Lost in Cyberspace: How do search engines handle Arabic queries? *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, (Winnipeg, 2004)

[7] W. Chung, Y. Zhang, Z. Huang, G. Wang, T. Ong, and H. Chen, "Internet Searching and Browsing in a Multilingual World: An Experiment on the Chinese Business Intelligence Portal (CBizPort)", *Journal of the American Society for Information Science and Technology*, 2004, 55(9), pp. 818–831.

[8] M. Porter, "An algorithm for Suffix Stripping", *Program*, 1980, 14(3), pp. 130-137.

[9] T. Z. Kalamboukis, "Suffix stripping with Modern Greek", *Program*, 1995, 29(3), pp. 313–321.

[10] G. Tambouratzis, and C. Carayannis, "Automatic corpora-based stemming in Greek", *Literacy and Linguistic Computing*, 2001, 16, pp. 445-466.

[11] Ntais, G., *Development of a stemmer for the Greek language, MSc Thesis*, Stockholm University 2006, www.dsv.su.se/~hercules/papers/

[12] Triantafyllidis, M., *Modern Greek Grammar* Institute M Triantalyllidis, 1941 (in Greek)

[13] Lazarinis, F., *Combining Information Retrieval with Information Extraction, MSc Thesis*, University of Glasgow, 1997.

[14] J. Savoy, "A Stemming Procedure and stopword List for General French Corpora" *Journal of the American Society for Information Science*, 1999, 50(10) pp. 944-952.

[15] F. Zou, F. Wang, X. Deng, S. Han, "Automatic Identification of Chinese Stop Words", *Research on Computing Science*, 2006, 18, pp. 151-162.

[16] F. Lazarinis, "Engineering and utilizing a stopword list in Greek web retrieval", *Journal of the American Society for Information Science* (in press)

Engineering and Utilizing a Stopword List in Greek Web Retrieval

Fotis Lazarinis
Technological Educational Institute of Mesolonghi, New Buildings, Mesolonghi 30200, Greece.
E-mail: lazarinf@teimes.gr

The main aim of the article is the presentation of the construction process of a stopword list for a non-Latin language and the evaluation of the effect of stopword elimination from user queries. The article presents the phases of engineering a stopword list for the Greek language as well as the problems faced and the inferences deduced from this procedure. A set of 32 authentic queries are proposed by users and are run in Google with and without the stopwords. The importance of eliminating the stopwords from the user queries is then evaluated, in terms of relevance, in the top-10 results from Google.

Introduction

Typical text-retrieval systems eliminate stopwords from both queries and index files (Baeza-Yates & Ribeiro-Neto, 1999). Stopwords are the terms which appear very frequently in documents, and thus their discriminatory value is low for them to be useful index terms (van Rijsbergen, 1979; Salton & McGill, 1983). It has been claimed that a word which appears in 80% of the documents in a document collection is useless for purposes of retrieval (Salton & McGill, 1983). Therefore, such words are eliminated during the indexing and querying phases. Usual candidates of the stopword list are articles, prepositions, and conjunctions, although specific nouns, verbs, or other grammatical types could be of low importance in terms of information retrieval in specific domains.

Furthermore, the elimination of stopwords reduces the index file and speeds up the retrieval procedure. In Web retrieval systems, although removal of stopwords is not extensively supported, one can easily realize that some query terms do not influence the retrieval procedure whatsoever. For instance, the queries "I won a nobel" and "won a nobel" retrieve exactly the same Web documents in Google. Some search engines even state explicitly that specific words have been eliminated from the query as they are too frequent. Stopword lists have been engineered for the English language since the "ancient" years of information retrieval (IR). A list of 425 stopwords is presented in Frakes and Baeza-Yates (1992). A slightly different English stopword list can be found in Fox (1990). The SMART system uses an augmented listing of items as its stopword list (Buckley, 1985).

With the advent of the Internet and its multilingual user base, there is a growing interest in facilitating the retrieval process of non-English users. One of the factors that might increase the effectiveness of text-retrieval systems in non-English searching is the removal of stopwords. In this article, we present the engineering procedure of a stopword list for the Greek language. The difficulties arising when processing non-Latin text are discussed, and the stopword list is presented. This stopword list is applied in a Web retrieval experiment to test its significance and its added value to the retrieval process.

Related Work

Stopword lists have been constructed for most of the major European languages.[1] In a study by Savoy (1999), the construction process of a stopword list for the French was analyzed. This list has been semi-automatically created based on term frequency and on careful manual elimination of certain words from the list. These terms, although quite frequent, could not be considered as stopwords because they carry significant information. It seems like the document collection used for identifying the potential stopwords was restricted to a specific domain (e.g., politics). This could explain why the words "Président" and "France," for instance, were highly ranked.

Received October 29, 2006; revised December 24, 2006; accepted December 25, 2006

 ● Published online 17 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20648

[1]See http://snowball.tartarus.org and http://www.ranks.nl/tools/stopwords.html

The automatic construction of an English stopword list based on a complex statistical model was discussed in Tsz-Wai Lo, He, and Ounis (2005). Their model assigns weights on each term using the Kullback–Leibler divergence measure (Coverand & Thomas, 1991). The authors claimed that computing a stopword list with their so called term-based random sampling approach decreases the required computational effort; however, they also mentioned that their produced stopword list is slightly worse than the classical stopword lists constructed on term frequency. The authors concluded that the experimental results demonstrate that a more effective stopword list could be derived by merging Fox's (1990) classical stopword list with the stopword list produced by their proposed approach.

Chinese stopword identification was discussed in Zou, Wang, Deng, and Han (2006). Chinese text tokenization is more difficult than in other natural languages since the word boundaries are not well defined. Therefore, the authors employed a segmentation algorithm first and then built a statistical model for engineering the stopword list. This statistical model is primarily based on calculating the term frequencies of the words in a given collection. The frequencies are normalized based on the documents' lengths, and then the probability of a word being a stopword is calculated.

A Spanish common-word catalog is incorporated in SMART (Buckley, 1985) and has been employed in some Cross Language Evaluation Forum (CLEF) IR experiments (Méndez Díaz, Vilares Ferro, & Cabrero Souto, 2005); however, the construction process and the benefits of stopwords are not discussed in either study. Thus, the importance of purging the stopwords only can be inferred, as the primary scope of the research articles is either the presentation of an IR system or the application of natural-language-processing techniques in IR.

Two further studies discussed the effect of stopword elimination in Greek Web retrieval and in utilizing search engines of e-shops (Lazarinis, 2005, 2007a). These studies focused on the capabilities of search engines in Greek retrieval, and included some initial and intuitive explorations on the effect of removing some common words from one user query and measuring the number of relevant documents thereafter. Nevertheless, an increase in accuracy was reported in both cases in the top-ranked documents.

In this article, we will discuss the phases of constructing a stopword list for Greek. The process of tokenizing the Greek texts and of ranking the words is presented. Additionally, our work differs from the previously discussed studies in that it tries to realize the effects of stopword elimination in Greek Web retrieval by performing an experiment in Google, using authentic user queries.

Engineering the Greek Stopword List

The steps of engineering the Greek stopword list are graphically depicted in Figure 1. The process initiated with the assembly of a text collection and ended with the calculation of the frequencies of Greek terms appearing in the collection. The following sections will analyze each phase and discuss the problems that arose during the construction process.

Text Collection

Initially, we needed a domain-independent document collection. Using automated tools, we downloaded 5,124 HTML documents from the Web. After removing the HTML tags and the embedded scripts, the size of the text collection was 12.22 MB. The size of the resulting 5,124 text documents varies from 1 KB to approximately 50 KB. The documents were from five general-purpose newspapers, one computer-related magazine, three conference proceedings related to public affairs, medicine, and education, respectively, and from one computer science educational book. We consider this collection as domain independent since the documents come from various sources and concern various topics.

Tokenization

The next step was the segmentation of each text file. Unlike in the case of the Chinese language (Zou et al., 2006), the word boundaries are concrete in Greek text. Spaces and other delimiters such as the comma, full stop, exclamation mark, question mark, and a few other characters separate words and other strings. During this procedure, we realized that several non-Greek punctuation marks are used in Greek text. For instance, the English question mark (?) and the English quotation marks (“ ”) are used. The Greek equivalents are; and ‹‹ ››.

During the tokenization procedure, the non-Greek strings also were removed. Thus, English words or other Latin words were eliminated; however, by eliminating the Latin words, we realized that some terms which, deceptively, looked Greek also were removed. For example, the word “ΑΒΑΚΑΣ” (abacus) seems perfectly encoded in the Greek alphabet term; however, our tokenizer removed it because it was considered a Latin encoded word. When this word was transformed to lowercase, then instead of the Greek word “αβακας”, the semi-Greek–semi-Latin term “abakaς” appeared. These terms were mostly in uppercase letters because several Latin uppercase letters are identical to the respective Greek

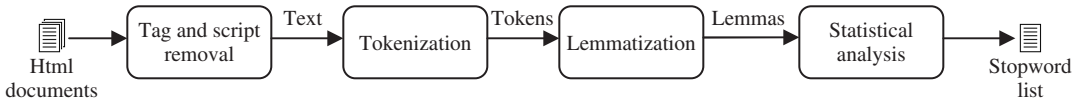


FIG. 1. Logical view of the stopword creation process

DOI: 10.1002/asi

uppercase letters. In lowercase, though, the alphabets are quite dissimilar. The semi-Greek–semi-Latin encoded terms appear at the beginning of sentences or near English text as a result of user negligence. Users forget to switch their keyboard to Greek mode after typing English text, and they type letters using English characters until they see a character which is clearly non-Greek. Then they switch the mode of their keyboard and continue to type the rest of the term without first removing the Latin encoded characters. Therefore, although externally identical, internally the pseudo-Greek terms are encoded in two ISO codes. In total, 1,932 such terms appeared and accounted for 0.11% of the total tokens.

Note that this observation probably affects the effectiveness of Greek-supporting search engines. Query terms appearing in a deceptive Greek-encoding mode cannot retrieve relevant documents. On the other hand, erroneous index terms cannot match to the query terms. Indeed, this problem exists in Greek Web retrieval. For instance, we ran the single term queries "ΑΠΟΤΕΛΕΣΜΑ" (αποτελεσμα) and "ΑΠΟΤΕ–ΛΕΣΜΑ" (aποτελεσμα) in Google. Both words mean "result" but, as their lowercase versions indicate, the first word is indeed in Greek while in the second instance the first letter is in English. The first query retrieved approximately 3,790,000 potentially relevant documents while the second retrieved 10,100 Web pages. The number of retrieved documents is huge in the erroneous instance of the word, and this clearly leads Google to exclude several relevant documents from its results.

Taking into account the previous observations, we expanded the delimiter list to include the English punctuation marks and created a routine which changes the mixed coded words to Greek. The tokenization process was rerun, and 1,734,053 terms were produced. Of these terms, as explained, the 0.11% was the semi-Greek–semi-Latin terms, which were now fixed. The tokenization process was exhaustive, meaning that only regular words compose the 1,734,053 word list. On average, each text contained 338.42 words.

Lemmatization

Lemmatization is the normalization of the form of a word to a form that is used as the headword in a dictionary, glossary, or index. This normalization is important to cluster the lemmas. Lemmatization techniques already have been applied in conjunction to stemming in non-English IR experiments (Korenius, Laurikkala, Järvelin, & Juhola, 2004). In our case, lemmatization is restrained only to normalizing single terms. Greek is a language with conjugations and accent marks. Therefore, terms should be normalized before they are used in text-processing experiments (Lazarinis, 2006).

Currently, the lemmatization process is restricted to change all the lowercase tokens to uppercase. Additionally, accent marks are removed. Some exceptions were introduced in this procedure. For instance, the term "ή" is the English "OR" while "η" is a feminine article. In this case, the accent mark was not removed when the word was capitalized. Some of these exceptions were realized in "a trial-and-error" method.

The normalization procedure is especially important in the case of the acute accent. In our tests, it was clear that several instances of specific words are erroneously written without acute accents. For example, the word "Ευρωπαϊκή" (European) is frequently typed without the umlaut (i.e., "Ευρωπαική"). Several other terms are typed without accents, many of which are potential stopwords. For present study, we wanted only to identify the stopwords, so we did not evaluate other techniques for calculating the approximation of lemmas or for identifying spelling errors such as n-grams (Zamora, Pollock, & Zamora 1981). Greek stemming algorithms (Kalamboukis, 1995; Tambouratzis & Carayannis, 2001) also could be considered as an alternative to lemmatization, although their results probably would be harder to utilize in our case since some stems are identical to stopwords. For instance, the stem for the verb "ΔΕΝΩ" (tie) is "ΔΕΝ," which is identical to the stopword "ΔΕΝ" (NOT).

After completing this phase of the engineering process, 77,913 unique lemmas were identified. On average, each of the 5,124 text files used for constructing the stopword list contained 15.21 unique words.

Statistical Analysis

For identifying the common words, we calculated the frequency of each of the 77,913 terms produced in the previous phase. Thus, we followed the term-frequency (tf) approach for constructing our stopword list (Fox, 1990; Savoy, 1999). Although some other techniques for identifying the terms with low discriminatory value based on more complex statistics have been proposed (Tsz-Wai Lo et al, 2005; Zou et al, 2006), most of the currently existing stopword lists are based on term frequency only. Table 1 lists the 20 more frequent words, their frequency, and their English equivalents.

These top-20 words occur 524,678 times within the 1,734,053 lemmas. Thus, they account for 30.26% of the total lemmas. By removing these stopwords, the size of the

TABLE 1. Top 20 words with their frequencies.

Word	TF	Word in English	Word	TF	Word in English
Και	61,372	And	Για	23,480	For
Το	46,068	The	Τα	22,082	The
Να	39,653	To	Είναι	21,805	Is, Are
Του	36,725	Of	Των	18,736	Of
Η	31,754	The	Σε	18,572	At, In, To, Into
Της	27,868	Of	Ο	16,955	The
Με	26,492	With	Οι	15,786	The
Που	26,234	Where	Στο	15,426	To, At
Την	24,616	The	Θα	14,172	Will
Από	23,481	From	Τη	13,401	The

DOI: 10.1002/asi

index file could be reduced approximately by 30%. The occurrences of the first 99 common words are 765,812, which is 44.16% of the total lemmas.

The Greek language is grammatically more complex than the English language. The Greek language has conjugations and more complex words than do the English terms. That is why in Table 1 the terms "To", "H", "Την", "Τα", "Ο", and "Τη" were translated to "The" in English. These terms are articles in various cases (e.g., nominative, genitive, etc.), in singular or plural, and concern the three genders (i.e., masculine, feminine, neuter). In total, there are 18 articles, some of which are the same for the masculine and feminine genders. These articles are used in defining the gender of nouns and adjectives.

Finally, we selected the first top-99 words to structure our stopword list.[2] The research presented in the current article is the first attempt to build a stopword list and to explore its effect in the retrieval of Greek Web pages. Consequently, we wanted to be cautious in the selection of stopwords. After the initial 99 words, the frequency of the words diminishes dramatically, and so it is not capable of classifying a word as common. More specifically, the frequency dropped from 1,844 in Term 99 to 422 in Term 100. The stopword list is presented in the appendix. All alternative forms for each entry are maintained in the stopword list. For example, the 10th entry is actually composed of the terms "Από," "από," "ΑΠΟ," and "απο" (from) to cover all the acceptable forms of the term and even to compensate for usual minor grammatical errors (e.g., in "απο," the accent mark is omitted).

Stopwords in Greek Web Retrieval

As explained in the Introduction, our purpose was not only to construct a stopword list but also to evaluate how this affects retrieval of Greek Web documents. Therefore, we constructed a set of authentic queries with the aid of end users and ran it in Google with and without the stopwords. Our intention was to assemble a set of realistic Greek queries which did or did not contain one or more stopwords.

Several inferences and research questions then could be applied on this sample. For instance, if most of the queries did include stopwords, then this is a strong argument toward the importance of constructing a stopword list and its utilization in Web search engines or in other local search systems of e-shops. By running the queries with and without stopwords, the effect in the retrieval of relevant documents also could be measured, and the question of how accuracy is affected if stopwords are automatically removed from user Web queries could be answered. Finally, the process could be evaluated in terms of speed. Would the elimination of stopwords speed the retrieval process?

[2]Our stopword is available online at http://rea.teimes.gr/~lazarinf/gr_stop.htm and at www.lazarinis.gr/gr_stop.htm

Sample Queries

To assemble the sample queries, we asked 13 users to provide two real queries. Additionally, we utilized the six sample queries used in another study of Greek Web searching (Lazarinis, 2007b). The queries in this study were suggested by users participating in the evaluation of the capabilities of Greek-supporting international and local search engines. None of our query providers or the participants of the previous study were aware of the usefulness of stopword elimination or the purpose of our experiments, in an effort to allow them to form their queries in an unbiased way. All the reported experiments and their results realized during a 2-hr search-engine-related seminar frequented by 13 users. Users had medium computer-handling capabilities. and 6 of them were using search engines for the first time. The rest were using search engines in an occasional mode. The participants were high-school or University graduates.

The following list presents the 32 queries of the users in alphabetical order. Stopwords, based on the list presented in the appendix, are underlined. As can be seen, 20 of 32 (62.50%) queries contain one or more stopwords. The total number of words in these queries is 105, which means 5.25 words on average per query. The total number of stopwords in these 20 queries is 41 (i.e., 2.05 stopwords per query on average). In other words, 41 of the 105 words (39.05%) exhibit low discriminatory value. These statistics prove that stopwords are used quite often in query formulation. Hence, their importance in Web retrieval needs to be studied. Additionally, note that the total number of words of all 32 queries is 136, meaning that on an average each query contains 4.25 words. Thus, the length of the queries increases, probably unnecessarily, when articles, prepositions, and other common words are included. This may lead to a series of problems ranging from increased retrieval time to a possible drop in the relevance of the retrieved documents.

1. Αεροδρόμιο Ελευθέριος Βενιζέλος
2. αποτελέσματα των δημοτικών
3. Αυτοκίνητα μεταχειρισμένα σε τιμές ευκαιρίας
4. δημος αθηνων
5. Δημοτικές εκλογές 2006
6. εγκατάσταση σε νέο κτήριο και ηλεκτρονική οργάνωση
7. Εγκύκλιος για τις μεταθέσεις καθηγητών από ΠΥΣΔΕ σε ΠΥΣΔΕ
8. Εθνική πινακοθήκη Αθήνας
9. εκπαίδευση και ΜΜΕ
10. ενοικιαζόμενα δωμάτια καρπενήσι
11. εξ αποστάσεως εκπαίδευση
12. Ευρωπαϊκό δικαστήριο
13. ζώα που ζουν στο νερό και στην ξηρά
14. ιστορια της τεχνης
15. Μορφές Ρύπανσης Περιβάλλοντος
16. νησιά του Αιγαίου
17. Ο κόσμος των επενδύσεων
18. Ο ΡΟΛΟΣ ΤΟΥ ΠΑΡΑΜΥΘΙΟΥ ΚΑΙ ΤΗΣ ΜΑΡΙΟΝΕΤΑΣ ΣΤΗΝ ΑΝΤΙΜΕΤΩΠΙΣΗ ΤΩΝ ΜΑΘΗΣΙΑΚΩΝ ΔΥΣΚΟΛΙΩΝ

DOI: 10.1002/asi

19. Οδυσσέας Ελύτης
20. ΟΛΥΜΠΙΑΚΟΣ
21. ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
22. Περιοχές για μεταθέσεις καθηγητών στην περιοχή Άρτας
23. προβλήματα υγείας από τα κινητά τηλέφωνα
24. Προκηρύξεις για θέσεις εργασίας στην Αθήνα
25. Πως να βγάλω αχινό
26. Σελίδα της ΑΕΚ
27. σεμινάρια ενηλίκων στα ΚΕΕ
28. συνταγές για στιφάδο
29. Τεστ για την πιστοποίηση των εκπαιδευτικών
30. τεστ για το ECDL
31. τιμές για δίκλινα στο Καρπενήσι
32. τοπικές εφημερίδες αιτωλοακαρνανίας

Another conclusion drawn by this list of authentic user queries is that questions are formulated in an ad hoc manner. This means that most users type their query as they would type it in a document or as they would express it to a librarian. They use stopwords and type the terms either in uppercase letters or in title or sentence case. Sometimes, they even type queries in lowercase letters without accent marks (e.g., Queries 4 and 14). This ad hoc manner originates from the style an individual chooses to express his or her information needs in natural language. Search engines should be aware of these differences in the technical method chosen to express the information need and should therefore focus on the content rather than on the form of the query. This is especially important in non-Latin and non-English Web retrieval as certain documents where the terms appear in a different form than the query terms (e.g., upper- or lowercase, with or without accent marks) will not be retrieved. This is true at least for Greek Web retrieval, and probably is true in other more morphologically complex languages such as French, Serbian, or Russian.

Running the Queries

To realize the effect of stopwords in Web retrieval, we asked the 13 users to run in Google the 20 queries containing stopwords. Each user had to run one or two queries, then rerun the same queries without the stopwords. Thus, users were given one or two queries which included stopwords and the same queries without the stopwords. Participants were asked to evaluate the relevance of the first 10 results presented in Google. They had to report the number of retrieved documents, as indicated by Google, and the number of relevant documents in the top-10 results based, naturally, on their beliefs. Each query typed as it was suggested (i.e., in lower- or uppercase letters and with or without accent marks. These evaluation experiments were performed during a 2-hr IT course in mid-September 2006.

Table 2 presents the number of retrieved documents and the relevance estimates for the first 20 results. The number of retrieved documents is the number of results that Google displays on the header of its page, and is an approximation of the retrieved pages. Here, it is used as an indication of how stopwords influence the retrieval. Nevertheless, the most important measure of the negative or positive effects of stopword elimination is the relevance of the results. It has been argued that the first 10 or 20 results returned by a search engine hold the highest possibility to be viewed by users

TABLE 2. Retrieved pages and relevance of queries with and without stopwords.

	Stopwords		No stopwords	
Query	No. of retrieved pages	No. of relevant docs. in top 10	No. of retrieved pages	No. of relevant docs. in top 10
2	139,000	9	145,000	10
3	234	5	234	6
6	10,000	6	10,000	6
7	35	3	76	6
9	465,000	4	555,000	4
13	508	7	984	10
14	421,000	10	460,000	10
16	334,000	10	387.000	10
17	71,300	4	76,000	8
18	0	0	0	0
22	77	2	79	3
23	38,700	5	39000	7
24	55,500	7	59,800	10
25	30	0	63	0
26	643,000	4	564,000	4
27	708	4	709	7
28	721	7	784	9
29	806	3	818	5
30	11,400	4	11,500	7
31	121	3	182	4
Average	109,607	4.85	115,561.45	6.30

DOI: 10.1002/asi

(Silverstein, Henzinger, Marais, & Moricz, 1998). Therefore, as has been done in other studies (e.g., Chu & Rosenthal, 1996), the relevance was measured on the top-10 results of the ranked pages.

As can be seen, in almost all cases, more pages are retrieved when stopwords are omitted. The average number of retrieved pages also proves it. However, there is one case (Query 26) where the number of documents retrieved by Google diminishes when stopwords are omitted. Query 26 is a three-word query which tries to retrieve the Web page of a football team. In this case, we examined some of the results at the final positions of the rank, returned when the query is run with stopwords. Some of these results contain only the stopwords "της," which are erroneously considered relevant. It seems that Google acts like the unix's grep utility in this case using one of the query terms and thus retrieves more documents.

As explained, the increase of retrieved documents when common words are excluded from user queries is simply an indication of the influence of the stopwords in Greek Web retrieval. A valid criterion is the relevance of the documents. Table 2 shows that relevance increases or remains unaltered when stopwords are missing. Both the average number of relevant documents retrieved and the individual query instances demonstrate this. This is true in Query 26 as well. While relevance remains on the same level in 2- or 3-word queries (e.g., Queries 2, 14, 16), it significantly increases in longer queries containing more low discriminatory words (e.g., Queries 13, 22, 24, 29).

The queries were run in Google, which has an exceptionally fast searching mechanism. But in other search engines, the number of words may play an important role in the time required to retrieve the possibly relevant documents. For example, Lazarinis (2007b) showed that Yahoo (www.yahoo.com) and Anazitisis (www.anazitisis.gr), a native Greek search engine, require more time than does Google in Greek queries. Query 24 is a six-word query, and two of these words are stopwords. The six-word query in Yahoo needs 4 s to retrieve some pages. If we run the query without the stopwords, then only 2 s are required. In Anazitisis, the time required when stopwords are included is 1 min 14 s. When stopwords are not included, only 15 s are needed. These calculations were performed with the aid of the built-in utilities of Opera's Internet browser. These results clearly demonstrate that searching is faster when stopwords are removed from user queries. This result, along with the increase in relevance, makes elimination of stopwords a desirable feature in Greek Web retrieval and probably in other natural languages as well.

Discussion

This article presents the engineering phases of a stopword list for Greek. Initially, a set of HTML documents from various sources and domains were assembled. The Greek text included in these documents was processed to produce a list of valid terms. The frequency of these terms was calculated, and the top-99 words which appear numerous times in the text collection were used in our list. During the engineering process, some unexpected problems appeared which influence processing of Greek documents and might affect Web searching as well. The most important problem is that several pseudo-Greek words appear in Greek texts. These words consist of Greek and Latin encoded characters, which confuses retrieval systems since they cannot match to the regular form of the word. Additionally, we realized that stopwords appear in different forms (i.e., upper- or lowercase forms and sometimes without accent marks). Therefore, the stopword list should contain alternative forms for each entry.

The first top-20 common words form 30.26% of the total lemmas, which is in accordance with other studies (Baeza-Yates & Ribeiro-Neto, 1999). Therefore, one could significantly decrease the index size of an information retrieval system; however, by removing them during the indexing phase, the exact matching option offered by most search engines will not be supported. Something that could be exploited in Greek Web retrieval though, is the alternative forms of stopwords. As explained, the same stopword may appear with or without intonation and in upper- or lowercase. Instead of having to keep multiple stopword forms in the index file, stopwords could be normalized into one form. This could result in a small compression of the index file, but more significantly, it also could allow matching of user queries with documents regardless of the form of stopwords.

Web experimentation with 32 real-user-constructed queries illustrated that users do use stopwords in their queries. Moreover, the inclusion of stopwords leads to exclusion of some relevant documents, at least in the top-10 results. The average number of relevant documents was significantly improved once stopwords were excluded by user queries. This increase could be of utmost importance in e-shop catalog searching via local search engines. Lazarinis (2007a) showed that some searches fail; that is, they retrieve zero relevant products because user queries are capitalized or contain some words, usually stopwords, which are not part of the product's name. Normalized forms of the queries where query terms are in sentence or title case and do not include stopwords could be suggested or automatically rerun in these occasions.

As seen from the list of user queries, stopwords are used in alternative forms. This observation makes necessary our practice of storing alternative forms of the stopwords. Removal of stopwords from user queries could result in a shorter retrieval process. Time saving is important in cases of retrieval systems with modest performance in terms of speed.

The previous experiments reveal, as anticipated, that elimination of stopwords offers certain advantages over Web retrieval; however, note that most of the stopwords used in the sample queries belong to the first half of the stopword list. Although the words in the second half of the list presented in the appendix exhibit high frequency, their utilization in queries may be low. A new research direction would be the construction of stopword lists according to user query-formulation methods. This could allow Web retrieval

DOI: 10.1002/asi

systems to eliminate only those words which are overused in both documents and queries. This tactic could possibly result in better stopword-elimination practices. To study and construct stopword lists for this method, a large set of authentic queries would be needed.

In summary, stopword elimination has proven beneficial in traditional IR experiments and systems; however, their importance, primarily in terms of relevance, has not been extensively studied in non-English Web retrieval. More experiments are needed to construct optimized stopword lists and to evaluate their effects on retrieval efficiency and effectiveness.

Acknowledgments

I thank Professor John Tait of Sunderland's University for constructive comments on my article.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. Reading, MA: Addison Wesley, ACM Press.

Buckley, C. (1985). Implementation of the SMART Information Retrieval System. Tech. Rep. No. TR85–686, Cornell University, Ithaca, NY. Retrieved October 1, 2006, from ftp://ftp.cs.cornell. edu/pub/smart/

Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. Proceedings of the annual Conference for the American Society for Information Science, Baltimore (pp. 127–135).

Coverand, T.M., & Thomas, J.A. (1991). Elements of information theory. New York: Wiley.

Fox, C. (1990). A stop list for general text. ACM-SIGIR Forum, 24, 19–35.

Frakes, W., & Baeza-Yates, R. (1992). Information retrieval: Data structures and algorithms. Englewood Cliffs, NJ: Prentice Hall.

Kalamboukis, T.Z. (1995). Suffix stripping with modern Greek. Program, 29(3), 313–321.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. Proceedings of the 13th ACM International Conference on Information and Knowledge Management, Washington, DC (pp. 625–633).

Lazarinis, F. (2005). Do search engines understand Greek or user requests "sound Greek" to them? Open Source Web Information Retrieval Workshop in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology, France (pp. 43–46).

Lazarinis, F. (2006). Automatic extraction of knowledge from Greek Web documents. DIR'06 Information Retrieval Workshop, 33–37. Delft: The Netherlands.

Lazarinis, F. (2007a). Evaluating the searching capabilities of Greek e-commerce Web sites evaluating. Online Information Review Journal (in press).

Lazarinis, F. (2007b). Web retrieval systems and the Greek language: Do they have an understanding? Journal of Information Science. CA: Sage (in press).

Méndez Díaz, E., Vilares Ferro, J., & Cabrero Souto, D. (2005). COLE experiments at QA@CLEF 2004 Spanish Monolingual Track. Multilingual Information Access for Text, Speech and Images, 3491, 544–551. Germany: LNCS Springer-Verlag.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50(10), 944–952.

Silverstein, C., Henzinger, M., Marais, J., & Moricz, M. (1998). Analysis of a very large Alta Vista query log. Tech. Rep. No. 1998–014, COMPAQ Systems Research Center, Palo Alto, CA.

Tambouratzis, G., & Carayannis, C. (2001). Automatic corpora-based stemming in Greek. Literacy and Linguistic Computing, 16, 445–466.

Tsz-Wai Lo, R., He, B., & Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. DIR'05 Workshop (17–24). Utrecht, The Netherlands.

van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.

Zamora, E.M., Pollock, J.J., & Zamora, A. (1981). The use of trigram analysis for spelling error detection. Information Processing and Management, 17(6), 305–316.

Zou, F., Wang, F.L., Deng, X., & Han, S. (2006). Automatic identification of Chinese stop words. Research on Computing Science, 18, 151–162.

DOI: 10.1002/asi

Appendix

Και	(And)	Ήταν	(Was)
Το	(The)	Ενός	(One)
Να	(To)	Πολύ	(Very)
Του	(Of)	Όμως	(But, Yet)
Η	(The)	Κατά	(Against)
Της	(Of)	Αυτή	(She)
Με	(With)	Όταν	(When)
Που	(Where)	Μέσα	(In)
Την	(The)	Οποίο	(Who)
Από	(From)	Πως	(How)
Για	(For)	Έτσι	(Such)
Τα	(The)	Στους	(To, At)
Είναι	(Is, Are)	Μέσω	(Through)
Των	(Of)	Όλα	(Everything)
Σε	(At, In, To)	Καθώς	(Such as)
Ο	(The)	Αυτά	(These)
Οι	(The)	Προς	(Towards)
Στο	(To, At)	Ένας	(One)
Θα	(Will)	Πριν	(Before)
Τη	(The)	Μου	(My)
Στην	(To, At)	Όχι	(No)
Του	(The)	Χωρίς	(Without)
Τους	(The)	Επίσης	(Also)
Δεν	(No, Not)	Μεταξύ	(Between)
Τις	(The)	Μέχρι	(Until)
Ένα	(One)	Έναν	(One)
Μια	(One)	Μιας	(One)
Ότι	(That)	Αφού	(Since)
Ή	(Or)	Ακόμα	(Yet)
Στη	(To, At)	Όπου	(Where)
Στα	(To, At)	Είχε	(Had)
Μας	(Us)	Δηλαδή	(That is)
Αλλά	(But)	Τρόπος	(Manner)
Στον	(To, At)	Όσο	(As long as)
Στις	(To, At)	Ακόμη	(Yet)
Αυτό	(This)	Τόσο	(So much)
Όπως	(Like, As)	Έχουμε	(Have)
Αν	(If)	Ώστε	(So)
Μπορεί	(Maybe)	Αυτές	(Them)
Μετά	(After)	Γιατί	(Why)
Σας	(Your)	Πάνω	(On)
Δύο	(Two)	Τότε	(Then)
Τι	(What)	Τώρα	(Now)
Ως	(Until)	Κάτι	(Something)
Κάθε	(Every)	Άλλο	(Another)
Πρέπει	(Have to)	Μην	(Do not)
Πιο	(More)	Εδώ	(Here)
Οποία	(Who)	Είτε	(Either)
Μόνο	(Alone)	Μη	(Do not)
Ενώ	(While)		

DOI: 10.1002/asi

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0033-0337.htm

Received 26 September 2006
Revised 18 December 2006
Accepted 18 January 2007

Forming an instructional approach to teach web searching skills to non-English users

Fotis Lazarinis
Technological Educational Institute of Mesolonghi, Mesolonghi, Greece

Abstract

Purpose – Locating information on the internet is an important skill in the Information Society. Some recent studies showed that searching using non-English terms is a more demanding task than searching in English. Based on these observations, this paper aims to apply the Instructional System Design (ISD) methodology to analyse, design and implement a training course for Greek users. This instructional approach considers the explanation of the internal search engine intelligence and inefficiencies with respect to non-English natural language as its basic structural element.

Design/methodology/approach – Based on the ISD methodology, the tasks that needed to be trained as a web searcher were identified and a six-phase instructional sequence was constructed. The instructional methodology is evaluated with the aid of students in an authentic environment.

Findings – The evaluation revealed that learners who followed the structured approach and were aware of the search engines' limitations relating to the Greek language performed better in the web searching experiments.

Originality/value – The instructional methodology described can be applied in any course which aims at teaching basic web searching skills. The instructional approach presented can also be adapted to other non-English languages.

Keywords Information retrieval, Search engines, Worldwide web, Information literacy, Adult education, Greece

Paper type Research paper

1. Introduction

Searching for and locating information on the Internet is an important skill in the Information Society (Schlein, 2002) and a valuable tool for learning (Large and Beheshti, 2000). Previous studies showed that finding information on the Internet requires a variety of e-skills (Nachmias and Gilad, 2002). These e-skills vary from the ability to use search engines, to the key ability to transform the information need into an appropriate search query, and to the ability to browse through the retrieved set of documents. Each of these classes of skills can be further analysed, making information hunting a complex task.

Web information retrieval is more demanding when the query terms are in a language which exhibits considerable morphological variance like Greek or Russian or Hebrew which are not based on the Latin alphabet (Kolliakou, 1996; Bar-Ilan and Gutman, 2005; Lazarinis, 2005a). For instance, it was shown that searching using Greek terms is more challenging than in English Web searching (Lazarinis, 2005a; Lazarinis, 2005b). Greek searchers must be aware of the lack of ability of international and even local Greek search engines to value the characteristics of the Greek language and to handle properly queries typed in Greek. In the Greek language accents are used on lower case letters only. If a word is presented in capitals then accents are omitted.

Program: electronic library and information systems
Vol. 41 No. 2, 2007
pp. 170-179

0033-0337
DOI 10.1108/00330330710742935

Based on this difference, it was shown that queries with capital letter, or words with accent marks produce different results than those where the same words are unaccented and in lower case. In other words it is the form of the query terms which differentiates the rank of relevant documents and not the content of the query.

Based on the previous observations, in order to utilise successfully search engines one has to possess some knowledge of the internal intelligence and the weaknesses of search engines and therefore one needs to be trained in a disciplined way. In this paper we propose and evaluate a methodological approach based on Instructional System Design (ISD) (ISD, 2006). This method aims at providing learners with the required knowledge to access a search engine, to formulate their queries, to evaluate quickly the results and to navigate in the retrieved set of documents. Our approach originates from constructivism learning theory (Piaget, 1950) and is a mixture of learning by example and action learning.

2. Instructional System Design (ISD): an overview

ISD provides a means for sound decision making to determine the who, what, when, where, why, and how of training. ISD is divided into five phases which are briefly described below:

(1) *Analysis*:
- analyse system (department, job, etc.) to gain a complete understanding of it;
- compile a task inventory of all tasks associated with each job (if needed);
- select tasks that need to be trained (needs analysis);
- build performance measures for the tasks to be trained;
- choose instructional setting for the tasks to be trained, e.g. classroom, on-the-job, self study, etc.; and
- estimate what it is going to cost to train for the tasks.

(2) *Design*:
- develop the learning objectives for each task, to include both terminal and enabling objectives;
- identify and list the learning steps required to perform the task;
- develop the performance tests to show mastery of the tasks to be trained, e.g. written, hands on, etc.;
- list the entry behaviours that the learner must demonstrate prior to training; and
- sequence and structure the learning objectives, e.g. easy tasks first.

(3) *Development*:
- list activities that will help the students learn the task;
- select the delivery method such as tapes, handouts, etc.;
- review existing material so that you do not reinvent the wheel;
- develop the instructional courseware;
- synthesise the courseware into a viable training programme; and
- validate the instruction to ensure it accomplishes all goals and objectives.

(4) *Implementation*:
- create a management plan for conducting the training; and
- conduct the training.

(5) *Evaluation*:
- review and evaluate each phase (analyse, design, develop, implement) to ensure it is accomplishing what it is supposed to;
- perform external evaluations, e.g. observe that the tasks that were trained can actually be performed by the learner on the job; and
- revise training system to make it better.

3. An instructional approach based on ISD

Our aim is to teach basic web searching skills to people who already know how to use a web browser and are familiar with the concepts of the Internet and the Web. Thus we confine our efforts to adult learners and to high school students who already have a range of basic e-skills. By basic e-skills we mean the ability to handle efficiently an operating system, a word processor, a spreadsheet application, a web browser and e-mail software.

3.1 Analysis

To gain a complete understanding of the requirements of the course we asked 20 adult learners to search information about three topics relevant to public affairs. All users were aware of these topics which is an important prerequisite when an individual tries to ask questions about a subject. The participants were asked to write down the queries they devised for each topic before running them and hand them to the instructor. These queries were then qualitatively analysed.

From this first three-query test, it was clear that most users were unaware of the use of search engines and of how they could search. Also, we noticed that almost 90% could not complete the given assignment nor were able to discover information for at least one topic. Each problem faced was recorded so as to compile a list of all the difficulties, resulting either from user misconceptions or lack of knowledge. The difficulties confronted, related to formulating queries and retrieving relevant documents, arise from the following points:

- Difference between "broad" and "narrow" queries.
- Difference between upper and lower case query terms.
- The importance of accent and other intonation marks in searching.
- The function of stopwords which are not automatically removed as in English web searching.
- The significance of suffixes and especially of the final sigma which is not used only in the plural form as in English, e.g. both "Υπολογιστής" (singular form, nominative case) and "Υπολογιστή" (singular form, genitive and accusative case) mean Computer and "Υπολογιστές" (plural form, nominative case) means Computers. In Greek, suffixes result from conjugations of verbs and declension of adjectives and nouns and even first and last names.

Broad queries are those composed by one or two general terms which have several meanings in different contexts or do not adequately describe the information need (e.g. Athens). Narrow queries are those which consist of more than one, usually specialised, terms which describe the information request in a better manner (e.g. Athens Georgia United States). The other topics are related mostly to the inability of search engines to value all the attributes of non-Latin based languages.

From this initial experiment it was apparent that the previously mentioned user difficulties should be taken into consideration in training and that training should happen in a laboratory with the use of computers and carefully designed examples.

3.2 Design

The main learning objectives of the training course are to provide learners with the ability to:

- access a web retrieval system;
- formulate and refine queries;
- evaluate the retrieved documents based on the summaries and on the highlighted matching terms; and
- visit the retrieved set of documents.

In this first phase we have not dealt with advanced searching capabilities, such as Boolean searching, or other options that the international search engines support.

Figure 1 shows the learning steps, identified during the design phase, required to teach learners how to use search engines.

Following these steps teachers will be able to communicate their knowledge to their students. As can be seen, navigation of the result set could lead the learning procedure to step backwards to either of the previous two stages. If no results are retrieved then additional explanations and queries should be given to users and thus this outcome could push the learning procedure to a previous step of the sequence of learning steps.

3.3 Development

At this stage specific activities should be listed and the courseware should be constructed. The following list of activities is scheduled in our instructional approach:

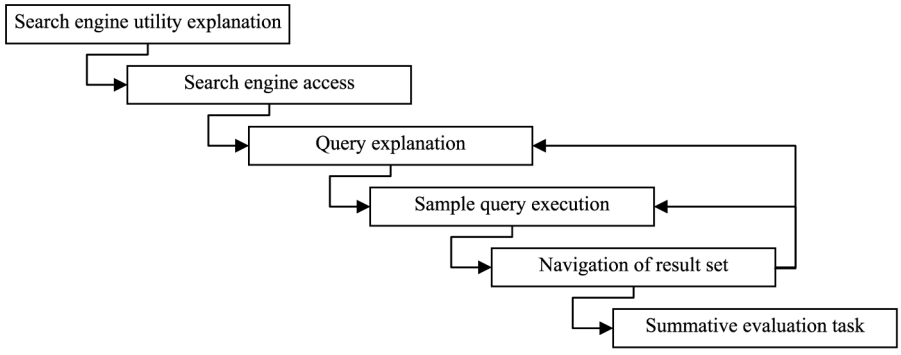


Figure 1.
Learning steps required to teach learners how to use search engines

(1) Explain the importance and the functions of search engines.

(2) Access the Greek version of Google at www.google.gr.

(3) Explain how to enter queries and how to initiate searching.

(4) Explain the various elements of the result list (e.g. link of relevant document, summary).

(5) Explain how to evaluate quickly the retrieved documents based on the summaries and on the highlighted matching terms.

(6) Visit the first result.

(7) Go back and visit the second result.

(8) Explain the concept of broad and narrow queries.

(9) Run an example, e.g. "Ολυμπιακοί αγώνες" (Olympic games) and "Ολυμπιακοί αγώνες 2004" and ask users to observe that the number of results and the highly ranked results differ between the two runs as shown in Figures 2 and 3.

(10) Explain the difference between upper and lower case query terms.

(11) Ask users to run the query "ΟΛΥΜΠΙΑΚΟΙ ΑΓΩΝΕΣ 2004" which retrieves a different set of documents, at least in the first 10 results as can be seen in Figure 4.

(12) Explain the importance of accent and other intonation marks in searching.

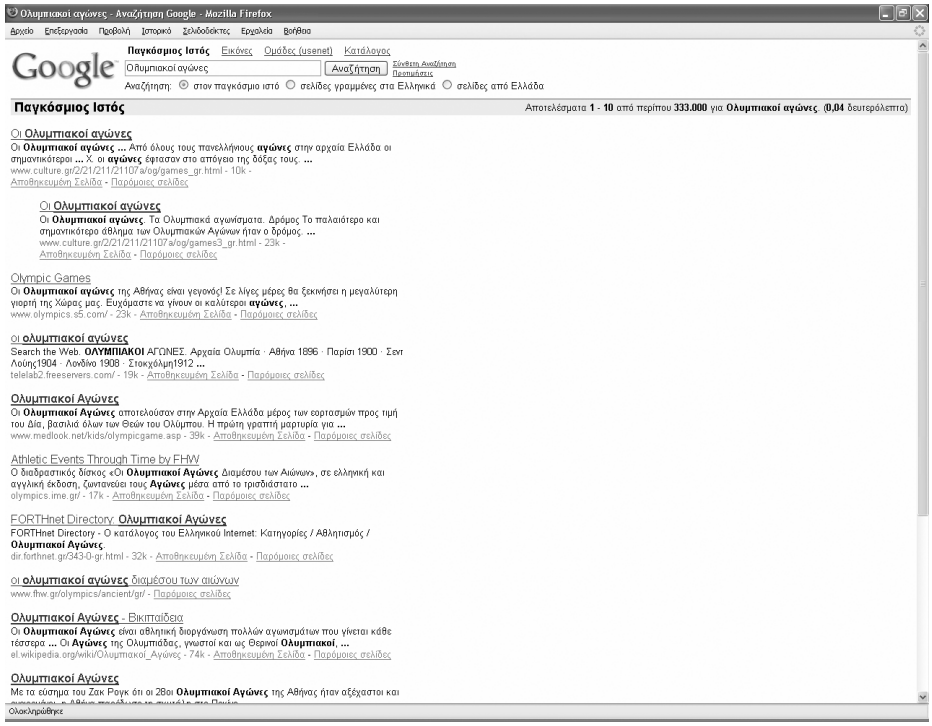


Figure 2.
Search results for
Ολυμπιακοί αγώνες

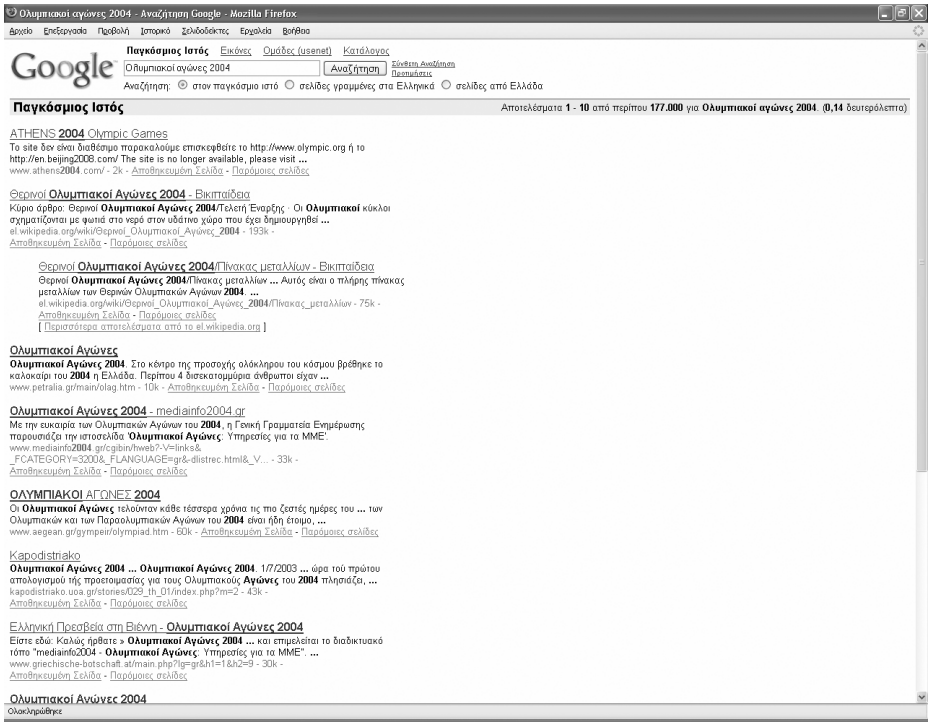


Figure 3.
Search results for Ολυμπιακοί αγώνες 2004

(13) Ask users to run the queries "Ευρωπαϊκό δικαστήριο" (European court of law) and "Ευρωπαικό δικαστήριο". In the second query the umlaut was not used. Both queries retrieve the same results but are ranked differently.

(14) Explain the function of stopwords (e.g. articles, prepositions).

(15) Run the examples "Προβλήματα υγείας από τα κινητά τηλέφωνα" (Health problems of mobile phones) and "Προβλήματα υγείας κινητά τηλέφωνα" (Health problems mobile phones). These two examples retrieve and rank documents differently.

(16) Explain the significance of suffixes.

(17) Guide users to run queries "κάρτα γραφικών" (graphics card) and "κάρτες γραφικών" (graphics cards). These queries produce different ranks of documents.

We believe that these tasks will help learners to realise fully the capabilities of search engines and some of their deficiencies related to the handling of non-English queries. At this point we should underline that we restricted our experiments to Google since it is the only search engine which values some of the attributes of the Greek language and provides a Greek interface to the search engine.

These activities were validated against the questions raised in the analysis phase and the learning steps identified in the design stage.

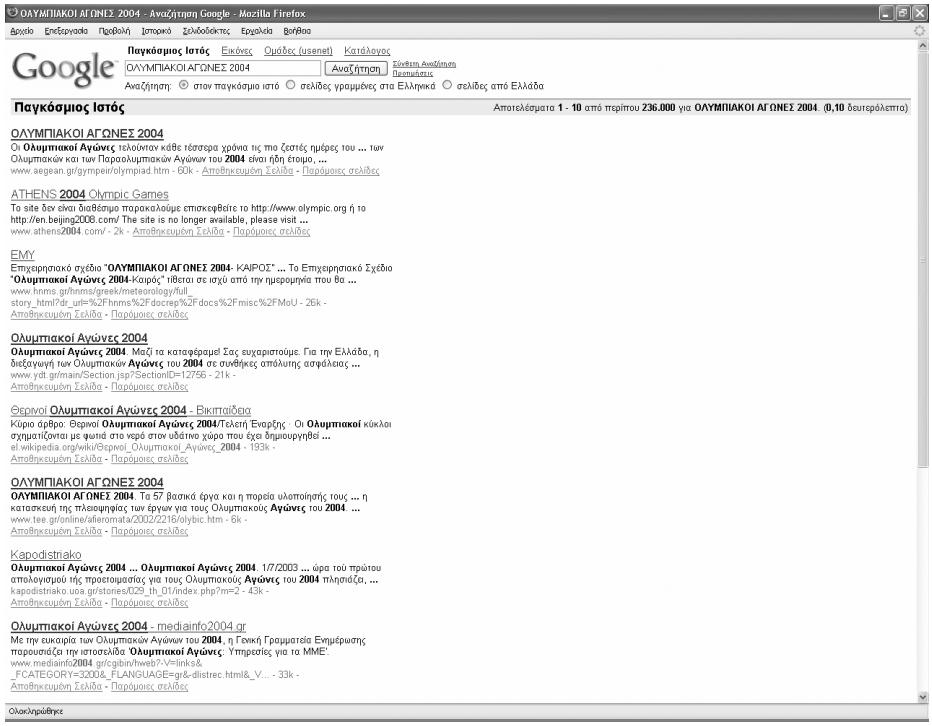


Figure 4.
Search results for ΟΛΥΜΠΙΑΚΟΙ ΑΓΩΝΕΣ 2004

3.4 Implementation

The activities described in the previous section were taught in two classes: one a high school class of 19 students and one an adult class of 28 learners. Learners were familiar with the concept of the Web and they could access web pages. Each class session lasted for two hours. During the first hour students were instructed on how to use the search engines, following the steps and activities described previously. In the second part of the teaching, an assignment was handed out and completed by learners.

To compare the efficiency of our method concerning issues related to Greek language searching we carried out a further set of two classes, but this time leaving out activities 10 to 17 described in section 3.3. This group of learners consisted of a class of 20 high school students and a class of 26 adult learners.

3.5 Evaluation

Estimation of the success of each teaching technique relied on an evaluation task. The task was distributed to each participant and composed of two sets of queries in Greek. The first set consisted of four one-word or two-word queries. These words were general purpose computer related terms (e.g. binary system) of which the participating student groups were aware. Two queries were in capital letters without intonation (i.e. accents) and the other two were in lower case letters with intonation. The second group of queries consisted of descriptions of the information need rather than specific queries. So in this case participants had to construct their own queries.

Table I refers to the first set of queries. Participants were asked to return at least six relevant URLs for each query. Students who were taught the whole list of activities are symbolised as SG1 and AG1, while SG2 and AG2 form the second learner group with the condensed list of activities.

As seen in Table I all participants were able to discover relevant documents for queries typed in lower case letters with intonation. Since the query terms referred to focused, and uniquely identifiable computer science terminology, Google ranked highly the truly relevant results. For the other two queries, typed in capital letters, some participants could not successfully complete the task. This is because some of the first ranked results were not relevant, so students had to devise ways of refining their query in order to discover relevant pages. Some of them, and especially those in the second learner group, were unable to formulate and execute alternative queries.

Another observation made, which was not formally measured though, is the task's completion time. The second student group needed more time to complete the task than learners who were aware of the issues related to the Greek language.

After the first experiment, the students were asked to complete the second task and to return at least three relevant URLs. Table II shows that a high percentage of the first group accomplished the task. All of them were able to discover three relevant Web locations for at least one query. On the other hand, the vast majority of adolescent and adult students who had no idea about stopwords, suffix removal and capitalisation were unable to complete fully their assignment. A small number of them were able to discover relevant web sites for only one query and most of them for none.

In both cases high school students performed better than the adult participants, as they were more competent computer users and could type queries and assimilate and evaluate the retrieved web pages faster.

	Completion of lower case queries with accents %	Completion of upper case queries without accents %
SG1 ($n = 19$)	100	100
SG2 ($n = 20$)	100	70
AG1 ($n = 28$)	100	96.43
AG2 ($n = 26$)	100	52.63

Notes: SG1, SG2: Student Groups; AG1, AG2: Adult Groups

Table I.
Evaluation data of the first query set

	Full task completion %	Completion of at least one query %
SG1 ($n = 19$)	94.74	100
SG2 ($n = 20$)	15	30
AG1 ($n = 28$)	85.71	100
AG2 ($n = 26$)	7.69	19.23

Table II.
Evaluation data of the second query set

4. Summary

In this paper we proposed a methodological teaching approach for searching information on the web applicable both to English and to other spoken languages with inflections and intonation. This principled teaching methodology relied on the ISD methodology. During the analysis phase we identified the tasks that needed to be included in training and in the later phases we constructed a six-phase instructional sequence of tasks (Search engine utility explanation, Search engine access, Query explanation, Sample query execution, Navigation of result set, Summative evaluation task). Then a list of activities was constructed which takes into account the searching behaviour of international search engines.

Learners who followed the principled instructional approach successfully accomplished their assignments. On the other hand, students who had not been properly instructed on how to use search engines and students who were not aware of the deficiencies of search engines related to non Latin queries performed worse and in several occasions could not retrieve relevant documents.

The problems identified in this study and the instructional approach presented, are applicable to other non English languages as well. For example, the queries "Libreria Roma" (bookshop Rome), "Librerie Roma" (bookshops Rome) and "Librerie di Roma" (bookshops of Rome) retrieve different results in Google. In Yahoo the retrieved set of documents differs between the queries "Università di Roma" (University of Rome) and "Universita di Roma". The German queries "das Wasser", "des Wassers" and "die Wässer" (kinds of potable water) retrieve different results. Also the queries "Wasser" and "die Wasser" produce different ranks in both Yahoo and Google. In other words the article "die" significantly influences the retrieval process. In complex languages, like German or other European languages or even Asian languages, the proposed instructional approach is important as it will help users refine their queries and eventually retrieve more relevant Web pages.

The main conclusion drawn is that searching the internet is not an easy task. Teachers should follow disciplined approaches to equip their students with all the necessary abilities and knowledge so as to utilise successfully searching systems. This is even more important in cases of searches in specific domains (Meskó, 2003). That is why a few techniques such as educational games (Halttunen and Sormunen, 2000) and visualisation tools (Brusilovsky, 2002) have been used to support teaching of information retrieval. When teaching advanced retrieval techniques or specialised searches in specific databases, more techniques and approaches need to be devised so as to help learners.

A final argument is that non-English speaking users need to be more creative and knowledgeable than English searchers when using search engines. This conclusion should be taken into account when designing courseware. Teachers should re-engineer their teaching methodology to take into consideration some of the basic internal search engine characteristics in order to help their students.

References

Bar-Ilan, J. and Gutman, T. (2005), "How do search engines respond to some non-English queries?", *Journal of Information Science*, Vol. 31 No. 1, pp. 13-28.

Brusilovsky, P. (2002), "Web-based interactive visualization in an information retrieval course", *Proceedings of ED-MEDIA'2002 –World Conference on Educational Multimedia,*

Hypermedia and Telecommunications, AACE, Denver, CO, pp. 198-203, available at: www2.sis.pitt.edu/ ~ peterb/papers/EDMED02IR.pdf

Halttunen, K. and Sormunen, E. (2000), "Learning information retrieval through an educational game: is gaming sufficient for learning?", *Education for Information*, Vol. 18 No. 4, pp. 289-311.

ISD (2006), *Instructional System Design*, available at: www.nwlink.com/ ~ donclark/hrd/sat.html

Kolliakou, D. (1996), "Definiteness and the make-up of nominal categories", *Edinburgh Working Papers in Cognitive Science*, Vol. 12, pp. 121-64.

Large, A. and Beheshti, J. (2000), "The web as a classroom resource: reactions from the users", *Journal of the American Society for Information Science*, Vol. 51 No. 12, pp. 1069-80.

Lazarinis, F. (2005a), "Evaluating user effort in Greek web searching", *10th Pan Hellenic Conference: Advances in Informatics, University of Thessaly, Greece*, pp. 99-109.

Lazarinis, F. (2005b), "Do search engines understand Greek or user requests 'sound Greek' to them?", *Open Source Web Information Retrieval Workshop, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, France*, pp. 43-6, available at: www.emse.fr/OSWIR05/2005-oswir-p43-lazarinis.pdf

Meskó, E. (2003), "Teaching information retrieval in the chemistry curriculum", *Chemistry Education: Research and Practice*, Vol. 4 No. 3, pp. 373-85.

Nachmias, R. and Gilad, A. (2002), "Needle in a hyper stack: searching for information on the World Wide Web", *Journal of Research on Technology in Education*, Vol. 34 No. 4, pp. 475-86.

Piaget, J. (1950), *The Psychology of Intelligence*, Routledge, New York, NY.

Schlein, A. (2002), *Find it Online*, Facts on Demand Press, New York, NY.

Corresponding author
Fotis Lazarinis can be contacted at: lazarinf@teimes.gr