Improving text search performance with grammar support

Damir Krstinić, Mr.Sc. and Ivan Slapničar, Ph.D.
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
University of Split
R. Boškovića b.b., 21000 Split, Croatia
E-mail: damir.krstinic@fesb.hr, ivan.slapnicar@fesb.hr
Phone: +385 21 305 617

Abstract— **As a result of a fast growth in the amount of digital data available, automated techniques for indexing and search have been developed based on a vector space model. Great majority of modern Information Retrieval systems are oriented to the English language, with relatively simple grammar rules. For languages with complex grammatical rules, lack of grammar support introduces a drawback in performance of these systems. We have designed an experimental web search engine with language grammar support based on Ispell data format. As part of this work, Croatian language files have been developed. First results of experiments with the developed system have confirmed improvement achieved by implementing language grammar.**

I. Introduction

DEVELOPMENT of digital technologies and Internet has provided a way for a fast information exchange. In an environment where almost any information is accessible, the problem of extracting relevant knowledge from enormous amounts of available data arises. To provide users with the tool for locating and extracting relevant data, number of techniques were developed for automatic indexing of textual materials.

Modern information retrieval systems are based on the concepts of a vector space [2]. Data is modeled as a term-by-document matrix, where each column of the matrix represents a vector associated with one document. Each component of the vector is associated with a particular term, where a nonzero value represents occurrence of the term in the document. These values are often weighted to emphasize the importance of distinguishing terms. To compare the user query with the documents in the collection, the query must be coded in the same vector space defined by the dictionary of terms. Geometrical interpretation of query matching is finding the document vectors closest to the query vector according to some measure [1]. Common measure of similarity of the user query and the documents in the list is the cosine of an angle between the query and document vectors.

For large documents collections, each document generally uses only a small subset of the entire dictionary of terms generated for a given collection, and most of the vector elements are zero. The technique of *Latent Semantic Indexing (LSI)* ([2], [3], [4]) uses truncated Singular Value Decomposition to project high dimensional documents and query vectors into a low dimensional subspace. Projecting high dimensional vectors onto the low dimensional space has an effect of grouping similar vectors and emphasizing similarities not visible in the original vector space. Additionally, lowering the rank of the term-by-document matrix lowers the computational cost of query matching and memory needed for storing document vectors.

As a dictionary defining the vector space in which documents are coded, a set of all terms found in the entire collection is often used. This is especially the case for large collections with documents from various fields and thematics. In the process of data indexing, each new term found generates a new row in the term-by-document matrix. In the languages with a complex set of grammatical rules, the same word could have many different forms, resulting in several rows of the matrix representing the same word. As a result, comparing the user query with the collection will not yield all of the relevant data. If LSI method is applied to this matrix, important connections and similarities will not be extracted. Besides, additional computational power and memory is needed to store and process generated data.

In order to improve the performance of the *Information Retrieval (IR) System*, basic knowledge of the language grammar should be implemented. We have developed an experimental web search engine with language grammar support. The developed search engine is based on the vector space principles, and generated term-by-document matrix is processed with LSI method. As a format for a grammar support, *Ispell* [7] data format was selected. Croatian grammar support was developed with the dictionary of approximately 20000 terms and more then 100 grammatical rules. When rules are applied to the dictionary, list of 350000 different words is generated. Using this dictionary in compact form, significant improvement in web search results for Croatian language is achieved. The document index created with this technique is more consistent [1] and querying the database yields better results.

The rest of the paper is organized as follows: Results of implementing grammar support in the Information Retrieval system are presented in the following section. In section 3 performance of the developed search engine is compared with results obtained using *Google* [5] and *Yahoo* [6], followed by the conclusion in the last section.

TABLE I

Query "znanost genetika" (science genetics)

Search results with grammar support
0.58480 `http://www.klik.hr/vijesti/2003/08/13/0002003.html` *Courtney Love - unuka Marlona Branda!* *Kao rockerica se s bendom Hole i nije nešto proslavila, a kao glumica ... jasno i zašto - u pitanju je genetika...*
0.56330 `http://www.klik.hr/naslovnica/znanost/200108210002025.html` *Svijet je izgubio razvojne stanice* *Velika većina, ako ne i sve, embrijske razvojne stanice koje su dostupne istraživačima iz javnog sektora ...*
0.54189 `http://www.klik.hr/vijesti/2001/04/27/0002024.html` *Otkriven genetski uzrok neplodnosti* *Znanstvenici vjeruju kako su pronašli gen koji determinira hoće li žensko tijelo odbaciti oplodeno jajašce ili ...*

Search results without grammar support
0.74962 `http://www.klik.hr/vijesti/2003/08/13/0002003.html` *Courtney Love - unuka Marlona Branda!* *Kao rockerica se s bendom Hole i nije nešto proslavila, a kao glumica ... jasno i zašto - u pitanju je genetika...*
0.59492 `http://www.klik.hr/naslovnica/lovesex/200403090002003.html` *Seksi sličice - istraživanje terena* *Naravno da muškarci imaju senzibilnije mozgove nego žene - da ne bude zabune - samo kad se radi o ...*
0.56498 `http://www.klik.hr/naslovnica/hotcool/200307180002006.html` *Tražili ste, izdržite!* *Kad malo bolje razmislite, jeste li sigurni da baš želite gledati muškarce u suknjama? Njemački časopis Max ...*

TABLE II

Query: "povijest avijacije" (aviation history)

Search results with grammar support
0.99239 `http://www.klik.hr/naslovnica/scitech/200312170002004.html` *Jesu li braca Wright zbilja poletjeli prvi?* *Moda jesu, a mozda i nisu, jer Brazilci imaju svog pionira avijacije kojem pripisuju prvi uspješan let...*
0.79265 `http://www.klik.hr/naslovnica/scitech/200312050002004.html` *Hoce li Amerikanci opet na Mjesec?* *Prema prvim najavama, George Bush bi ovog mjeseca mogao najaviti nove letove u svemir i pojacani program...*

Search results without grammar support
0.99432 `http://www.klik.hr/naslovnica/scitech/200312170002004.html` *Jesu li braca Wright zbilja poletjeli prvi?* *Moda jesu, a mozda i nisu, jer Brazilci imaju svog pionira avijacije kojem pripisuju prvi uspješan let...*
0.51489 `http://www.klik.hr/naslovnica/znanost/200206200002034.html` *Siromašni češe umiru od raka nego bogati* *Pedesetih i šezdesetih godina prošlog stoljeća, bogati su Amerikanci češe umirali od raka nego siromašni, no ...*

II. Grammar support in experimental web search engine

USING our experimental web search engine, experiment was conducted in which document index was created for web site*Klik* (`http://www.klik.hr`). The indexing process was started twice, one time with grammar support, using developed Croatian grammar, and second time without grammar support. In this experiment, the number of documents was limited to 3000. Both indexes were tested with the same set of queries and results were compared.

As expected, for majority of queries results were better when search was performed on the document index created using grammar support. For several queries, results were comparable, or the same in both cases. After inspecting dictionaries generated in the indexing process, it was found that this was the case when some

TABLE III

QUERY "EKONOMIJA" (ECONOMY))

Experimental web search engine with grammar support
http://www.klik.hr/danas/novac/200109110002064.html *Teroristički napad mogao bi gurnuti svjetsko gospodarstvo u recesiju* *Njujorška burza našla se u potpunom rasulu nakon što su teroristi s dva zrakoplova simbol američkog ...* http://www.klik.hr/naslovnica/novac/200201040002038.html *Euro u Hrvatskoj* *Uvodenje eura kao jedinog sredstva gotovinskog plaćanja u zemljama Europske unije od 1. siječnja 2002...* 0.66297 http://www.klik.hr/naslovnica/novac/200203060002034.html *Cijena sirove nafte najviša u posljednjih pet mjeseci* *Porast cijene potaknut je očekivanjima da će Rusija nastaviti suradnju s OPEC-om (Organization of ...*
Google results
http://www.klik.hr/vijesti/2001/01/22/0002094.html *Osnovne činjenice o euru* *što je euro, koje sve države koriste euro, tko procjenjuje njegovu vrijednost, koja je uloga Europske centralne ...* http://www.klik.hr/vijesti/2004/03/01/0002002.html *Oscar: provale i zahvale* *Samo je jedan dobitnik (za kratki igrani film) toliko trkeljao da ga je morala prigušiti glazba, a unaprijed ...* http://www.klik.hr/naslovnica/dossier/200403250002014.html *Ishlapila i Coca-Cola?* *Nakon fijaska s flaširanom vodom iz pipe, dionice Coca-Cole su na burzi dostigle najnižu vrijednost još od ...*
Yahoo results
http://www.klik.hr/naslovnica/novac/200201090002053.html *Senzacionalni rezultati istraživanja Instituta za javne financije: Siva ekonomija u Hrvatskoj ispod 10 posto* *Mjereno metodom nacionalnih računa (kojom se računa domaći bruto proizvod), tzv. neslužbeno ...* http://www.klik.hr/naslovnica/novac/200204080002065.html *Hrvati zaraduju milijune na trgovini drogom* *Hrvati godišnje od trgovine drogom zarade 400 milijuna kuna, od prostitucije 300, a trgovanja ljudima više ...* http://www.klik.hr/vijesti/2001/01/22/0002094.html *Osnovne činjenice o euru* *što je euro, koje sve države koriste euro, tko procjenjuje njegovu vrijednost, koja je uloga Europske centralne ...*

TABLE IV

QUERY "RAZVOJ AVIJACIJE" (DEVELOPMENT OF AVIATION)

Experimental web search engine with grammar support
http://www.klik.hr/naslovnica/scitech/200312170002004.html *Jesu li braca Wright zbilja poletjeli prvi?* *Moda jesu, a mozda i nisu, jer Brazilci imaju svog pionira avijacije kojem pripisuju prvi uspješan let...* Same result obtained with a query "avijacija razvoj" (aviation development)
Google and *Yahoo*
Asinkroni kompjutori *Znanstvenici su razvili novu generaciju hardwarea i softwarea na temelju jednostavnijeg dizajna iz pedesetih ...* No documents were found for queries "avijacija razvoj", "povijest avijacije" or "avijacija povijest"

or all of the words used in the query were not a part of the initial Croatian dictionary of terms marked with the appropriate grammar rule. In some cases, querying the document index created with no grammar support returned no results, even if documents with appropriate content were found in the collection, but words from the query were used in different grammatical form.

Results of querying the collection with the query `"znanost genetika" (science genetics)` are given in table I. In the upper part of the table, the results of text search with grammar support are given, followed by the results obtained using the document index created without grammar support. In both cases, first three documents with greatest similarity measure (cosine of the angle between the query and the document vectors [2]) are given. For this experiment, a document was considered relevant for the given query if this value was greater than 0.5. Each document is presented with its title and fraction of the document text (all in Croatian), with similarity measure printed by the document address.

In both cases, the best result (first document in the list) is the same, while others differ. All three documents found using the index created with grammar support have content relevant to the query, while second and third document found with no grammar support have a little or no connection to the thematics defined by the query. Another interesting fact is that the similarity measure for the first result in the list is greater when the search is performed without grammar support. This is due the fact that words from the query in the exact grammatical form as in the query appear only in a small number of documents. As a result, those documents are highly ranked in the results list. When search is performed with grammar support, all grammatical forms of the words are recognized, and ranking of the particular document with the word from the query is lower.

In table II the results for the query `"povijest avijacije" (history of aviation)` are given. In this case, only two results were found with similarity over 0.5 for both indexes. As in previous case, the first result is the same, with thematics relevant to the user query. When search is performed with grammar support, second document found is about space flights and NASA, which can be considered relevant to the query. When search is performed without grammar support, second document found has no connection to the thematics defined by the query.

Another test was conducted with the query `"avijacija povijest" (aviation history)`, which is basically the same query, but one word from the query `(avijacija)` is in different grammatical form. In this case, results returned when querying document index created with grammar support are the same as in the previous case, while querying database without grammar support returned no results. This suggests that word `"avijacija"` (this is basic form of this noun) in the documents was found only in the form `"avijacije"` (genitive). For web indexing and search system without grammar support, two different forms of the same word represent different terms, while system with grammar support can recognize all different grammatical forms of the same word.

III. Comparing with commercial search engines

PERFORMANCE of the experimental search engine with grammar support was compared with results obtained by widely used search engines *Google* and *Yahoo.* As in the first experiment, domain `klik.hr` was used as a test case. Indexing process was started on 25. March 2004. and 17075 documents were found in the named domain. Created document index was tested with a set of queries. *Google* and *Yahoo* search engines were tested with the same set of queries with search process limited to the documents from the domain `klik.hr`

In most cases, results obtained by our search engine were comparable with the results from *Google* and *Yahoo.* Sets of documents obtained for a particular query usually differ from one search engine to another, at least in ordering of documents in the results list, due to differences in an indexing and search algorithms.

In a number of cases, results returned by our search engine were better when compared to the results obtained by *Google* or *Yahoo.* In the tables III and IV results for two queries are presented.

In table III results are given for the query `"ekonomija" (economy)`. First three documents obtained by our experimental search engine with grammar support are given first, followed by the results obtained by commercial search engines. Although results differ for different search engines, all documents found can be considered relevant for the given query, except second best result obtained by *Google.* Thematics of this document is Oscar Movie Award, but word `"ekonomija"` is used in this document in its exact form as in the query, resulting in high ranking of this document in the search results. Without grammar support, all other forms of this word are recognized as separate terms and documents using other forms are not considered valid search results. With the fact that in the Croatian language nouns can have up to 15 different forms, and verbs up to 40, it is obvious that the lack of grammar support can result in a significant drawback in search engine performance. Even for the English language, with relatively simple grammatical rules, lack of grammar support can result in a drawback of search engine performance [1].

Next, search engines were tested with queries "povijest avijacije" and "avijacija povijest" (search results for this query with grammar support are given table II), and no documents were found with commercial search engines. Results for the similar query "razvoj avijacije" (development of aviation) are given in table IV. For this query, our search engine returned the same document as for the query "povijest avijacije". The same result was obtained with a slightly modified query "avijacija razvoj" (aviation development), with word "avijacija" in its basic grammatical form.

For the query "avijacija razvoj" *Google* and *Yahoo* returned no results, while for the query "razvoj avijacije", both search engines returned the same document. This document is about development of a asynchronous computers and has no connection to the development of aviation. However, appearance of both words from the query in the document in the exact grammar form as in the query resulted in ranking this document as only valid result for the given query.

IV. Conclusion

In our analysis we have shown that modern automated information retrieval systems have no knowledge of language structure and syntax. These systems are based on English language, which has relatively simple grammatical rules. For languages with more complex grammar, lack of grammar support can introduce a significant drawback in query matching. With this fact in mind, experimental web search engine with support for language grammar was designed. Our search engine is based on the vector space model and Latent Semantic Indexing algorithm. As part of this work, grammar rules for the Croatian language are implemented, and the dictionary is created with about 20000 words marked with appropriate rule, expanding the dictionary to over 300000 words.

Influence of language grammar on the search results was tested by indexing the same set of documents with and without grammar support. As expected, superior results of query matching were achieved when search was performed with grammar support. Improvement in the search result for a particular query is proportional to the number of terms in the query that can be found in the dictionary of terms marked with a grammatical rule.

Performance of the developed search engine on Croatian language documents was compared with two widely used search engines *Google* and *Yahoo*. In most cases, the results of our search engine were comparable to those of *Google* and *Yahoo*. This experiment confirmed lack of grammar support in conventional search engines, as search results strongly depend on a grammatical form of words appearing in a query. For some queries *Google* and *Yahoo* returned no results, although documents with relevant thematics are present in the collection and have been found by the search engine with grammar support. In other cases, documents have been highly ranked in the result list if a word from the query was used in exact grammatical form as in the query, although content of the document does not fit into the thematics defined by the query.

Considering the fact that Croatian grammar and dictionary used in the experiments covers only a small subset of the dictionary and grammar rules, achieved results are very promising. Additional improvement could be achieved by adding new grammar rules and new words to the dictionary. In fact, at this moment a new version of the Croatian grammar is prepared together with an updated dictionary with about 40000 words marked with an appropriate rule.

Data format for grammar and dictionary files used in the developed search engine is based on the *Ispell* data format. *Ispell* files are available for great number of languages widely used on the Internet. This way, the search engine can easily be configured for any language with available *Ispell* files, resulting in better performance than conventional search engines.

As the main purpose of the developed search engine is experimental, with the main goal to confirm hypothesis of influence of language grammar on performance of information retrieval systems, additional improvements can be achieved by optimization of the indexing and search algorithms.

References

[1] Damir Krstinić, Ivan Slapničar, *Web indexing and search with local language support*, in: Proceedings of SoftCOM 2003, pp. 488-492, Split, 2003.

[2] Michael W. Berry, Zlatko Drmač, Elizabeth R. Jessup, *Matrices, Vector Spaces and Information Retrieval*, SIAM Review, Volume 41, Number 2, pp. 335-362, 1999.

[3] Michael W. Berry, Susan T. Dumais, Gavin W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review, Volume 37, Number 4, pp. 573-595, 1995.

[4] Parry Housband, Horst Simon, Chris Ding, *On the Use of Singular Value Decomposition for Text Retrieval*, Proc. of SIAM Comp. Info. Retrieval Workshop, 2000.

[5] Google search engine, http://www.google.com.

[6] Yahoo search engine, http://www.yahoo.com.

[7] *Ispell, interactive spell-checking program* http://www.gnu.org/software/ispell/ispell.html