Indexing of textual databases based on lexical resources: A case study for Serbian

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović

Indexing of textual databases based on lexical resources: A case study for Serbian | Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović | Semantic Keyword-based Search on Structured Data Sources : First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, September 8-9, 2015. Revised Selected Papers | 2015 | |

10.1007/978-3-319-27932-9_15

http://dr.rgf.bg.ac.rs/s/repo/item/0000870

Indexing of textual databases based on lexical resources: A case study for Serbian

Ranka Stanković[1], Cvetana Krstev[2], Ivan Obradović[1], and Olivera Kitanović[1]

[1] University of Belgrade, Faculty of Mining and Geology,
ranka@rgf.bg.ac.rs, ivan.obradovic@rgf.bg.ac.rs,
olivera.kitanovic@rgf.bg.ac.rs
[2] University of Belgrade, Faculty of Philology,
cvetana@matf.bg.ac.rs

Abstract. In this paper we describe an approach to improvement of information retrieval results for large textual databases by pre-indexing documents using bag-of-words and Named Entity Recognition. The approach was applied on a database of geological projects financed by the Republic of Serbia in the last half century. Each document within this database is described by metadata, consisting of several fields such as title, domain, keywords, abstract, geographical location and the like. A bag of words was produced from these metadata using morphological dictionaries and transducers, and named entities within the metadata were recognized using a rule-based system. Both were then used for indexing documents and ranking was based on *tf_idf* measure. Evaluation of ranked retrieval results based on data obtained by pre-indexing are compared to results obtained by informational retrieval without pre-indexing with Precision-Recall Curve, showing a significant improvement in terms of Mean Average Precision measure (MAP).

1 Introduction

Three basic problems related to Information Retrieval (IR) are the presentation of document content, the presentation of information needs and the comparison of these two representations. Presentation of documents as a rule contains metadata about the document, such as title, abstract, location or, in case of indexing of content contained in a textual database, the primary key of the document in the database. If the search is performed by scanning textual documents, then their additional representation is not required. In order to increase efficiency, especially in the case of large collections, a formal representation surrogate of each document is usually formed. Based on these representations, during the preparatory phase an index of the collection of documents is formed, which is then used in the search phase. Automatic assigning of surrogates is usually done by extracting and selecting terms (words) that appear in the text of documents. To that end, many natural language processing (NLP) methods and techniques are used: determining the boundaries of sentences, tokenization, stemming, tagging, recognition of nominal phrases and named entities and, finally, parsing. [4]

Finding and ranking of relevant documents on basis of the index is realized using the model of approximate matching, based on the frequency distribution of terms and documents. Two basic approaches are the vector space model, based on weight coefficients of terms, and the probabilistic model, based on relevance feedback. [14]

In this paper we describe the motivation for developing an IR system in geological domain, current approach based on text scanning, its shortcomings, the improvement developed using NLP methods, and the evaluation of this improvement.

2 Motivation

The constant increase of geological data and information in Serbia was not accompanied by adequate development and introduction of modern information technologies until recently. Thus the management of the various geological documentation has been organized on the principle of traditional libraries and archives, where obtaining specific information was difficult or time consuming. Usage analysis of the geological investigations results, stored in numerous archives and document libraries, showed that they were inefficiently used because of inadequate organization, limited access and general lack of readiness for the use of modern information technology.

Consequently, the Ministry of Natural Resources and Environment Protection, now the Ministry of Mining and Energy of the Republic of Serbia, launched in 2004 the project of the Geological Information System of Serbia, which has been developed in several phases over the past decade. The aim of developing such an information system was to establish an object-oriented database for digital archiving of geologic data, and provide a modern and effective information basis for carrying out all activities related to planning, design and decision-making in the field of geology.

Within the project a web portal[3] was established that allows quick and easy access to geological data and information in the field of general geology, exploration of mineral deposits, hydrogeology and engineering geology. Users of published information, professionals or ordinary citizens, can use this geo-portal for search and access to information they need.

The content on the portal can be grouped into several categories: cartographic content, multimedia, dictionaries and textual databases. The "core" is the whole information system of the Geological Dictionary (Thesaurus) containing about 4,000 geological terms described by definitions, of which about 3,000 have a translation into English. The most important cartographic content is: the basic geological map, maps of national parks, map of endangered groundwater bodies, geomorphological map, map of exploration-mining fields, and within multimedia content the most prominent are the gallery of photos and movies, geoheritage, and jeweler mineral resources. The web portal supports access to applications for

[3] http://geoliss.mre.gov.rs; search of fund documentation http://geoliss.mre.gov.rs/index.php?page=fodib

search of textual database (catalogs): projects, archival documents and bibliographies, library of geological projects documentation and exploration-exploitation approvals for water and solid mineral resources.

One of the textual databases is the database of geological projects documentation that contains metadata, namely structured descriptions of over 4,900 geological projects financed by the Republic of Serbia from 1956 to the present day, with: title, year, location, name of the company that developed the project, the authors, abstract, keywords, prospects, application of mineral resource and possibilities for its use, field works, geomechanics, mining works, geodesic works, and prospective exploration.

3 Present Solution

Search of textual databases (which is in use for several years) is derived from the basic model, which is based on the input of keywords, single or multi-word units (MWU), that can be combined into Boolean expressions. In addition to general search which goes through all fields in the relevant tables, the search can also be performed using specific criteria. For example, if the user chooses the criterion *mineral raw material*, then only the following fields in the database are taken into consideration: title, field, keywords, and abstract, while for *location* criterion other fields are searched: municipality, county, name of the cartographic sheet, location and chronological number of the document, and the sheet signature. The search system takes into account the field where the keyword was found as this information is used for document ranking.

The users express their information need by selecting the criteria and entering keywords, where they can add any number of criteria, of the same or different type. The criteria are linked by conjunction and when there are more keywords within one criterion a disjunction is generated. For example, if the user is interested in projects that deal with the research of "gold in the Bor region", then *mineral resource* is selected as one criterion and the words *zlato* 'gold' and *Au* are defined for search, and the system will search for any of these two words. If another criterion is added, for example *location*, then keywords which define the desired location are entered, e.g. *Bor* (the name of a city) or *Borski okrug* 'Bor county'. If we want a MWU to be treated as a whole, then it is entered under quotation marks.

The search is performed by scanning the text of appropriate fields with given keywords while word boundaries are not taken into consideration. This can partially solve the problem of the rich morphology that characterizes Serbian, as a language belonging to the South-Slavic Language family. For instance, scanning with *lignit* 'lignite' will also retrieve inflected forms *lignita*, *lignitu*, *lignitom*, etc.

Search results are ranked on the basis of weight factors assigned to individual fields in the function of search criteria. Given the information available in the database, it is possible to add criteria and fine-tune the ranking output based on the number of occurrences of a keyword or phrase and the sum of the weight factors of the search fields. Namely, each search criterion fits several different entities and their attributes within the database, and each entity/attribute includes weight factors that determine the relevance of the appearance of the resource within the result set. One example can illustrate this: *location* is one search criterium. When searching with this crterion the weights of some fields are: Municipality 8, County 7, Title 4, Keywords 3, Abstract 2, Appendices 1. Thus, if search is performed with the location criterion *Bor*, documents in which *Bor* occurs in Municipality field will be better ranked than those in which it occurs in the Abstract field. The ranking of results is performed in descending order according to the total sum of the product of weight factors and the number of occurrences of the corresponding keywords.

A specific feature of Serbian is the common use of two alphabets: Cyrillic and Latin, so a user can initiate her/his search using any of the two, and the method on the server automatically expands the search query with the other, while the results are shown in the original script, and that is Cyrillic.

Query processing on the server side expands the query by creating a matrix of key words, fields that are searched, and weight factors, and then translates this query into SQL (Structured Query Language) form. The query generated in such a way searches the text of the subset of attributes in the database that correspond to the selected criteria of search.

4 The Improved Solution

One of the problems of full text search in Serbian is its rich morphology, where the keyword for search is always entered in the first person singular, while in the texts that are searched it can occur in different inflectional forms.

For languages such as Serbian, some kind of normalization of morphological forms has to be performed both for document indexing and query processing. One soultion is to use stemmers. For Serbian, work on several stemmers was reported: a stemmer as a part of a larger system for information retrieval, PoS tagging, shallow parsing and topic tracking [9], a stemmer and lemmatizer based on suffix stripping [5], the same basic idea being used in the stemmer presented in a later paper [11]. The only stemmer available for analysis is the last one since its code is given in the paper. However, although the author claims accuracy of 92% it was evaluated on a very small text (522 words) so its reliability is not confirmed. Also, as Hiemstra states [3] “Stemming tends to help as many queries as it hurts.” The other possibility is statistical lemmatization for which we could have used the TreeTagger trained for Serbian that was used for the lemmatization of the Corpus of Contemporary Serbian [16]. However, this lemmatizer was trained on a corpus that differs significantly from our collection, and additionally it does not take into account MWUs.

The approach described in this paper bases lemmatization on morphological electronic dictionaries and finite state transducers for Serbian [6].

4.1 Used Resources

Lexical Resources. The resources for natural language processing of Serbian consisting of lexical resources and local grammars are being developed using the finite-state methodology as described in [1], [2]. The role of electronic dictionaries, covering both simple words and multi-word units, and dictionary finite-state transducers (FSTs) is text tagging. Each e-dictionary of forms consists of a list of entries supplied with their lemmas, morphosyntactic, semantic and other information. The forms are, as a rule, automatically generated from the dictionaries of lemmas containing the information that enable production of forms. For this purpose almost 1,000 inflectional transducers were developed. The system of Serbian e-dictionaries covers both general lexica and proper names and all inflected forms are generated from 135,000 simple forms and 13,000 MWU lemmas. Approximately 28.5% of these lemmas represent proper names: personal, geopolitical, organizational, etc.

Another lexical resource that is being developed for Serbian is Wordnet (SWN), whose development started in 2001 as part of the Balkanet project. In its initial phase, all languages featured in Balkanet followed the approach similar to EuroWordnet, namely, developing monolingual wordnets interconnected through an interlingual index (ILI) [17]. The SWN is continuously growing [12] and today has more than 22 thousand synsets linked to the Princeton WordNet 3.0 through ILI. During the development of the SWN, special attention was given to certain conceptual domains — emotions — and scientific domains — biological species, biomedicine, nutrition, religion, law, linguistics, literature, librarianship, etc. Geology is insufficiently covered — SWN presently contains only 157 synsets from this domain.[4]

Named Entity Recognition. According to [13] the term “Named Entity” (NE) usually refers to names of persons, locations and organizations, and numeric expressions including, time, date, money and percentage. Recently other major types are being included, like “products” and “events”, but also marginal ones, like “e-mail addresses” and “book titles”.

The NE hierarchy in our Named Entity Recognition (NER) system consists of five top-level types: persons, organizations, locations, amounts, and temporal expressions, each of them having one or more levels of sub-types. Our tagging strategy allows nesting, which means that a named entity can be nested within another named entity, e.g. a persons name within an organization name, like in *<org>Institut za vodoprivredu “<persName>Jaroslav Černi</persName>” </org>* ‘Institut for waterpower engineering Jaroslav Černi’.

The Serbian NER system is a handcrafted rule-based system that relies on comprehensive lexical resources for Serbian described in the previous subsection. For recognition of some types of named entities, e.g. personal names and locations, e-dictionaries and information within them is crucial; for others, like temporal expressions, local grammars in the form of FSTs that try to capture a

[4] http://resursi.mmiljana.com/Default.aspx

variety of syntactic forms in which a NE can occur had to be developed. However, for all of them local grammars were developed that use wider context to disambiguate ambiguous occurrences as much as possible [7]. These local grammars were organized in cascades that further resolve ambiguities [10]. NER system was evaluated on a newspaper corpus and results reported in [7] showed that F-measure of recognition was 0.96 for types and 0.92 fot tokens.

For the purpose of indexing, we applied our NER system to title and abstract fields of our geological structured data. The whole collection consist of 4,902 documents, 2,880,229 tokens (900,403 simple word forms). Almost all documents contained at least one NE — in only 61 (1.24%) not a single NE was recognized. On the average, 4 NEs of all types were recognized per document, with as many as 47 NEs for one of them. For indexing we used only three top level types: personal names, locations and organizations and their distribution is presented in Table 1.

Table 1. Distribution of three top-level NEs: persons, locations and organizations

NE type	Frequency	Average per doc	% of the text
person	11,991	2.45	1.33
location	49,414	10.08	5.49
organization	2,882	0.59	0.32
total	64,287	13.11	7.14

4.2 The Architecture of the New System

Indexing of metadata database on the projects is done so that for each document a text is generated of all the fields and records in the database, where the title, discipline and cartographic sheet are given extra weight. Two types of such “representative items” or indexes that are used for search are generated: a bag of words and named entities.

- The bag of words implies the representation of the document by a set of ungrammatical words — in our case nouns, adjectives, adverbs and acronyms — followed by their frequencies. Thus, the text is lemmatized and lemmas (simple and multi-word) are extracted and their frequency is calculated. In that way 12,204 simple lemmas (with 450,418 occurences) and 271 MWUs (with 6,525 occurences) were extracted.
- Recognized NEs that belong to 3 selected types — location, organization, persons — are entered in the index. Figure 1 represents one document from our collection in which recognized NEs are highlighted: locations in blue, persons in pink, and organizations in light green.

Determination of term weights is a complex process and there are numerous models, the most used being: *idf* based on the term frequency in the document, probabilistic, which includes in addition relevance weights, *tf_idf* which takes into account the number of documents in which the term appears, tfc_tfc which modifies the formula for ranking with cosine normalization, tfc_nfc which uses a normalized tf factor for terms of the query (because it has been shown that different mapping of the vector space of documents and queries is more efficient), *lnc_ltc* where the linear function is replaced by the logarithm, and finally the *lnu_ltu* which uses the document length and the average length of documents instead of cosine measure for normalizing length. [3]

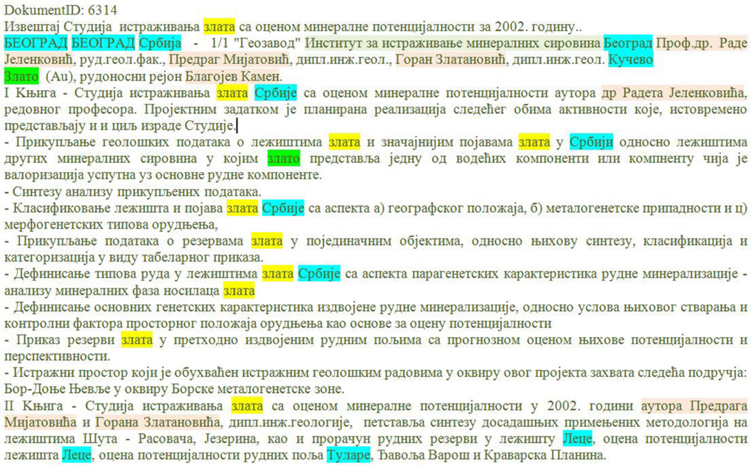
The improved ranking uses *tf_idf* measure that is based on frequencies of words allocated to the text, text length, and the document frequency [8]. Indexing is performed in following steps:

1. Generating a D_i text from several records and fields in the database related to a particular document or project;
2. Lemmatizing and Part-Of-Speech tagging of all texts D_i, where $i = 1, \dots N$ and N is the size of text collection;
3. Recognizing NEs and assigning the chosen types to documents;
4. Selecting ungrammatical words T_{ij} and calculating their frequencies n_{ij} for each D_i, $i = 1, \dots N$;
5. Calculating the relative frequency tf_{ij} for each term T_{ij} in a text D_i as n_{ij}/l_i where l_i is the length of the text in the number of simple words;
6. Calculating document frequency df_j as the number of documents in the collection in which the term T_j appears, and the acceptable indicator of term value as a document discriminator as $log(N/df_j)$;
7. Calculating the combined measure $tf\_idf = t_{ij} \cdot log(N/df_j)$.

In the search stage the similarity of the search query vector and the document are determined as follows:

1. the query is analyzed, tokenization is performed (separating into words, where a MWU within quotation marks is treated as one word);
2. for each document and for each word in the query the weight *td_idf* is found;
3. the similarity between the query and the document is ranked based on the sum of weights for all words in the query.

Document in Figure 1 is about a project that deals with gold. When searching with the keyword *zlato* 'gold' the old system ranks the document as 125th with general search and as 84th when searching in the category *mineral deposits* because the keyword matches only that particular form of the word (two matches highlighted in green in Figure 1). The new system ranks it as the first because the *td_idf* of this term for this document is calculated on the basis of its frequency $n_j = 12$ (ten additional matches are in Figure 1 highlighted in yellow). Due to its simple pattern matching the old system finds the positive match *zlatonosna žica* 'gold vein' when searching with *zlato*, which the new system misses (*zlato* and *zlatonosan* are two different lemmas); however, for the same reasons, the old system has the negative match *Zlatokop* (the name of one settlement), which the new system again misses.



DokumentID: 6314
Извештај Студија истраживања злата са оценом минералне потенцијалности за 2002. годину..
БЕОГРАД БЕОГРАД Србија - 1/1 "Геозавод" Институт за истраживање минералних сировина Београд Проф.др. Раде Јеленковић, руд.геол.фак., Предраг Мијатовић, дипл.инж.геол., Горан Златановић, дипл.инж.геол. Кучево
Злато (Au), рудоносни рејон Благојев Камен.

I Књига - Студија истраживања злата Србије са оценом минералне потенцијалности аутора др Радета Јеленковића, редовног професора. Пројектним задатком је планирана реализација следећег обима активности које, истовремено представљају и и циљ израде Студије.

- Прикупљање геолошких података о лежиштима злата и значајнијим појавама злата у Србији односно лежиштима других минералних сировина у којим злато представља једну од водећих компоненти или компненту чија је валоризација успутна уз основне рудне компоненте.
- Синтезу анализу прикупљених података.
- Класификовање лежишта и појава злата Србије са аспекта а) географског положаја, б) металогенетске припадности и ц) мерфогенетских типова оруднења,
- Прикупљање података о резервама злата у појединачним објектима, односно њихову синтезу, класификација и категоризација у виду табеларног приказа.
- Дефинисање типова руда у лежиштима злата Србије са аспекта парагенетских карактеристика рудне минерализације - анализу минералних фаза носилаца злата
- Дефинисање основних генетских карактеристика издвојене рудне минерализације, односно услова њиховог стварања и контролни фактора просторног положаја оруднења као основе за оцену потенцијалности
- Приказ резерви злата у претходно издвојеним рудним пољима са прогнозном оценом њихове потенцијалности и перспективности.
- Истражни простор који је обухваћен истражним геолошким радовима у оквиру овог пројекта захвата следећа подручја: Бор-Доње Њевље у оквиру Борске металогенетске зоне.

II Књига - Студија истраживања злата са оценом минералне потенцијалности у 2002. години аутора Предрага Мијатовића и Горана Златановића, дипл.инж.геологије, петставља синтезу досадашњих примењених методологија на лежиштима Шута - Расовача, Језерина, као и прорачун рудних резерви у лежишту Леце, оцена потенцијалности лежишта Леце, оцена потенцијалности рудних поља Туларе, Ђавоља Варош и Криварска Планина.

Fig. 1. One document dealing with the gold.

5 Evaluation

The goal of evaluation was to assess the efficiency of the old and new search method. The evaluation was performed over the entire collection of documents and a set of 10 information needs, represented by respective queries. For example, the first informational need "gold in Bor and the surrounding area" is converted into formal Boolean query *(zlato OR Au) AND (Bor OR Borski okrug)* '(gold OR Au) AND (Bor OR Bor district)'. For evaluation standard measures were used, namely Precision $P = tp/(tp + fp)$, Recall $R = tp/(tp + fn)$, and F-measure $F = 2 \cdot P \cdot R/(P + R)$, where tp – *true positive* is the number of relevant documents retrieved, fp – *false positive* is the number of non-relevant documents retrieved, and fn – *false negative* is the number of relevant documents that were not retrieved. During the evaluation ranked responses were offered to users, and the measures P and R were calculated for sets containing the first i choices offered, where $i \in [1, 120]$ [8]. In this way, curves showing the dependency between precision and recall for all 10 queries were obtained, as illustrated in Figure 2 for the abovementioned query. The charts show that the search precision of the old system is significantly better among first-ranked documents, while the recall is much better with the new system: among the first 80 documents the old system had 25 relevant responses, and the new had 39.

If the Interpolated Average Precision for 11 levels of recall $0.0, 0.1, 0.2, ..., 0.9, 1.0$ is calculated, a comparative graph is obtained of the relationship between precision and recall presented in Figure 3. The same procedure was applied for all 10 information needs (queries) and the results are presented in Table 2. In the columns *New AP* and *Old AP* the Average Precision $AP = \sum_{k=1}^{n} P(k)\Delta(k)$ is given for a particular query using the new and the old system, where $n = 120$ is the number of retrieved documents, $P(k)$ is the precision in the intersection

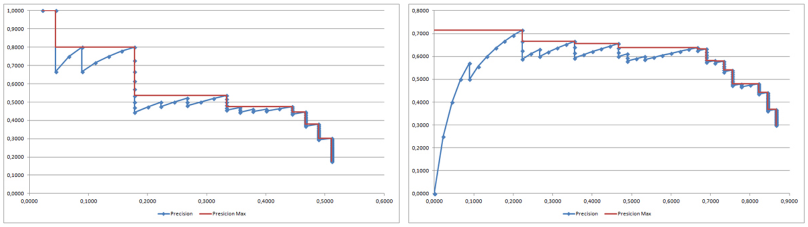


Fig. 2. Precision-recall curve for the first query *(zlato OR Au) AND (Bor OR Borski okrug)*: left, retrieval wihout index; right, retrieval with index.

point k, and $\Delta(k)$ is the change in recall from items $k-1$ to k. The results show that for 4 queries AP was higher in the old system, while for 6 queries AP was higher in the new one. Mean Average Precision (MAP) is more than 11% higher for new system in comparison to the old one. One of the reasons for small precision of some queries is the unbalanced length of documents. Namely, some are very short and the difference in length goes up to 1:20.

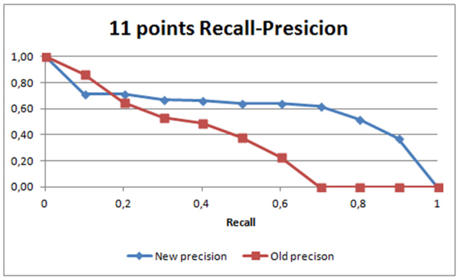


Fig. 3. 11-point Interpolated Average Precision for the first query for the old and new system.

The evaluation, as well as the analysis of the results showed that the old system achieves good results when searching with terms that occur as discipline and mineral resources classifiers, which are thus listed in the appropriate fields in the same form in which the users formulate their query (e.g. the nominative singular). However, a large number of other forms cannot be found by scanning the text, for example, the form *zlata* (genitive singular) cannot be aligned with the query keyword key *zlato* (nominative singular). The disadvantage of the system based on text scanning which affects the precision is especially visible when short words that could be parts of other words are used in a search. As explained before, the old system does not require alignment with whole words precisely in order to recognize at least some inflectional forms. This problem is generated by query keywords such as the chemical symbol for gold "Au" or the name of the company "NIS". In such cases, the accuracy of the new system is significantly greater (see queries 8 and 9 in Table 2).

Table 2. Average Precision per query and Mean Average Precision (MAP) for the old and the new system. A space in a query stands for an OR operator, a semicolon for an AND operator (relevant for the old system).

No.	Query	New AP	Old AP	Diff.
1	*zlato*; *Au*; *Bor*; *Borski okrug* gold; Au; Bor; Bor county	0.503	0.302	0.201
2	*opekarska keramička glina*; *Geozavod Geoinstitut* brick ceramic clay; Geozavod	0,465	0,334	0,131
3	*opekarska keramička*; *glina*; *Geozavod Geoinstitut* brick ceramic; clay; Geozavod	0.465	0.84	0.81
4	*poplava plavljenje izlivanje* flood flooding spills	0.638	0.446	0.192
5	*nestabilna padina* unstable slopes	0.505	0.521	-0.015
6	*izvorište zagadjenje*; *Obrenovac Lazarevac Lajkovac* wellspring pollution; Obrenovac Lazarevac Lajkovac	0.189	0.428	-0.239
7	*klizište* landslide	1.000	0.193	0.807
8	*geofizički karotaž*; *Naftagas* geophysical logging; Naftagas	0.235	0.596	-0.361
9	*geofizički karotaž*; *Naftagas NIS* geophysical logging; Naftagas NIS	0.118	0.113	0.004
10	*ugalj lignit*; *Kostolac Požarevac* coal lignite; Kostolac Požarevac	0.285	0.732	-0.447
	MAP	0.440	0.395	0.045
	Improvement			11.49%

The evaluation revealed that some NEs referring to the same entity occur in various forms which can deteriorate search results. For instance, the name of one person that documents refer to frequently occurs in 12 different forms:[5] *Dr Petar Petrović*, *Mr Petar Petrović*, *Prof. dr Petar Petrović*, *Prof. Dr Petar Petrović*, *dr Petar Petrović*, *Mr. Petar Petrović*, *PETROVIĆ PETAR*, *Dr Petar Petrović, docent*, *Mr Petar Petrović, asistent*, *docent Dr Petar Petrović*, *Mr. Petar Petrović, asistent*, *Petrović*. The biggest problem of the new system are specific technical terms that are not found in electronic dictionaries as well as quite a number of typographical errors in the document collection. However, this shortcoming can be rectified by correcting errors (based on the list of words unrecognized by the

[5] The name of the person was de-identified for privacy reasons.

vocabulary), and by continuous enhancement of the vocabulary by adding new words.

6 Conclusion and Future Work

The evaluation showed that our simple solution has its advantages: besides being simple to apply, it performs well for certain types of queries. However, although the new solution based on pre-indexing already outperforms it, its main advantage is that it can be improved and there are various means to do that:

- Enriching morphological e-dictionaries of simple words and MWUs by terms from geological domain;
- Addapting NERs to the new domain and text type (project rather than newspapers) and adding named entity normalization;
- Experimenting with different term weight measures;
- Experimenting with different comparison of document representation and information need representation.

Further research will be done by applying the new solution to other textual databases, as well as by applying a geodatabase for visualization of location of recognized named entities. An analysis of queries in the full sentence form is planned, which would eliminate stop words — prepositions, followed by lemmatization to produce a bag of words for the query. Finally, the integration of created indexes will enable the realization of a query expansion by adding synonyms from available resources, such as the geologic dictionary [15] for terminological query terms and WordNet for more general terms.

Acknowledgement. This research was supported by the Serbian Ministry of Education and Science under the grant #47003.

References

1. Courtois, B., Silberztein, M.: Dictionnaires électroniques du français. Larousse, Paris (1990)
2. Gross, M.: The use of finite automata in the lexical representation of natural language. In: Gross, M., Perrin, D. (eds.) Electronic Dictionaries and Automata in Computational Linguistics, Lecture Notes in Computer Science, vol. 377, pp. 34–50. Springer Berlin / Heidelberg (1989), http://dx.doi.org/10.1007/3-540-51465-1_3
3. Hiemstra, D.: Using language models for information retrieval. Taaluitgeverij Neslia Paniculata (2001)
4. Jackson, P., Moulinier, I.: Natural language processing for online applications: Text retrieval, extraction and categorization, vol. 5. John Benjamins Publishing (2007)
5. Kešelj, V., Šipka, D.: A Suffix Subsumption-Based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources. INFOtheca 9(1–2), 23a–33a (May 2008)

6. Krstev, C.: Processing of Serbian - Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade, Belgrade (2008)
7. Krstev, C., Obradović, I., Utvić, M., Vitas, D.: A System for Named Entity Recognition Based on Local Grammars. J Logic Computation 24(2), 473–489 (2014)
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press (2008)
9. Martinović, M.: Transfer of Natural Language Processing Technology: Experiments, Possibilities and Limitations Case Study: English to Serbian. INFOtheca 9(1–2), 11a–21a (May 2008)
10. Maurel, D., Friburger, N., Antoine, J.Y., Eshkol, I., Nouvel, D., et al.: Cascades de transducteurs autour de la reconnaissance des entités nommées. Traitement Automatique des Langues 52(1), 69–96 (2011)
11. Milosevic, N.: Stemmer for Serbian language. CoRR abs/1209.4471 (2012), http://arxiv.org/abs/1209.4471
12. Mladenović, M., Mitrović, J., Krstev, C.: Developing and Maintaining a WordNet: Procedures and Tools. In: Orav, H., Felbaum, C., Vossen, P. (eds.) Proceedings of the Seventh Global Wordnet Conference, GWC 2014. pp. 55–62. Tartu, Estonia (2014)
13. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. In: Sekine, S., Ranchhod, E. (eds.) Named Entities: Recognition, Classification and Use, pp. 3–28. John Benjamins Pub. Co., Amsterdam/Philadelphia (2009)
14. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1983)
15. Stanković, R., Trivić, B., Kitanović, O., Blagojević, B., Nikolić, V.: The Development of the GeolISSTerm Terminological Dictionary. INFOtheca 12(1), 49a–63a (August 2011)
16. Utvić, M.: Annotating the Corpus of contemporary Serbian. INFOtheca – Journal of Informatics & Librarianship 12(2), 36a–47a (2011)
17. Vossen, P.: EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Boston (1998)