Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović

Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources | Ranka Stanković, Cvetana Krstev, Ivan Obradović, Olivera Kitanović | Trans. Computational Collective Intelligence - Lecture Notes in Computer Science | 2017 | 26 |

10.1007/978-3-319-59268-8_8

http://dr.rgf.bg.ac.rs/s/repo/item/0001870

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Transactions on Computational Collective Intelligence, TCCI XXVI	
Series Title		
Chapter Title	Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources	
Copyright Year	2017	
Copyright HolderName	Springer International Publishing AG	
Corresponding Author	Family Name	**Stanković**
	Particle	
	Given Name	**Ranka**
	Prefix	
	Suffix	
	Division	Faculty of Mining and Geology
	Organization	University of Belgrade
	Address	Belgrade, Serbia
	Email	ranka@rgf.bg.ac.rs
Author	Family Name	**Krstev**
	Particle	
	Given Name	**Cvetana**
	Prefix	
	Suffix	
	Division	Faculty of Philology
	Organization	University of Belgrade
	Address	Belgrade, Serbia
	Email	cvetana@matf.bg.ac.rs
Author	Family Name	**Obradović**
	Particle	
	Given Name	**Ivan**
	Prefix	
	Suffix	
	Division	Faculty of Mining and Geology
	Organization	University of Belgrade
	Address	Belgrade, Serbia
	Email	ivan.obradovic@rgf.bg.ac.rs
Author	Family Name	**Kitanović**
	Particle	
	Given Name	**Olivera**
	Prefix	
	Suffix	
	Division	Faculty of Mining and Geology
	Organization	University of Belgrade
	Address	Belgrade, Serbia

Email olivera.kitanovic@rgf.bg.ac.rs

Abstract

Large collections of textual documents represent an example of big data that requires the solution of three basic problems: the representation of documents, the representation of information needs and the matching of the two representations. This paper outlines the introduction of document indexing as a possible solution to document representation. Documents within a large textual database developed for geological projects in the Republic of Serbia for many years were indexed using methods developed within digital humanities: bag-of-words and named entity recognition. Documents in this geological database are described by a summary report, and other data, such as title, domain, keywords, abstract, and geographical location. These metadata were used for generating a bag of words for each document with the aid of morphological dictionaries and transducers. Named entities within metadata were also recognized with the help of a rule-based system. Both the bag of words and the metadata were then used for pre-indexing each document. A combination of several tf_idf based measures was applied for selecting and ranking of retrieval results of indexed documents for a specific query and the results were compared with the initial retrieval system that was already in place. In general, a significant improvement has been achieved according to the standard information retrieval performance measures, where the InQuery method performed the best.

Improving Document Retrieval in Large Domain Specific Textual Databases Using Lexical Resources

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

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

Abstract. Large collections of textual documents represent an example of big data that requires the solution of three basic problems: the representation of documents, the representation of information needs and the matching of the two representations. This paper outlines the introduction of document indexing as a possible solution to document representation. Documents within a large textual database developed for geological projects in the Republic of Serbia for many years were indexed using methods developed within digital humanities: bag-of-words and named entity recognition. Documents in this geological database are described by a summary report, and other data, such as title, domain, keywords, abstract, and geographical location. These metadata were used for generating a bag of words for each document with the aid of morphological dictionaries and transducers. Named entities within metadata were also recognized with the help of a rule-based system. Both the bag of words and the metadata were then used for pre-indexing each document. A combination of several *tf_idf* based measures was applied for selecting and ranking of retrieval results of indexed documents for a specific query and the results were compared with the initial retrieval system that was already in place. In general, a significant improvement has been achieved according to the standard information retrieval performance measures, where the InQuery method performed the best.

1 Introduction

Advancements in database technology provide nowadays for management of large quantities of heterogeneous data. However, besides this engineering challenge—efficient management of such data, the "exploding world of Big Data" poses yet another, semantic challenge—finding and meaningfully combining information that is relevant to a user query [1]. Large textual databases, that is, large collections of textual documents are an example of big data, which pose three basic problems to Information Retrieval (IR): the representation of document content, the representation of user information needs and the comparison of these two representations.

N.T. Nguyen et al. (Eds.): TCCI XXVI, LNCS 10190, pp. 1–24, 2017.
DOI: 10.1007/978-3-319-59268-8_8

In general, if the response to a user query related to a collection of documents is performed by keyword search of these documents, then a preprocessing phase for additional representation of document content is not necessary. However, when these collections reach a certain volume, simple keyword search through the entire collection becomes time consuming and inefficient, as the recall grows but precision is likely to drop considerably. Thus in the case of large collections, an additional representation of the document content is generated, formally referred to as the document surrogate, with the aim of increasing document retrieval efficiency, through a better matching of user needs and retrieval results.

Document surrogates typically consist of metadata about the document, such as title, abstract, author and the like, as well as of keywords which denote document content. Surrogates can also contain an abstract and/or a snippet, a relevant text fragment. The content of a document surrogate, or its part, can be generated automatically by extracting and selecting specific terms (words) from the document text. Language processing methods and techniques developed within the field of digital humanities are used for completing this task. They provide for determining the boundaries of sentences within the document text, tokenization, stemming, tagging, recognition of nominal phrases and named entities and, finally, parsing [10].

Based on document representations by surrogates in large collections of textual documents, a preprocessing of documents usually takes place, in which an index of the collection of documents is formed, to be used for search and retrieval purposes. Relevant documents are retrieved and ranked upon a user query on basis of this index, using an approximate matching model, such as the vector space model, based on weight coefficients of terms, or the probabilistic model, based on relevance feedback [22].

When language processing methods and techniques are used for generating a document surrogate, they rely heavily on lexical resources, which is especially important in the case of languages with rich morphology, such as Serbian, and South-Slavic languages in general. Although Serbian belongs to the group of less-resourced languages, in which comprehensive lexical resources and language technology tools are still lacking or have not reached full maturity, it is safe to say that the current level of achievement is not negligible. According to the META-NET extensive survey performed in [20], some important lexical resources for Serbian were developed (corpora and e-dictionaries), as well as applications for basic language processing (tokenization, Part-Of-Speech (POS) tagging, morphological analysis), information retrieval and extraction [26].

Several successful applications of Serbian language resources and tools in tasks related to document indexing, retrieval and classification have been reported. A system for PhD dissertation metadata and full-text search was developed at the University of Novi Sad. It uses an index based on Lucene that integrates a Porters stemmer for Serbian [9]. Furlan and associates [5] also use Porter's stemmer and min-max normalization for logarithm of tf_{log} (the log-number of times the given word appears in a document) for calculating semantic similarity of short texts. Graovac [6] applies lexical resources for

Serbian—morphological dictionaries and the WordNet—for text categorization. Mladenović and associates [18] use the same resources for document-level sentiment polarity classification using maximum entropy modeling. Zečević and Vujičić-Stanković [27] apply various language-identification tools to distinguish Serbian among other closely related languages.

In this paper we describe an application of lexical resources and language tools for solving a big data problem, namely improvement of document retrieval from the database of geological projects financed by the Republic of Serbia. To that end documents in this large textual database are indexed using simultaneously two methods: by generating a bag-of-words and by named entity recognition for each document. In the next section we give an outline of the textual database of geological projects and the initial document retrieval system based on text scanning, along with its shortcomings. The improved system developed using lexical resources and language tools is described in Sect. 3, while the evaluation of this improvement is given in Sect. 4, followed by some concluding remarks.

2 The Database of Geological Projects

2.1 Motivation

Although the volume of geological data and related information on various geological phenomena in Serbia has been growing rapidly in the last decades, it was not accompanied by an adequate development and implementation of modern information technologies. Until recently, management of this growing body of geological documentation relied on traditional methods based on libraries and archives, which often made the task of obtaining specific information difficult or time consuming. Geological data have been collected for years, using different methods, and stored in various formats, seldom structured, most often in textual form. A very small part of these data was transformed in machine readable and structured format. The better part is still in paper form, and thus subject to decay or loss. An analysis of the way geological research results, stored in numerous archives and document libraries, were used, showed that this usage was inefficient, due to inadequate organization, limited access and general lack of readiness for introducing modern information technology. Thus a comprehensive digitalization is needed, which is expected to be intensified in the forthcoming years [23].

As a result of this analysis, the then Ministry of Natural Resources and Environment Protection (now Ministry of Mining and Energy of the Republic of Serbia), launched in 2004 the project of the Geological Information System of Serbia (*GeolISS*), which has been developed in several phases over the past decade. The aim of this information system was primarily to establish an object-oriented database for digital archiving of geologic data in the field of general geology, exploration of mineral deposits, hydrogeology and engineering geology, as a modern and efficient information basis for planning, design and decision-making in the geological domain.

Within the *GeolISS* project a web portal[1] was developed with the aim to provide quicker and easier access to geological data and information. Users, both professional and lay, can use this geo-portal to search and access information available within *GeolISS* database. The content on the portal is grouped into several categories: cartographic content, multimedia, dictionaries and textual databases. The "core" of *GeolISS* is the Geological Dictionary (Thesaurus) containing 5,152 geological terms described by definitions, of which 4,839 have a translation into English. The cartographic content includes a general geological map, maps of national parks, map of endangered groundwater bodies, geo-morphological map, map of exploration-mining fields, while the most prominent multimedia content are the gallery of photos and movies, geoheritage, BEWARE GIS web portal for interactive landslide data management, hazard and risk analysis, and jeweler mineral resources. Textual databases, also known as catalogs, consist of projects, archival documents and bibliographies, library of geological projects documentation and exploration-exploitation approvals for water and solid mineral resources. Within the geo-portal access to applications for document search and retrieval is available.

A textual database of special importance is the database of documentation related to over 5,500 geological projects financed by the Republic of Serbia from 1956 to the present day.[2] For each project a structured description is available in the form of a project summary containing the following metadata: title, year, project location, company that developed the project, authors, abstract, keywords, geological field, prospects, application of mineral resource and possibilities for its use, field works, geomechanics, mining works, geodesic works, and prospective exploration. Each project summary also contains approximately 30% of the content of the projects itself, obtained basically by removing pictures, maps, tables, and the like, offering thus a reasonably accurate representation of the textual content of the geological project. Future plans include digitalization and full text archiving of the project content, followed by the implementation of the approach described in this paper to this future full text database.

2.2 The Initial Solution for Document Retrieval

The initial solution for searching the textual databases in *GeolISS* (which is in use for several years) is based on user queries consisting of keywords, single or multi-word units (MWU), which can be combined into Boolean expressions. When a MWU is used as a keyword, it must be entered under quotation marks in order to be treated as a whole. A general search, which looks for keyword matches in all metadata fields in the summary report is available, but the user can also perform a faceted search using additional criteria. These criteria restrict

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

[2] Actually, almost 9,000 geological projects were financed in this period, but some of them were lost, some are not open to general public, and for some only basic data exists.

the search to specific metadata fields. For example, if the user chooses the facet *mineral resource* then only the following fields in the database are taken into consideration: title, field, keywords, and abstract. Likewise, the *location* facet searches for keyword matches in another group of fields: municipality, county, name of the cartographic sheet, location and chronological number of the document, and the sheet signature. When a match is found, the search system registers the corresponding metadata field, and this information is subsequently used for document ranking.

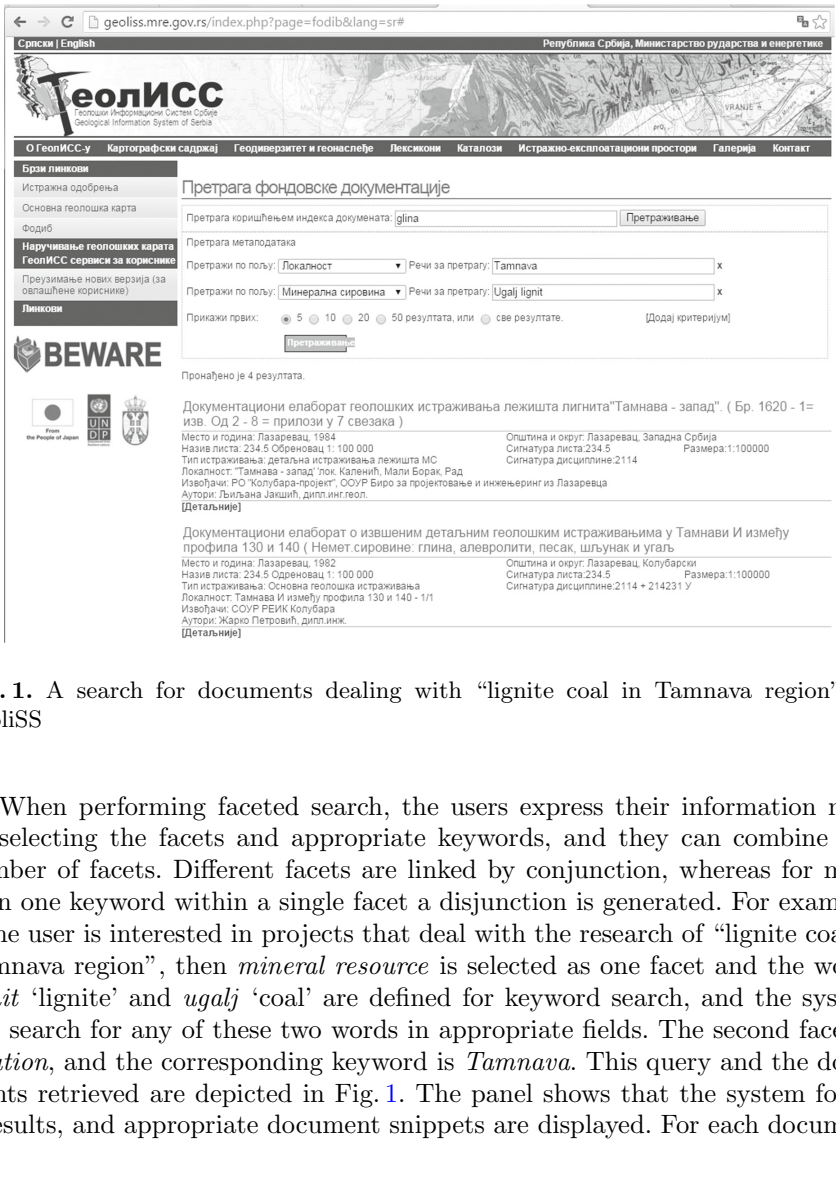

Fig. 1. A search for documents dealing with "lignite coal in Tamnava region" by GeoliSS

When performing faceted search, the users express their information need by selecting the facets and appropriate keywords, and they can combine any number of facets. Different facets are linked by conjunction, whereas for more than one keyword within a single facet a disjunction is generated. For example, if the user is interested in projects that deal with the research of "lignite coal in Tamnava region", then *mineral resource* is selected as one facet and the words *lignit* 'lignite' and *ugalj* 'coal' are defined for keyword search, and the system will search for any of these two words in appropriate fields. The second facet is *location*, and the corresponding keyword is *Tamnava*. This query and the documents retrieved are depicted in Fig. 1. The panel shows that the system found 4 results, and appropriate document snippets are displayed. For each document

there is also a link enabling the user to look into further details on the retrieved document. The user can chose the maximum number of retrieved documents to be displayed on one page, and if the recall exceeds this number, a pagination is generated and the user can review the recall page by page.

A specific feature of Serbian is the common use of two alphabets: Cyrillic and Latin, and consequently the system allows the user to initiate her/his search in either of them, automatically expanding the search query with the other. However, search results are displayed in the original alphabet used in the documents, and that is Cyrillic.

Keyword search in the initial system is performed by scanning the text of appropriate fields with given keywords in which word boundaries are not taken into consideration. This partially solves the problem of rich morphology that is characteristic for Serbian. For instance, scanning with *lignit* will also retrieve inflected forms *lignita*, *lignitu*, *lignitom*, etc. However, this solution is also a source of many false retrievals, especially when keywords are very short and can be found as substrings in many unrelated document words, like acronyms (e.g. *SO* derived from *skupština opštine* 'municipality council') and symbols of chemical elements (e.g. *Ca* for calcium).

For each field searched within a specific facet, a weight factor is assigned. For example, when the facet 'location' is used in a query, the weights of corresponding metadata fields are: Municipality 8, County 7, Title 4, Keywords 3, Abstract 2, Appendices 1, whereas for the facet "mineral resource", the fields searched and their weights are the following: Title 8, Signature 6, Keywords 4, Abstract 2, Appendices 1. System administrators have the possibility of adjusting these parameters, which are not hardcoded in the software solution, but rather registered in a database and accessible to users.

Table 1. A matrix of key words, fields that are searched, and weight factors for query "lignite coal in Tamnava region".

DocId	Abstract	Appendices	Keywords	Municipality	Title	Total
577			**4**	**8**	**12**	**24**
lignit					8	8
Tamnava				8	4	12
ugalj			4			4
578			**3**	**8**	**8**	**19**
Tamnava			3	8		11
ugalj					8	8
8823				**8**	**8**	**16**
Tamnava				8		8
ugalj				8		8
6255	**2**		**4**			**6**
Tamnava	**4**					2
ugalj			4			4

Query processing on the server side expands the faceted query by creating a matrix of key words, fields that are searched, and weight factors (see Table 1), 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 search facets. Ranking of retrieved documents is performed in descending order of the total sum of weight factors for all fields in which the keywords from the faceted search were found. The keyword matrix and the ranking of retrieved documents for the faceted query related to "lignite coal in Tamnava region" are depicted in Table 1.

The initial system achieved good results with faceted search in which keywords in metadata fields appeared in the same form in which they were used in the query (usually the nominative singular), but a large number of other forms could not be found. For example, if the keyword *Tamnava* (used in the nominative) appeared in the database in the locative form as in *u Tamnavi* 'in Tamnava' then the initial system would not be able to recognize it. As already mentioned, the initial system did not match whole words precisely in order to recognize at least some inflectional forms. However, we also pointed out that this became a serious disadvantage affecting precision when short words that could be parts of other words were used in a search. In order to improve the way the search mechanism copes with Serbian rich morphology an upgrade of the initial system was developed.

3 The Improved Solution

The initial solution has been used from 2008, while the development of the improved solution started at the beginning of 2015 and is available for use since June 2015. Besides tackling the problem of morphology this new solution also offers new features that enable the user to evaluate the retrieved documents.

As we have already pointed out, a Serbian keyword in a search query is almost always entered in the nominative singular, while in the texts that are searched it can occur in different inflectional forms. Thus, for languages such as Serbian, some kind of normalization of morphological forms has to be performed both for document indexing and query processing. One solution 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 [15], a stemmer and lemmatizer based on suffix stripping [11], the same basic idea being used in the stemmer presented in a later paper [17].

The only stemmer available for practical use is the last one since its code is available from the paper itself. 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 [8] "Stemming tends to help as many queries as it hurts." The other possibility is statistical lemmatization for which TreeTagger trained for Serbian is available, already used for lemmatization of the Corpus of Contemporary Serbian [25]. However, this lemmatizer was trained on a general corpus that differs significantly from domain corpora, such as our textual database of geological projects, and additionally it does not take into account MWUs.

The approach to lemmatization in developing an improved solution for searching the textual database of geological projects described in this paper is based on morphological electronic dictionaries and finite-state transducers for Serbian [12].

3.1 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 [3,7]. 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.

Named Entity Recognition. According to [19] 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 toponym within an organization name, like in *<org>Institut za gradjevinarstvo<top>Subotica</top></org>* 'Institute for civil engineering Subotica'.

The Serbian NER system is a handcrafted rule-based system that relies on comprehensive lexical resources for Serbian. 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 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 [13]. These local grammars were organized in cascades that further resolve ambiguities [16]. NER system was evaluated on a newspaper corpus and results reported in [13] showed that F-measure of recognition was 0.96 for types and 0.92 for tokens.[3]

[3] Tokens are all occurrences (in this case, NEs) in a given texts, types are different occurrences.

Table 2. Distribution of three top-level NEs: persons, locations and organizations.

NE type	Frequency	Average per doc	% of the text
Person	14,817	2.69	1.14
Location	65,724	11.93	5.05
Organization	3.492	0.63	0.27
Total	84.003	15.25	6.45

For the purpose of indexing, we applied our NER system to title and abstract fields of our geological structured data. The whole collection consisted of 5,510 documents (1,302,521 simple word forms). Almost all documents contained at least one NE—in only 71 (1.29%) not a single NE was recognized. On the average, 11 NEs of all types were recognized per document, with as many as 102 NEs for one of them. One of documents contained 25 different NEs. For indexing we used only three top level types: personal names, locations and organizations and their distribution is presented in Table 2. Nested NEs were also used for indexing, e.g. toponym *Subotica* in "Institute for civil engineering Subotica".

3.2 The Architecture of the New System

Indexing of documents on geological projects is done so that for each document a text is generated of all the fields and records in the project summary database, where the title and geological subdomain are given extra weight, 3 and 2, respectively. Two types of "representative items" or indexes that are used for search are generated: a bag of words and named entities, which are equally treated in document indexing. Figure 2 presents the architecture of the new system, where the left side depicts the preprocessing phase for document indexing, based on lexical resources and language technology tools, and the right side the query processing including calculation of similarity between information need represented by user query and document representations. 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 generated for each document is lemmatized and noun lemmas (simple and multi-word) are extracted and their frequency is calculated. In that way a total of 12,790 simple lemmas (with 647,303 occurrences) and 271 MWUs (with 9,219 occurrences) were extracted from all documents on geological projects. The bag of words generated for each particular document represents its first index.

The second index is generated from recognized NEs that belong to 3 selected types—location, organization and persons. Figure 3 represents one document from our collection in which recognized NEs are highlighted—toponyms are underlined, personal names (with roles) are underlined with a double line, organization names are framed. Determination of weights for terms within the indexes of a document is a complex process and there are numerous models, the most used being: tf based on the term frequency in the document index, idf which

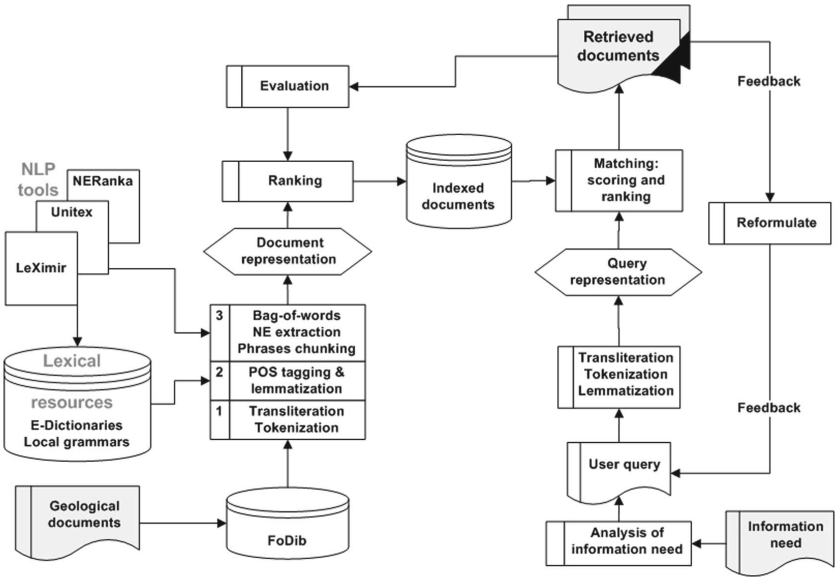

Fig. 2. The architecture of the improved system.

Прелиминарна техноекономска студија услова и могућности експлоатације лежишта угља Черевић Нови Сад, Фрушка Гора, Рударски институт Београд. Ову студију је требало урадити на основу података из Елабората о прорачуну резерви угља Ц2 категорије између Беочина и Баноштора. Институт за грађевинарство Суботица, Миливој Макар, дипл. инж. руд. Беочин. Истраживано подручје налази се на северним падинама Фрушке Горе, између Беочина на истоку и Баноштора на западу. Угљени слојеви се даље настављају према западу до Корушке. Студија о хидрогеолошким истраживањима у зони између Параћина и Главице у циљу отварања новог изворишта (I фаза). Београд Др Војислав Томић, доцент, Невен Крешић, дипл.инж.

Fig. 3. A document showing NER results.

takes into account the number of documents in which the term appears, tf_idf which is the combination of these two, probabilistic, which includes in addition relevance weights, tfc_tfc which modifies the formula for ranking with cosine normalization, tfc_nfc which uses a normalized tf factor for terms of the query (as 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 [8].

The improved system ranking uses several measures, starting with tf_idf measure based on frequencies of words allocated to the text, text length, and the document frequency [14]. Further development included modification of tf_idf with cosine normalization (tfc_tfc), tfc_nfc term weighting algorithm

with normalized tf factor for the query term weights, *lnc_ltc* measure where l stands for weights with a logarithmic tf component, *lnu_ltu* where normalization is based on the number of unique words in text, as well as several measures used in InQuery and Okapi systems.[4] Authors in [4] report on term weighting experiments with a linear combination of retrieval clues, one of the form $\alpha + \beta \cdot tf + \gamma \cdot idf + \delta \cdot tf \cdot idf$. The best performance was achieved when $\alpha = 0.4$, $\beta = \delta = 0$ and $\gamma = 0.6$.

Indexing is performed in the following steps:

1. Generating a text (D_i) from records and fields of the project summary related to a particular project document, where $i = 1, \ldots N$ and N is the size of the document collection;
2. Lemmatizing and POS tagging of all D_i texts;
3. Recognizing NEs and assigning the chosen types to documents;
4. Selecting ungrammatical words t_{ij} for each D_i and calculating:
   (a) n_{ij} as frequencies of t_{ij},
   (b) average and maximum term frequency (avg_tf_i, max_tf_i),
   (c) number of unique words in document (no_uw_i) and document length (l_i),
   (d) 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,
   (e) normalized term frequency $ntf = \log(tf + 0.5)/\log(max\_tf + 1)$,
   (f) $H\_inquery = (max\_tf \leq 200 \ ? \ 1 \ : \ 200/max\_tf)$, penalty for long documents,
   (g) $K\_okapi = k_1 \cdot ((1 - k_2) + k_2 \cdot l_i/avg\_dl)$ where avg_dl is the average document length, $k_1 = 1.2$, $k_2 = 0.75$ length normalisation;
5. Creating a dictionary of the whole document collection from all words selected in Step 4. For each term T_k in the document collection, $k = 1, \ldots M$, where M is the size of the dictionary of document collection:
   (a) calculating document frequency df_k as the number of documents in the collection in which the term T_k appears,
   (b) calculating the acceptable indicator idf_k of term value as a document discriminator as $\log(N/df_k)$ (*lnc_ltc* algorithm uses for *ltc* expression $idf\,1_k = \log((N + 1)/df_k)$),
   (c) $nidf_k = idf_k/\log(N)$,
   (d) $idf\_okapi = \log((N - df_k + 0.5)/(df_k + 0.5))$;
6. Calculating the document vector combined measure:
   (a) $tf\_idf = tf \cdot idf$, and vector intensity int_tf_idf,
   (b) $ltf = 1 + \log(tf)$ and vector intensity int_ltf,
   (c) $tfc = tf\_idf/int\_tf\_idf$,
   (d) $lnc = ltf/int\_ltf$,

[4] InQuery, an indexing and retrieval "engine" is developed at the Center for Intelligent Information Retrieval (CIIR), College of Information and Computer Sciences, University of Massachusetts Amherst [2]. The Okapi system was originally developed at the Polytechnic of Central London in the early 1980's and later developed at City University London and Microsoft Research [21].

(e) $lnu = (ltf/(1 + \log(avg\_tf)))/((1 - s) + s \cdot (no\_uw))$, with the constant value $s = 0.25$,
(f) $w\_inquery = 0.4 + 0.6 \cdot (b \cdot H\_inquery + (1 - b) \cdot ntf) \cdot nidf$, $b = 0.5$,
(g) $w\_okapi = (k_1 + 1) \cdot tf/(K\_okapi + tf)$.

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) followed by calculating:
   (a) maximum term frequency max_tf and number of unique words no_uw from query,
   (b) $nfc = (0.5 + 0.5 \cdot tf/max\_tf) \cdot lema\_idf$,
   (c) $ltc = (1 + log(tf)) \cdot lema\_idf1$,
   (d) $ltu = ltc/((1 - s) + s \cdot no\_uw)$;
2. for each document and for each word in the query depending on the selected method the weight is calculated, where d stands for document and q for query:
   (a) 'tfc_nfc': $d\_tfc \cdot q\_nfc$,
   (b) 'lnc_ltc': $d\_lnc \cdot q\_ltc$,
   (c) 'lnu_ltu': $d\_lnu \cdot q\_ltu$,
   (d) '$inquery$': $d\_w\_inquery \cdot q\_tf$,
   (e) '$okapi$': $d\_w\_okapi \cdot q\_idf\_okapi$.
3. the similarity between the query and the document is ranked based on the sum of weights for all words in the query.

Документациони елаборат геолошких истраживања лежишта лигнита "Тамнава - запад". Обреновац 1:100 000 "Тамнава - запад" лок. Каленић, Мали Борак, Рад РО "Колубара-пројект", ООУР Биро за пројектовање и инжењеринг из Лазаревца... исклињавања слојева угља, јер су растојања између истражних радова била два пута мања од растојања дата за поједине категорије резерви угља. Током 1983-84. год. изведене су укупно 62 бушотине, просечне дубине 82,11 м. Извештај о резултатима истражних радова минералних сировина у Тамнавском басену - шира околина Уба за 1984. годину - Уб Бој Брдо и Гредина.

Fig. 4. Examples of sources of erroneous hits.

Document in Fig. 4 has been retrieved for the faceted query with keywords "Tamnava OR Ub" (two mining sites) in the *location* facet and "ugalj OR lignit" (coal or lignite) keywords in the *mineral resource* facet. The initial system would identify *lignita* (as a substring of the genitive form of *lignit*) and *Tamnava* as matches (underlined by a thicker single line) but would fail to identify *uglja* (genitive form of *ugalj* underlined by a double line), which is, however, identified by the improved system. As we have already mentioned, the initial system identification method was based on 'like' function, which means that it searched the entire content within a field, regardless of the position of the search string, which

solves some problems but introduces other. Thus, for example, the initial system falsely matched the keyword *Ub* in the words *Kolubari* and *dubini* (the substring *ub* is framed in these words). As a result, in the first 50 matches offered by the initial system only the first document pertains to *Ub*, six pertain to *Tamnava*, and all the rest are a result of "false matches" of the keyword *Ub*. Such errors do not occur in the improved system, but "false matches" cannot be completely avoided, as for example in the case of *so* 'salt', which is a mineral resource, but also the acronym for *skupština opštine* 'municipality council', which appears in a number of documents. However, the initial system finds a string *so* in as many as 1,842 documents since it is a frequent syllable in Serbian, while the faceted search retrieves only 169 of them, some of which refer to a municipality council, others to salt. Faceted search in general improves ranking within the initial system. Namely, if a "general search" is performed instead of a faceted search for the query "Tamnava or Ub and coal or lignite" the document "Coal of the Tamnava basin in the vicinity of Ub", which was ranked first in the faceted search, drops to the 21st place, with irrelevant documents being better ranked due to "false match" in a greater number of fields, given that the general search looks into all fields in the database, not only those related to a specific facet.

4 Evaluation

The improved system includes an evaluation module, which allows a logged user to evaluate the relevance of retrieved documents after the search has been completed.[5] Figure 5 depicts the evaluation panel: when the query and the search

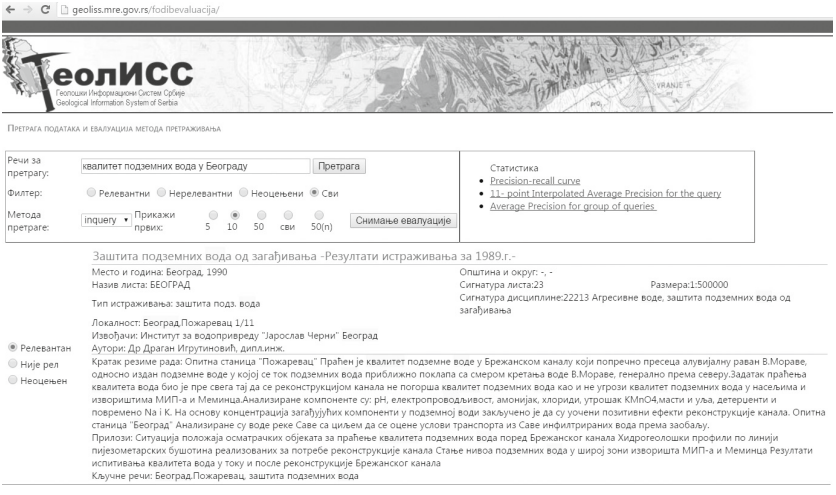

Fig. 5. The web panel for evaluation of retrieved documents by different methods.

[5] Available at http://geoliss.mre.gov.rs/fodibevaluacija/.

method are specified, results of a previous search can optionally be filtered. In order to alleviate this task, the system highlights the text in which keywords from the search were found.

From the same web page access is being enabled to pages with statistical data pertaining to evaluation, such as the precision-recall curve or 11-point Interpolated Average Precision for the query [22]. Comparison of several queries cross linked with different ranking methods can also be represented with the Average Precision graph. In this section we will present and describe these graphs.

The goal of evaluation was to assess and compare the efficiency of the initial and improved search methods. The evaluation was performed over the entire collection of documents and a set of 55 information needs, represented by respective queries. For query selection the log of the existing system was used, while also consulting geologist about their most common information needs. It turned out that most frequent requests are for a mineral resource type like copper, gold, coal, optionally at some location, or a geological event like landslide or earthquake. For evaluation standard measures were used, namely precision $P = tp/(tp + fp)$, recall $R = tp/(tp + fn)$, and $F1$-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 $i \in [1, 50]$ [14]. In this way, curves showing the dependency between precision and recall for all 55 queries were obtained, as illustrated in Fig. 6 for the query "geothermal energy in spas". It should be noted that results in this section are presented and

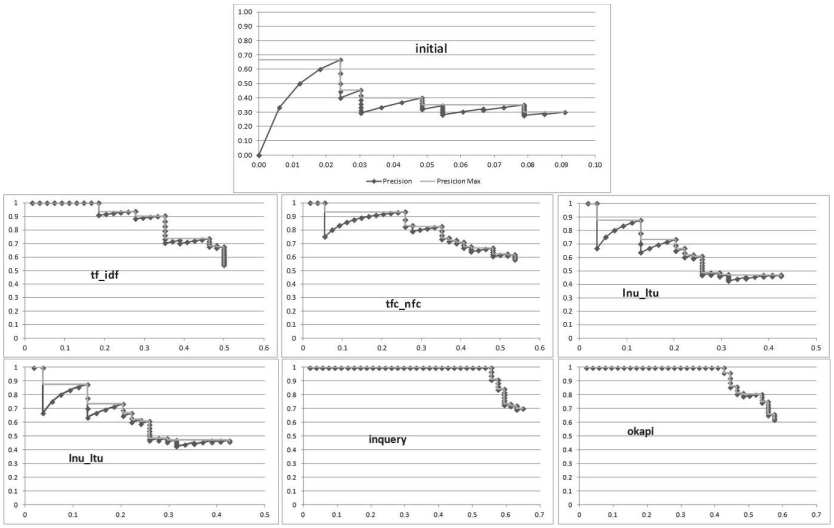

Fig. 6. Precision-recall curve for the query "geothermal energy in spas".

discussed for six out of seven available indexing methods. Namely the results of the tfc_tfc method were omitted as they were identical to the results of the tfc_nfc method. The precision of the initial system is in general better among first-ranked documents than in the case of the improved system, with the exception of the InQuery method, while the recall is better with the improved system.

Although the precision-recall curves offer a pretty good insight into system performance, it is often necessary to generate some sort of concise information, or even a single number. The usual approach is the 11-point Interpolated Average Precision, obtained by calculating the interpolated precision for each query over 11 recall values of $0.0, 0.1, 0.2, \ldots, 1.0$, and then calculating their average. When the Interpolated Average Precision for 11 levels of recall is calculated, a comparative graph is obtained of the relationship between precision and recall. The procedure was applied for all 55 information needs (queries). This set of queries has then been sorted in ascending order of the difference between the best performing ranking method for indexed documents InQuery and the performance of the initial method, and the results are depicted in Fig. 7.

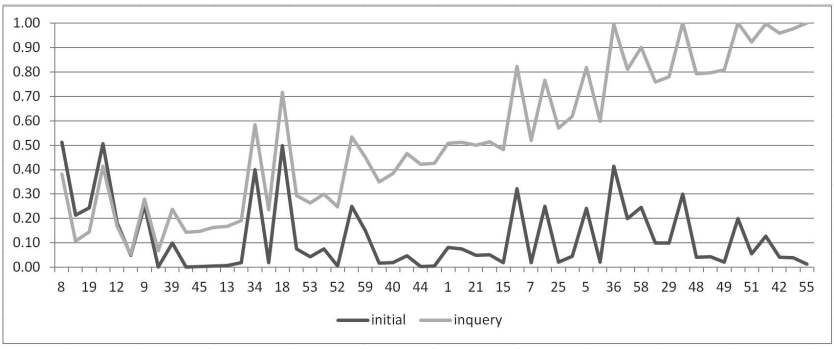

Fig. 7. Differences between initial method and InQuery indexing sorted in ascending order.

We shall now discuss two sets of five queries each, where differences between the initial and improved system are greatest, once in favor of the former and once in favor of the latter, as well as a set of 10 queries where the improvement is achieved, but on a moderate scale. The results will be presented in tables, showing the information need in the first column, the corresponding faceted query keywords in the second and the query surrogate, that is, linguistically preprocessed information need in the third. For example, within this preprocessing, the information need *kvalitet podzemnih voda u Beogradu* 'quality of underground waters in Belgrade' is transformed by the improved indexed based system into a set of lemmas, recognized by lexical analysis. Prepositions, conjunctions and the like are omitted, e.g. the preposition *u* 'in' in this query. The noun *kvalitet* 'quality', adjective *podzemni* 'underground' and toponym *Beograd* have unambiguous lemmas, whereas *voda* can correspond to nouns 'water' and 'platoon', but also

to a form of the verb *vodati* 'to lead someone'. The system also recognizes the MWU *podzemne vode* 'underground waters'.

Table 3. Five queries for which the initial system performed better than the improved.

8	**information need**: *geofizički karotaž Naftagas* — **faceted query**: *geofizički karotaž*; *Naftagas* — **indexed query**: *geofizički*; *karotaž*; *NAFTAGAS* — 'geophysical logging Naftagas'
22	**information need**: *dolomit "Institut za puteve" "Institut za istraživanja i ispitivanja"* — **faceted query**: *Dolomit*; *Institut za puteve*; *Institut za istraživanja i ispitivanja* — **indexed query**: *dolomit*;*istraživanje* — 'dolomite "The Highway Institut"'
19	**information need**: *Dunav Sava podzemna* — **faceted query**: *Dunav Sava*; *podzemna* — indexed query: *Dunav*;*Sava*;*Savo*; *podzemni* — 'Danube Sava underground'
14	**information need**: *rezerve uglja u Sjeničkom basenu* — **faceted query**: *rezerve uglja*; *Sjenički basen* — **indexed query**: *rezerva*; *ugalj*; *uglja*; *sjenički*; *basen* — 'coal reserves in Sjenica basin'
12	**information need**: *izvorište zagadjenja Obrenovac Lazarevac Lajkovac* — **faceted query**: *izvorište zagadjenja*; *Obrenovac Lazarevac Lajkovac* — **indexed query**: *izvorište*; *zagadjenje*; *Obrenovac*; *Lazarevac*; *Lajkovac* — 'source of pollution Obrenovac Lazarevac Lajkovac'

Table 3 shows examples of queries in which the initial system performed better than the improved. This is mostly due to the fact that the improved system often erroneously selects documents in which a toponym is specified, by failing to discern between toponyms in names of institutions that have produced the document, e.g. *Rudarski institut Beograd* 'Mining Institute in Belgrade' and toponyms pertaining to the location of the geological site to which the document refers. In the initial version of the improved system a problem was also encountered related to the lack of domain specific geologic terms, such as *karotaž* 'logging'. This problem is being solved by continuous enrichment of e-dictionaries.

In the columns of Table 4 the Average Precision $AP = \sum_{k=1}^{n} P(k)\Delta(k)$ is given for queries in Table 3 using the current system and all indexing methods except tfc_tfc, where $n = 50$ is the number of retrieved documents, $P(k)$ is the precision in the intersection point k, and $\Delta(k)$ is the change in the recall from item $k-1$ to item k. The comparison of all applied methods for 5 queries for which the initial method achieved better results is given in Fig. 8 (Table 4).

Table 5 depicts a set of queries in which a moderate improvement has been achieved in comparison to the initial system, ranging from 0.36 to 0.46. It should be noted that for the query *ležište bakra* 'copper deposit' the improved system assigns two nouns, *bakar* 'copper' and *bakra* 'copper cauldron', but the results are nevertheless better in comparison to the initial system, which does not recognize *bakra* as the genitive of *bakar* at all (Table 6).

Table 4. Comparison of indexing methods for queries in Table 3.

Id	*Difference inquery initial*	*initial*	*tf_idf*	*tfc_nfc*	*lnc_ltc*	*lnu_ltu*	*inquery*	*okapi*
8	−0.13	0.51	0.232	0.144	0.158	0.128	0.381	0.315
22	−0.11	0.21	0.124	0.19	0.108	0.131	0.108	0.137
19	−0.10	0.24	0.168	0.176	0.268	0.235	0.146	0.168
14	−0.09	0.51	0.087	0.233	0.168	0.05	0.414	0.249
12	−0.01	0.18	0.197	0.02	0.049	0.015	0.168	0.169

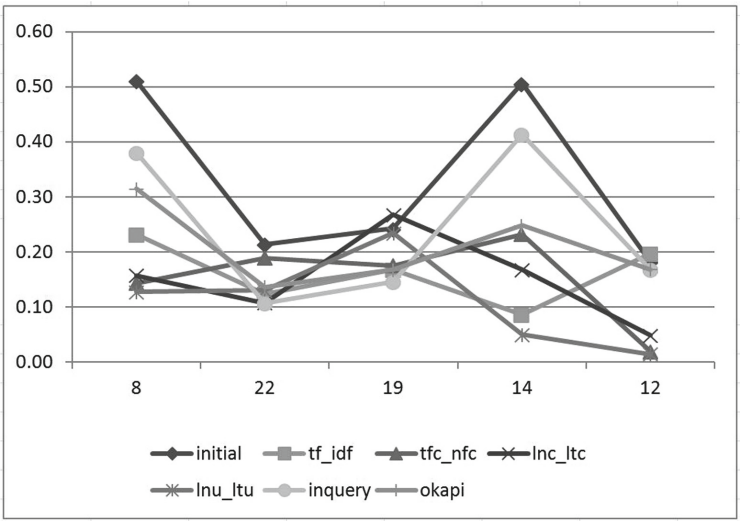

Fig. 8. Comparison of all indexing methods for queries in Table 3.

Finally, Table 7 outlines a set of queries for which a considerable improvement has been made in comparison to the initial system ranging from 0.87 to 0.99 (Fig. 9). In the query *ispitivanje tla u Nišu* 'soil analysis in Niš' linguistic preprocessing eliminates the preposition *u* 'in' unambiguously recognizes the nouns *ispitivanje* 'analysis' and *tlo* 'soil', whereas *Nišu* is recognized as the genitive of the toponym *Niš*, but also the accusative of the noun 'niche' and the third person plural of the verb 'to swing'. Nevertheless, the obtained results are considerably better than the results of the initial method.

Figure 11 depicts the MAP for all 55 queries over the six ranking methods for indexed documents and the initial method. The worst performing ranking method *lnu_ltu* has an almost three times better performance than the initial method, whereas the results for the best performing method InQuery are over four times better than for the initial method.

Table 5. Five queries for which the improved system achieved a moderate improvement.

40 **information need**: *nafta bušotina Braničevo* — **faceted query**: *nafta*; *bušotina*; *Braničevo* — **indexed query**: *nafta*; *bušotina*; *Braničevo* — 'oil drill-hole Braničevo'
27 **information need**: *gravimetrija geofizika* — **faceted query**: *Gravimetrija*; *geofizika* — **indexed query**: *Gravimetrija*; *geofizika* — 'gravimetry geophysics'
44 **information need**: *istraživanja u kolubarskom basenu* — **faceted query**: *istraživanje*; *"kolubarski basen"* — **indexed query**: *istraživanje*; *kolubarski;basen* — 'exploration in Kolubara basin'
35 **information need**: *rudno telo* — **faceted query**: *"rudno telo"* — **indexed query**: *rudni*; *Rudno*; *telo* — 'ore body'
1 **information need**: *zlato Au Bor Borski okrug* — **faceted query**: *zlato*; *Bor*; *"Borski okrug"* — **indexed query**: *zlato*; *Bor*; *Borski okrug* — 'gold Au Bor Bor region'
46 **information need**: *ležiste* — **faceted query**: *"ležiste bakra"* — **indexed query**: *ležište*; *bakar*; *bakra* — 'deposit'
21 **information need**: *klizište Umka Geozavod* — **faceted query**: *"klizište Umka"*; *GEOZAVOD* — **indexed query**: *klizište*; *Umka*; *GEOZAVOD* — 'landslide Umka Geozavod'
37 **information need**: *arteski bunar* — **faceted query**: *"arteski bunar"* — **indexed query**: *arteski*; *bunar* — 'Artesian well'
15 **information need**: *kvarcni pesak bušotine* — **faceted query**: *"Kvarcni pesak"*; *bušotina* — **indexed query**: *kvarcni*; *pesak*; *bušotina* — 'quartz sand drill-hole'
4 **information need**: *poplava plavljenje izlivanje* — **faceted query**: *poplava*; *plavljenje*; *izlivanje* — **indexed query**: *poplava*; *plavljenje*; *izlivanje* — 'flood flooding outpouring'

Table 6. Comparison of indexing methods for queries in Table 5.

Id	*Difference inquery initial*	*initial*	*tf_idf*	*tfc_nfc*	*lnc_ltc*	*lnu_ltu*	*inquery*	*okapi*
40	0.36	0.02	0.352	0.255	0.394	0.337	0.383	0.263
27	0.42	0.05	0.351	0.407	0.459	0.409	0.466	0.468
44	0.42	0.00	0.202	0.406	0.417	0.364	0.423	0.458
35	0.42	0.00	0.329	0.346	0.462	0.310	0.426	0.479
1	0.43	0.08	0.133	0.105	0.306	0.164	0.508	0.465
46	0.44	0.08	0.307	0.240	0.400	0.269	0.513	0.344
21	0.45	0.05	0.020	0.055	0.082	0.098	0.500	0.536
37	0.46	0.05	0.171	0.449	0.373	0.090	0.514	0.511
15	0.46	0.02	0.371	0.426	0.385	0.321	0.483	0.38
4	0.50	0.32	0.802	0.834	0.845	0.843	0.822	0.856

The evaluation, as well as the analysis of the results confirmed that the initial system achieves good results when searching with terms that occur as discipline

Table 7. Five queries for which the improved system achieved a considerable improvement.

51 **information need**: *ispitivanje tla u Nišu* — **faceted query**: *"ispitivanje tla";Niš* — **indexed query**: *ispitivanje*; *tle*; *tlo*; *nihati*; *Niš*; *niša* — 'soil analysis in Niš'
47 **information need**: *konturno bušenje* — **faceted query**: *"konturno bušenje"* — **indexed query**: *konturni*; *bušenje* — 'contour drilling'
42 **information need**: *životna sredina zaštita* — **faceted query**: *"životna sredina"; zaštita* — **indexed query**: *životan*; *sredina*; *zaštita* — 'environment protection'
33 **information need**: *klizište Barajevo* — **faceted query**: *"klizište Barajevo"* — **indexed query**: *klizište*; *Barajevo* — 'landslide Barajevo'
55 **information need**: *gama zračenje* — **faceted query**: *"gama zračenje"* — **indexed query**: *gama*; *zračenje* — 'gamma radiation'

Table 8. Comparison of indexing methods for queries in Table 7.

Id	*Difference inquery initial*	*initial*	*tf_idf*	*tfc_nfc*	*lnc_ltc*	*lnu_ltu*	*inquery*	*okapi*
51	0.87	0.05	0.678	0.642	0.704	0.301	0.922	0.766
47	0.87	0.13	1.000	1.000	1.000	0.994	0.996	0.998
42	0.92	0.04	0.594	0.584	0.741	0.408	0.959	0.948
33	0.94	0.04	0.588	0.734	0.793	0.464	0.976	0.892
55	0.99	0.01	0.860	0.986	0.986	0.952	1.000	1.000

and mineral resources facets, which are thus listed in the appropriate fields in the same form in which the users formulate their query (e.g. the nominative singular). It also confirmed the disadvantage of the system based on text scanning which affects the precision when short words that could be parts of other words are used in a search, such as the acronym of company "NIS". The average precision for query "geophysical logging Naftagas NIS" (obtained by adding "NIS" the acronym of the Naftagas company to query "geophysical logging Naftagas"), drops from 0.512 to 0.255 for the initial solution, while results for methods based on indexing are improved. The opposite happens for queries that contain terms that are missing from e-dictionaries, e.g. *karotaž* 'logging' or terms for which inflected forms are missing, e.g. the plural form *ugljevi* of *ugalj* 'coal' that is specific to the mining terminology. The graph presented in Fig. 12 gives average precisions for one query ("geothermal energy in spas"). The comparison of current system and new indexing methods shows that the InQuery method is superior for majority of queries over current system and other indexing methods.

The comparison of several methods for queries with greatest difference between highest and lowest value (in favor of improved methods) is given in Fig. 10. It is obvious that for the selected queries the performance of the initial method is significantly worse, whereas among the methods of the improved system InQuery is the most successful one. InQuery also gave best results measured by "Precision at k". This measure, giving precision at specific levels of retrieved

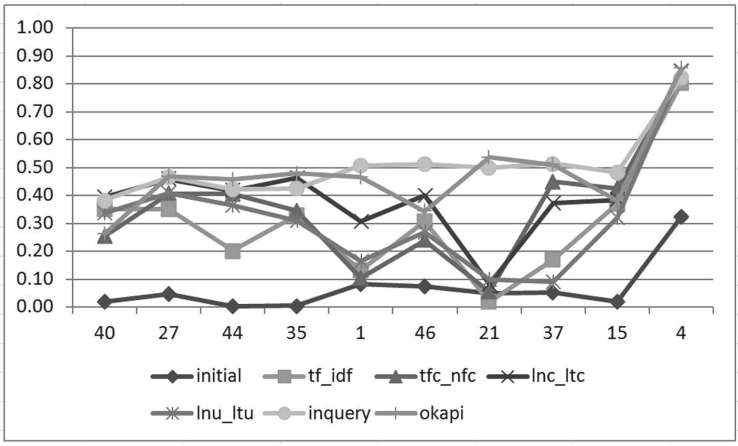

Fig. 9. Comparison of all indexing methods for queries in Table 5.

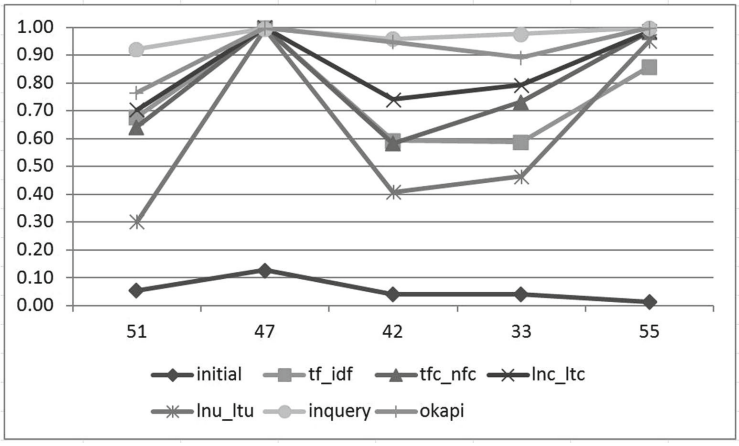

Fig. 10. Comparison of all indexing methods for queries in Table 7.

results (e.g. first 5, 10, 20, etc.) is also often taken into account, as users often look only at the certain number of best ranked results. It is especially the case when web search is performed, where a certain number of retrieved documents is given on each page, and users take only the first page into consideration. The advantage of this measure is that it does not require an estimate of the size of the set of relevant documents, but its disadvantage is that it is the least stable of the commonly used evaluation measures and that it does not average well, since the total number of relevant documents for a query has a strong influence on "Precision at k". [14] Fig. 13 depicts the average precision for the first 5, 10, 20, 30, 40, and 50 retrieved documents for all 55 queries, or information needs, where InQuery performs best, followed by Okapi.

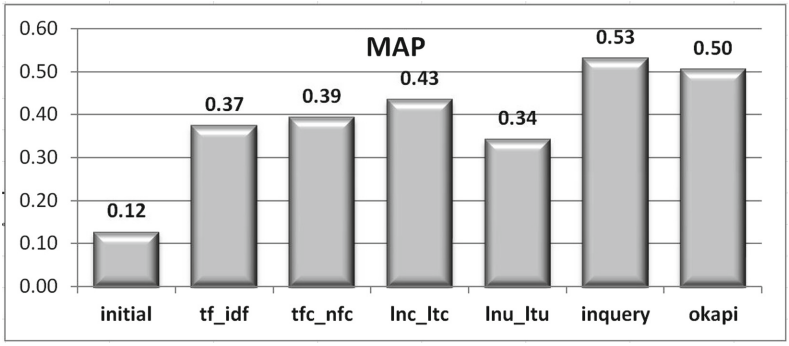

Fig. 11. Mean Average Precision (MAP) of all 55 queries for the initial and improved system with different ranking methods.

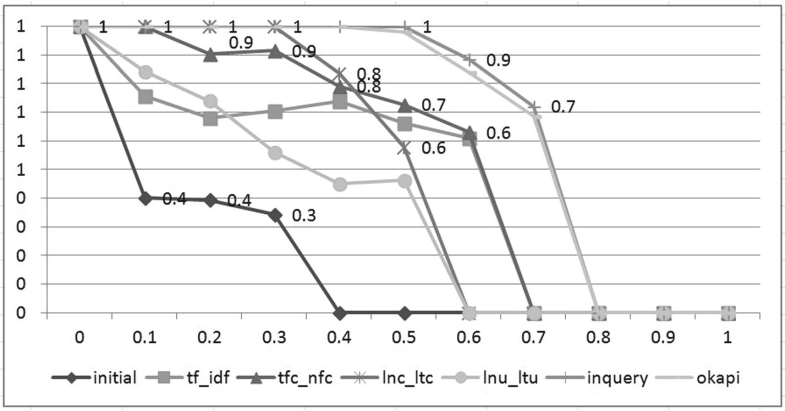

Fig. 12. Average precision for the query "geothermal energy in spas".

Evaluation results presented here are publicly available.[6] According to one of the evaluators the initial system only "looks nice" but its evaluation is cumbersome, and the improved system is by all means more precise. Another evaluator remarked that although the initial system performs well for some simple queries, it often gives completely erroneous results for complex ones. If assessed, the initial system would have a mark of 2 and the improved a 6, in comparison to Google Search with a mark of 10.

[6] For the initial system at http://geoliss.mre.gov.rs/fodibevaluacija/statistika.php, the improved system at http://geoliss.mre.gov.rs/fodibevaluacija/statistika-index.php for individual queries, and for the entire sets of queries at http://geoliss.mre.gov.rs/fodibevaluacija/statistika-all-methods.php.

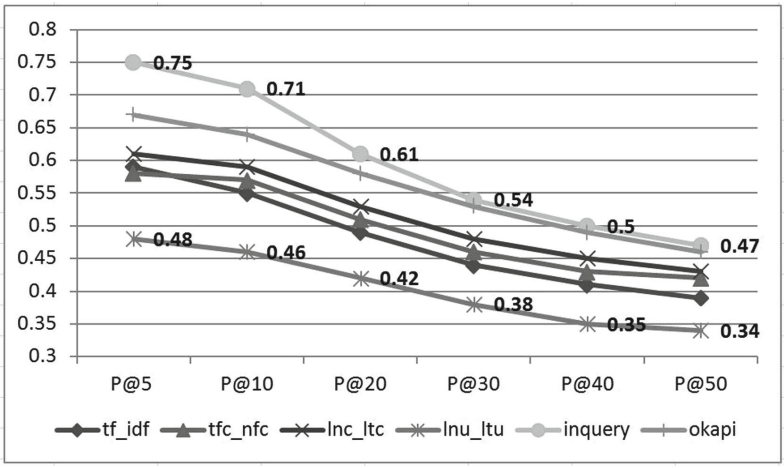

Fig. 13. The average precision for levels of recall ranging from 5 to 50 for all 55 queries.

5 Conclusion and Future Work

The evaluation results show that an improvement has indeed been achieved by introducing document indexing in the system. However, the initial solution, besides being simple to apply, outperformed the improved system for certain types of queries. Hence, future research should look into possibilities of combining these two methods, namely faceted search with linguistic preprocessing of queries and document indexing.

There are also several other courses of action towards further improvement. In order to minimize the number of unrecognized words within a query, morphological e-dictionaries of simple words and MWUs must be continuously updated with new geological terms. As for named entities, they will be adequately introduced into faceted search, along with the implementation of their normalization. The name entity recognition system will be improved to perform better in the geological domain, as it was initially developed for recognizing named entities in journal articles. Various combinations of weight measures for terms will be explored in order to reach an optimal combination. Finally, presentations of the information need and the document will be revisited, looking for the most efficient ways they can be matched.

Other textual as well as spatial databases and data collections will be used in the future for assessing the flexibility of the proposed solution, along with the introduction of geodatabases for visualization recognized toponym locations. Query expansion is also under consideration, where query terms would be supplemented with related terms, such as hyponyms or hypernyms, with the help of available resources, such as the geologic dictionary [24] for terms from domain terminology and WordNet for more general terms.

One of the shortcoming of the improved system is the absence of the possibility for fine-tuning the search, based on an analysis of results with or without specific keywords. Improvement of the software solution for filtering and extraction of search result is also needed. Another possible improvement would be the use of relevant queries for re-ranking the retrieved documents, where the weights of relevant documents would be included in the metrics. This is a variation of the Okapi method, which is pretty demanding for the average user, but makes sense for some frequently repeated queries, or queries that are repeated at certain time intervals.

Acknowledgement. This research was supported by the Serbian Ministry of Education and Science under the grant #47003 and KEYSTONE COST Action IC1302. The authors would like to thank the anonymous reviewers for their helpful and constructive comments. We are also grateful for the time and effort our young colleagues from the Faculty of Mining and Geology Biljana Lazić, Dalibor Vorkapić and Nikola Vulović invested in evaluating the document retrieval results.

References

1. Bizer, C., Boncz, P., Brodie, M.L., Erling, O.: The meaningful use of big data: four perspectives-four challenges. ACM SIGMOD Rec. **40**(4), 56–60 (2012)
2. Callan, J., Croft, W.B., Harding, S.: The inquery retrieval system, pp. 78–83 (1992)
3. Courtois, B., Silberztein, M.: Dictionnaires électroniques du français. Larousse, Paris (1990)
4. Croft, W.B., Smith, L.A., Turtle, H.R.: A loosely-coupled integration of a text retrieval system and an object-oriented database system. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 223–232. ACM (1992)
5. Furlan, B., Batanović, V., Nikolić, B.: Semantic similarity of short texts in languages with a deficient natural language processing support. Decis. Support Syst. **55**(3), 710–719 (2013)
6. Graovac, J.: Wordnet-based serbian text categorization. INFOtheca **14**(2), 2a–17a (2013)
7. Gross, M.: The use of finite automata in the lexical representation of natural language. In: Gross, M., Perrin, D. (eds.) LITP 1987. LNCS, vol. 377, pp. 34–50. Springer, Heidelberg (1989). doi:10.1007/3-540-51465-1_3
8. Hiemstra, D.: Using language models for information retrieval. Taaluitgeverij Neslia Paniculata (2001)
9. Ivanović, D., Milosavljević, G., Milosavljević, B., Surla, D.: A CERIF-compatible research management system based on the MARC 21 format. Inf. Knowl. Manag. **44**(3), 229–251 (2010)
10. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction and categorization, vol. 5. John Benjamins Publishing, Amsterdam (2007)
11. 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 (2008)
12. Krstev, C.: Processing of Serbian - Automata. University of Belgrade, Belgrade, Texts and Electronic Dictionaries. Faculty of Philology (2008)

13. Krstev, C., Obradović, I., Utvić, M., Vitas, D.: A system for named entity recognition based on local grammars. J. Logic Comput. **24**(2), 473–489 (2014)
14. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
15. Martinović, M.: Transfer of natural language processing technology: experiments, possibilities and limitations case study: English to Serbian. INFOtheca **9**(1–2), 11a–21a (2008)
16. 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)
17. Milosevic, N.: Stemmer for Serbian language. CoRR abs/1209.4471 (2012). http://arxiv.org/abs/1209.4471
18. Mladenović, M., Mitrović, J., Krstev, C., Vitas, D.: Hybrid sentiment analysis framework for a morphologically rich language. J. Intell. Inf. Syst. 1–22, to appear
19. 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 Publishing Company, Amsterdam (2009)
20. Rehm, G., Uszkoreit, H. (eds.): META-NET White Paper Series. Springer, Heidelberg (2012). http://www.meta-net.eu/whitepapers
21. Robertson, S.E., Walker, S.: Okapi/Keenbow at TREC-8. In: TREC, vol. 8, pp. 151–162 (1999)
22. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1983)
23. Stanković, R., Prodanović, J., Kitanović, O., Nikolić, V.E.: Development of the Serbian geological resources portal. In: Proceedings of 17th Meeting of the Association of European Geological Societies, pp. 61–65 (2011)
24. Stanković, R., Trivić, B., Kitanović, O., Blagojević, B., Nikolić, V.: The development of the geolissterm terminological dictionary. INFOtheca **12**(1), 49a–63a (2011)
25. Utvić, M.: Annotating the corpus of contemporary Serbian. INFOtheca - J. Inform. Librariansh. **12**(2), 36a–47a (2011)
26. Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., Stanojević, M.: Srpski jezik u digitalnom dobu - The Serbian Language in the Digital Age. In: Rehm and Uszkoreit [20] (2012). http://www.meta-net.eu/whitepapers
27. Zečević, A., Stanković-Vujičić, S.: Language identification–the case of Serbian. In: Pavlović-Lažetić, G., Krstev, C., Vitas, D., Obradović, I. (eds.) Natural Language Processing for Serbian – Resources and Applications, pp. 101–112. Faculty of Mathematics, University of Belgrade. http://jerteh.rs/wp-content/uploads/2015/05/Zecevic.pdf

Author Queries

Chapter 8

Query Refs.	Details Required	Author's response
AQ1	Please confirm if the corresponding author is correctly identified. Amend if necessary.	
AQ2	Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city and country names in affiliations "1 and 2". Please check and confirm if the inserted city and country names are correct. If not, please provide us with the correct city and country names.	
AQ3	Please check and confirm if the inserted citations of "Tables 6 and 8" and "Fig. 9" are correct. If not, please suggest an alternate citations.	