

NABU: A Macedonian Web Search Portal

Igor Janevski
ITEA Solutions, Ltd.
Skopje, Macedonia
i.janevski@itearesearch.com

Kristijan Takasmanov
Faculty of Management and Information Technologies,
Skopje, Macedonia
{kristijan.takasmanov,jovan.pehcevski}@mit.edu.mk

Jovan Pehcevski

Abstract

The advent of Google and other modern search engines makes the wealth of information that could be discovered on the web universally accessible. The information retrieval techniques used by these search engines are mostly effective when applied on English web collections; however, there are many challenges that need to be addressed when using these engines on non-English web collections. In this paper, we present Nabu - a web metasearch engine that, for a given user query, customizes the results obtained from an underlying search service with the aim of providing effective retrieval on Macedonian web collections. In addition to the main web search service, Nabu also offers services such as news, hardware, image and video search, which makes it a fully-featured Macedonian web search portal. Results of our analysis of the query logs and usage statistics, coupled with feature comparisons to other Macedonian web search engines, demonstrate that the services offered by Nabu are being effectively used and increasingly adopted by the Macedonian web users.

1. Introduction

Worldwide Internet usage has grown rapidly in recent years, particularly in non-English speaking regions. For example, the number of Internet users in Macedonia has increased by 109% in the period between 2005 and 2007 [5]. Even more rapid growth has been observed for the online populations in Latin America and Middle East [3], which creates demand for customized web searching in non-English languages. However, existing web search engines, such as Google [2], may be unable to meet this because they primarily serve English speaking users.

Web search in a foreign language is closely related to the grammatical rules of that language, thus making the English nature of rules for searching mostly inappropriate [4]. In most cases, in English, a word in

infinitive is a substring of the expanded word (work » works, worked etc.). However, in Macedonian, the stem of an expanded form of a word may have a different meaning. For example, the word “работа” (*rabota* » *eng.* work) has the stem “rabot”, which in Macedonian means something else – “the edge”, derived from the word “раб” » *eng.* edge. On the other hand, the word “висок” (*visok* » *eng.* tall) follows the rules for English search and its morphological forms are accomplished by adding suffixes to the end of the word. As a third example, the word “медитација” (*meditacija* » *eng.* meditation) has the stem “medit” which does not actually mean anything in the Macedonian language.

Table 1. Examples of Cyrillic/Latin transliteration

Word	Transliteration	Translation	Explanation
<i>n.</i> работа	rabota	work	Work
<i>v.</i> работам	robotam	work	I work
<i>v.</i> работиш	robotish	work	You work
<i>v.</i> работи	roboti	works	He works
<i>n. pl.</i> работи	roboti	things	Near things
<i>v.</i> работите	robotite	work	As in where do You work?
<i>n. pl.</i> работите	robotite	things	Those things
<i>n. pl. aug.</i> работитата	robotishtata	things	Those huge things over there
<i>adj.</i> висок	visok	tall	He is tall
<i>adj. pl.</i> високи	visoki	tall	They are tall
<i>adj. f.</i> висока	visoka	tall	She is tall
<i>n.</i> медитација	meditacija	meditation	Meditation
<i>v.</i> медитирање	meditiranje	meditating	The act of performing meditation

Table 1 lists these and similar word examples in Macedonian, along with their corresponding transliterations, translations and explanations. The following conclusions can be derived from the table:

- The stem of the word may be either a non-existing word or a completely different word.
- The grammatical forms in Macedonian language are not always formed by using suffixes to a word in infinitive; rather, some

ending letters might be dropped or replaced with others.

- In specific cases, there are many existing heteronyms to a word - words written the same, but having different meanings and pronunciations.
- No fixed number of letters subtracted of the word's end guarantees a valid resulting stem.

To address these challenges, we have developed Nabu (www.nabu.mk) - a web metasearch engine that, for a given user query, customizes the results obtained from an underlying search service with the aim of providing effective retrieval on Macedonian web collections. The Nabu project is an excellent example of an academic collaboration with the IT industry in Macedonia. During the Nabu development, a large number of algorithms appropriate for searching on Macedonian document collections were developed and tested. Because of their language-specific rules, these algorithms provide more accurate results as opposed to those provided by the modern search engines. This approach is rather practical as most of the Macedonian web users are accustomed to creating only simple searches; indeed, rather small fraction of these users possess knowledge for creating complex searches such as using an "OR" query, using time and domain restrictions, searching in documents' titles, or combining Cyrillic and Latin searches [5]. It is therefore important to widen the exploitation of these algorithms beyond standard web search and apply them on local news, hardware prices, governmental and education pages, and image and video search.

2. Nabu web search services

This section presents the services offered by the Nabu web search portal. It particularly focuses on services offering innovative aspects which, when applied on Macedonian web collections, emphasize Nabu's uniqueness compared to other search engines.

2.1. Web search

Figure 1 depicts the Nabu web metasearch engine, showing a typical data flow starting from the user's query and ending to the point where the results are displayed. To address the issues resulting from the Macedonian language complexity, an approach called *fast stem* is used in the web search process.

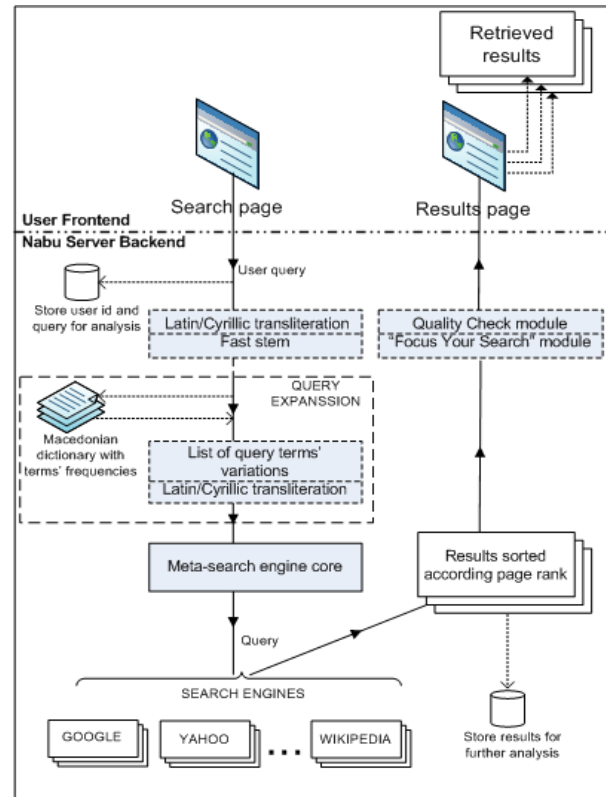


Figure 1. Nabu web metasearch engine

The *fast stem* is an algorithm that has a purpose of finding the stem of an arbitrary chosen word in fast and safe manner. Speed relevance is emphasized so as to avoid longer search response time. It operates by gradually removing one letter at a time from the end of a word and checking its stem options. The algorithm discontinues if more than 32 words are found by using the actual word as a stem. This is due to both the limitations of the underlying search service and the reasonable amount of query expanded words when considering the average number of possible words that can be derived from a stem in Macedonian language.

All words originate from an internal dictionary consisted of Macedonian words collected from a set of representative repositories. The words are sorted by their frequencies of occurrence in these repositories. The quality of the fast stem algorithm largely depends on this dictionary. Given the inexistence of an official source of all forms of Macedonian words, the dictionary is being continually expanded so as to produce an almost complete list of words that could branch from a particular word stem.

Since not all words from the query expansion process need to be included, a frequency sort in descending order is performed. This approach results in the best guess of possible variations of the word that the user is looking for. Successful cases are

“*rabotishtata*”, which would end up to “*raboti*” because there are no in-between stems, as well as “*visok*” and “*meditiranje*”. Even though these rules may be inapplicable for specific scenarios, the intention is that trading speed for quality and having the searched word in the query expansion would enable effective retrieval and better user satisfaction.

Some search engines covering the Macedonian spoken area incorrectly assume that a Macedonian web page is the one written in Cyrillic that belongs to the Macedonian *.mk* domain. However, there are many Macedonian web pages transliterated in Latin as well as many Macedonian web sites which do not belong to the *.mk* domain. Accordingly, Nabu uses *transliteration* for each word that resulted from the query expansion process thus equalizing the search regardless of the alphabet used. A problem addressed during transliteration is mapping of a Cyrillic letter in either one or two English letters – (ќ » [k, kj], “ш” » [‘s’, ‘sh’]).¹ Furthermore, the search is not only performed on *.mk* domains, but it also exploits Macedonia region search covering pages written in Macedonian language or connected to other Macedonian pages.

Advanced web search. There are specific scenarios, such as named entities search, where query expansion fails to provide effective retrieval. To cater for such cases, a mechanism called *quality check* (QC) is employed in Nabu to examine the obtained search results and to check whether the word searched by the user appears in the resulting snippets. A specific search is considered passing the quality check when the searched word is within an acceptable percentage of the obtained results. The TF/IDF algorithm [1] is applied to detect whether the word is important for the displayed results. By considering snippets to represent documents, the importance of the word t_i within the document d_j is calculated by taking its word frequency (tf) and multiplying it with the inverse of the document frequency (idf):

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} ; idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Here, $n_{i,j}$ is the frequency of occurrence of the word t_i in d_j , D is the set of all documents (snippets), whereas the tf denominator is the sum of frequencies of all words in d_j and the idf denominator is the number of documents containing the word t_i .

¹ Some parts of Macedonian web documents are written using Macedonian fonts mapped to the English alphabet, instead of using UTF or Cyrillic support. The characters “[{}~@\]”, occasionally found in some web documents, are thus treated by Nabu as letters.

The final score for every word t_i is multiplied by N/P , where N is the number of displayed results and P the position of the snippet, hence valuing more the words occurring at the upper snippets. Additional consideration is also taken into account when the word occurs in the title of the snippet. By using QC it can be easily assessed whether a search is successful. If a search did not pass the QC, it is possible to search the same query without query expansion by using the “-“ prefix for each word in the query.

It is interesting to note that in most cases the query expansion fails with named entities or verbs that are also names of companies or other named entities. For example, if “*plivanje*” (swimming) is the searched word, the first result is “*pliva*” (swim). However “*pliva*” is the name of a medicine company having better page rank than the pages concerning swimming. It would be useful to have automated mechanism of detecting such situations and deciding when to employ query expansion. Currently, the choice whether to perform query expansion is left to the user.

In addition to QC, Nabu also provides another tool, *focus your search* (FYS), which is used to discover words related to the user query. The FYS tool analyzes the results, extracts a list of most relevant words from the search results, and subtracts this list from the expanded query word list. The result is a collection of most relevant words that at the same time do not represent variations of the exact searched word. As opposed to query expansion, the FYS tool performs quite satisfactory for named entities and vague searches. For example, if the searched word is “*al*”, the FYS tool would return the related Macedonian words *program*, *news* and *television*, as *A1* is a Macedonian TV and Internet News medium.

2.2. News search

There is a plethora of Internet news services in Macedonia serving news feeds on a 15 minute interval. These services either do not provide searching at all, or provide searching using plain database words match. Furthermore, some of these services are unclassified, whereas others are misclassified. Nabu, on the other hand, is capable of combining news from various sources, classifying the unclassified and misclassified news, finding similar news by analyzing their content, as well as detecting popular news.

News gathering. Nabu continuously gathers news from various RSS feeds and HTML pages, updates its local database and performs word stemming of these news. The stemming process uses the *rule based stemming* (RBS) algorithm. The rule based stemming is a set of rules and procedures that, when applied, return

a valid stem of a word. The main difference between the fast stemming and the rule based stemming algorithm is that RBS is guided by rules extracted from the Macedonian grammar that can even be applied to words that do not exist in the internal dictionary.

Based on grammar analysis and experiments, around 100 rules covering most cases for word stemming are produced. This kind of stemming has the advantage of automatically annotating the transformations of the words (plural – singular, male – female etc). After the news words are stemmed, they are stored in the database for further analysis.

News classification. While the service is running, the already classified news are gathered and assigned to one of the eight predefined news classes. The annotated news and their corresponding stemmed words are stored in the database from where after a certain period of time a *Naïve Bayesian* training data set is generated. The training data set contains stemmed words and their probabilities to be in one of the eight predefined news classes. The resulting class for an unclassified news item is then defined by the combined probability of each word contained by the news item.

Nabu also determines the textual similarity between two news items by using an algorithm that assigns a similarity value to the items that needs to be greater than an empirically determined threshold for the items to be similar. This approach has proven to be quite effective, and more importantly, highly efficient, considering the need for frequent classification of large amount of news items.

News popularity. After calculating news similarity, the popularity of a news item is determined by taking the number of items to which this particular news is similar. For example, if there are ten similar news items to the news item X and five similar news items to the news item Y, then X is considered to be more popular than Y. By using this simple sorting of the news similarity model, a selected number of highly ranked news are shown on top of the news service web page and placed under the category of popular news.

2.3. Hardware search

Nabu also facilitates searching for hardware component parts from local retailers and allows for comparing prices for identical components. A problem concerning this type of search is the unclassified hardware components in the sources obtained on a daily basis. There are pricelists obtained from companies that have properly annotated their offered hardware components. Using these annotations, the

training set is created to automatically classify new unclassified components. Here, a combination of Naïve Bayesian and TF/IDF is applied mainly due to the fact that an identical word can appear in a number of classes and at the same time can have different values assigned to each class. For example, the Bayesian part of the calculations would imply that Logitech is almost certainly related to Mouse, but not to a CPU. The TF/IDF part of the calculations allows lowering the importance of some common words for classes in which other common words would be more important. Such example is the common word “128MB” having greater importance within the Graphic Cards class than in the Mouse class. The components are then classified and presented to the users in a manner allowing them to easily identify the lowest prices.

2.4. Other web search services

Nabu offers a wealth of other useful services for Macedonian web users. These include government services search, including searches throughout government’s tenders and publicly available employment openings by restricting the time frame within the actual search; educational search, accomplished through searching in educational sites’ titles, presentations, notes, and lectures; and image and video search services.

3. Usage statistics and query log analysis



Figure 2. Nabu’s usage statistics

Analyzing the query logs of a search engine is helpful in obtaining usage statistics that could reveal the way users interact with the search engine, the frequency of their visits, the level of user satisfaction, and so on. Figure 2 depicts the usage statistics of the Nabu web search portal, showing that recently the portal is being increasingly adopted by the Macedonian web users. User privacy and anonymity is guaranteed by avoiding associations between their personal information and searched queries.

4. Feature comparisons

Currently, apart from Nabu, two alternative Macedonian web search engines, Najdi and Pogodok, offer their services. Whilst Nabu relies on a meta-search engine core for obtaining the search results, the two alternatives have their own implementations of web crawling and page indexing mechanisms.

The first Macedonian search engine, Najdi (www.najdi.org.mk), was established in 2004 by Petar Kajeovski. Besides web and blog search services, it offers content search over a limited collection of more than 340 digitalized Macedonian books. The relevance of the search results produced by Najdi is solely based on the word frequency (tf), without considering the general importance of the word (idf) or employing page-ranking algorithms. Identification of stemming and query expansion utilization was difficult due to the missing or non-highlighted queried terms in the results' snippets. Nevertheless, after an exhaustive usage of Najdi search services it was evident that these features are not supported.

Pogodok (www.pogodok.com.mk) was launched by Interseek, Ltd. in 2005 as one of the series of localized search engines established in the Balkan region states. Pogodok is based on proprietary search engine that employs relevance ranking of results and support for Latin/Cyrillic transliteration, stemming and query expansion; however, our tests revealed that all variations of a stemmed word are not considered. It further tends to combine image results with standard web search results without clear separation as, for example, Google Universal Search does.

Table 2 shows a comparison of available features offered by the three search engines. It can be observed that Nabu has competitive advantage over its alternatives by delivering a wider choice of useful web services at users' disposal.

Table 2. Feature comparisons

<i>Feature</i>	<i>Nabu</i>	<i>Najdi</i>	<i>Pogodok</i>
<i>Web search features</i>			
Latin-Cyrillic transliteration	✓	✓	✓
Query expansion	✓	✗	✓
Stemming/Stop words support	✓/✓	✗/✓	✓/✓
Results' quality check	✓	✗	?
"Focus Your Search" feature	✓	✗	✗
Search support for different type of documents	✓	✓	✓
Search within specified domains	✓	✓	✓

Search according to page modification date	✓	✓	✓
<i>Other search services</i>			
Images/video files search	✓/✓	✓/✗	✓/✗
News search/classification /clustering	✓/✓/✓	✗/✗/✗	✗/✗/✗
Blog search	✓	✓	✗
Hardware search /classification	✓/✓	✗/✗	✗/✗
Government sites search	✓	✗	✗
Educational sites search	✓	✗	✗
Book search	✗	✓	✗
Email search	✗	✗	✓
Location/Map search	✓	✗	✗
Advertisements' search	✓	✗	✗

5. Conclusion and future work

Nabu is a fully-featured Macedonian web search portal that offers a wealth of innovative user services by providing effective retrieval on Macedonian web collections. Results from our analysis of the query logs and usage statistics, coupled with feature comparisons to other Macedonian web search engines, demonstrate that the services offered by Nabu are being used and increasingly adopted by the Macedonian web users.

In the future, we plan to carry out extensive user studies in order to reliably evaluate the effectiveness of the offered Nabu services. We also plan to offer personalized web search services by applying collaborative filtering on the query analysis data. This, we believe, would further improve the usability of the web search portal.

Acknowledgements We thank Vladimir Radevski for his valuable comments on earlier drafts of this paper.

6. References

- [1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", Addison Wesley, 1999.
- [2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", In *Proceedings of the 7th International Conference on World Wide Web*, Brisbane, Australia, 1998. pp. 107 – 117.
- [3] W. Chung, "Web searching in a multilingual world", *Communications of the ACM*, Volume 5, Number 5, 2008. pp. 32 – 40.
- [4] C. Peters et al., "Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007)", Lecture Notes in Computer Science, Volume 5152, Springer, 2008.
- [5] United States Agency for International Development (USAID), "Internet and computer usage survey in the Republic of Macedonia: Quantitative research", 2007. <http://www.mkconnects.org.mk/new/public/readings.php>.