QUERY FORMULATION IMPROVED BY SUGGESTIONS RESULTING FROM INTERMEDIATE WEB SEARCH RESULTS

Ladislav Martinsky and Pavol Navrat

Slovak University of Technology, Bratislava

ladislav.martinsky@gmail.com *and* navrat@fiit.stuba.sk

Abstract

Adequate query formulation for web search is a very important prerequisite for successful retrieval of desired results. One productive way to improve it that is used in most popular search engines is query suggestion. Common search engines are quite good at suggesting words to general, highly popular queries, but their suggesting capability deteriorates swiftly when more specific queries are involved. In this paper, an approach is proposed to suggest words that are derived from actual search results. The approach guarantees some words will be suggested for any query, not just for most general or popular ones.

Keywords: *Web search, query suggestion, personalization*

1 Introduction

Users – in fact particular kind of computer users that can be described as interested fellows - formulate their needs and thoughts to search engines by means of a query. Its interpretation is affected by various factors: length, information richness, type (information, navigational, transactional (Cho, 2005), (Levinson, 2004)) or context. All these properties can be modified or enhanced using statistical methods (Vanekova, 2010), techniques or by monitoring interested fellow's behaviour (Barla, 2010). Visible problem or possible obstacle is interested fellow's low willingness and attitude to sufficiently formulate the query which results in too general search results. Traditional approach in this case is an analysis of results and query refinement. This cycle can take a considerable amount of time, depending on the complexity of the domain and the information need.

Popular solution used in this case is query suggestion, known from common search engines like Bing (`www.bing.com`), Google (`www.google.com`) or Yahoo (`www.yahoo.com`). It is based on a suggestion of possibly helpful words to the query in real time. The way of suggesting is non-obtrusive and motivates the interested fellow to further specify his/her information need. One of the main constraints associated with this solution is its limitation to statistically popular words. This can be observed when interested fellow starts typing more specific or complicated queries. Since the number of possible queries is unlimited, it is not practically possible to have all relevant suggestions for them generated and stored in advance. Without it, it is hard to see how suggesting in real time could be accomplished. From the interested fellow's point of view, when he/she submits query he/she does not receive any suggestion but still receives a lot of results, so why was there no suggestion? It means that data is available but not used for this purpose.

The main idea of the approach proposed in this paper is to provide query suggestion to any possible meaningful query for which the search engine returned at least one relevant result. The most important benefit is to relieve the interested fellow from having to browse and analyze offered results by himself/herself in case s/he is not entirely sure of what s/he is looking for or know how to specify it more precisely. The method is based on the automatic analysis of results from multiple search engines. Relevant words, helpful to the interested fellow are extracted during the analysis.

In this paper we show how a query reformulation based on suggestions originating from intermediate web search results can yield results that better reflect the interest of the interested fellow.

We begin in Section 2 by discussing current research directions related to query suggestion in the context of development of web applications serving as information providers, providing information resulting from web search determined by interested fellow’s query. There are many approaches on how to improve quality of results expressed either in terms of various measures (precision, recall) or simply in terms of interested fellow’s satisfaction with what s/he gets in some of the top results. We narrow our focus on approaches to some reformulation of the query and discuss also the limitations of this research.

In Section 3, we present our approach using results of initial search results returned by common search engines. We proceed in Section 4 with an outlook of information seeking with our Information Providing Agent. Section 5 brings a discussion on personalization.

Finally, in Section 6 we report on the method and results of our evaluation of our approach and then in Section 7 conclude with some general observations and recommendations for ongoing work.

2 Current Research

In this section we briefly explore the limitations of current state of the art based on some research works related to ours. Methods of query reformulation based on search results have been studied for several years. Often described as query expansion and pseudo relevance feedback related research can be traced down to the standard information retrieval area, e.g. (Mitra, 1998), (Xu, 1996). More recent works include (Kalmanovich, 2009), (Liu, 2011), attempts to improve the overall effectiveness of the query expansion by clustering documents.

2.1 Query Suggestion

Several of the known and widely used search engines employ query suggestion in one way or another. Google, Bing and Yahoo, to name just three known and widely used ones, offer this kind of assistance to the interested fellow, who plays the role of information seeker or consumer. At a closer look, however, there could be various ways of suggesting words to amend a query. In a static mode, suggestion is shown together with search results, not before they are produced. In that moment, interested fellow’s attention is usually directed to results, so the suggestion can easily be overlooked.

On the other hand, interactive query suggestion is generated and shown to the interested fellow while s/he is still typing the query. This approach encourages interested fellow to think about his/her real information need in an unobtrusive way. Interested fellow is free to ignore the suggestions, but making use of them can lead to a more specific and precise query. Figure 1 shows sample query suggestions from the three search engines as they are appearing while interested fellow types in the word “schwarzenegger”.

Search engine	Query suggestion
Google	letter, veto letter, soundboard, veto, movies, quotes, hidden message, memo, twitter
Ask.com	filmography, films, middle name, movies, governor
Yahoo	soundboard, movies, prank calls, arnold, pumping, governor of, sound clips, middle

Figure 1 Sample query suggestions for the query “schwarzenegger”

Similarly, figure 2 shows suggestions for the query “jaguar”.

Search engine	Query suggestion
Google	slovensko, xf, .sk, xj, xk, auto, zviera, bratislava, classic, cz
Ask.com	parts, facts, pictures, xf, cars, animals, information
Yahoo	2009 jaguar xf, cars, jacksonville jaguars, xk, parts, animal, s-type, xj220, dealers, pictures

Figure 2 Sample query suggestions for the query "jaguar"

It is interesting to note that Google's suggestions are localized to the language that is determined by the current location of the interested fellow. Ask.com and Yahoo, on the other hand, do not localize, but otherwise their suggestions seem similar.

The way of forming suggestions reflects several principles. One of the underlying ideas could be identification of statistically most successful and frequently used expressions in connection with the one that the interested fellow is currently writing. A weaker point of such approaches is a limited capacity of the resulting dictionary of expressions to be suggested. When the query is more specific, likelihood decreases that a suggestion is generated successfully. For example, when writing one's own name (ladislav martinsky), there were no words suggested at all. Still, search engine returned quite a big number of relevant results. Hence we see that the usual search engines are limited in their suggesting capabilities, obvious reason being the requirement of shortest possible response time.

The space of possible queries that any interested fellow can formulate is enormous, more precisely countable but practically unlimited. However, more important is that the world of things that any interested fellow can be curious about is also enormous. It would not be practical, in fact not even possible, for a search engine to have the best suitable query to suggest in stock for each interested fellow and for any of his or her interests. Hence suggestions must be formed to a large extent on the fly. Note that this does not exclude the possibility that at least some information that could be useful when forming suggestions is prepared (pre-computed, precompiled) and stored in advance. Also note that these contemplations imply suggestions that are in principle a function of the querying person since a different interested fellow can have a different thing in mind when writing possibly the same words in a query (Baeza-Yates, 2006). Hence our effort in this research includes personalization (Teevan, 2008), (Liu, 2004).

Making suggestions dependent on interested fellow's information need or desire requires that the suggesting engine is able to receive some information reflecting what interested fellow's information need is. Moreover, the information must be such that the engine can extract or derive from it the right suggestion. Several research works have shown that search results for a query given by the interested fellow could be such a copious source of information (Jiang, 2009), (Song, 2010).

In Dreher (2006) there is described an approach to assist query formulation by semantic word analysis using thesaurus based vector representation. The interested fellow (re-) formulates the query iteratively, but the algorithm relies on reference data typically represented by a thesaurus, which obviously must be created in advance.

In Ma (2010), an approach for query expansion is described where query expansion is understood as a process of adding words to the original query to improve retrieval performance. Their approach is based on the recent history of queries. First, they generate a historical query-click graph that records the clicks that were generated by URLs when an interested fellow input a query. Then, they transform it into a term-relationship graph, with nodes representing terms inside query logs and edges representing the following relationships between terms: two terms are related if the interested fellow input them in different queries, and clicked the same URL in their lists of results. For a given query, for each term in it, most similar terms from the graph are used for query expansion. Again, the approach needs precompiled graph data structures.

Broccolo (2010) emphasized the need of periodically updating, or rebuilding from scratch, models used in query recommendation, to keep up with the possible variations in the interests of users.

Li (2009) proposed a way to suggest queries but their purpose is somewhat different. Given an initial query and the suggested related queries, their search system concurrently processes their search results lists from an existing search engine and then forms a single list aggregated by all the retrieved lists.

Yang (2008) outlined briefly their idea of query suggestion based on search context representation consisting of search results (titles and snippets of web pages only) and query log sessions. Sahani (2006) proposed a web based measure for similarity of short text snippets and then evaluated it in a simple query suggestion application. It needs initially a repository of previously issued queries, culled from search engine logs and computes similarities between them and the current query to suggest those most similar ones to the interested fellow.

2.2 Using search engine results

In Chen (2007), a method is proposed there to correct spelling errors in query using actual search results. Spelling errors in queries can be corrected quite easily and successfully in domains for which there are available dictionaries or word databases. But the main problem lies in the nature of web search. Query can come from any area or domain. It can be composed of words from potentially unlimited sets of words or their grammatical forms in many different languages. It is practically impossible to create in advance a sufficiently comprehensive representation of possible spelling errors and their corrections. The idea of using search results stems from the fact that often search results contain not only correctly spelled words, but also their misspelled occurrences. Results of this research show statistically a very good improvement in spellchecking using web search results as background knowledge.

Using search results as main data source is proposed also in Levinson (2004) and Martinsky (2010). Authors analyze main parts like title, description or url address to categorize search results. This approach has proved itself as a fast and efficient way to improve results readability.

2.3 Personalization

Finding a right context for interested fellow in web search can be very difficult. The main problem is that due to shortness of query, there is very little information available. When interested fellow submits for example "operation" query, how do we know whether s/he means financial, surgical or mathematical operation? One way to approach this problem is via personalization. This process is based on acquiring and storing interested fellows' profiles or analyzing their actual activities.

3 Our Approach

Rationale

Let us consider the process of searching for information, as it usually takes place in an interaction between an interested fellow and some search engine. The interested fellow submits a query composed of very few words, realizing this is just a beginning of the search. The reason is that s/he is well aware of the fact that the search engine would not understand the query written in natural language, in a natural way as a simple question, anyway. This is psychologically a good starting point for the search engine: its client, the interested fellow understands that the searching process will be an iterative one. The search engine needs precisely this, since it is not able to produce a precise and recalling answer in one step. Our point is that it should not behave as if it was able to. The first list of results is not the final answer. It is usually too imprecise, too little recalling at best. At worst, the results may completely miss the intention of the interested fellow. The search engine should act as if continuation of their interaction with the interested fellow was the most natural thing. Of course, it should offer something interesting to the other party, to keep him/her involved. Our idea is to suggest query enhancements for the interested fellow to choose from. The interested fellow chooses one of them by a very simple act – just a single click – and hence initiates another round of searching for results. The search engine returns another list of results, hopefully more precise, more recalling, containing documents with information that answers the implicit original information request of the interested fellow. If not, another round in this interaction loop takes place.

Main idea of our approach is to find words enhancing the actual query automatically and to suggest a few such alternatives to the interested fellow. Words enhancing the actual query are the result of an analysis of the search results by both the interested fellow and the search engine.

Search engine analyses each document from the list of results, producing the usual vector of words. Interested fellow browses through the list of results and clicks on some of them. This is interpreted as a sign of interest.

Proposed Approach

Our approach identifies four main data handling operations:

1. Data elicitation – gathering and storing words
2. Data processing – transformation to unique form
3. Data evaluation – evaluating and sorting
4. Presentation – providing suggested words to interested fellow

The operations form a four step body that is the core of the information seeking/providing process. It can be repeated several times.

The whole process starts when interested fellow finishes typing the query. An information providing agent (IPA) implemented as a small prototype application working according to our approach submits it to data sources and gathers data, which is processed and evaluated. The final chosen words, which are most relevant to the query, are presented, via suggestion, to interested fellow. Clicking on any of suggested words restarts the whole process with a modified query.

Data elicitation

The operation of data elicitation is responsible for gathering and storing words from search results. It is devised to:

1. produce words or texts which are related to query.
2. differentiate and evaluate words on the basis of certain characteristics.
3. be accessible via internet with acceptably low latency.

Results that are returned by contemporary popular search engines mostly do possess some degree of relevance to almost any meaningful query. Basic structure of the search result also provides a good starting point for ordering individual words in terms of importance. Because the space available in a result is very short, search engines use advanced techniques to compose this information wisely. Based on subjective judgment, search results are a very good source of information related to query even without analyzing the actual web sites that they describe.

Three search engines were chosen as data sources for the purpose of experimentation. Google and Yahoo represent a more classical type. The results that they return include title, description and url address of the referenced web page. Some of Yahoo results also include words related to the page. Third data source is DMOZ ODP (`www.dmoz.org`) representing a human-edited directory of web pages. Very important feature is presence of the category of the page and often also the category of the query. This is very valuable information when the objective is to determine words for suggestion.

Other possible data sources may be for example Bing, Delicious (`www.delicious.com`), Wikipedia (`www.wikipedia.org`) or any other source of information, which is compatible with requirements of this operation. The operation produces *search result objects*.

Data processing

Search results represent heterogeneous texts. Main task of data processing operation is to transform search results into a set of different homogenized data units – *word objects*. Word object is uniquely identified by a base form of word. Text processing involves some simple transformations, such as converting letters to lowercase, removing punctuation, redundant spaces and stop words, and stemming.

Removing stop words and stemming can be complicated tasks due to the fact that they are both language specific. Methods applicable to English texts are by definition not the same as their counterparts applicable to some other language. Surprisingly perhaps, we employed one for stop word removal for texts in Slovak language with satisfactory results. Similarly, instead of stemming, we applied a simple technique of removing vowels. In the future, they should be replaced with more advanced techniques and methods.

Figure 3 shows important characteristics attributed to a word with their values stored in word object.

Characteristic	**Description**
Position	Where the word appears in the result: title, description, url address, etc.
Spread	Density of occurrence of the word across different results.
Dominance	Percentage of occurrence to all the words.
Distance	Number of words between the word and a word from query.

Figure 3 Important characteristics stored in word object

Data evaluation

Most important part is selecting the relevant and helpful words, which will be suggested to interested fellow. In practical terms, our intention is to select some 10-20 words. From the nature of the problem it follows that the task is to select them from approximately 300 words. Because words are just strings of letters bearing no semantics, we need some sort of other characteristics on which the evaluation will be based. We use the characteristics shown in figure 3. They are heuristic in nature. Each characteristic is devised in such a way that it can be evaluated and as a result, a numerical value can be attributed to it (see figure 4). The range of values for each characteristic is calibrated to interval 1-10. The heuristic evaluation should rate higher those words that are hopefully more relevant to intention as expressed in the query.

Characteristic	**Description**
Position	Words are rated based on position where they appear. Positioning in title contributes for example 4 points, in description 3 points, etc. The more points, the higher rating.
Spread	If word appears in many results, than it may be important in relation to the query. The more occurrences in different results, the higher rating.
Dominance	If word occurrences dominate to all other words, then the dominating word may be important. The more occurrences, the higher rating.

Distance	If word appears close to a word from the query, then there is chance, that it can effectively enrich it. The closer to query word, the higher rating.

Figure 4 Rationale for individual characteristics rating

It may be interesting to mention, that our original interpretations of spread and dominance were different prior to experimentation. One of the considerations was for example to include as many contexts for a query as possible. A word that appears just in one result seemed to us to be a stronger referent of a particular context than a word that appears in, let's say, a dozen of results. Experiments have not endorsed this consideration. Uniquely appearing words do not bear any more significant relevance to the query.

After we made the crucial design decisions regarding the choice of word (suggestion candidate) characteristics, it was necessary to devise rules for evaluating them. We devised some set of rules that allows ordering the set of words according to their relevance to the query. Our design rationale was to devise the rules as simple as possible, since our hypothesis has been that even with relatively simple operations it is possible to achieve a significant improvement in query formulation resulting in better (according to some measure) results returned.

The rules are again heuristic in nature. They do not guarantee any improvement but, as our experiments will show, they often lead to a significant improvement. The results match more closely the interested fellow's intention as expressed, due to query language and other limitations, inherently imperfectly, by one or a few words. The rules are implicitly expressed in descriptions of how to evaluate the particular characteristics of a word (see figure 5).

Our approach relies also on some limited personalization. It is possible to formally view the effect of personalization as a kind of word characteristic, albeit not directly related to query results. Personalization allows evaluating a word according to its inclusion in the interested fellow's profile. Formally, we include personalization rating as the fifth characteristic. Its evaluation can be up to 10 points. Thus we have now five characteristics of a word. Maximum possible evaluation for a word is 50.

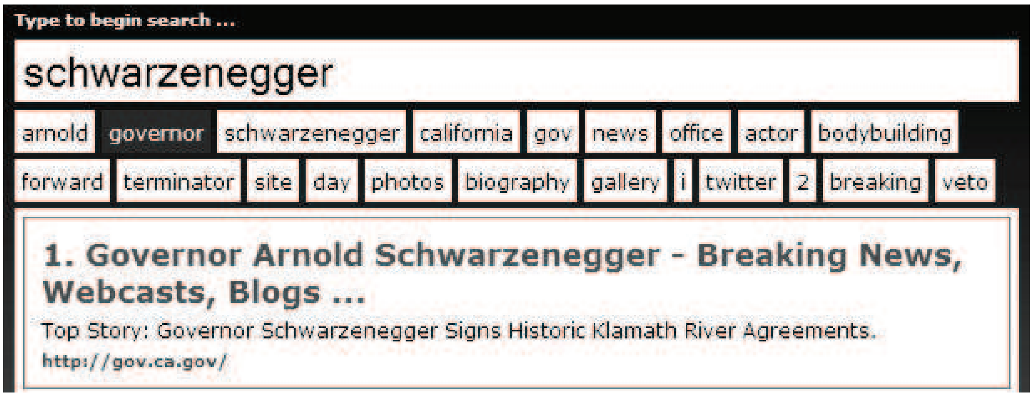
Characteristic	Evaluation
Position	Parts of the results identified by their position have different importance. • **Dmoz categories, Yahoo relevant expressions.** These items come from manually created or controlled assignments of a meaning. At the same time, they are directly linked to the given query. These properties rank these items among the most important ones. (**3 and 4 points**) • **Title** – authors or search engines tend to formulate them as short apposite strings directly related to the query or the context. Due to the fact that their length is limited and their composition requires including truly relevant expressions, item at this kind of position gets second best rating. (**2 points**) • **Description** – lowest rated position due to the fact, that here there are usually higher numbers of words many of which may not be related to the query. (**1 point**) **Evaluation rules:** • Maximum points that a word can earn is **14 points**. • For each occurrence of a word at a position, the word earns the corresponding number of points.

	- Resulting sum is mapped onto interval 1-10, where 10 is the best rating.
Spread	**Evaluation rule:** - A word earns 10 points if it is included in the highest number of results, earns 9 points if it is included in the second highest number of results and so on down to one point.
Dominance	**Evaluation rules:** - A word with the highest relative occurrence count earns 10 points, it earns 9 points if its relative occurrence count is the second highest etc.
Distance	**Evaluation rules:** - The less distant a word is from the given query the more likely it is related to it. - A word earns 10 points if its distance to query is 1, it earns 9 points if its distance to query is 2 etc.

Figure 5 Evaluation of word characteristics

Presentation

The last step of process is delivering the chosen words to interested fellow. A single web page is used for this purpose with top search bar, panel for suggestion words and results list (see figure 6).



schwarzenegger

1. Governor Arnold Schwarzenegger - Breaking News, Webcasts, Blogs ...

Top Story: Governor Schwarzenegger Signs Historic Klamath River Agreements.

http://gov.ca.gov/

Figure 6 of suggested words to interested fellow within the web page

When interested fellow clicks on one of the suggested words (in figure 6, s/he is clicking on the word governor), the query is immediately augmented with this word and search restarts with such an augmented query. Interested fellow can easily specify additional words with only one mouse click.

4 Outlook of information seeking with our Information Providing Application

Let us give an of outlook how an interested fellow could typically interact with an information provider that acts according to the idea that we propose. For experimental purposes we created a prototype information providing agent (IPA) that implements the query suggestion capabilities. A diagram showing its operation is shown in figure 7.

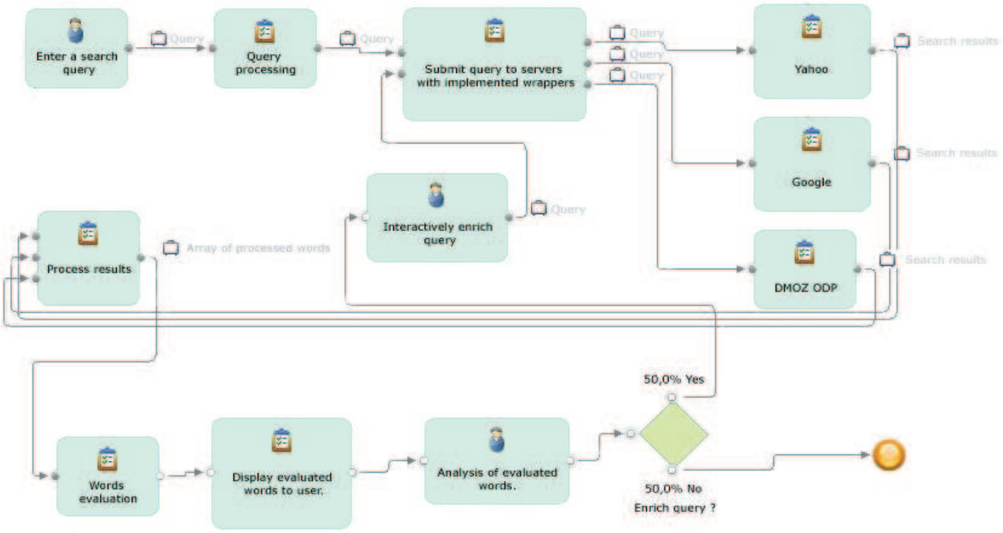
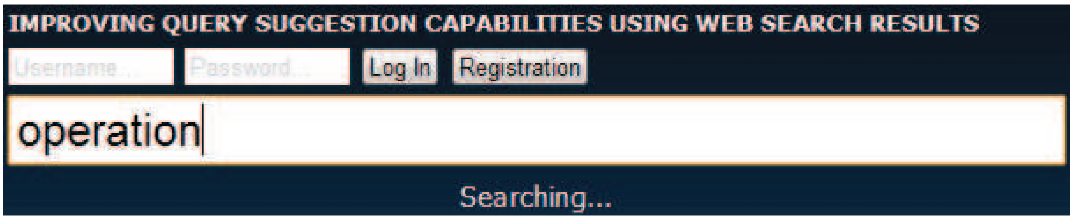


Figure 7 Operation of our information providing application

Initial point of the information seeking/providing endeavour is when interested fellow writes a query. The query is in *statu nascendi*, and intentionally so, since the interested fellow is aware that the IPA will come up with suggestions. Here, it consists of just one word “operation“ (see figure 8).

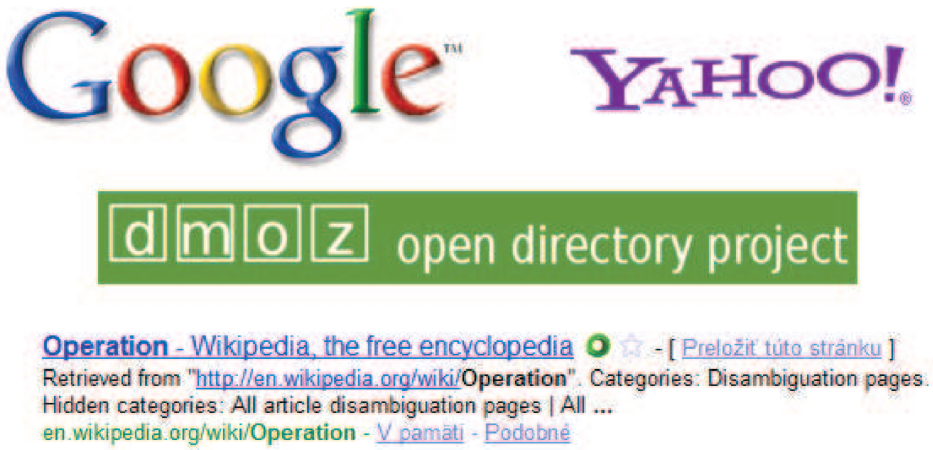

IMPROVING QUERY SUGGESTION CAPABILITIES USING WEB SEARCH RESULTS

operation

Searching...

Figure 8 Interested fellow begins information seeking by writing initial query “operation“

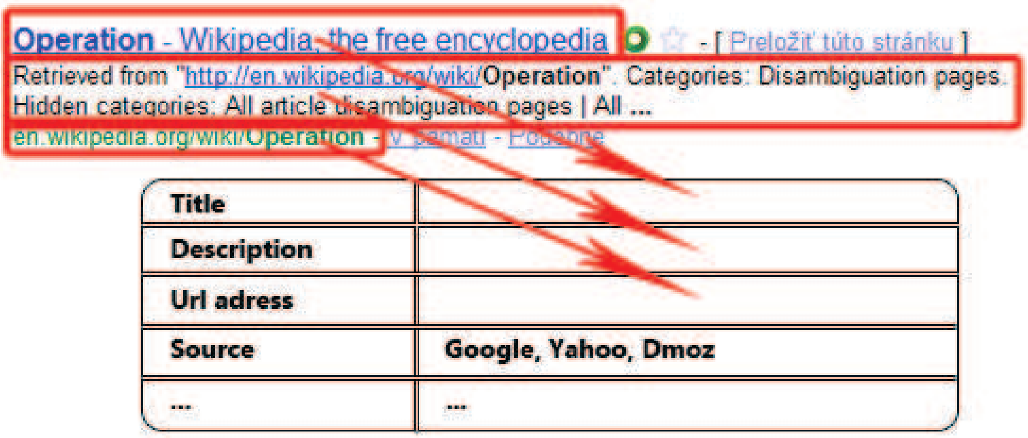
The IPA submits the query to three common search engines (see figure 9) in order to receive search results that will be subjected to analysis.



Operation - Wikipedia, the free encyclopedia - [Preložiť túto stránku]
Retrieved from "http://en.wikipedia.org/wiki/Operation". Categories: Disambiguation pages. Hidden categories: All article disambiguation pages | All ...
en.wikipedia.org/wiki/Operation - V pamäti - Podobné

Figure 9 Submitting the initial query to three known search engines

The search results are first processed to yield their object representation (see figure 10).



Operation - Wikipedia, the free encyclopedia - [Preložiť túto stránku]
Retrieved from "http://en.wikipedia.org/wiki/Operation". Categories: Disambiguation pages. Hidden categories: All article disambiguation pages | All ...
en.wikipedia.org/wiki/Operation - V pamäti - Podobné

Title	
Description	
Url adress	
Source	Google, Yahoo, Dmoz
...	...

Figure 10 Processing search results into object representation

IPA then forms a set of objects that represent single words from the set of objects of results (see figure 11).

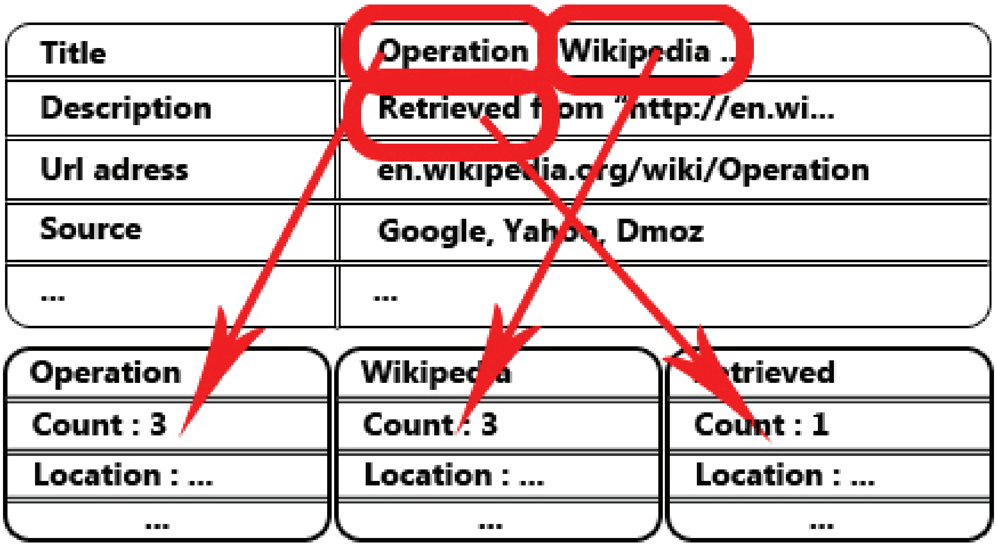
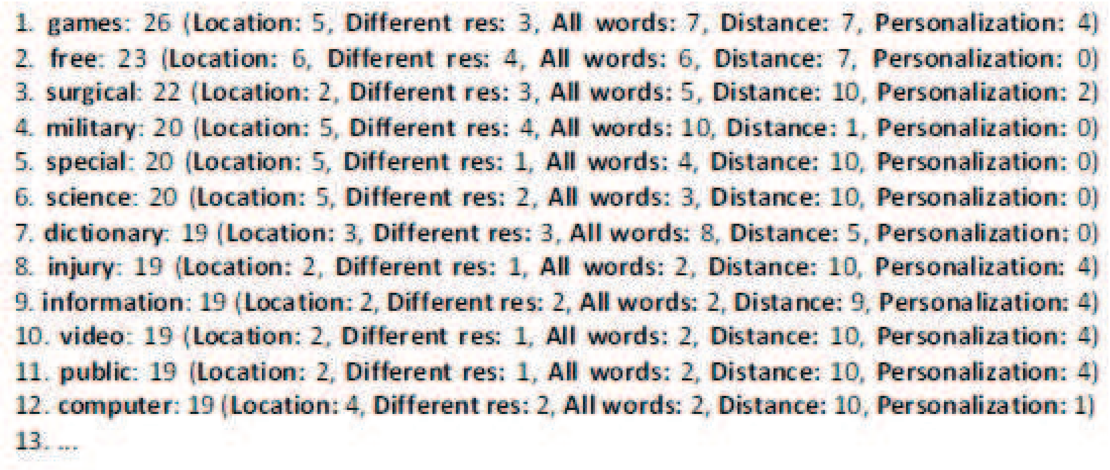


Figure 11 Forming word objects from a result

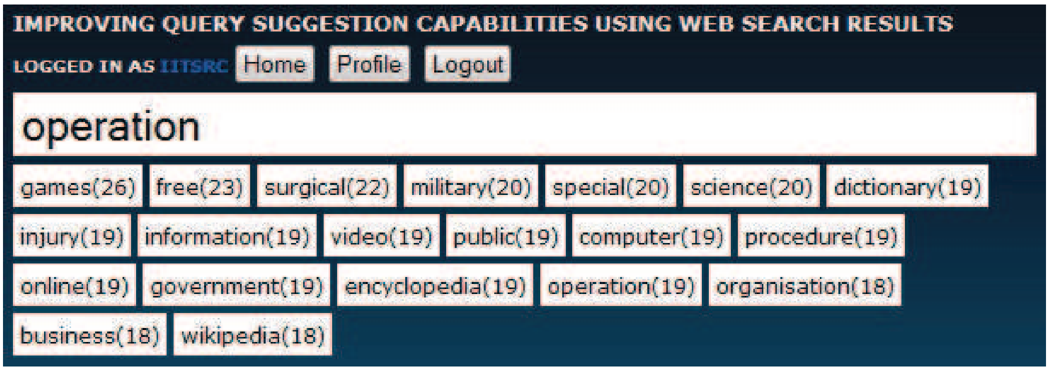
Single words are evaluated based on the chosen characteristics (see figure 12).



```
1. games: 26 (Location: 5, Different res: 3, All words: 7, Distance: 7, Personalization: 4)
2. free: 23 (Location: 6, Different res: 4, All words: 6, Distance: 7, Personalization: 0)
3. surgical: 22 (Location: 2, Different res: 3, All words: 5, Distance: 10, Personalization: 2)
4. military: 20 (Location: 5, Different res: 4, All words: 10, Distance: 1, Personalization: 0)
5. special: 20 (Location: 5, Different res: 1, All words: 4, Distance: 10, Personalization: 0)
6. science: 20 (Location: 5, Different res: 2, All words: 3, Distance: 10, Personalization: 0)
7. dictionary: 19 (Location: 3, Different res: 3, All words: 8, Distance: 5, Personalization: 0)
8. injury: 19 (Location: 2, Different res: 1, All words: 2, Distance: 10, Personalization: 4)
9. information: 19 (Location: 2, Different res: 2, All words: 2, Distance: 9, Personalization: 4)
10. video: 19 (Location: 2, Different res: 1, All words: 2, Distance: 10, Personalization: 4)
11. public: 19 (Location: 2, Different res: 1, All words: 2, Distance: 10, Personalization: 4)
12. computer: 19 (Location: 4, Different res: 2, All words: 2, Distance: 10, Personalization: 1)
13. ...
```

Figure 12 Evaluation of single words based on the chosen characteristics

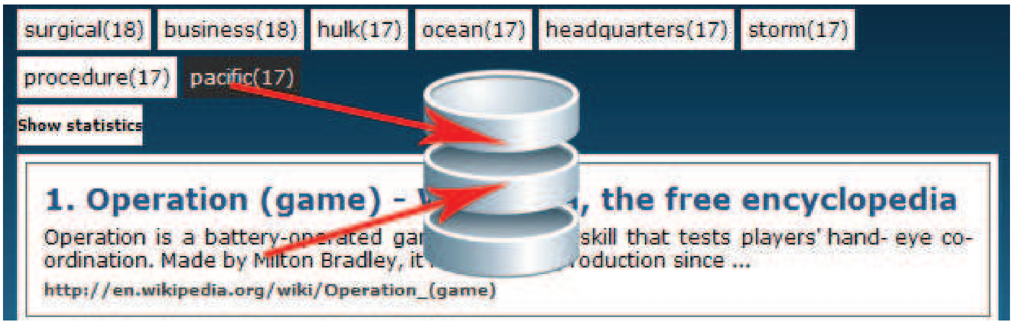
The evaluation of words induces their ordering. IPA suggests the ordered words to the interested fellow (see figure 13).



operation

Figure 13 Suggesting selected ordered words to the interested fellow

If the interested fellow is logged in, all of his/her actions are recorded in his/her profile and it can be used in subsequent processing (Figure 14).



1. Operation (game) - W , the free encyclopedia

Operation is a battery-operated ga skill that tests players' hand- eye co-ordination. Made by Milton Bradley, it roduction since ...

http://en.wikipedia.org/wiki/Operation_(game)

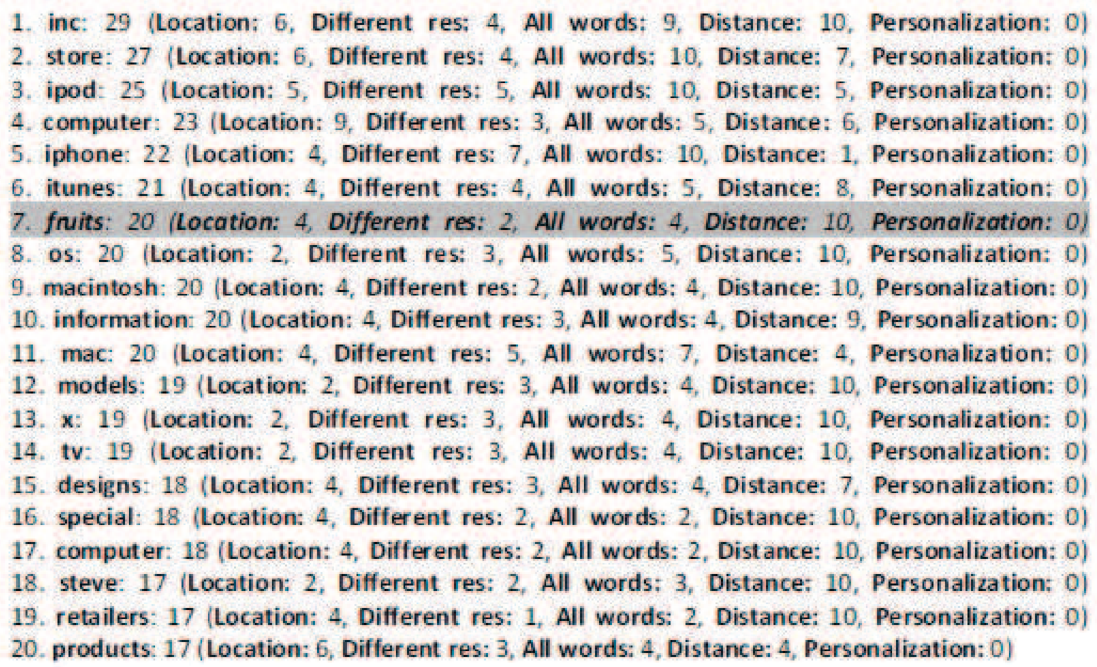
Figure 14 Recording the interested fellow's action in his/her profile

By clicking on one of the suggested words, or by augmenting the query with some other word, the interested fellow initiates the next cycle of the interaction with the IPA.

5 Discussion on personalization

To begin discussion on personalization and its effects in our approach, it is helpful to present a simple example for illustration. Let us consider a sample query “apple“. It is a word that can have several meanings (eg computer company and fruit) that are utterly different and such can be the interested fellow's intentions.

In figure 15 there is shown a list of the first 20 words (in the format showing their characteristics) that are suggested to the interested fellow who submitted the query apple but who has not signed in yet. Thus it is result of suggestion without personalization.

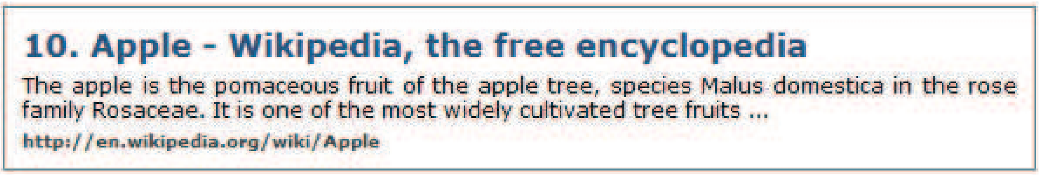


```
1. inc: 29 (Location: 6, Different res: 4, All words: 9, Distance: 10, Personalization: 0)
2. store: 27 (Location: 6, Different res: 4, All words: 10, Distance: 7, Personalization: 0)
3. ipod: 25 (Location: 5, Different res: 5, All words: 10, Distance: 5, Personalization: 0)
4. computer: 23 (Location: 9, Different res: 3, All words: 5, Distance: 6, Personalization: 0)
5. iphone: 22 (Location: 4, Different res: 7, All words: 10, Distance: 1, Personalization: 0)
6. itunes: 21 (Location: 4, Different res: 4, All words: 5, Distance: 8, Personalization: 0)
7. fruits: 20 (Location: 4, Different res: 2, All words: 4, Distance: 10, Personalization: 0)
8. os: 20 (Location: 2, Different res: 3, All words: 5, Distance: 10, Personalization: 0)
9. macintosh: 20 (Location: 4, Different res: 2, All words: 4, Distance: 10, Personalization: 0)
10. information: 20 (Location: 4, Different res: 3, All words: 4, Distance: 9, Personalization: 0)
11. mac: 20 (Location: 4, Different res: 5, All words: 7, Distance: 4, Personalization: 0)
12. models: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
13. x: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
14. tv: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
15. designs: 18 (Location: 4, Different res: 3, All words: 4, Distance: 7, Personalization: 0)
16. special: 18 (Location: 4, Different res: 2, All words: 2, Distance: 10, Personalization: 0)
17. computer: 18 (Location: 4, Different res: 2, All words: 2, Distance: 10, Personalization: 0)
18. steve: 17 (Location: 2, Different res: 2, All words: 3, Distance: 10, Personalization: 0)
19. retailers: 17 (Location: 4, Different res: 1, All words: 2, Distance: 10, Personalization: 0)
20. products: 17 (Location: 6, Different res: 3, All words: 4, Distance: 4, Personalization: 0)
```

Figure 15 List of the first 20 suggested words without personalization

In the list, highlighted is the word that links the word apple to its meaning as a kind of fruit. In the pure, not personalized state, this interpretation is the seventh on the list. It is also to be noted that most of the other suggested words link apple to the computer company or its products.

As a next step, after having signed in, the interested fellow may click on the item of the results shown in figure 16.



10. Apple - Wikipedia, the free encyclopedia

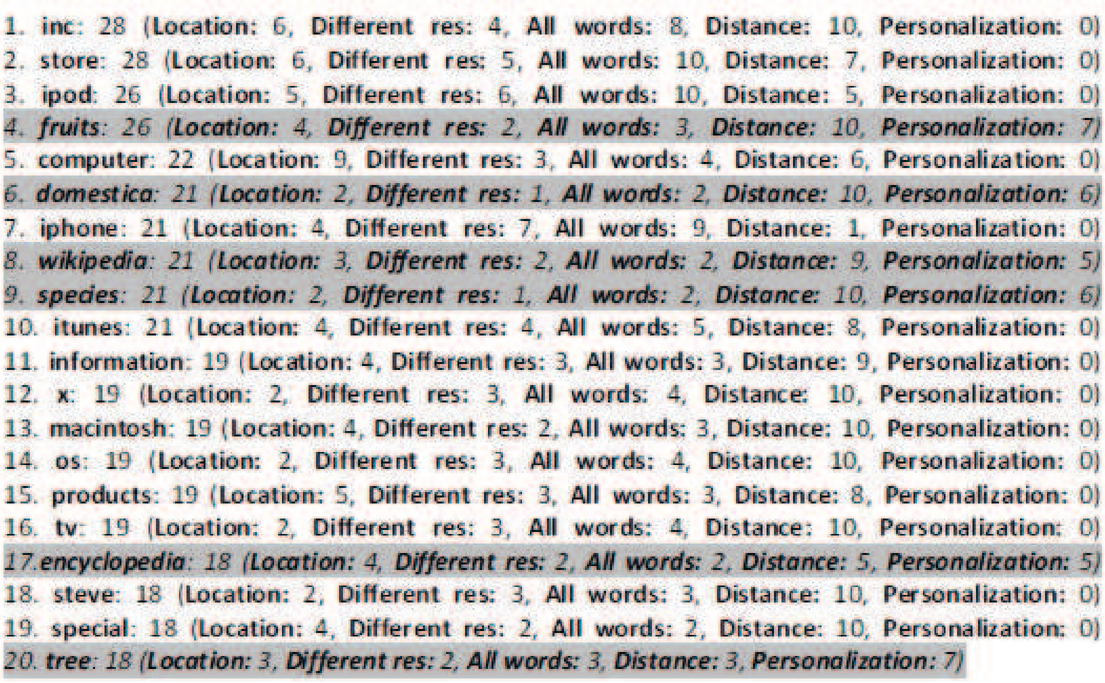
The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family Rosaceae. It is one of the most widely cultivated tree fruits ...

http://en.wikipedia.org/wiki/Apple

Figure 16 Results item selected by the interested fellow

As the content of the item suggests, it clearly refers to the meaning of apple as a fruit. It contains words as “fruit“, “tree“, “species“, etc. All the words from this item are inserted in the interested fellow’s profile.

In figure 17 an updated list of suggestions is shown. Words from the interested fellow‘s profile are now interspersed in the updated list according to their evaluations. When the query “apple“ is repeated, “fruits“ is higher on the list due to personalization. Three new, fruit related words are now also on the list: “domestica“, “species“, “tree“.



```
1. inc: 28 (Location: 6, Different res: 4, All words: 8, Distance: 10, Personalization: 0)
2. store: 28 (Location: 6, Different res: 5, All words: 10, Distance: 7, Personalization: 0)
3. ipod: 26 (Location: 5, Different res: 6, All words: 10, Distance: 5, Personalization: 0)
4. fruits: 26 (Location: 4, Different res: 2, All words: 3, Distance: 10, Personalization: 7)
5. computer: 22 (Location: 9, Different res: 3, All words: 4, Distance: 6, Personalization: 0)
6. domestica: 21 (Location: 2, Different res: 1, All words: 2, Distance: 10, Personalization: 6)
7. iphone: 21 (Location: 4, Different res: 7, All words: 9, Distance: 1, Personalization: 0)
8. wikipedia: 21 (Location: 3, Different res: 2, All words: 2, Distance: 9, Personalization: 5)
9. species: 21 (Location: 2, Different res: 1, All words: 2, Distance: 10, Personalization: 6)
10. itunes: 21 (Location: 4, Different res: 4, All words: 5, Distance: 8, Personalization: 0)
11. information: 19 (Location: 4, Different res: 3, All words: 3, Distance: 9, Personalization: 0)
12. x: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
13. macintosh: 19 (Location: 4, Different res: 2, All words: 3, Distance: 10, Personalization: 0)
14. os: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
15. products: 19 (Location: 5, Different res: 3, All words: 3, Distance: 8, Personalization: 0)
16. tv: 19 (Location: 2, Different res: 3, All words: 4, Distance: 10, Personalization: 0)
17.encyclopedia: 18 (Location: 4, Different res: 2, All words: 2, Distance: 5, Personalization: 5)
18. steve: 18 (Location: 2, Different res: 3, All words: 3, Distance: 10, Personalization: 0)
19. special: 18 (Location: 4, Different res: 2, All words: 2, Distance: 10, Personalization: 0)
20. tree: 18 (Location: 3, Different res: 2, All words: 3, Distance: 3, Personalization: 7)
```

Figure 17 List of 20 suggested words after personalization

6 Evaluation

It is not easy to evaluate and experimentally verify methods like the proposed one. Frequently, only some properties of the method can be subjected to experimental evaluation and the evaluation often is not a comprehensive one. But it is always difficult to evaluate some method that involves humans, in particular human computer interaction. One of the usual approaches is to engage a number, not just one, of interested fellows who are able and willing to participate in experiments. The aim is to solicit their opinions. In the case of our method, it is possible to conjecture, quite safely, that the method suggests in some way words that are determined by the interested fellow's profile. The hypothesis is, however, that (at least some of) the suggested words are indeed relevant to the query viewed as an imperfect expression of interested fellow's intended informational interest. Deciding about relevancy is best done by the interested fellow himself/herself.

We devised a simple experiment that includes an interested fellow. The intention was to involve a greater number of interested fellows willing to participate in experiments. A simple questionnaire was devised where the interested fellow who participates in the experiment indicates relevancy of the suggested words.

Experiments were arranged as groups of three queries in each of the following six categories:

- **general** (euro, Beatles, weather)
- **computers and technologies** (mail, computer, printer)
- **games** (counter-strike, need for speed, poker)
- **specific** (personalized web search, sorting algorithm, design patterns)
- **sport** (soccer, hockey, cycling)
- **Internet** (facebook, youtube, google)

For each query, our method suggested 10 words. Interested fellows then indicated in the questionnaire if the word is relevant and augmented the meaning of the particular query.

There were 71 interested fellows involved. We could not rule out the possibility that some participants filled in the questionnaire more or less by random ticking. In order to deal with this possibility, we introduced three thresholds. Each suggested word received certain number of ticks from interested fellows. For each query the total number of interested fellows who evaluated it is known. This is also a maximum value that a suggested word can receive as a suggestee for a given query. Let us assume there were x_q interested fellows submitting a query q and evaluating the words $w_{q,1}$, $w_{q,2}$,...,$w_{q,10}$, suggested by the method for relevancy. Let us further assume $y_{wq,i}$ interested fellows indicated, that the word $w_{q,i}$ is relevant to query q. This number is at most $x_{q.}$ Fraction $y_{wq,i}/x_q$ gives the percentage of those interested fellows who indicated $w_{q,i}$ as relevant to q from all interested fellows who considered q. To minimize effects of possible random ticking, we introduced a “trustworthiness” threshold p in the following sense: considering threshold p, only words $w_{q,i}$ for which the fraction $y_{wq,i}/x_q$ is at least p are accepted as having been indicated as relevant by the interested fellows.

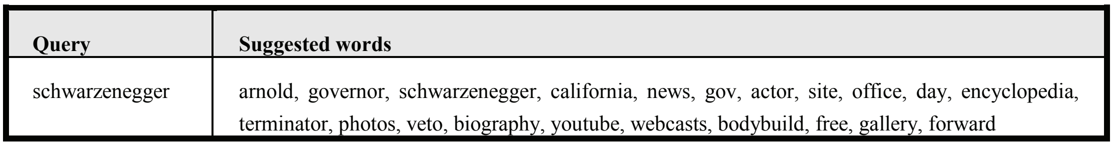
We evaluated the experiments considering three thresholds: 20%, 30% a 40%. So for example, if a query was submitted by 27 interested fellows then for a suggested word its indication to be accepted as relevant required at least 5 indications for the 20% threshold, at least 8 indications for the 30% threshold, at least 11 indications for the 40% threshold. Overall results across all queries and all suggested words expressed as average values of relevancy as indicated by the interested fellows, are as follows:

- threshold 20%: **73,33 %** (7 out of 10 suggested words are relevant)
- threshold 30%: **62,22 %** (6 out of 10 suggested words are relevant)
- threshold 40%: **49,44 %** (5 out of 10 suggested words are relevant)

It should be noted that the average suggestion success rate decreases as the threshold increases. In the case of 20% threshold there are 7 out of 10 words that were suggested by our method and these words are indeed relevant as indicated by interested fellows in our experiments. Considering some 300-400 words extracted from all the results returned by search engines, the rate could be viewed as quite satisfactory. The rate means that among those 10 words suggested by our method, which selected them from some 300-400 words, included are 73,33 % relevant ones.

One of the most important performance indicators for the kind of service that we envisage is response time. Acceptable limit is determined mostly by physiological and psychological factors, which squeeze the response time to one or two seconds at most. Anything above it would be considered by the interested fellow as too slow, detracting the attention. Part of the response time is the time needed for processing. To have some clue on viability of our approach, we performed experiments with our prototype IPA. Average processing time measured for 10 suggestions was 1.89 seconds. This value may vary depending on server, data sources or word processing techniques. Nevertheless values under 2 seconds are still very promising especially when we realize that it was achieved by a prototypical implementation whose primary purpose has not been performance efficiency. One of the possible sources of speed improvement is parallel access to data sources.

As a comparison to sample results in figure 1, figure 18 shows suggestions by our method.



Query	Suggested words
schwarzenegger	arnold, governor, schwarzenegger, california, news, gov, actor, site, office, day, encyclopedia, terminator, photos, veto, biography, youtube, webcasts, bodybuild, free, gallery, forward

Figure 18 Suggestions for query “schwarzenegger” from presented approach

7 Conclusions and Future Work

In this paper we proposed a method for automatic suggestion of words to improve query formulation. The method relies on results of initial searches by three known search engines.

Search engines like Google, Yahoo or Ask.com are quite good at suggesting words related to a given query for well known and commonly used queries. As the query becomes more specific the suggesting deteriorates or fails. Our method attempts to remedy this weakness of current suggesting solutions. Our method is simple and does not attempt to replace the current approaches. Our method brings noticeable improvement, but at a cost. Whereas the known suggesting search engines are able to make suggestions instantly, our method needs some extra time to allow retrieving search results, analyzing them and producing a list of suggestions. Currently, we are able to suggest with less than 2 seconds of response time, which can be considered just within limits of acceptance. This delay is a price for greatly increased completeness of suggestion operation. Our method produces some suggestion for any query, the only limitation being queries for which search engines are not able to return any documents. Dealing with this particular case would require a different approach.

Our approach differs from usual approaches also in another way. The above-mentioned known approaches tend to suggest words, that when appended to the original query make it more complete (e.g., completing full name of a car, full title of a movie). Our method often succeeds in suggesting words that refine the meaning of the original query. Let us consider for example an original query “operation”. Ordinary search engine suggests “Operation Flashpoint”. Our method is able to suggest, among others, also words that help distinguish various alternative meanings of the word, e.g. “military“, “financial“, “surgical“. This could be viewed as a shift towards more sophisticated suggestions.

Our approach has been intentionally geared towards simple solutions. In particular, our current level of personalization is obviously almost trivial comparing with what has already been achieved in that area. We do not employ any persistent interested fellow’s model that could be gradually improved. However, even a very simple idea can yield a noticeable improvement in query suggestion, which has been our research goal.

Considerable work still remains to be carried out in this area in general, and specifically, in improving the approach as proposed by us. In particular, it would be useful to tackle the following issues:

- Queries in other languages (than English). Actually, the method works also with queries in other languages (in particular, we focus on Slovak language). However, its effectiveness deteriorates. The method relies on the usual techniques of deleting stop words and stemming, which have been designed to work primarily for English documents. Moreover, the volume of documents in English is greater than in any other language by orders of magnitude, so there is less to search in and retrieve from.
- Reordering of results**.** The method produces a personalized list of ordered words that reflects interested fellow’s interests. The words could be used to reorder the results as returned by the search engine.
- Recognition of named entities. Currently, any word is treated as an independent entity. However, this is not appropriate when an entity occurs named by two or more words (e.g., San Marino or São Tomé and Príncipe or Saint Vincent and the Grenadines). There are known techniques to recognize named entities that could be employed so that such entities would remain intact.
- Making use of other data sources. It is possible to open access to any other source that is capable of providing results with a certain

structure that allows inferring relative importance of the words.

- Refining word characteristics. Currently, we use 4 characteristics and a set of simple evaluation rules. All this is heuristic in nature. Other characteristics could be considered and rules could be fine tuned to increase relevancy of the set of suggested words.

Strohmayer (2009) proposed to refocus search engines from letting interested fellows guess arbitrary words from the set of documents they seek to retrieve to encourage the interested fellows to express their original search intent in a more unambiguous and natural way.

Acknowledgements

This work was supported by the Scientific Grant Agency of Slovakia, grant No. VG1/0508/09, by the Agency for Supporting Research and Development supporting the project Tradice, grant No. APVV -0208-10 and it is a partial result of the Research & Development Operational Program for the project PerConIK, ITMS code: 26240220039, co-funded by ERDF.

References

Baeza-Yates, R., Calderón-Benavides, L., and González-Caro, C. (2006), "The Intention Behind Web Queries", in *String Processing and Information Retrieval, Lecture Notes in Computer Science*, (F. Crestani, P. Ferragina and M. Sanderson eds.), Vol. 4209: 98-109.

Barla, M. (2011), "Towards Social-based User Modeling and Personalization", *Information Sciences and Technologies Bulletin of the ACM Slovakia*, **3** (1): 52-60.

Broccolo, D., Frieder, O., Nardini, F.M., Perego, R., and Silvestri, F. (2010), "Incremental Algorithms for Effective and Efficient Query Recommendation", *Lecture Notes in Computer Science*, *String Processing and Information Retrieval* **6393**: 13-24.

Chen, Q., Li, M. and Zhou, M. (2007), "Improving Query Spelling Correction Using Web Search Results", in *Proceedings of EMNLP-CoNLL 2007*: 181-189.

Cho, J., Lee, U. and Liu, Z. (2005), "Automatic Identification of User Goals in Web Search", in *International World Wide Web Conference*, ISBN:1-59593-046-9: 391-400.

Dreher H. and Williams, R. (2006), "Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering", *Lecture Notes in Computer Science*, *Flexible Query Answering Systems,* **4027**: 282-294.

Jiang, S., Zilles, S. and Holte, R. (2009), "Query Suggestion by Query Search: A New Approach to User Support in Web Search", in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Vol. 01* (WI-IAT '09), Vol. 1. IEEE Computer Society, Washington, DC, USA: 679-684.

Kalmanovich, I.G. and Kurland, O. (2009), "Cluster-based query expansion", in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*, ACM, New York, NY, USA: 646-647.

Levinson, D. and Rose, D. E. (2004), "Understanding User Goals in Web Search", *International World Wide Web Conference,* ISBN: 1-58113-844-X: 13-19.

Li, L., Xu, G., Zhang, Y. and Kitsuregawa, M. (2009), "Enhancing Web Search by Aggregating Results of Related Web Queries", *Web Information Systems Engineering - WISE 2009, LNCS,* (G. Vossen, D. D.E. Long and J. Xu Yu, eds.), **5802**: 203-217.

Liu, F., Meng, W. and Yu, C. (2004), "Personalized Web Search For Improving Retrieval Effectiveness", *IEEE Transactions on Knowledge and Data Engineering* ISSN: 1041-4347: 28-40.

Liu Sheng, O.R., Ma, Z. and Pant, G. (2007), "Interest-Based Personalized Search", *ACM Transactions on Information Systems (TOIS)*, ISSN: 1046-8188, **25** (1), Article 5.

Liu, Z., Natarajan, S. and Chen, Y. (2011), "Query expansion based on clustered results", *Proc. VLDB Endow.*, **4** (6): 350-361.

Ma, Y., Lin, H. and Jin, S. (2010), "A Revised SimRank Approach for Query Expansion", *Lecture Notes in Computer Science, Information Retrieval Technology*, **6458**: 564-575.

Martinsky, L. (2010), "Improving Query Suggestion Capabilities Using Web Search Results", in *Proc. Informatics and Information Technologies Student Research Conference IIT.SRC 2010,* Slovak University of Technology, Bratislava: 3–5.

Mitra, M., Singhal, A. and Buckley, C. (1998), "Improving automatic query expansion", in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*, ACM, New York, NY, USA: 206-214.

Sahami, M. and Heilman, T.D. (2006), "A web-based kernel function for measuring the similarity of short text snippets", in *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, ACM, New York, NY, USA: 377-386.

Song, Y. and He, L. (2010), "Optimal rare query suggestion with implicit user feedback", in *Proceedings of the 19th international conference on World wide web* (WWW '10), ACM, New York, NY, USA: 901-910.

Stamou, S., Kozanidis, L., Tzekou, P. and Zotos, N. (2009), "Ontology-Driven Personalized Query Refinement", *Journal of Web Engineering*, **8** (2): 113-153.

Strohmaier, M., Kroll, M. and Korner, C. (2009), "Intentional query suggestion: making user goals more explicit during search", in *Proceedings of the 2009 workshop on Web Search Click Data* (WSCD '09), ACM, New York, NY, USA: 68-74.

Teevan, J., Dumais, S.T. and Liebling, D.J. (2008), "To personalize or not to personalize: modeling queries with variation in user intent", in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '08), ACM, New York, NY, USA: 163-170.

Tvarozek, M. (2011), "Exploratory Search in the Adaptive Social Semantic Web", *Information Sciences and Technologies Bulletin of the ACM Slovakia*, **3** (1): 42-51.

Vanekova, V. (2010), "Preferential Querying for the Semantic Web", *Information Sciences and Technologies Bulletin of the ACM Slovakia*, **2** (2): 137-148.

Wang, G.T., Xie, F., Tsunoda, F., Maezawa, H. and Onoma, A.K. (2002), "Web Search with Personalization and Knowledge", in *Proceedings of the Fourth IEEE International Symposium on Multimedia Software Engineering (MSE '02)*, IEEE Computer Society, Washington, DC, USA: 90-97.

Xu, J. and Croft, W.B. (1996), "Query expansion using local and global document analysis", in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96)*, ACM, New York, NY, USA: 4-11.

Yang, J., Cai, R., Jing, F., Wang, S., Zhang, L. and Ma, W. (2008), "Search-based query suggestion", in *Proceeding of the 17th ACM conference on Information and knowledge management* (CIKM '08), ACM, New York, NY, USA: 1439-1440.