Computing and Informatics, Vol. 38, 2019, 1341–1374, doi: 10.31577/cai_2019_5_1341

AUTOMATIC QUERY REFINING BASED ON EYE-TRACKING FEEDBACK

Alena Martonova, Jozef Marcin, Pavol Navrat
Jozef Tvarozek, Gabriela Grmanova

Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
e-mail: alena.martonova@stuba.sk

Abstract. This paper presents a new method named AQueReBET, which automatically refines a query set by an information seeker searching on the web. A revelation of the intention of an information seeker who is running a search can bring a significant improvement to the search process and to browsing as well. It is practically impossible to acquire such intention by the explicit indication (feedback) due to the fact that web browsing takes place in real time. Therefore the intention must be determined in some other way. We hypothesize that it can be approximated by means of the implicit feedback preferably in the form of data from an eye tracker and mouse. We propose a method which automatically refines a seeker's search query, and thus we can offer documents with higher relevance, decrease the number of query reformulations and increase the seeker's satisfaction. The query refinement is based on an analysis of gaze data from an eye tracker and on groupization. In the proposed method, we calculate word-level importance based on term frequency, term uniqueness (tf-idf) and total fixation duration within the subdocument (word's snippet in search results).

Keywords: Web search, query refinement, eye-tracking, groupization, implicit feedback

Mathematics Subject Classification 2010: 68-U35, 68-M11

1 INTRODUCTION

Nowadays, most information seekers rely on the web when searching for documents to find the desired information. However, typical web search engines expect text queries, which are rather imperfect ways of expressing the intention of an information seeker. Prompting the information seeker for additional information beyond the query itself could be seen as asking too much. Therefore, one of the current research challenges is to derive as much useful information as possible from so called implicit feedback, which can be collected unobtrusively [17] – based on an information seeker's interaction with the search engine. There are various kinds of implicitly provided data that information seekers generate while interacting, such as their choice of a particular snippet to read in more detail or even a track of their gaze. In this paper, we investigate how an information seeker's eye gaze data acquired from an eye-tracking device can be used for refining the seeker's query during his/her searching session. An eye tracker provides potentially a very comprehensive immediate feedback unobtrusively, without any explicit questions that often tend to be perceived as annoying and also without demanding the seekers to perform searching in some specific (artificial) way. Experimental results [15] show that eye tracking is a valuable real-time implicit source of information about what the user is searching for and that it can be used for real-time user interface adaptation.

The initial query is often reformulated during a typical web search. Approximately one third of all formulated queries are composed by gradually reformulating an initial query [31, 4]. The relevance of the initially received documents could be quite low, implying the necessity of explicit query reformulation by the information seekers themselves.

The aim of our research is to improve search outcomes by reducing the need for explicit query reformulation and increasing relevancy of the offered documents. We attempt to achieve this objective by acquiring and utilizing data obtained from an eye-tracker and using it to implicitly reformulate the query. Furthermore, we propose to complement this by so-called groupization (i.e. the data are used also from other previous users with the same or very similar search intention) that can provide useful additional information to interpret an information seeker's intention.

Our approach is based on the assumption that if we recognize the web seeker's intention, we should be able to offer more relevant documents in response to the initial query. Since we are not able to detect exactly the line the seeker is reading by eye tracking, we focused on larger areas such as search result snippets (snippet is a web-page element, which contains a short text representing one of the results in search engine result page). We detect the snippet on which the seeker fixates his/her gaze within a search engine's results page (hereinafter SERP). We extracted specific words out of these snippets, which can possibly give us a strong clue on how to refine the initial search query to get more relevant documents to his/her search intention.

The rest of the paper is structured as follows. In the next section, we present important related results as published in the current literature. Section 3 presents our

proposed method for refining an information seeker's query based on the analysis of his/her gaze movements and enhanced also with the groupization method. Experiments are presented and discussed in Section 4. The paper concludes in Section 5, where suggestions of future work are briefly discussed.

2 RELATED WORK

Query refinement [7, 5] and groupization methods play a key role in our work. Let us take a look at some current and related approaches in this area. Many existing works are focused on how the information seekers work with the SERP and how they select the relevant documents. It was already established that seekers are not very keen to provide explicit feedback [27], e.g. in the form of some kind of a relevance feedback mechanism or explicit query reformulation. Instead, various forms of implicit feedback are to be preferred. One direction of research is to use mouse-clicking data as implicit feedback to reformulate a query [13]. The gaze of seekers who reformulate a query was studied by Eickhoff et al. [8] and Umemoto et al. [32]. Li et al. [19] tracked the gaze to acquire relevance of retrieved images to be used for query expansion in an image retrieval system. However, we focus on text documents in the present research, but note that studies have emerged recently that broaden the applicability of eye gaze analysis and enhance the analysis itself by incorporating also e.g. pupil dilatation [21]. A seeker's gaze as a source of implicit feedback for information retrieval is a topic of research [6, 29]. Granka et al. [10] explored the behaviour of users during web searches and established the minimum fixation duration for web searches, which represents the minimum time the seeker is looking at relevant information. It was experimentally defined as 200 to 300 milliseconds. Various methods how to estimate an information seeker's search intention based on eye-tracking data were explored [31]. They represent search intention in a table created from a set of pairs – term and its weight. Weights of terms (TermScore) were computed using various functions combining quantities such as the number of times that a seeker looked at a term and the term frequency. Others investigated whether word relevance to a seeker's current intent could be inferred from the text and his/her eye movements [20]. It has already been shown that gaze-based feedback can be used to expand a query [5]. A particular simple data acquired from eye tracking, i.e. attention time, was used to recommend new online items [34].

Research led by Buscher [5, 6] is the most similar to ours, both considering the method and the way it is evaluated. Our method differs mainly by the choice of tf-idf (see below) as the base formula that is gradually expanded. On the other hand, our approach was to make additional emphasis on term frequency combined with total fixation duration on the snippet (Equation (3)).

White et al. [30] addressed methods of groupization. They were looking for seekers with similar features to obtain more relevant links to their queries. Current trends in devising groups are as follows:

- similarity based on link clicks: determined by three ways:
  1. match the URL clicked on,
  2. match the domain of the URL clicked on, and
  3. consistency between the categories of topics of the URL clicked on;
- syntactic similarity – similarity of the querying (keywords);
- semantic similarity – even if the queries are not similar based on syntactic similarity, they can be similar in their meanings.

Another area of active research in which researchers are engaged, is obtaining text via eye tracking. Methods for extracting text have been explored by Biedert et al. [3]. They are working on the interesting specific problem of adjusting eye tracking errors when reading a structured text to automatically position the cursor at a proper place. We decided not to follow this line of research, since our aim is different.

Eye fixation coordinates are correlated with mouse cursor positions thus facilitating considerations of various behavioural patterns – reading, hesitation, scrolling, clicking [11, 26]. Therefore, it may very well be possible that some of the results obtained by a method employing eye gaze tracking, and particularly by the proposed method, could be achieved or at least approximated by using mouse cursor data, which is more accessible for commercial search engines. On the other hand, one of their conclusions claiming that the cursor approximates the gaze is misguided. Other than that of the mouse cursor, eye tracker provides data on the actual viewed location. In our research, we attempt to make use of this additional data to achieve new insights in the automatic query refinement.

3 PROPOSED METHOD FOR REFINING

Our method – automatic query refining based on eye-tracking feedback (AQueReBET) – primarily deals with word table creation (a similar data structure is sometimes referred to as an "intention vector" [25], which however could be misleading, since a vector is conventionally considered to be a one-dimensional structure). The table is created by selecting appropriate words from useful elements of pages (mostly snippet areas, see Figure 2) scanned by the gaze during a seeker's web search. Such a table can facilitate refining the initial query by better (i.e., more relevant) words and thus offer the seeker results, which better fit the seeker's intention. Our hypothesis is: Some of the information obtained by tracking a seeker's gaze, can, after suitable processing, be helpful in providing search results that better reflect their intention. Unlike the methods based on mouse tracking, our method does not need to wait until the seeker moves the mouse cursor within the open page, the seeker clicks within the open page, the seeker opens any other page, or the seeker chooses some additional words to pose a new query.

While devising such a method, three assumptions emerged:

1. The longer the user's fixation within the snippet area, the higher the probability that it is closer to his/her intention than other snippets (the strongest criterion). Here, we have been inspired by Maglio et al. [22] who use eye-gaze information to help disambiguate user interests.
2. The higher the term frequency in watched snippets, the higher the probability that the term is closer to the user's intention. Here we have been inspired by Umemoto et al. [31] who proposed several formulas which include term frequency and their multiplication or division and also various weighting.
3. The more unique the terms, the better distinguishing ability can be obtained between similar or major intentions. Here we have been inspired by [5, 6, 8] who proposed formulas that include term frequency and inverse document frequency.

We found these criteria by experimenting with Equation (3) using data from our preliminary experiments. A more detailed description of them can be found in a later subsection of this paper. The overview of the method is shown in Figure 1.

3.1 AQueReBET Method

Our method is fully automated and thus requires data only from the seeker's gaze within the first SERP and from the seeker's input query. To clearly interpret the process we need to represent the data in a usable form. We decided to use a table τ_x to represent text x (x could be a query, a snippet, or a web page) by a set of pairs consisting of a word w and its importance $im_x(w)$ (we start with word importance within the text x but in further steps we will recalculate it to a wider context):

$$\tau_x = \{(w_i, im_x(w_i) \mid i \in 1, 2, \ldots, n\} = \begin{bmatrix} w_1 & w_2 & \ldots & w_n \\ im_x(w_1) & im_x(w_2) & \ldots & im_x(w_n) \end{bmatrix} \quad (1)$$

where $n = |\tau_x|$ is the number of different words chosen from text x, w_i is the i^{th} word, and $im_x(w_i)$ is the importance of $w_i \mid i \in \{1, 2, \ldots, n\}$ in text x. Initially (before processing the data) all unique words are chosen from text x and the importance of each word is equal to its multiplicity (term frequency) in text x: $im_x(w_i) = tf_x(w_i) \mid i \in \{1, 2, \ldots, n\}$.

The method consists of 6 steps shown in Figure 1. In Step 1, keywords from the initial web search query are selected and the initial table τ_{in} is created:

$$\tau_{in} = \begin{bmatrix} w_1 & w_2 & \ldots & w_n \\ 1 & 1 & \ldots & 1 \end{bmatrix}. \quad (2)$$

For example, if the query is "monk" (e.g. with the aim to find more information about monks in a monastery and their ascetic lives), then

$$\tau_{in} = \{(\text{monk}, 1.0)\} = \begin{bmatrix} \text{monk} \\ 1 \end{bmatrix}.$$

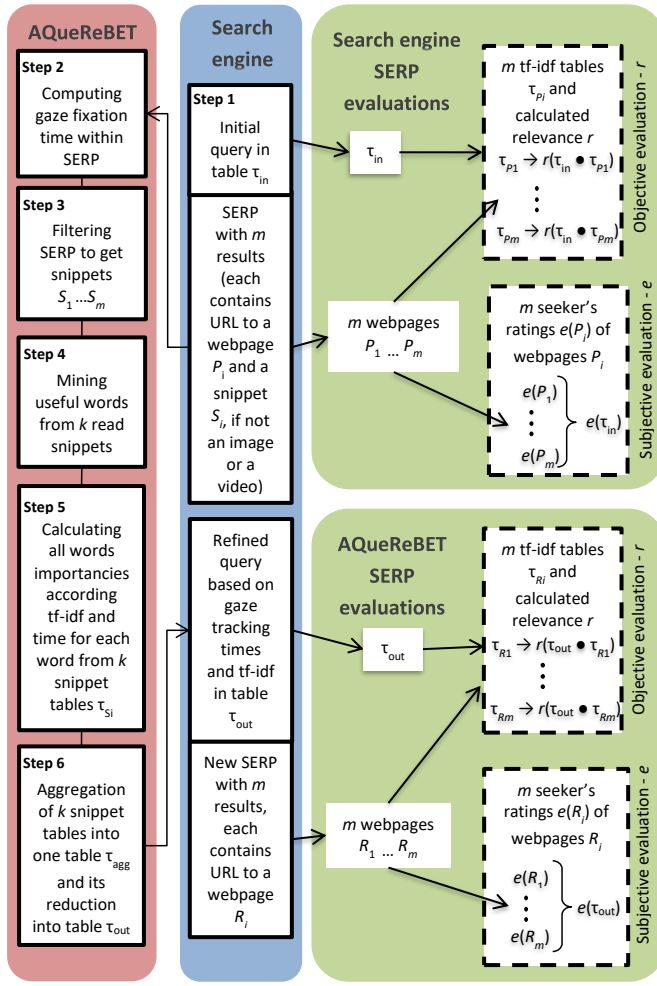


Figure 1. Process of initial query refinement (with new snippets suggested) within AQueReBET and system of its objective and subjective evaluation both search results – with and without AQueReBET

Step 2: *Web seeker's gaze acquisition.* Web search is an interactive experience, typically implemented using dynamic web technologies (using Ajax, DOM rewriting), which is very hard to track accurately using off-the-shelf eye tracking analysis software. To obtain a seeker's gaze in a dynamic web environment, we employ the data collection infrastructure [24] developed at our University User eXperience Research Centre. Raw data from the eye tracker is processed into normalized coordinates of eye position and analysed based on the underlying web page (see Step 2

in Figure 1). The data is subsequently sent to a browser plugin and enriched with XPath data.

Step 3: *Filtering out unnecessary data from SERP to get snippets.* Our data are from SERP within the google.com domain and subdomains (an example of SERP for "monk" is shown in Figure 2).

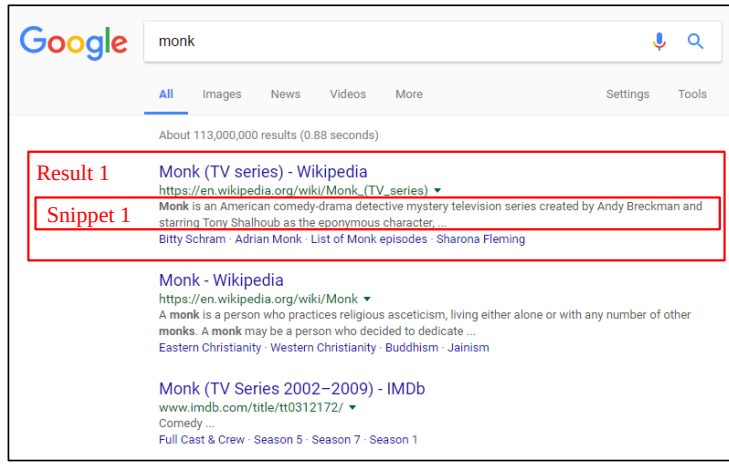


Figure 2. Google's SERP for query "monk", only first three results (snippets) shown

Since we are interested only in snippets and these pages contain also other data, we need to find these snippets within SERP, and to record only snippet areas. From the implementation point of view, the most important part of this process is to filter out the data that do not carry snippets. We used HTML Document Object Model elements (DOM elements), which enabled us to manipulate with HTML elements in DOM. By analysing DOM elements in Google's SERP the page is divided into several sections that contain or do not contain keywords. In this regard it is appropriate to filter out both too large DOM elements, which involve a number of snippets and too small elements that do not carry any relevant information. As a result of this filtration we obtained only the data from all the SERP snippets $S_1, \ldots, S_m$ (see Step 3 in Figure 1, these m snippets are linked to web pages $P_1, \ldots, P_m$). The individual data records we get from the first step contain XPath information, from which we could obtain the element and its value (the snippet text). The seeker's gaze data (the snippets texts) are processed in the following steps in order to obtain additional words on the seeker's intention (also called intent or interest), further refining the initial seeker's query.

Step 4: *Mining the relevant words from relevant snippets and creating their tables.* After that, we filter out from all m snippets only the k relevant ones, $k \leq m$ (see Step 4 in Figure 1) – relevant are those which are gaze inspected, i.e. have at least one seeker's fixation (in eye tracker fixation I-VT filter we set the velocity

threshold to 30 degrees/second; this gave us the seekers' minimum fixation time in the interval 200–500 ms). Then we filter out all irrelevant words from all gaze-inspected snippets $S_1, \ldots, S_k$ – we remove all stop words using a modified Wordnet dictionary and perform lemmatization of words (nouns and adjectives) to receive a base form (lemma) of the corresponding word. This bears much of the meaning of the word, contrary to stemming. The importance of irrelevant words is set to 0 and are left out of table representation. An example of the gaze inspected snippet S in SERP is e.g. the second snippet: *"A monk is a person who practices religious asceticism, living either alone or with any number of other monks. A monk may be a person who dedicate ..."*. Then after performing the above mentioned operations, the set of relevant snippet words is table τ_S. The importance of each word w_i is equal to its multiplicity (term frequency) $tf_S(w_i)$ within the processed snipped S.

$$\tau_S = \begin{bmatrix} \text{monk} & \text{person} & \text{religious} & \text{asceticism} & \text{living} & \text{number} & \text{monks} \\ 2 & 2 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Step 5: *Performing tf-idf analysis of the snippets.* Having a set of snippets, we need to find words which are specific (unique) in them. If a seeker looks at a particular snippet, the specific word that it contains likely contributes to the expression of the seeker's intention. Those specific words are obtained by applying the td-idf formula [14]. Therefore, in this step we perform tf-idf for all the relevant snippet tables $S_1, \ldots, S_k$ from the previous step (see Step 5 in Figure 1). Its result is tf-idf value for each word in each snippet set: $tfidf_S(w_i)$. This will help us to recognize the relevance of words from gaze inspected snippets $S_1, \ldots, S_k$ compared to all snippets in SERP $S_1, \ldots, S_m$. But the importance of the words should not be determined only by their multiplicity $tf_S(w_i)$ as the main criterion and tf-idf value $tfidf_S(w_i)$ as the second criterion, but also by the relative time the seeker spent on the snippets with their gaze t_S (relative to the averaged time spent on one snippet within SERP). Finally, the importance of each word from table τ_S is given by a combination of three contributing factors: the user's total fixation time within the snippet area (see Figure 2), term frequency in watched snippets and the level of term uniqueness, and calculated by Equation (3).

$$im_S(w_i) = tf_S(w_i) * (tfidf_S(w_i) + t_S) \tag{3}$$

where

$$tfidf_S(w_i) = tf_S(w_i) * idf_S(w_i) = tf_S(w_i) * \log \frac{|\{S_j \in \text{SERP}\}|}{|\{S_j \in \text{SERP} : w_j \in \tau_{S_j}\}|}.$$

Let us suppose, in our example, the dwelling time was 1.2 seconds, averaged dwelling time for all snippets in SERP was 1 second, thus $t_S = 1.2\,\text{s}/1\,\text{s} = 1.2$. Then our

example of table τ_S has changed importance values as follows:

$$\tau_S = \begin{bmatrix} \text{monk} & \text{person} & \text{religious} & \text{asceticism} & \text{living} & \text{number} & \text{monks} \\ 2*(0.2+1.2) & 2*(0.22+1.2) & 1*(0.11+1.2) & 1*(0.02+1.2) & 1*(0.06+1.2) & 1*(0.03+1.2) & 1*(0.3+1.2) \end{bmatrix}.$$

And the resulting values of word importance for the snippet S are:

$$\tau_S = \begin{bmatrix} \text{monk} & \text{person} & \text{religious} & \text{asceticism} & \text{living} & \text{number} & \text{monks} \\ 2.8 & 2.84 & 1.31 & 1.22 & 1.26 & 1.23 & 1.5 \end{bmatrix}.$$

Our example demonstrates the calculation of word importance only for one snippet, but the seeker can gaze on more of them: $S_1, \ldots, S_k$. Therefore, the final step is to aggregate tables $\tau_{S_1}, \ldots, \tau_{S_k}$ into table τ_{agg}, which contains the union of all words from tables $\tau_{S_1}, \ldots, \tau_{S_k}$ and their importance:

$$im_{agg}(w_i) = \sum_{j=1}^{k} im_{S_j}(w_i). \tag{4}$$

If a word w_i is not in table τ_{S_j} then $im_{S_j}(w_i) = 0$.

Step 6: *New query definition.* The next step is SERP enrichment with snippets (and corresponding links to pages) that are not present in the original SERP and contain more relevant pages/documents/information sources. Based on the final table τ_{agg} (see Step 6 in Figure 1), we start a new search (as a background process). Since table τ_{agg} can consist of many words, only the most important ones have to be chosen (too many words in a query to a search engine are counterproductive). For example, the new query would be set to the four most important words from table τ_{agg}, resulting to τ_{out}:

$$\tau_{out} = \begin{bmatrix} \text{person} & \text{monk} & \text{monks} & \text{religious} \\ 2.84 & 2.8 & 1.5 & 1.31 \end{bmatrix}.$$

Although the table τ_{out} is reduced to four pairs, it is still a refinement of the initial query represented by table τ_{in}, since it is often shorter. We determined the number of pairs empirically simply by using different queries and four worked out the best. By performing a new search using this new query, we get new SERP with new snippets with links to web pages $R_1, \ldots, R_m$.

3.2 Groupization Method

In general, a groupization is one of the several methods used to improve a personalized web search [16, 28] defined as *"combining an individual's data with that of other related people to enhance the performance of personalized search"*.

In our work the groupization is used in a very specific way. Its purpose is to enhance our AQueReBET method in its last step (see Step 6 in Figure 1). It

is performed by using various methods, but still includes the index of similarity. Groups of seekers are formed by two techniques that are based on:

- syntactic similarities with the initial query from the table τ_{in}.
- syntactic similarities with a refined query from the table τ_{out}.

The use of groupization is particularly important due to the fact that a seeker can search for a completely different thing than the one that was written in the initial query. Because of this fact, refinement of a seeker's query using only information elicited from eye tracking data may in some cases have only a limited effect. By also involving the groupization we attempt to remedy such situations. To involve groupization essentially amounts to involving some data from previous experience. In the context of our method, we use groupization to provide results that other seekers adopted as relevant for their query before or during a search session.

We identified a need for implementation of two groupization types, where groups are formed based on the initial queries and the refined query.

1. Initial query from table τ_{in}: After calculation of relevance (Equation (5)) for each page P_i,
2. Refined query from table τ_{out}: After creating a new query defined by table τ_{out} and calculation of relevance (Equation (6)) for each page R_i.

The groupization module in AQueReBET writes one or several of the most relevant pages/documents into a database for the initial or the refined query. This creates a group with a certain query and pages/documents that our system evaluated as relevant for it. We insert only distinct records into the database. If some group already exists in the database, we only update the set of relevant pages/documents. In subsequent search sessions, we attempt to assign a querying seeker to a group that is related to his/her initial query and acquired words by semantic similarity. In return, he/she is provided with the most relevant documents from the group that have been gathered in the previous sessions. If none of the existing groups is relevant enough (i.e. the query is not similar to any group or the level of similarity is too low), a new group is created. Hereafter AQueReBET with a turned on groupization module is AQueReBET + G and with a turned off groupization module is AQueReBET.

4 EXPERIMENTS

To perform experiments evaluating our proposed method requires a specific approach, one of the reasons being that we have not found any other similar systems published in a way that would allow an effective comparison. Experiments that we completed so far deal with the evaluation of our method that refines seekers' queries. We conducted two types of evaluations: Automatic evaluation and evaluation by seekers.

4.1 Automatic Objective Evaluation (Relevance Evaluator)

Automatic evaluation of web pages addressed by snippets without involving seeker's gaze. It is quite possible that a snippet S_i (result in SERP) may not accurately reflect the content of its destination page P_i. It is appropriate to analyse each destination page separately. Since we already have the initial seeker's query (table τ_{in}), we can analyze the relevance of individual pages addressed by a snippet from SERP in the context of this query (see Figure 1, objective evaluations r in Results without AQueReBET). The process of evaluation begins very similarly to the one with snippets. We perform tf-idf on the content of each destination page P and get its table τ_P (the time and aggregation is skipped). To compute relevancy r of each page P_i (examples in Table 1) we use the initial table τ_{in} as follows:

$$r(\tau_{in} \cdot \tau_{P_i}) = \sum_{\forall w \in \tau_{in} \cdot \tau_{P_i}} im_{P_i}(w) \tag{5}$$

where $\tau_{in} \cdot \tau_{P_i}$ represents a set of words, which belong to both τ_{in} and τ_{P_i} and $im_{P_i}(w)$ is the importance of word w from tf-idf analysis on page P_i.

Original documents P_i	norm. rel.
`https://en.wikipedia.org/wiki/Monk_(TV_series)`	0.03
`http://en.wikipedia.org/wiki/Monk`	1
`http://www.imdb.com/title/tt0312172/`	0
`http://www.tv.com/shows/monk/`	0
`http://www2.usanetwork.com/series/monk/`	0
`http://www.newadvent.org/cathen/10487b.htm`	0.48
`https://www.facebook.com/monk`	0
`http://us.battle.net/d3/en/class/monk/`	0
`http://www.battle.net/wow/game/class/monk`	0
Suggested documents R_i	norm. rel.
`http://en.wikipedia.org/wiki/Asceticism`	0.24
`http://en.wikipedia.org/wiki/Monk`	0.24
`http://dictionary.reference.com/browse/ascetic`	0.29
`http://www.vocabulary.com/dictionary/ascetic`	1
`http://stgeorgegreenville.org/OurFaith/Articles/Asceticism-Rossi.html`	0.06
`http://www.huffingtonpost.com/...`(very long URL)	0.06
`http://www.monasteryofstjohn.org/...`(very long URL)	0
`http://orthodox.cn/patristics/300sayings_en.htm`	0.02
`http://www.cgg.org/...`(very long URL)	0.17
`https://books.google.sk/...`(very long URL)	0.25

Table 1. Example URLs of web pages P_i and R_i with their normalised relevance (Equation (7)). Normalized relevance computation does not involve seeker's gaze.

The second document P_2 has the highest normalised relevancy because it is the closest to the seeker's intention from among the original set of documents P_i. There is another document with a quite high normalised relevance of 0.48 and indeed, it at least partially fits the seeker's intention, too. The remaining documents deal with entirely different things (e.g., a TV series, a computer game).

Automatic evaluation of new web pages addressed by new snippets. To measure relevancy of m new pages/documents R_i (examples in Table 1), we perform a new tf-idf analysis of these pages in the context of the new query from the table τ_{out} (see Figure 1, objective evaluations r in Results from AQueReBET). For computing the table τ_{R_i} we determine the tf-idf importance of individual words on page R_i. We use the same process as for pages P_i: We download the content of pages $R_1, \ldots, R_m$, then calculate the tf-idf analysis of these pages to get tables $\tau_{R_1}, \ldots, \tau_{R_m}$, compute $\tau_{out} \cdot \tau_{R_i}$ and finally we sum up the importance of queried words for each page R_i to get its relevance $r(\tau_{out} \cdot \tau_{R_i})$:

$$r(\tau_{out} \cdot \tau_{R_i}) = \sum_{\forall w \in \tau_{out} \cdot \tau_{R_i}} im_{R_i}(w). \tag{6}$$

Normalisation. For each relevance value of page P_i (alternatively page R_i from a new SERP) we also perform normalisation into interval $[0, 1]$ using feature scaling:

$$||r(\tau_{in} \cdot \tau_{P_i})|| = \frac{r(\tau_{in} \cdot \tau_{P_i}) - \min_{j \in \{1,\ldots,m\}} r(\tau_{in} \cdot \tau_{P_j})}{\max_{j \in \{1,\ldots,m\}} r(\tau_{in} \cdot \tau_{P_j}) - \min_{j \in \{1,\ldots,m\}} r(\tau_{in} \cdot \tau_{P_j})}. \tag{7}$$

Examples of calculated normalised relevance are in Table 1 for both P_i and R_i. A page with normalised relevance equal to 1 represents the most relevant document and a page with 0 represents the least relevant document. We introduce normalisation under the assumption that search results are at least partially different. In the very unlikely case that all the results in SERP have the same relevance, the normalisation would not work. On the other hand, we need to normalise, because we need to compare:

- First, after normalisation we can compare the relevance of different queries. This would be otherwise difficult, since different queries may generate widely differing relevance, e.g. in one case in range 0–1 000, in another case 0–5. To compare them without normalisation would be misleading. Here, the fact that the pages' relevance values are distributed similarly for any query is very helpful.
- Second, we need to normalise relevance to be able to compare it with seekers' evaluations (each seeker ranks both P_i and R_i pages – for more details see the next section). This comparison ensures that we calculated relevance correctly. Correctly calculated relevance can be used to extend the AQueReBET method, e.g. by automatic SERPs browsing (the next SERPs for a set

query) and filter out only the pages with the best relevance, order them and offer them to a seeker.

- Third, the above-mentioned normalisation helps other researchers to compare their results with ours.

4.2 Subjective Evaluation by Seekers

Eye tracker settings. The experiment was conducted in the User eXperience Research Centre at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava. The Centre consists of two eye-tracking-enabled laboratories. The data collection was performed in the laboratory for detailed research of user experience using a Tobii TX300 eye tracker. The Tobii TX300 is a high precision remote eye tracker with average accuracy 0.4 deg (under ideal conditions), processing latency 1.0 to 3.3 ms, total system latency at most 10 ms, and blink recovery time 10 to 165 ms. The eye tracker is able to compensate for large head movements enabling unobtrusive research. Gaze data was acquired in binocular mode at 300 Hz sampling rate. Participants took part in the study separately. Each participant was comfortably seated at 65 cm eye distance from the eye tracker, calibrated and verified using the live viewer.

Participants. 22 participants took part in this second evaluation. All of them were university students, 6 females and 16 males. Each participant received five queries with specific search intent, and subsequently evaluated (through explicit ratings) the relevance of documents that were obtained with and without using our method. Following that, we tested the impact of using the proposed groupization approach on respondent's satisfaction with documents' relevance. The last step of the experiments was to compare calculated documents' relevance produced by our system versus evaluation of the same documents' assessed by the respondents. We divided our participants into two equally-sized groups. Both groups consisted of 11 participants. All of the participants queried Google with the given intentions. Participants in the first group received results from Google and then from AQueReBET, participants in the second group received results from Google and then from the AQueReBET with the groupization on. Similarly, those 22 participants were divided into 2 groups based on their IT skills: IT-participants with higher search skills (mostly students of information technology related study programs), and non-IT-participants with lower web search skills. Both groups consisted of 11 participants. In the IT group 5 participants had turned on the groupization in AQueReBET and 6 off; in the non-IT group, 6 on and 5 off.

Experiment settings. We allocated a time slot of 30–40 minutes for each of the 22 participants (seekers). The time slot includes eye tracker calibration, an explanation of experiments and queries, and answering seeker's questions. Seekers searched answers for the queries from Table 2 during their sessions with specific search intent with the aim to find the most relevant pages/documents for the

particular query. Seekers were allowed to set a given search query into a search bar of `www.google.com` search engine in a private window of the Google Chrome browser and look through the search results.

Query ID	Initial Query	Complexity (Level)	Interest/Intent
Q1	Monk	Understand (2)	Find more information about monks in monasteries and their ascetic lives.
Q2	Major	Remember (1)	What is the meaning of major in music in the context of the musical scale?
Q3	Hockey stick	Evaluate (4)	Where to buy a hockey stick?
Q4	Amnesiac band*	Analyze (3)	Retrieve information about Radiohead's album Amnesiac. However, there also exists a musical band called Amnesiac.
Q5	Australian second world war	Analyze (3)	Service records details of Australian soldiers who fought in the Second World War.

Table 2. Queries and interests (* initial query for Q4 is intentionally wrong to find out if AQueReBET can correct it and give the correct pages in SERP)

They first looked at the results from Google and if they were unsatisfied, they could click on any of them and read the new page/document. They could even amend the original query. After a few seconds, based on the data collected from the seeker's gaze, AQueReBET offered them its search results (suggested additional relevant pages/documents) and they were allowed to do with it the same as with the first one from Google. Finally, they rated both sets of results. The order of presenting the Google and the AQueReBET results cannot be changed, since results from AQueReBET depend on results from Google, which have to be seen first. The AQueReBET evaluations have to be done by the same person, who saw the Google SERP first. The fixed order can, of course, shift our results, but since both SERPs contained a mixed order of right and wrong results, this keeps the seeker's evaluations very closely level to the seeker's objectivity. In later results, we consider the seekers' evaluations as not shifted.

Query determination. Queries were selected based on several aspects: query ambiguity, query complexity in context of length, cognitive complexity of intention and its domain [2, 33], and the number of relevant documents retrieved by the initial SERP provided by Google. Most of the participants did not have any prior knowledge about any query subject. In a few cases, they did have some prior knowledge but we found their prior knowledge was unrelated to the evaluated scenario and we asked them to ignore it to reduce the bias as much as possible. Q4 is a special query, using which we wanted to evaluate the case with a partly wrong initial query ("Amnesiac band" instead of "album Amnesiac" –

see Table 2) but using the their gaze data, the system was able to correct it and provide the relevant output.

Questionnaires (seekers' subjective evaluations). Seekers subsequently evaluated through explicit ratings documents retrieved by the initial SERP (provided by Google in response to the initial query) and those retrieved by our enhanced SERP (provided by Google in response to a query refined by AQueReBET), and after that they answered a questionnaire. In this questionnaire they filled the suggested relevance score e as a value from 0 (absolutely irrelevant) to 10 (absolutely relevant) for each page/document (see Figure 1: subjective evaluations – e in Results without AQueReBET and in Results from AQueReBET). They also indicated an overall satisfaction rate of the suggested documents for each query. For results provided by Google in response to a query refined by AQueReBET, we shall use the formulation "provided by AQueReBET." We wish to emphasize that the role of the Google search engine is twofold. It was quite natural to use it as the "base" search engine upon which to build our extension. But despite our efforts to find results of similar research that could be used for a direct comparison, we found none and therefore have chosen Google also as a system to compare with – of course within the limited scope of our goals.

Data collection reliability. Eye tracking data collection can be unreliable when used with consumer or lower precision research eye trackers, or when the experimental setup is not congruent with the eye tracker's capabilities. The size and spacing of AOIs (areas of interest) on which the metrics are computed need to be large enough so that the gaze points are not misattributed to a wrong AOI. In our case, we used a very robust experimental design:

1. we used relatively large areas of interest – the whole snippets in results page,
2. we employed simple gaze metric (total fixation duration), and also
3. we used a high precision eye tracker (Tobii TX300) that is a time proven robust device that would be able to provide reliable measurements even on word level, or saccadic movements which we did not analyze.

Data quality for each participant was verified by the experiment moderator.

4.3 Metrics

There are many different metrics used to measure the success or effectiveness of information retrieval (through explicit seeker's ratings). We choose the following as they fit our aim (method evaluation) and they are also used by other authors, what allows us to compare to them.

The discounted cumulative gain (DCG), was introduced by [12]. They explained that DCG reflects the fact that "the greater the ranked position of a relevant document, the less valuable it is for the user, because the less likely it is that the user

will ever examine the document due to time, effort, and cumulated information from documents already seen." Moreover, they introduce normalised discounted cumulative gain (nDCG), to be able to compare different DCG curves (e.g. those that use different ranges of seekers' evaluations). These metrics are determined using the following formulas:

$$DCG@k(Q_j) = \sum_{i=1}^{k} \frac{2^{||e(R_i)||} - 1}{\log_2(i+1)} \quad \text{and} \quad nDCG@k(Q_j) = \frac{DCG@k(Q_j)}{\sum_{i=1}^{k} \frac{1}{\log_2(i+1)}} \tag{8}$$

where $||e(R_i)||$ represents a normalised satisfaction rate of a page/document R_i (either provided by Google or by AQueReBET). Our seekers rate pages using an 11 point Likert's scale – $e(R_i) \in \{0, 1, 2, \ldots, 10\}$. Ratings are afterwards normalised (divided by 10) into the interval $[0, 1]$, where 0 means no satisfaction with the document and relevance 1 means the maximum satisfaction with the document relevance. The number k represents the number the first k results in SERP for one set query Q_j. The denominator in nDCG metrics represents the norm – the ideal performance, where all the evaluations are equal to 1. In later evaluations we use also averaged nDCG@k:

$$DCG@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} DCG@k(Q_j) \quad \text{and} \quad nDCG@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} nDCG@k(Q_j) \tag{9}$$

where $|Q|$ is the number of different queries.

We also calculated mean average precision (MAP), which is similar to nDCG, since both have a maximum equal to 1 and both decrease with each irrelevant result in SERP. Unlike nDCG, MAP uses only binary classification (0 for irrelevant result and 1 for relevant result) and for a set of queries is the mean of the average precision scores for each query [31, 23, 5]:

$$MAP@k(Q_j) = \frac{1}{k} \sum_{i=1}^{k} \text{Prec}@i(Q_j) \text{ and } MAP@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} MAP@k(Q_j) \tag{10}$$

where $|Q|$ is the number of different queries, $\text{Prec}@i(Q_j)$ is computed as the fraction of relevant documents within the top i results for query Q_j. Since our seekers did not use binary classification, we had to set the threshold from which the results are relevant and the rest had to be irrelevant. During our pilot tests we noticed that seekers used the following criteria to divide the scale in 5 parts:

- the first (values equal to 10) – exactly what was seeker looking for,
- the second (values 7, 8, 9) for relevant results,
- the middle part (values 4, 5, 6) for partly relevant results,
- the fourth (values 1, 2, 3) for irrelevant results and

- finally the last one (values equal to 0) for totally irrelevant result.

This division is noticeable in our histograms (see Figure 3), where one of the local maximums is usually at 7 or 8 and the other at 4 or 5. This division is similar also to other authors (e.g. [27]). Based on other studies (including [5] we decided to use the border value number 4, since from this value there are at least partly relevant documents. It means if a seeker evaluated a result by $e \geq 4$, we calculated with it as with a relevant result (in binary classification precision equal to 1 = positive) and if the evaluation $e < 4$, then we calculated with it as with an irrelevant result (in binary classification precision equal to 0 = negative). Another metric used to evaluate a query is the reformulation necessity called $Rfactor$. It takes into account the number of times that a query had to be reformulated on average. It is determined using the following formula:

$$\text{Rfactor} = \frac{\sum_{i=1}^{|Q|} \text{Refor}(Q_i)}{|Q|} \tag{11}$$

where $\text{Refor}(Q_i)$ is the number of necessary reformulations of initial query Q_i until the relevant results are achieved. If the query Q_i has not been reformulated by the seeker even once, $Refor(Q_i) = 0$. $|Q|$ represents the number of initial queries (in our case, $|Q| = 5$).

Satisfaction with the whole result generated by the initial or a refined query evaluated (= rated) by i^{th} seeker is $e(\tau_{int})$ or $e(\tau_{out})$, respectively. It was a subjective evaluation, where seekers used the same Likert's scale afterwards normalised to interval $[0, 1]$. This number should have been in correlation with nDCG.

4.4 Comparison of Automatic Evaluation and Evaluation by Seekers

To compare our automatic evaluation with the evaluation done by seekers, we used standard metrics: $accuracy = \frac{TP+TN}{P+N}$, $recall = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$, and $F = 2 * \frac{precision*recall}{precision+recall}$, where TP or "true positive" is the number of correctly classified relevant documents, TN or "true negative" is the number of correctly classified irrelevant documents, P represents the number of relevant documents and N the number of irrelevant documents, FN or "false negative" is the number of incorrectly classified relevant documents and FP or "false positive" represents the number of incorrectly classified irrelevant documents. In the context of a web search, these metrics are calculated from the displayed results on SERP (mostly 8, 9 or 10 of them). To calculate these metrics, we used our calculated normalised relevance $||r|| \in [0, 1]$ and seekers' evaluations $e \in \{0, 1, \ldots, 10\}$. To compute these metrics we need to convert both in binary classification. As we already explained, we set for seekers' evaluations the borderline number 4. For relevance $||r||$ it was a bit complicated, so we calculated the trend line between e and $||r||$ from our pilot tests. It came out that border $||r|| = 0.1$ corresponds to border $e = 4$. It means that if AQueReBET relevance value $||r|| < 0.1$, the result is negative, if $||r|| \geq 0.1$, positive.

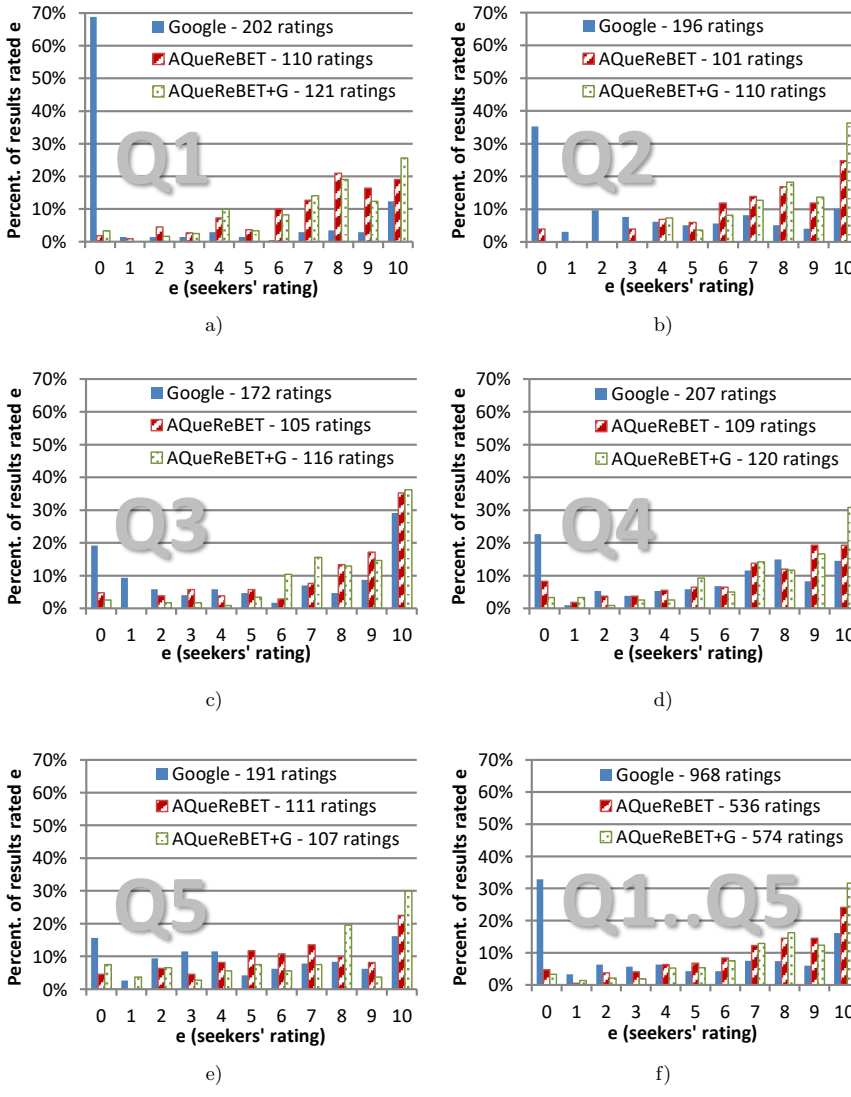


Figure 3. Normalised distribution (histogram) of seekers' ratings (how they rated results from SERP) for query a) Q1, b) Q2, c) Q3, d) Q4, e) Q5 and f) all five queries Q1, Q2, Q3, Q4, Q5 together. Comparison between results obtained using the Google engine (blue columns), our AQueReBET system (red columns) and our system enhanced by groupization – AQueReBET + G (green columns). All ratings e appear on axis x from $e = 0$ – irrelevant information source, to $e = 10$ – exactly the information source I wanted.

It should be noted that we do not present any correlation between ratings by seekers and AQueReBET. When attempting to compute it, we realised that the statistical distributions of the respective ratings are very different (cf. Table 1 and Figure 3) and therefore it is not appropriate to calculate the correlations.

5 RESULTS OF EXPERIMENTS

In this section, we describe the results of some of our experiments. We try to compare the achieved results with those achieved by other authors. This is not entirely possible due to different sources of data and different outputs of the corresponding methods. Therefore, any claim we shall make regarding superiority of any of those approaches is of limited validity only. In particular, when results of our method are better than results of another method, it may be due to the method itself or due to the experiment set up and we are not able to tell which is the case.

5.1 Google as a Baseline Versus AQueReBET Evaluated by Seekers

The primary goal of the performed experiments is to falsify or endorse the hypothesis that when we have the data from a seeker's gaze, we can suggest more relevant documents to the seeker and reduce the need for reformulation of the initial query.

To find out, we let all the seekers rate relevance of each document (see Figure 1 subjective evaluations – *e*), either provided by Google, by AQueReBET or by AQueReBET + G. At first, we compared these three response providers using the DCG@k, nDCG@k metrics for k from 1 to 9 (Figures 4 c) and 4 d)); to show the difference between individual queries we added also DCG@5(Qi) and nDCG@5(Qi) (Figures 4 a) and 4 b)).

Graphs in Figures 4 a) and 4 b) show that values of measures DCG and nDCG depend quite strongly on the type of query. Generally, however, we observe that our answer provider AQueReBET gives better results as the baseline (Google). In some cases, groupization is able to yield further, albeit slight improvement. Graphs in Figures 4 c) and 4 d) show how values of DCG and nDCG change with an increasing k. We note that the DCG measure in Figure 4 c) is somewhat harder to read due to the fact that the ideal maximum value changes with k. The nDCG measure of the baseline gives the smallest value for $k = 1$. This is caused by the ambiguity of queries, which are too short. In case of AQueReBET, all queries have been refined, so there is not such a steep increase of values from $k = 1$ to $k = 2$. Both AQueReBET curves show statistically significant improvement over the baseline for every k ($p < 0.05$, Wilcoxon signed-rank test). For example, for $k = 5$ we have a significant improvement from $38\,\% \pm 17\,\%$ (Google) to $74\,\% \pm 13\,\%$ (AQueReBET) alternatively $78\,\% \pm 14\,\%$ (AQueReBET + G) representing a rise by 36 alternatively 40 percentage points as well as 1.9 alternatively 2.1 times increase.

When comparing nDCG@5 of SERPs provided by AQueReBET and AQueReBET + G for the whole dataset, groupization improved it by only approximately

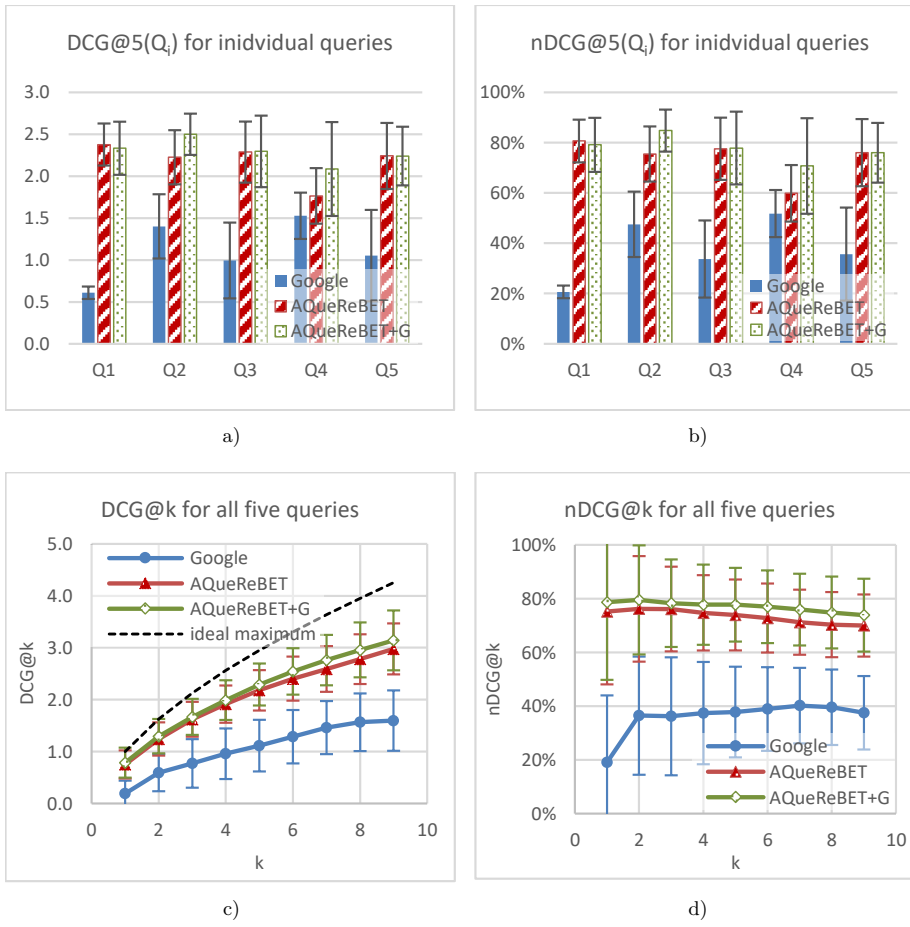


Figure 4. a) DCG@5 for individual queries, b) nDCG@5 for individual queries, c) DCG@k for all five queries and d) nDCG@k for all five queries averages with standard deviations when Google, AQueReBET, AQueReBET + G is used

4 percentage points. However, the difference is not a statistically significant improvement (Mann-Whitney Test for Two Independent Samples, with $\alpha = 0.05$, two tailed, p-value $= 0.11$ did not refute the hypotheses about the same means). This means that most of the times the refined query (and consequently its SERP) was already good enough without using groupization. In such circumstances, the groupization cannot bring tangible improvements. We hypothesize that this value should increase more significantly by enlarging the groupization database over time.

The difference in the individual averages of DCG@5 and nDCG@5 (see Figures 4 a) and 4 b)) shows that the results for individual queries are diverse and thus should be evaluated separately, too. Their diversity is also visible in Figures 3 a), 3 b), 3 c), 3 d) and 3 e), minor differences between the number of ratings are due to the exclusion of invalid results, e.g. lack of rates, and so on. Nevertheless, we found that AQueReBET, by enriching the web search by data collected from seekers' gaze, improves the relevance of documents retrieved significantly ($p < 0.05$) except the results for Q4.

Let us discuss all the queries one by one: The reason for such considerable DCG@5 and nDCG@5 improvement for Q1 and Q2 dwells probably in their ambiguity. These queries have more than one interpretation and therefore Google itself was not able to provide SERP with pages, which fit to the meaning the seeker had in mind (see the highest first column for Google in Figures 3 a) and 3 b)). Similarly, there is significant improvement for Q3 and Q5 because it is hard to guess from the set keywords what the user's exact intention is. Q5 is definitely the most complex query, therefore it is interesting how the averaged nDCG@5(Q5) increased from $36\,\% \pm 16\,\%$ to $76\,\% \pm 13\,\%$ (for more details see Figure 3 e), Figure 4 b)). Q4 did not pass the border value $\alpha = 0.05$ probably because of the low number of pairs in a sample and also the specific type of query – it is the one with intentionally partly wrong input. Its averaged nDCG@5(Q4) of the initial SERP (provided by Google) is $53\,\% \pm 10\,\%$ (sample1) respectively $51\,\% \pm 9\,\%$ (sample3), what seems high for a partly incorrect query, but AQueReBET provided SERP with an even higher $60\,\% \pm 11\,\%$ (statistically insignificant increase, $p = 0.11$) and $71\,\% \pm 19\,\%$ (when groupization involved, statistically significant increase, $p < 0.05$, see Figure 4 a), for more details see Figure 3 d)).

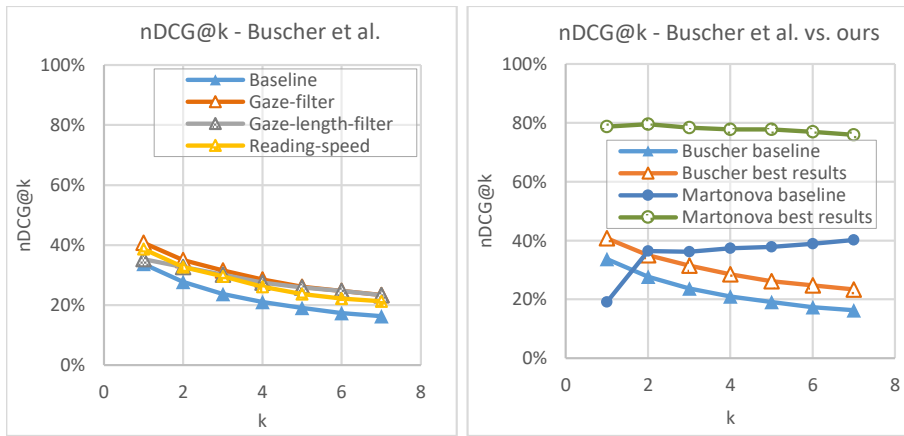
When interpreting these results, one should keep in mind three facts:

- The Google result page for the same query may slightly vary, since the experiment was performed over several days.
- When different seekers input the same initial query to Google, it may lead to different refined queries for AQueReBET, since individual seekers' gaze patterns can lead to different tables (with words and their importance) for each seeker. Because of the same reason the suggested pages and documents on the AQueReBET result page vary for different seekers and even if there are some the same, they have different relevance.
- Two different seekers rate the same SERP slightly differently (subjective evaluation). This is obviously more likely in the case of a greater number of seekers. Thus the maximal achievable nDCG (or any other metrics) for any query shall most likely not be 100 %. We assume the maximal achievable nDCG is in the top decile.

It would be beneficial, if not compulsory, to compare our results with some representative works of others who have dealt with a similar problem. For our work, it is instructive to make a comparison, e.g. with the influential work of [5]. It

should be noted, however, that there were differences in experiments. Contrary to ours, their users browsed whole documents (our users read snippets only). Also, the users' tasks were different.

Since Buscher et al. [5] evaluated the DCG metric based on values from the set of ratings $\{0, 1, 2, 3\}$ whereas we used the set $\{0, 0.1, 0.2, \ldots, 1\}$ it was necessary to perform a normalisation. Thus we at first normalised their DCG values to nDCG (see Figure 5 a)) to be able to compare it with our results. In Figure 5 b) one can see that values for respective baseline cases both stay within the 20–40 % interval. However, there is a difference with respect to k: while theirs decreases, ours increases. We assume that tendency of the nDCG@k measure depends on the kind and ambiguity of the given query. In the case of our baseline, queries were mostly ambiguous. As far as the problem itself is concerned, from some abstract view it is possible to treat both classes of problems, i.e. the one in [5] and ours as essentially comparable. This could be claimed while noting that they actually were solving a slightly different problem with a different data set. They also processed data from the eye tracker in a slightly different way. Still, an elementary comparison is possible. They achieved (for $k = 5$) a 1.37 fold improvement, we achieved a 2.06 fold improvement. Their maximum value of nDCG@k is 40.8 % (for $k = 7$), whereas ours is 79.5 % (for $k = 2$). This is, even respecting the 13.7 % standard deviation, considerably better. [6] performed a very similar experiment achieving DCG@10 = 9.15 which corresponds to nDCG@10 = 29.2 % for a read length of 150 characters. Snippets are about 150 characters long. Our result of nDCG@9 = 70 % makes us 2.4 times better.



a) b)

Figure 5. a) nDCG@k for all variants (calculated from their DCG@k [5] and b) nDCG@k comparison of [5] base line with their best results and our base line with our best results

Eickhoff et at. [8] used Mean Reciprocal Rank (MRR) metric:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{12}$$

where $rank_i$ is the rank position of the first relevant document for the i^{th} query and Q is the set of evaluated queries) attaining values 0.80 and 0.86, compared with 0.79 MRR value for Buscher et al. [5]. In our method, MRR metric for relevant results (seeker's evaluation value 7 or higher) is 0.87 for AQueReBET and 0.91 for AQueReBET + G.

Umemoto et al. [31] is another related work which used nDCG to evaluate their results. They achieved 81.6 % and our best value of nDCG@2 = 79.5 ± 20.3 %. The difference is probably not statistically significant. The main difference in evaluation is that they considered up to 15 terms, whereas we considered 4 words and their rating scale had 3 degrees, but our had 11 degrees. On the other hand, we received a better MAP metric (more details see below).

We also evaluate Prec@k and MAP@k metrics for k from 1 to 9 (Figures 6 c) and 6 d)); to show the difference between individual queries we added also Prec@5(Qi) and MAP@5(Qi) (Figures 6 a) and 6 b)). In Figures 6 a) and 6 b) one can see that both Prec and MAP measures depend, similarly to Figure 4, on the type of query. The effect of groupization is also similar, i.e. sometimes it may improve, but sometimes it may worsen the results of the AQueReBET. In Figures 6 c) and 6 d), the curve for the baseline case initially exhibits a similar steep increase for similar reasons as before (cf. our comments to Figure 4). All in all, however, all curves in Figure 6 look better in the sense they are closer to 100 %. This is caused by the binary evaluation scale of these metrics. A binary evaluation tends to supress small differences that can be observed in DCG and nDCG metrics (we used an evaluation with 11 different values there). As a consequence, differences between results with and without a groupization are almost invisible and look almost identical. But here, too, one can see that both AQueReBET curves show statistically significant improvement over the baseline for every k ($p < 0.05$). For example, for Prec@5 we have an improvement from 50 % ± 27 % (Google) to 91 % ± 14 % (AQueReBET) alternatively 95 % ± 10 % (AQueReBET + G), which represents a 1.8 alternatively 1.9 times increase. For for MAP@5 we have an improvement from 44 % ± 28 % (Google) to 93 % ± 15 % (AQueReBET and the same for AQueReBET + G), which represents an increase of 2.12 times. When comparing AQueReBET and AQueReBET + G, it is obvious, the difference is statistically insignificant. The results for individual queries (see Figures 6 a) and 6 b)) resemble the results for nDCG, which were already discussed.

In an attempt somehow to compare our results with those of [5], we note that their absolute MAP for baseline is 46.6 % (depending on the used variant from 29.7 % to 54.3 %) and absolute MAP of their method is 55.9 % (depending on the used variant from 39.3 % to 66.7 %), which represents an increase of 1.20 times.

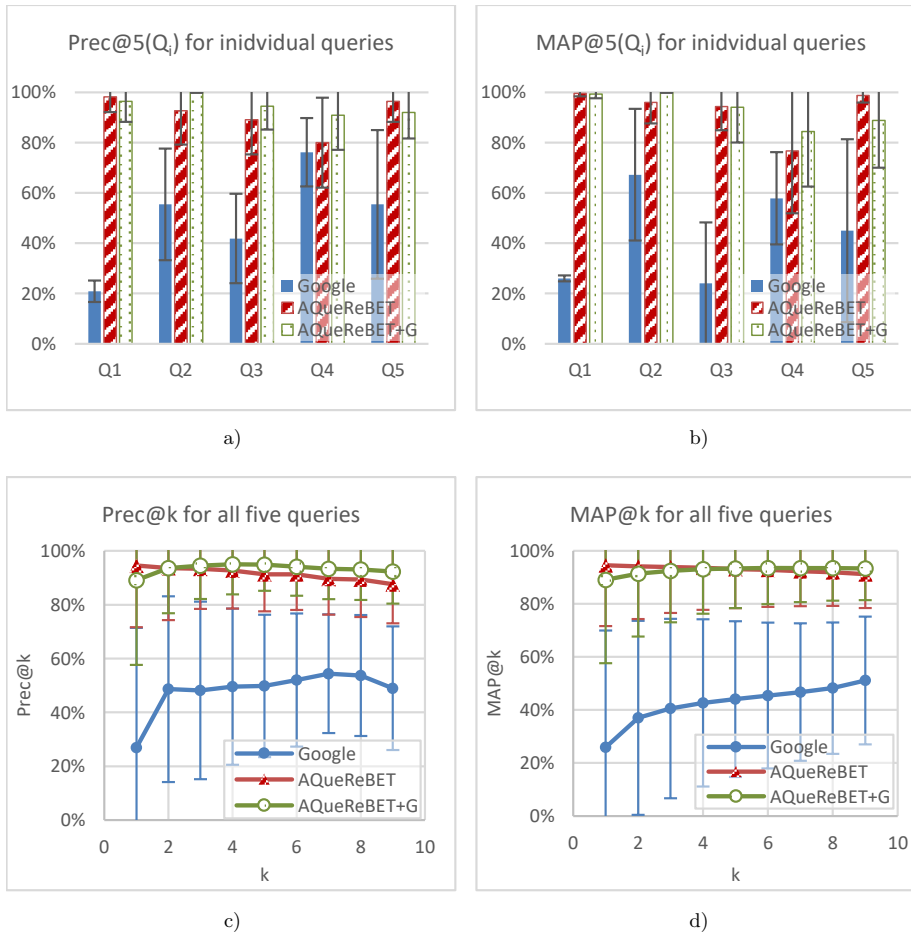


Figure 6. a) Prec@5 for individual queries, b) MAP@5 for individual queries, c) Prec@k for all five queries and d) MAP@k for all five queries averages with standard deviations when Google, AQueReBET, AQueReBET + G is used

Since our and their baseline results are very close, we consider our results considerably better, whether absolute improvement (our 93 % vs. their 55.9 %) or relative improvement (our 2.12 times vs. their 1.20 times increase). [6] experimented in a very similar way achieving MAP@10 = 73.6 % for read length of 150 characters. Snippets are about 150 characters long. Our result of 93 % makes us 1.26 times better. [31] achieved up to MAP = 65.2 % and our best value of MAP@1 = 94.5 ± 22.9 %. However, these results have been achieved under slightly different conditions so they can serve for a rough comparison only. The main difference in evaluation is that they considered up to 15 terms whereas we considered

4 words and their binary mapping of user evaluations uses a different threshold than does ours.

	Rfactor		Satisfaction		Time [s]	
	avg	stdev	avg	stdev	avg	stdev
AQueReBET	25.5 %	9.3 %	76.0 %	4.9 %	66.1	32.6
AQueReBET + G	21.8 %	10.8 %	80.2 %	6.3 %	47.1	13.6

Table 3. Averages and standard deviations of selected metrics for AQueReBET and AQueReBET + G for all five queries together

In Table 3 Rfactor, Satisfaction and time metrics are shown. These are not evaluated for Google, but we can see, that in all of them the AQueReBET + G gives slightly better results than AQueReBET: Rfactor is 3.7 percentage points better, overall satisfaction 4.2 percentage points better and Time 18.7 s shorter.

The AQueReBET average time of presenting the suggested documents was 66.1 (47.6 seconds for AQueReBET + G). Zhu and Mishne [35] presented that the average time for choosing a document as relevant in a web search is approximately 46.15 seconds. That is, provided that the eye tracker calibration is completed, our solution needs only approximately 20 s respectively 1.5 seconds more to show suggested documents. But this strongly depends on calibration, internet connection and seekers' search stereotypes.

5.2 Accuracy of Automatic Evaluation

The power of the AQueReBET system lies not only in the detection of gazed snippets but also in the system's ability to correctly anticipate the seekers' ratings. To evaluate the correctness of the system's calculations we used standard relevance measures – F-measure, precision, recall and accuracy. We calculated them by comparing the seeker's evaluation (rating) e and our system's calculated normalised relevance $||r||$ (see Figure 1 for subjective and objective evaluations).

We should like to note that the former three metrics give higher values with an increasing share of true positives. It is therefore to be expected that results from Google achieve lower values than those acquired by applying our method. On the other hand, the latter accuracy metric reflects only the volume of correctly identified results; it increases with the increasing sum of true positives and true negatives. Provided our automatic evaluation is correct, accuracy shall be the same regardless of the search engine used. This is the reason why, in this part, we present the combined results of all three methods. However, also partial results according to the input query and the combined results for all five queries are included there. We should also note that the achievable maximum for all four metrics can practically never be 100 %, since seekers can differ in their subjective ratings.

Averages with standard deviations of all four relevance metrics for individual queries using different search engines can be found in column graphs in Figure 7.

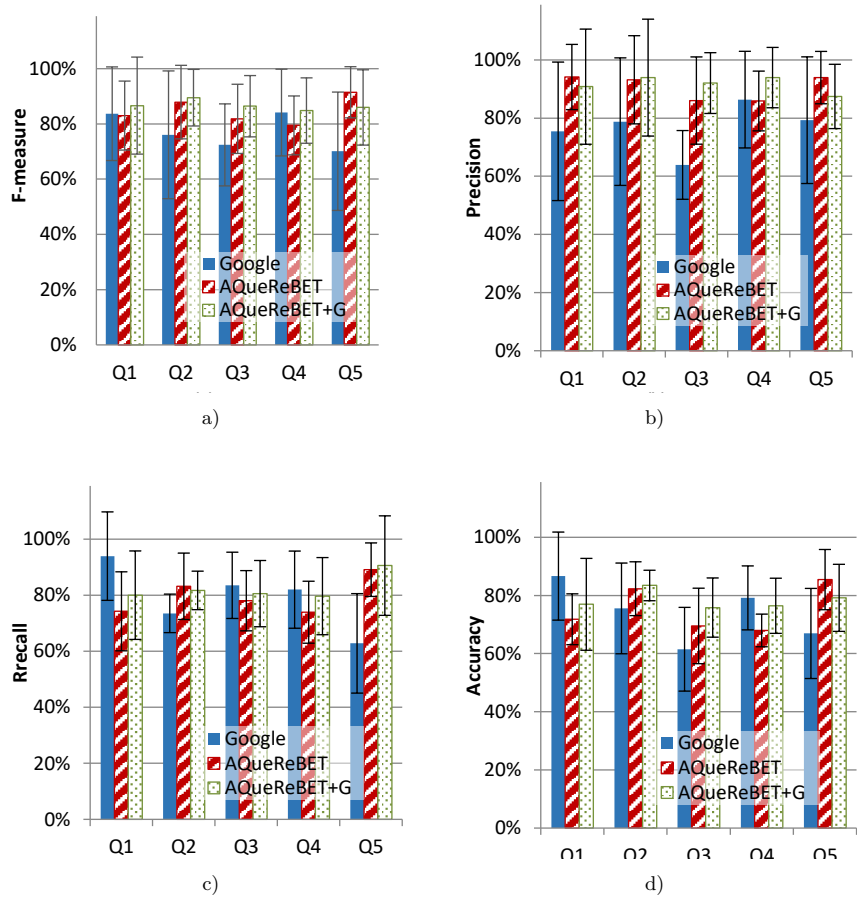


Figure 7. a) F-measure, b) precision, c) recall and d) accuracy averages with standard deviations for queries Q1, Q2, Q3, Q4 and Q5 when Google, AQueReBET, AQueReBET + G is used

When comparing these measures between the individual queries (see Figures 7 a), 7 b), 7 c) and 7 d), we can see that it is not generally the case that the AQueReBET has better averages than Google, or that with groupization it is mostly better. The small difference made by groupization has the same reason as with nDCG – most of the times the refined query (and consequently its SERP) was already good enough without using groupization. When calculated, the statistical significance of differences in averages between Google, AQueReBET and AQueReBET + G, many of them were not significant (F-measure for Q1 and Q4, precision for Q1, Q2 and Q4, recall for Q2, Q3, Q4 and accuracy for Q2 and Q4, same for AQueReBET and AQueReBET + G). It was probably caused by the low number of samples (only 10

or 11 for individual queries, which is not enough for such wide standard deviations and such close averages).

When taking all queries together, F-measure and precision gave us a significant difference ($p < 0.05$) between Google and AQueReBET either with or without groupization. E.g. for F-measure, there was improvement from 71 % ± 17 % (Google) to 84 % ± 10 % (AQueReBET) and from 78 % ± 15 % (Google) to 85 % ± 11 % (AQueReBET + G), which is 13 respectively 7 percentage points improvement. As expected, values for accuracy are nearly the same: 71 % ± 17 % (Google Sample1), 77 % ± 16 % (Google Sample3), 75 % ± 12 % (AQueReBET) and 78 % ± 11 % (AQueReBET + G) – the Wilcoxon test approved that the differences between averages are not significant.

Another point of view gives us results calculated without the search engine differentiation. Values for all search engines together, for individual queries as well as for all of them, are in Table 4 (calculated directly from *TP*, *TN*, *FP*, *FN* counts). The differences between individual queries showed that the queries could give us slightly different results. When summing up all queries together, F-measure, precision, recall and accuracy are in all cases greater than 75 % (see Table 4, grey row), what we consider to be the good quality of our automatic evaluator.

	Rating	TP	TN	FP	FN	F-Measure	Precision	Recall	Accuracy
Q1	431	206	140	35	50	82.9 %	85.5 %	80.5 %	80.3 %
Q2	407	220	103	22	62	84.0 %	90.9 %	78.0 %	79.4 %
Q3	391	241	24	67	59	79.3 %	78.2 %	80.3 %	67.8 %
Q4	435	262	67	36	70	83.2 %	87.9 %	78.9 %	75.6 %
Q5	407	227	78	36	66	81.7 %	86.3 %	77.5 %	74.9 %
$\sum$	2 017	1 156	412	196	307	82.1 %	85.5 %	79.0 %	75.7 %

Table 4. Number of seekers' ratings, number of true positive, true negative, false positive and false negative identifications of our relevancy evaluator, calculated metrics (recall, precision, accuracy and F-measure) for all five queries together and separately as well (not differentiating a type of used search engine)

5.3 Comparison of IT and NonIT Seekers

In the last experiment we compared the results for IT and nonIT seekers' groups. Figure 8 depicts the comparison of nDCG obtained by Google, AQueReBET and AQueReBET + G. As an interesting fact, it appears that groupization significantly improves relevance of documents/pages mainly for the nonIT group (see Figure 8 queries Q1, Q2, Q3 and partly Q4). The results indicate that groupization is helpful especially for seekers who have lower web searching skills. It seems that it is precisely the type of skills that are conveyed to those seekers by the groupization. We also hypothesize that groupization is most effective especially in cases when intentions behind queries are similar across the group. The statistical significance of the difference between IT and nonIT groups was not evaluated since the groups had

only 5 and 6 participants and moreover, an improvement was not observed for all queries.

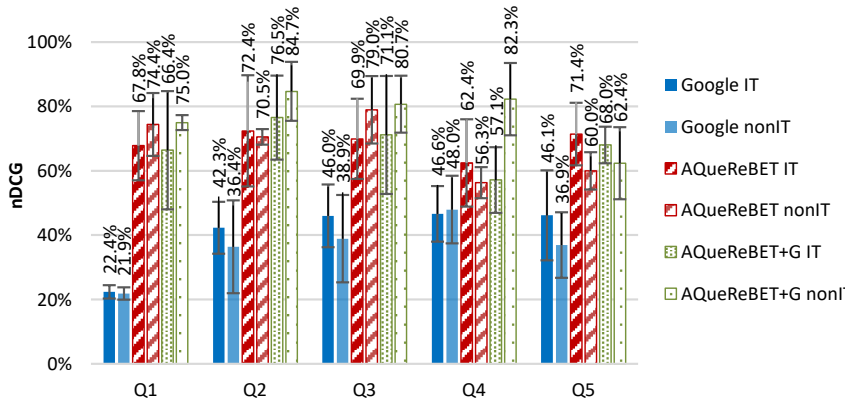


Figure 8. IT and nonIT comparison of nDCG metric for results provided by Google, AQueReBET and AQueReBET + G

We are aware of the desirability of comparison with related works. Since there are only a few works in this area and there is no common adopted methodology or dataset, it is not possible to compare directly different approaches and their results. We hope, that we have fulfilled the need for comparison at least partially by adopting Google as a baseline (which is and will be available to any other researcher in the future).

6 CONCLUSIONS

In this paper we described a new method called AQueReBET – automatic query refinement based on eye-tracking. This method provides web seekers during their web search (using their implicit feedback) with more relevant documents. We chose the way of extracting possible new words from the snippets in the page of SERP, where each word-level importance is calculated based on the term frequency, term uniquness (tf-idf) and total fixation duration within the snippets. We evaluated the proposed approach in a study with 22 participants. The results support our hypothesis that the information obtained from the analysis of the seeker's gaze can improve the relevance of the pages/documents provided to the seeker. On average, our solution improved nDCG@5 of web searches from 38 % to 74 % alternatively 78 %, representing a rise by 36 alternatively 40 percentage points as well as a 1.9 alternatively 2.1 times increase. We managed to reduce the query reformulation rate to approximately 23.6 % on average. The most notable improvement was obtained

for the most ambiguous query Q1. Considering several related works on gaze-based feedback, improvement has also been reported by other authors [5, 8, 3, 31].

When making a quantitative comparison with [5], they achieved (for $k = 5$ and similar baseline) a 1.37-fold improvement. Their maximum value of nDCG@k is 40.8 %, whereas ours is 79.5 %. This is, even respecting the 13.7 % standard deviation, considerably better. They used also MAP metrics (mean average precision). We have an improvement from MAP@5 = 44 % to 93 %, representing a rise of 49 percentage points as well as 2.12 times increase. They improved their absolute MAP from 46.6 % to 55.9 % representing a rise of 9.3 percentage points as well as 1.20 times increase. Since our and their baseline results are very close, we consider our results considerably better. However, it should be noted that the datasets were not the same, and nor were the experiment setting and methodology.

Buscher et al. [6] performed a very similar experiment. We consider their results only regarding a read length of 150 characters because our snippets are cca 150 characters long. They achieved DCG@10 = 9.15 which corresponds to nDCG@10 = 29.2 %. Our result is nDCG@9 = 70 % which is 2.4 times better. They achieved MAP@10 = 73.6 %, our result of 93 % makes us 1.26 times better.

Umemoto et al. [31] also used similar metrics, in particular nDCG and MAP to evaluate their results. They achieved nDCG = 81.6 %, our best value of nDCG@2 = 79.5 %. The difference is probably not statistically significant. They achieved MAP = 65.2 % and we achieved 94.5 %, what is 1.45 times better. However these results have been achieved under slightly different conditions so they can serve for a rough comparison only. The main difference in evaluation is that they considered up to 15 terms whereas we considered 4 words and their rating scale had 3 degrees whereas ours 11, and their binary mapping of user evaluations uses a different threshold than does ours.

Eickhoff et al. [8] used MRR metric attaining values 0.80 and 0.86, compared with 0.79 MRR value for Buscher et al. [5]. In our method, MRR metric for relevant results (seeker's evaluation value 7 or higher) is 0.87 for AQueReBET and 0.91 for AQueReBET + G.

In the experiments, we also studied the effect of groupization. The groupization improved the nDCG@5 on average by 4 percentage points, although the improvement was not significant. In general, however, we see for this approach a potential for improvement. E.g. bearing in mind the quick response of our system, our calculations worked only with the first 10 Google results. If we would take more of them, the nDCG would be even higher. One of the challenges is that it would sometimes benefit from an ability to choose proper query words based on their meaning (semantics).

The other important result of our research, besides the above mentioned method, is also the automatic objective evaluator, which allows us to order the results in SERP and get such high results in nDCG, MAP and MRR. We determined its accuracy and obtained 75 %. We consider it high, keeping in mind that seekers' subjective evaluations differ a little bit and thus 100 % accuracy is not achievable.

Results of our comparison of IT vs nonIT groups suggest that groupization is helpful especially for seekers that have lower web searching skills. However, this requires further research with more participants and more queries.

It should be observed that our approach works when at least one of the results in SERP corresponds to a user's intention. In the opposite case, i.e., when all results do correspond, there is no need to apply our method. We identify these to be limitations of our work.

Future work and possible improvements lie in creating an ontology with which we could extract appropriate words in a better way. Some use Wikipedia as a source of semantics [9]. Using semantic links between words we would be able to remove unrelated words from the table, resulting in more accurate suggestions. Enhancing semantic approach by sentiment could provide possibly even better results [1].

Another direction of future work may be to add a fourth assumption (and incorporate it into our Equation (3)): The higher the importance of a word, the better position it has within the snippet – e.g. the closer to a queried word, the higher the importance.

A different possible line of research of utilization of eye tracking feedback could be inspired by recent progress in the related research on utilization of cursor movement data [18], where frequent subsequences called motifs are automatically discovered to be used to improve result relevance estimation and re-ranking.

Acknowledgments

This work was partially supported by the Scientific Grant Agency of Slovakia, grant No. VG 1/0667/18 and it is a partial result of the development project No. 002STU-2-1/2018 funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic. It was also supported by the Research and Development Support Agency, grant No. APVV-15-0508 and grant No. APVV-17-0267.

REFERENCES

[1] Abdi, A.—Shamsuddin, S. M.—Aliguliyev, R. M.: QMOS: Query-Based Multi-Documents Opinion-Oriented Summarization. Information Processing and Management, Vol. 54, 2018, No. 2, pp. 318–338, doi: 10.1016/j.ipm.2017.12.002.

[2] Anderson, L. W.—Krathwohl, D. R.—Bloom, B. S.: Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman, New York, 2001.

[3] Biedert, R.—Dengel, A.—Käding, C.: Universal Eye-Tracking Based Text Cursor Warping. Proceedings of the Symposium on Eye Tracking Research and Application (ETRA '12), ACM, 2012, pp. 361–364, doi: 10.1145/2168556.2168637.

[4] Bing, L.—Lam, W.—Wong, T.-L.—Jameel, S.: Web Query Reformulation via Joint Modeling of Latent Topic Dependency and Term Context. ACM Transactions on Information Systems, Vol. 33, 2015, No. 2, Art. No. 6, 38 pp., doi: 10.1145/2699666.

[5] Buscher, G.—Dengel, A.—van Elst, L.: Query Expansion Using Gaze-Based Feedback on the Subdocument Level. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08), ACM, 2008, pp. 387–394, Art. No. 6, 38 pp., doi: 10.1145/2699666.

[6] Buscher, G.—Dengel, A.—Biedert, R.—van Elst, L.: Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond. ACM Transactions on Interactive Intelligent Systems, Vol. 1, 2012, No. 2, Art. No. 9, 30 pp., doi: 10.1145/2070719.2070722.

[7] Carpineto, C.—Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys, Vol. 44, 2012, No. 1, Art. No. 1, 50 pp., doi: 10.1145/2071389.2071390.

[8] Eickhoff, C.—Dungs, S.—Tran. V.: An Eye-Tracking Study of Query Reformulation. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15), ACM, 2015, pp. 13–22, doi: 10.1145/2766462.2767703.

[9] Goslin, K.—Hofmann, M.: A Wikipedia Powered State-Based Approach to Automatic Search Query Enhancement. Information Processing and Management: An International Journal, Vol. 54, 2018, No. 4, pp. 726–739, doi: 10.1016/j.ipm.2017.10.001.

[10] Granka, A. L.—Joachims, T.—Gay, G.: Eye-Tracking Analysis of User Behavior in WWW Search. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04), ACM, 2004, pp. 478–479, doi: 10.1145/1008992.1009079.

[11] Huang, J.—White, R.—Buscher, G.: User See, User Point: Gaze and Cursor Alignment in Web Search. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), ACM, 2012, pp. 1341–1350, doi: 10.1145/2207676.2208591.

[12] Järvelin, K.—Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems, Vol. 20, 2002, No. 4, pp. 422–446, doi: 10.1145/582415.582418.

[13] Joachims, T.—Granka, T.—Pan, B.—Hembrooke, H.—Radlinski, F.—Gay, G.: Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. ACM Transactions on Information Systems, Vol. 25, 2007, No. 2, Art. No. 7, doi: 10.1145/1229179.1229181.

[14] Spärck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation, Vol. 28, 1972, No. 1, pp. 11–21, doi: 10.1108/eb026526.

[15] Kajan, R.—Herout, A.—Bednarik, R.—Povolný, F.: PeepList: Adapting Ex-Post Interaction with Pervasive Display Content Using Eye Tracking. Pervasive and Mobile Computing, Vol. 30, 2016, Issue C, pp. 71–83, doi: 10.1016/j.pmcj.2015.12.004.

[16] Kompan, M.: Group and Single-User Influence Modeling for Personalized Recommendation. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 6, 2014, No. 2, pp. 11–20.

[17] Kovárová, A.: Special Interaction Approaches and Their Impact on Usability. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 3, 2011, No. 3, pp. 14–25.

[18] Lagun, D.—Ageev, M.—Guo, Q.—Agichtein, E.: Discovering Common Motifs in Cursor Movement Data for Improving Web Search. Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14), ACM, 2014, pp. 183–192, doi: 10.1145/2556195.2556265.

[19] Li, Q.—Tian, M.—Liu, J.—Sun, J.: An Implicit Relevance Feedback Method for CBIR with Real-Time Eye Tracking. Multimedia Tools and Applications, Vol. 75, 2016, No. 5, pp. 2595–2611, doi: 10.1007/s11042-015-2873-1.

[20] Loboda, T. D.—Brusilovsky, P.—Brunstein, J.: Inferring Word Relevance from Eye-Movements of Readers. Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11), ACM, 2011, pp. 175–184, doi: 10.1145/1943403.1943431.

[21] Low, T.—Bubalo, N.—Gossen, T.—Kotzyba, M.—Brechmann, A.—Huckauf, A.—Nürnberger, A.: Towards Identifying User Intentions in Exploratory Search Using Gaze and Pupil Tracking. Proceedings of the 2017 Conference on Human Information Interaction and Retrieval (CHIIR '17), ACM, 2017, pp. 273–276, doi: 10.1145/3020165.3022131.

[22] Maglio, P. P.—Barrett, R.—Campbell, C. S.—Selker, T.: SUITOR: An Attentive Information System. Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI '00), ACM, 2000, pp. 169–176, doi: 10.1145/325737.325821.

[23] Manning, C. D.—Raghavan, P.—Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, 2008.

[24] Móro, R.—Daráž, J.—Bieliková, M.: Visualization of Gaze Tracking Data for UX Testing on the Web. Late-Breaking Results – Doctoral Consortium and Workshop Proceedings of the 25th ACM Hypertext and Social Media Conference (Hypertext 2014 – Extended Proceedings), CEUR Workshop Proceedings, Vol. 1210, 2014.

[25] Murata, M.—Toda, H.—Matsuura, Y.—Kataoka, R.: Query-Page Intention Matching Using Clicked Titles and Snippets to Boost Search Rankings. Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09), ACM, 2009, pp. 105–114, doi: 10.1145/1555400.1555419.

[26] Papoutsaki, A.—Laskey, J.—Huang, J.: SearchGazer: Webcam Eye Tracking for Remote Studies of Web Search. Proceedings of the 2017 Conference on Human Information Interaction and Retrieval (CHIIR '17), ACM, 2017, pp. 17–26, doi: 10.1145/3020165.3020170.

[27] Spink, A.—Jansen, B. J.—Cenk Ozmultu, H.: Use of Query Reformulation and Relevance Feedback by Excite Users. Internet Research, Vol. 10, 2000, No. 4, pp. 317–328, doi: 10.1108/10662240010342621.

[28] Teevan, J.—Ringel Morris, M.—Bush, S.: Discovering and Using Groups to Improve Personalized Search. Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09), ACM, 2009, pp. 15–24, doi: 10.1145/1498759.1498786.

[29] Tran, V. T.—Fuhr, N.: Using Eye-Tracking with Dynamic Areas of Interest for Analyzing Interactive Information Retrieval. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12), ACM, 2012, pp. 1165–1166, doi: 10.1145/2348283.2348521.

[30] White, R. W.—Chu, W.—Hassan, A.—He, X.—Song, Y.—Wang, H.: Enhancing Personalized Search by Mining and Modeling Task Behavior. Proceedings of the 22nd International Conference on World Wide Web (WWW '13), ACM, 2013, pp. 1411–1420, doi: 10.1145/2488388.2488511.

[31] Umemoto, K.—Yamamoto, T.—Nakamura, S.—Tanaka, K.: Search Intent Estimation from User's Eye Movements for Supporting Information Seeking. Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12), ACM, 2012, pp. 349–356, doi: 10.1145/2254556.2254624.

[32] Umemoto, K.—Yamamoto, T.—Nakamura, S.—Tanaka, K.: Predicting Query Reformulation Type from User Behavior. Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13), ACM, 2013, pp. 894–901, doi: 10.1145/2480362.2480534.

[33] Wildemuth, B. M.—Kelly, D.—Boettcher, E.—Moore, E.—Dimitrova, G.: Examining the Impact of Domain and Cognitive Complexity on Query Formulation and Reformulation. Information Processing and Management, Vol. 54, 2018, No. 3, pp. 433–450, doi: 10.1016/j.ipm.2018.01.009.

[34] Xu, S.—Jiang, H.—Lau, F. C. M.: Personalized Online Document, Image and Video Recommendation via Commodity Eye-Tracking. Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08), ACM, 2008, pp. 83–90, doi: 10.1145/1454008.1454023.

[35] Zhu, G.—Mishne, G.: ClickRank: Learning Session-Context Models to Enrich Web Search Ranking. ACM Transactions on the Web, Vol. 6, 2012, No. 1, Art. No. 1, 22 pp., doi: 10.1145/2109205.2109206.

Alena Martonova received her computer science M.Sc. degree from the Comenius University in Bratislava in 2004, and Ph.D. degree from the Slovak University of Technology in Bratislava in 2011. Her research interests involve human computer visualization and visualization. She is currently Assistant Professor at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

Jozef Marcin received his software engineering B.Eng. and M.Eng. degrees from the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava in 2013 and 2015, respectively. His research interests combine web applications, web search and eye tracking. He currently works as a software engineer in industry, he has developed infrastructure for IoT and Industry 4.0.

Pavol Navrat received his M.Sc. and Ph.D. degrees from the Slovak University of Technology in Bratislava in 1975 and 1983, respectively. He (co-)authored several books and numerous scientific papers. His research involves intelligent information systems, software engineering and artificial intelligence. He is currently Full Professor at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

Jozef Tvarozek received his computer science M.Sc. degree from the Comenius University in Bratislava in 2007, and Ph.D. degree from the Slovak University of Technology in Bratislava in 2011. In research, he is interested in eye-tracking and online learning of programming via active problem solving. He is currently Assistant Professor at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

Gabriela Grmanova received her computer science M.Sc. degree from the Comenius University in Bratislava in 1998, and Ph.D. degree from the Slovak University of Technology in Bratislava in 2005. Her research interests involve data mining, pattern recognition and mathematical modelling. She is currently Assistant Professor at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.