Proof of concept and evaluation of eye gaze enhanced relevance feedback in ecological context

Vaynee Sungeelee, Francis Jambon, Philippe Mulhem
vaynee.sungeelee@etu.univ-grenoble-alpes.fr,Francis.Jambon@imag.fr,Philippe.Mulhem@imag.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

The major method for evaluating Information Retrieval systems still relies nowadays on the "Cranfield paradigm", supported by test collections. This sheds light on the fact that human behaviour is not considered central to Information Retrieval. For instance, some Information Retrieval systems that need users feedback to improve results relevance can not completely be evaluated with classical test collections (since the interaction itself is not a part of the evaluation). Our goal is to work toward the integration of specific human behaviour in Information Retrieval. More precisely, we studied the impact of eye gaze analysis on information retrieval. The hypothesis is that acquiring the terms read by a user on the result page displayed may be beneficial for a relevance feedback mechanism, without any explicit intervention of the user. We have implemented a proof of concept which allows us to experiment with this new method of interaction with a search engine. The contributions of our work are twofold. First, the proof of concept we created shows that eye gaze enhanced relevance feedback information retrieval systems could be implemented and that its evaluation gives interesting results. Second, we propose the basis of a evaluation platform for Information Retrieval systems that take into account users behaviour in ecological contexts.

CCS CONCEPTS

• **Information systems** → **Query reformulation**; *Test collections*; **Users and interactive retrieval**.

KEYWORDS

Relevance feedback, eye tracking, user behaviour, ecological context, proof of concept.

1 INTRODUCTION

One fundamental concern in Information Retrieval (IR) raises the question: what makes documents relevant to an information need [15]. Since the 70's, the major method for evaluating Information Retrieval systems, and therefore checking if a system provides relevant documents, relies heavily on the "Cranfield paradigm" [12], supported by test collections such as TREC[1]. These collections consists of a set of documents, a set of queries, and assessments corresponding to relevance judgements. Queries are chosen and written by experts, whereas the relevance of documents are also evaluated by experts.

However, such approach does not considers specific aspects related to human (see [12]), and does not tackle Web searches:

- only the first few snippets (document excerpts) are really considered by a user looking at a Search Engine Result Page (SERP) [3];
- actual document relevance assessment by users is a sequential two stages process: a user first looks at snippets, and then may consult the corresponding documents [15]. This is not really consistent with classical assessment, where experts are passing through full documents to check relevance;
- the behaviour of users changes and adapts to the quality of a the search engine [3];
- a real life Web search usually does not consist of a single query, but is composed of a set of progressively manually refined queries [9].

Our goal is here to complement classical IR systems evaluation via test collections, by adding some of the specifics of human behaviour to the evaluation method. Formally speaking, our objective is to search for human behaviour indicators that could have a positive or negative impact on the efficiency of search engine at large, and to promote their usage in addition with test collections.

To do so, we develop an original instrumented platform that mimics a classic Web search engine. Such a platform is configurable to work with research (i.e. Terrier) and commercial (i.e. Qwant) search engines. The platform could also be tuned to implement adhoc snippet generator and relevance feedback engine. To analyse user behaviour, the platform could collect user's actions and his/her perceptions –via of the shelf eye tracking system– of the result page, at different levels of granularity. Moreover, the platform could be deployed simply, in a way to allow user evaluation at a large scale in ecological context.

The concept of "ecological context" is widely used in research on the design and evaluation of user interfaces. For instance [11] proposes the following definition: "*the ecological context is a set of conditions for a user test experiment that gives it a degree of validity. An experiment with real users to possess ecological validity must use methods, materials, and settings that approximate the real-life situation that is under study.*"

Our first implementation of this platform –described in this paper– is a mock-up of a search engine enhanced with eye gaze assisted relevance feedback. More specifically, the search engine analyse user visual behaviour and try to refine user search intention. Such specific IR system could not be evaluated with test collections only since the user's feedback is a key element used by the IR system to improve the relevance of the documents returned.

[1] https://trec.nist.gov

This paper is structured in 5 sections. In this first section we provide an introduction to the general scientific motivation for the platform. Then, in the next section, we present the eye gaze enhanced relevance feedback use case. In the third section, we describe the design and implementation of the proof of concept. Next, the fourth section provides and discusses the results obtained after testing this proof of concept with user experiments. Finally, we conclude and propose future work in the fifth section.

2 EYE GAZE ENHANCED RELEVANCE FEEDBACK

In classical Web information retrieval systems, a user's query is used to filter and sort a corpus of documents, returning a list of results ordered by decreasing order of relevance. This list of results is collectively called a Search Engine Results Page (SERP), where each result is usually composed of a summary of the corresponding document, called a Snippet.

As with any information retrieval system, a Web search engine does not provide only relevant results in a SERP. Reasons for this are numerous: polysemy of query, indexing of documents, matching, etc. In addition, it can be difficult for users to phrase what they are looking for until they see the results [2]. Usually, Web retrieval systems do not provide simple ways for a user to give his/her feedback during the search process. To improve this, some Web search engines use the clicks on the displayed snippets in a SERP as relevance clues.

Such user explicit feedback can be used for relevance feedback, by modifying the initial query, with the expected consequence of improving the relevance of documents returned by the system. This use is consistent to IR literature [10, 19] which indicates that user feedback tends to improve the overall quality of the search. However, asking for explicit feedback can be a burden to the user [6], especially in the context of Web search where users are looking for fast and simple interactions (i.e. by providing very short queries and by looking at very few results in SERPs).

Therefore, we believe that automatically interpreting at a fine grain user behaviour while reading a SERP, thus allowing precise implicit feedback, is a promising approach. More precisely, our hypothesis is that the analysis of user perceptions –via analysis of his/her eye movements– and actions while reading a SERP could be used to implement an effective relevance feedback mechanism. Our objective is to implement and validate this hypothesis in an ecological context.

Literature shows that the acquisition of eye movements is indeed an interesting means of personalising information retrieval. For instance, in a Critiquing-Based Recommender System, Chen and Wang [7] study shows the feasibility of inferring users' feedback based on their eye movements. Buscher *et al* [6] verified that analysing the display time of documents while scrolling provides valuable data for retrieval purposes. They also captured eye movements over single lines of text in documents. In another study [5] they show a relation between eye movement measures and user-perceived relevance of read text passages in documents. They also demonstrate the effect of using reading behaviour of documents as implicit relevance feedback. Their studies concluded that these methods can significantly improve retrieval performance. This raised the question of whether analysis on a finer grain –words instead of lines– could lead to even better results. More recently, Y. Chen *et al.* [8] proposed the analysis of documents at the word granularity and concluded that this level of analysis was in fact a good idea. However, at we said, they still deal with full documents and not snippets in SERPs.

Closer to our hypothesis is the work of Eickhoff *et al.* [9], in which users use a search engine to answer given questions, and reformulate their queries several times to refine them. Eye movements analysis showed that there is a close link between the words used to reformulate the query, and those read in the SERP. However, Eickhoff *et al.* work has an explanatory purpose, i.e. the reformulation process is performed by users themselves, and there is no relevance feedback mechanism proposed.

Our current research draws elements from our previous work in Albarede *et al.* [1], which involved studying how eye gaze information could be used to provide a relevance feedback mechanism at the granularity of words in SERPs. We used the following assumptions: (1) the word read for the longest duration undergoes a deeper cognitive process; (2) the last word read before clicking on a document is the one which triggers the decision to select that document. These assumptions lead to the definition of two corresponding metrics: (1) the word read for the longest duration in a snippet, and (2) the last word read by users. We associated with each word of SERP the notion of positive, neutral and negative feature that reflects the potential contribution of each term in the relevance feedback. In the experiment, users were asked to choose the most relevant snippet for given queries. We found that (a) the detection of terms read by users in snippet is sensitive to eye tracking system hardware performance and can be fairly precise with high end devices; (b) the last and mostly the longest read terms are relevant to assess the relevance of a document and (c) while positive terms could give interesting clues to the relevance of a document, the metrics that were proposed to use negative terms were inconclusive.

3 PROOF OF CONCEPT

Our previous research objectives in Albarede et al. [1] was to identify useful metrics from the analysis of user behaviour with a SERP, and to derive optimal parameter settings for relevance feedback. Our present objective is to demonstrate –by a proof of concept– that a search engine enhanced by implicit relevance feedback driven by eye gaze analysis could be implemented and used in ecological context. Such enhanced search engine is then be used to experimentally evaluate whether the quality of a search engine is improved in the real world with the metrics and optimal setup we identified in [1]. To our knowledge, there is currently no working information retrieval system application which implements an implicit relevance feedback mechanism assisted by the analysis of eye movements (at word level) described in the literature.

The core hypothesis of this Proof of Concept (PoC) is that eye gaze analysis can identify words to be used for the implicit relevance feedback mechanism. Therefore our objective is, on the one hand, to design and implement an application that mimics a Web search engine, and on the other hand, to analyse user interactions –actions and perceptions– with he the search engine user interface as a means for implicit feedback to reformulate user queries. We

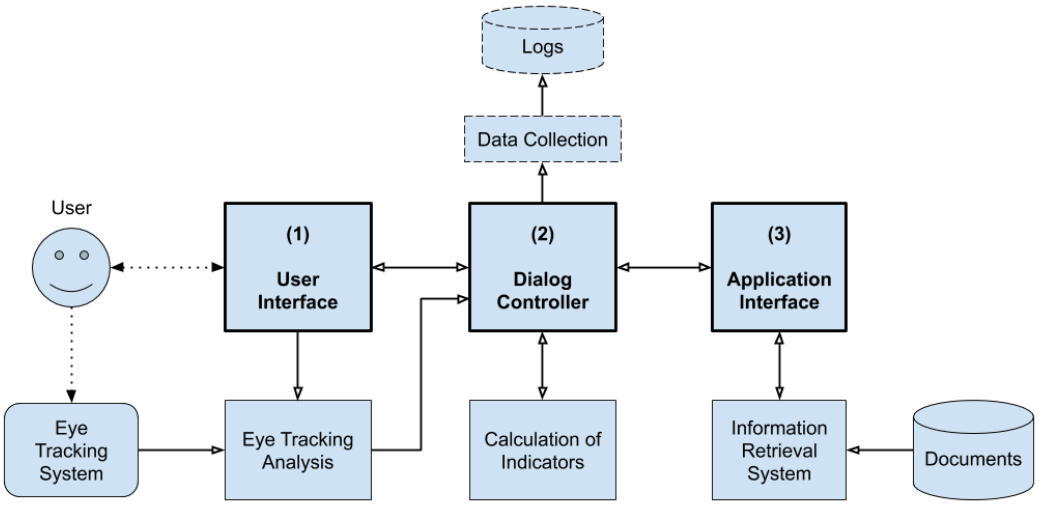

Figure 1: Software architecture overview of the application.

implement metrics derived from [1] to study their usefulness in an almost real context of information retrieval usage.

In short, we aim at evaluating whether the relevance feedback assisted by the analyse of eye gaze increases the precision of the expanded query results in the real world. The implementation of such a PoC raised questions that are tackled in the following subsections.

3.1 Software architecture overview

To structure the application, we make use of the Seeheim software architecture model [17], a reference model in Human Computer Interface domain. The Seeheim model is a high-level model which was designed for single user systems with a graphical user interface. It leads to an application whose components –the graphical user interface and/or the functional core– can easily be replaced if a different implementation is required. The Seeheim model splits the application into 3 main parts: the User Interface, the Dialog Controller and the Application Interface. The User Interface manages the user inputs and the outputs of the application. The Dialog Controller is a mediator between the user (via the User Interface) and the functional core (via the Application Interface) and is responsible for defining the structure of the exchanges between them. The Application Interface defines the software interfaces with the processes that can be initiated by the Dialog Controller.

Separating the User Interface from the rest of the application preserves the application from modifications of the User Interface (e.g. changing the display of snippets). Similarly, changes in the process (e.g. modification of the search engine) are hidden to the rest of the application by the Application Interface. At last, evolutions of user interaction (e.g. modification of the indicators used for relevance feedback) do not require significant changes to the rest of the application, since the Dialog Controller and the Application Interface live in their own separate components.

The PoC is implemented in Java as a standalone application. We have chosen not to develop a plugin in a Web browser to remain independent of the evolutions of Web browsers. The software architecture of the application (see fig. 1) is composed of the 3 main components of the Seeheim model: (1) the User Interface, (2) the Dialog Controller, and (3) the Application Interface. With the User Interface is associated an Eye Tracking Analysis component which role is to managed the eye gaze analysis. With the Dialog Controller are associated two separate modules, the Calculation of Indicators component which role is to refine the query, and the Data Collection component that store data (user actions and perceptions, queries, results) exchanged between User Interface and Application Interface in a log file for analysis purpose. These modules are detailed in the following subsections.

3.2 User Interface and Eye Tracking Analysis modules

The User Interface aims at mimicking classical Web search engines: this interface features a text field for the query and a "Search" button. Once the query is processed by the IR system, a SERP (composed of snippets) is displayed, with a "Refine" button at the right of each snippet (see fig. 2). For its implementation, we used the Java Swing graphical widget library.

The Eye Tracking Analysis module aims at detecting the zones viewed by the user. Any graphical element of the user interface, i.e. any widget, could potentially be defined as a zone. In our current implementation, zones only refer to text elements of the result page, i.e. snippets and their titles. Since the indicators are based on the semantic of texts, documents' URL were excluded. Zones are defined as rectangles around one word or a list of words (e.g. a entire snippet). Each word zone is represented by a bounding box around the word. For a list of words, the zone is defined as the union of the words it contains. The Eye Tracking Analysis module tracks theses zones, named Areas of Interest, in the SERP displayed by the User Interface (see fig. 3), and detects when user eye gaze is inside one of these zones. The Eye Tracking Analysis module is

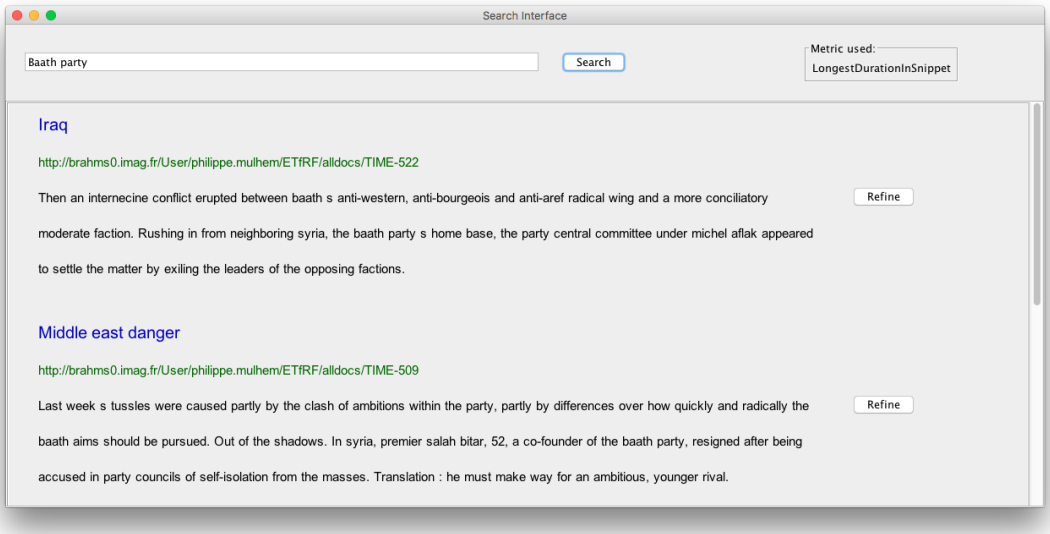

Figure 2: Simulated search engine Web page user interface.

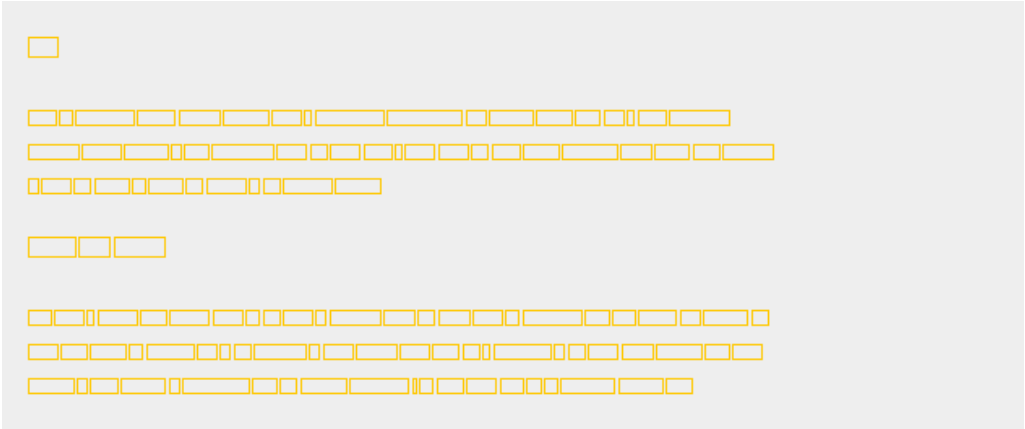

Figure 3: Areas Of Interest defined around words of figure 2.

also responsible for information exchange with the Eye Tracking System and with the Dialog Controller module. For exchanging messages between the Eye Tracking System and the application, we used the Usybus library which is based on Ivy library. Usybus is a framework that associates a type to messages so that devices know which messages to subscribe to, and Ivy[2] is a middleware which facilitates data exchange between applications on a network.

3.3 Dialog Controller, Calculation of Indicators and Data Collection modules

The Dialog Controller, as specified in the Seeheim model, has a central role in the application operation. This component is in charge of managing the sequence of communications between User Interface and Application Interface during query/results operations and is responsible of the dissemination of information to other modules.

The Calculation of Indicators component implements the algorithm for the estimation of the metrics. These metrics are used to determine the new query when the user request a refinement of the current results. We created an abstract class to manage the metric to be used. Subclasses are implemented for each metric. The name of the metric to use is provided as a String object, possibly from a configuration file and the corresponding class is initialised in the Dialog Controller when the application is launched. This allows better management of multiple metrics and makes them easier to test.

The Data Collection component is in charge of recording data for the experiments. In order to facilitate the testing process further, we

[2]https://www.eei.cena.fr/products/ivy/

can trace the program execution by creating two log files. The first log file keeps track of requests as they evolve during the refining mechanism of the application. The purpose this log is to keep a trace of expanded queries to analyse them with information retrieval evaluation measures. The second log file records viewed words and user actions. This latter contains a comprehensive view of the application execution and can be consulted for debugging purposes.

3.4 Application Interface and Information Retrieval System

The Application Interface component is responsible for establishing the network connection to the Information Retrieval System and contains data structures to represent the query and the corresponding results. This structure contains the query, with each result consisting of a document id, a title, an URL and a snippet. Data exchange between these modules are in XML defined by an internal DTD declaration.

We use the Terrier V4.0[3] [14] information retrieval platform, an open source platform which we adapted with Python and Perl add-ons to retrieve queries, generate snippets from documents and get back the response constituting the SERP in XML format. The snippet generation (see Algorithm 1), is inspired by [4, 18]: it consists in finding the text window of size *lmax*, from a document *doc*, that contains the larger amount of query terms *wqset*. This generator assumes that the interest of a snippet only depends on the query terms occurrences. Other additional elements, like the topical link between documents words and the query or the impact of the snippet for query disambiguation, may be used in the future.

Algorithm 1: Snippet generator (simplified).

```
Data: document source text : doc;
      query terms set : wqset;
      length max of snippet : lmax.
Result: The excerpt for the document source
initialization : wdoc ← split doc in words
p ← 0
curr_mscore ← 0
mp ← 0
while p < length(wdoc)-lmax do
    curr_score ← sum of query words occurrences in
     wdoc[p, p+lmax-1]
    if curr_mscore < curr_score then
        curr_mscore ← curr_score
        mp ← p
    end
    p++
end
Return wdoc[mp,min(length(doc),mp+lmax)]
```

4 USER EXPERIMENTS

The objectives of theses user experiments are twofold. First, we aim to make sure that the experimental configuration is technically effective, namely that the words are correctly detected by the eye movements analysis (i.e. "Functional tests"). Secondly, we want to evaluate whether the relevance feedback mechanism we propose effectively increases the relevance of the query results (i.e. "Relevance tests").

4.1 Experimental Setup

The experimental setup (see fig. 4) is composed of a classic desktop computer configuration with central unit, monitor, keyboard and mouse. The eye tracker device is attached under the screen, and does not have any impact on the natural interaction of the user with the search engine. Our PoC simulates modern Web search engine user interfaces to give users a real search engine experience. It is important to provide such an ecological setting to be able to analyse user behaviour with this new IR system.

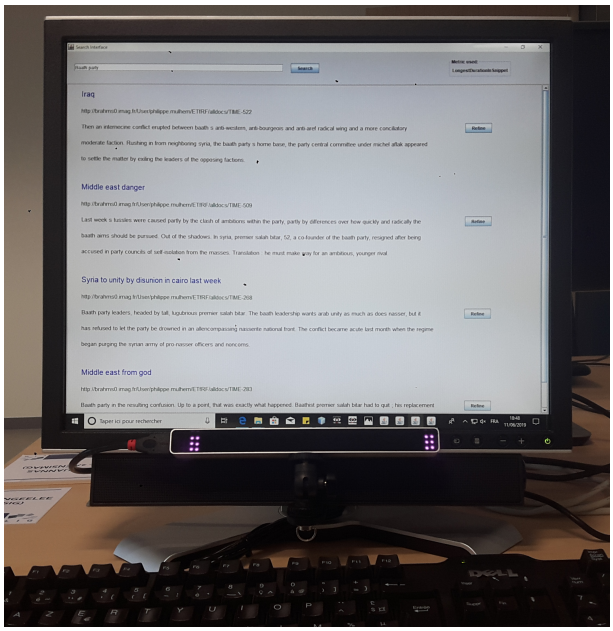

Figure 4: Experimental setup showing the monitor with the eye tracker device attached under the screen.

The query results are displayed as in typical SERPs, with a clickable URL which allows the user to consult documents (see fig. 2). The text of the SERP is in Arial 15 font as it was shown to allow good reading by the user and good detection by the eye tracking analysis [1]. The display parameters, such as font-style, font-size and text colours are customizable. We also provide a refining mechanism per snippet by adding a "Refine" button next to each snippet. After scanning the SERP, the user can identify a relevant snippet and refine his/her original query modification by clicking on "Refine". On our case, this adds a relevant word to the original query in the search bar and automatically relaunches the search. This relevant word is chosen based on a configurable metric, for instance: the longest fixation in the selected snippet, or the last fixation in the SERP.

[3] http://terrier.org

For the eye tracker device, we opted for the Eye Tribe ET1000[4], which is a low cost and popular device for human-computer interaction experiments. It has a sampling frequency of 30 Hz and an average accuracy between 0.5 to 1° of visual angle, which corresponds to an on-screen average error of 0.5 to 1 cm if the user sits about 60 cm away from the monitor. It has an acceptable precision for fixation analyses provided it is properly calibrated and tested in a proper setup. We used a 1280x1024 resolution 19" LCD monitor. The participants were instructed to keep a fixed position for best results and maintained a distance of 60-70 cm from the monitor during the experiment. The calibration error for participants varied between 0.37° and 0.48°, an acceptable range for reading research, with 0.5° being the maximum acceptable value [13].

The corpus of documents used is the TIME[5] collection, which consists of queries on articles from *Time magazine*. This collection is rather small, but adequate for a proof of concept. We experimented the following indicator: the longest fixation duration in the selected snippet. This indicator was founded to be effective by [1], enabling up to 87% of success in identifying positive words.

For the experiments, the task protocol is as follow: each participant is asked to run two different queries (chosen from a set of three possible queries); then, for each query, he/she looks through the SERP and chooses the snippet containing information he/she considered relevant to the query. The duration of an experiment is about 10 minutes per participant. The results of the experiments are then analysed thru the logs recorded by the Data Collection module.

We conducted experiments for a total of 9 participants. Due to eye tracking device limits (the device fails to detect eyes), only 7 of the participants were retained for analysis. We agree that, for now, the small number of participants does not allow us to conclude definitively on the question of whether or not eye tracking improves the relevance feedback mechanism of search engines, but is adequate for a first evaluation of this proof of concept.

4.2 Functional tests

The purpose of this first experiment was to evaluate the eye tracker ability to correctly detect the words users gazed at. Users were told to posed the query, and then to search in the SERP a specific word (target). Once they found it, they were advised to click immediately on the "Refine" button. Since searching a specific word involves a cognitive effort, using this protocol tends to simulate the user activity to look for words in snippets that could help him/her to asset the relevance of documents. The results (binary values) indicate whether or not the target was detected. For this experiment, the Calculation of Indicator module correctly detected the target for 3 out of 7 participants.

These limited results could be explained in two ways. From a technical point of view, the eye tracker device used, the Eye Tribe ET1000, is known to perform differently for a variety of participants and environmental conditions. For instance, the device is very sensitive to light conditions. Moreover, the tracking box, i.e. the area in which the user's head must stay to allow the detection of his/her eyes, is fairly small (30x40 cm). So, even though the participants usually sat attentively during the whole calibration, they could shifted position unconsciously during the actual experiments. This may have caused lost of gaze tracking. In addition, and on a behavioural point of view, some participants identified the target word and clicked on the "Refine" button before the end of reading of the word. In that case, this counteracted the metric used and the Calculation of Indicator module did not return the correct word. However, even if they are not particularly good, these results are fairly consistent with our previous findings [1] in which 3 out of 6 correct words have been detected with this eye tracker device.

4.3 Refinement tests

In order to evaluate the relevance of the expanded queries generated thanks to eye movements analysis, we used information retrieval evaluation measures based on recall and precision. The goal of this second experiment was to verify whether better results could be obtained with the expanded queries.

After posing the query, participants were asked to judge the most relevant result pertaining to their information need on the SERP. To do so, they have to look at one word which helped identify the most relevant result in a given snippet, and then click on the corresponding "Refine" button. Each user query is then expanded with the term which received the longest attention, gathered by implicit feedback. In case this term is already present in the original query, the next best term is considered.

To evaluate a query performance with respect to its initial performance before expansion, we use the following evaluation measures [16]: Precision at 5 (P@5), Precision at 10 (P@10) and Reciprocal Rank (RRank), each evaluating a different aspect of search engine performance. Thus, to compare relevance scores for the expanded query with the initial query, we calculate the measures above-mentioned and verify which ones yield an improvement.

P@5 corresponds to the number of relevant documents among the first 5 documents and P@10 corresponds to the number of relevant documents among the first 10: *Precision@k = (# of results @k that are relevant) / k*. The reciprocal rank is a statistical measure which takes the order of correctness into account and evaluates the result lists of a sample of queries. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer: 1 for first place, 1/2 for second place, 1/3 for third place and so on. We do not use measures such as Average Precision (AP) or Mean Average Precision (MAP), as they are not appropriate in our case because we only focus on the top results.

The results obtained are detailed for each of the three queries selected in table 1. The initial scores of each query –without expansion– is given on the first line for P@5, P@10 and Reciprocal Rank followed by their corresponding scores after expansion with the given term. A (+) sign next to each score denotes an improvement compared to the initial query's corresponding score. Similarly, (-) indicates a decrease and (=) indicates that the score has not changed.

As stated before, even if our experiments do not consider a large number of events to conclude, we believe that these results give interesting clues about the expected performance of this relevance feedback mechanism: 4 out of 7 experiments show improvements

[4]https://theeyetribe.com
[5]http://ir.dcs.gla.ac.uk/resources/test_collections/time

Query	Term added	P@5	P@10	RRank
Baath party	-	0.0000	0.0000	0.0164
	settle	0.0000 (=)	0.0000 (=)	0.0244 (+)
	self-isolation	0.0000 (=)	0.0000 (=)	0.0227 (+)
U.S. policy toward South Viet Nam	-	0.0000	0.1000	0.1429
	conference	0.0000 (=)	0.1000 (=)	0.1000 (-)
	Military	0.0000 (=)	0.0000 (-)	0.0714 (-)
	misinformed	0.0000 (=)	0.1000 (=)	0.1429 (=)
Ceremonial suicides of buddhists monks	-	0.0000	0.0000	0.0227
	automobile	0.2000 (+)	0.1000 (+)	0.3333 (+)
	school	0.2000 (+)	0.1000 (+)	0.2000 (+)

Table 1: Precision@5, Precision@10, Reciprocal Rank scores before/after expansion; A (+) sign denotes an improvement compared to the initial query's corresponding score, (-) indicates a decrease, and (=) indicates that the score has not changed.

after query expansion for at least one of the scores; and 2 out of 7 experiments show a degradation of performances.

We also note that the results are not uniform among the queries. The third query expansion has positive impact on all measures, which is not the case for the others. These finding are not really a surprise in the domain, in which many elements impact the quality of the result. So, it seems that the query strongly matters, may be due to the query topic, the query formulation, the snippet generation, the nature of the documents, etc. We do not have enough data here to clearly identify the element(s) that cause this disparity.

4.4 Comparison with our previous research presented in Albarede *et al.* 2019 [1]

In our previous research [1], we obtained significant higher results: it was found that up to 87% of words looked at in snippets were positive words, and we have obtained improvements in the relevance after query expansion for only 4 out of 7 experiments (57%). However, theses results should not be compared directly.

First of all, the element of comparison is not the same depending on the approach. In [1] we compared detections of positive terms, whereas for this PoC we have compared improvements in the relevance after query expansion. However, the two results could eventually be compared if we make the assumption that a positive word added to a query systematically improve the relevance of the results after query expansion. It is probably not always true, but on the contrary, non-positive words may also improve the relevance of the results after query expansion.

Moreover, the eye tracking device used in the two approaches is different. In [1] we were using a Tobii Pro X3-120[6] device, a professional class device, while for the PoC we use a Eye Tribe ET1000 device, a consumer class device. We have chosen this latter device for the PoC because this low cost eye tracker represents a class devices that could be used by end-users in an ecological context.

As in [1] we compared these two devices at a functional level in a preliminary experiment, it is possible to extrapolate from the latter results if the less efficient eye tracker device was used to detect positive words. If we assume that a positive word will actually improve the query, we can verify if [1] results matches our PoC results.

In [1] preliminary experiments showed that the ET1000 could detect a correct word for 3 out of 6 participants, where a X3-120 could detect 5 out of 6. To compute the percentage of positive words that could be expected for the ET1000, we multiply the results obtained for X3-120 by the ratio of detection performance for the ET1000 to detection performance for X3-120, obtained from [1] first experiment:

$$\begin{aligned}\text{Res(ET1000)} &= \text{Res}_{[Albarede\ et\ al.\ 2019]}\text{(X3-120)} \\ &\quad \times \frac{\text{Detect}_{[Albarede\ et\ al.\ 2019]}\text{(ET1000)}}{\text{Detect}_{[Albarede\ et\ al.\ 2019]}\text{(X3-120)}} \\ &= 87\% \times \frac{3/6}{5/6} \\ &= 52\%\end{aligned}$$

Such extrapolated results for the Eye Tribe gives 52% of correct word detection, which is close to the results we actually obtain ($4/7 \approx 57\%$, see subsection about Functional Tests). Even if we do not have enough data to draw a definitive conclusion, we estimate –subject to the limitations of our assumptions– that our results are consistent with our previous research [1].

Corollary, this means that with a more efficient eye tracking device, results about 87% of positive results, could probably be obtained, and as a consequence, better performance of eye gaze enhanced relevance feedback could be achieved.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented the implementation and the user evaluation of a proof of concept for a novel search engine enhanced with eye gaze assisted relevance feedback. To our knowledge, there is no similar implementation in the literature. We showed that: (a) there is a potential benefit of using eye gaze analysis as implicit relevance feedback method ; (b) the results we obtained are consistent with those we previously obtained in Albarede et al. [1] and so better performances could be expected in the future. Because of the limited number of participants (7 users) and tasks (2 out of 3 queries per participant), the results of this study are indicative only,

[6] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

but we noted a tendency for the number of relevant documents to increase after query expansion.

Our prototype and the experimental setup suffers from some limitations that could had negatively infer with the results we obtained. One technical limitation was that the participant had to keep his/her head relatively still to get good eye gaze detection with the eye tracking device we used. While this might not be a realistic setup for search engine use in ecologic context, the experiments showed that word detection is possible and could probably significantly be improved with more robust (better tracking box) and more precise (better angular precision) eye tracking devices. In a future implementation, we will consider replacing the Eye Tribe ET1000 with a more powerful eye tracker device, such as Tobii Pro X3-120 to yield better results.

Another limitation deals with the precision of eye gaze tracking. Most devices have an average accuracy between 0.5 to 1° of visual angle, which corresponds to an on-screen average error of 0.5 to 1 cm if the user sits about 60 cm away from the monitor. For short words and with usual character size, this spatial accuracy does not allow to make the distinction between two short words. This limitation have no simple answer since it is mostly linked to the human fovea size. Increasing screen and character size could be a solution, but these answers alterate the ecologic validity of the context. This is why a certain degree of uncertainty still remains in the detection of words read by users, and this aspect must be taken into account in the use of this technique.

In addition, the corpus of documents used –TIME collection– was not ideal for user tests, given that they cover ancient historical events only: users might not have been able to understand the context to these information needs. In a future user experiment, it would thus be desirable to have a collection that is of a more general and recent nature.

It will also be interesting to study and implement other metrics to test their usefulness in different information search contexts. Another possible pathway worth exploring would be testing new snippets generators, e.g. the generators provided by Terrier V5 or Apache Lucene[7] as the words in documents selected by the snippets generator may have a significant impact on words that could be viewed by users, and as a consequence, on metrics used.

Another experimental track could be to explore situations even closer to usual user searches on the Web, for instance when a user make multiple queries for the same information need topic.

We proposed here the basis of a modular platform for the evaluation of information retrieval systems that take into account both user behaviour and classical test collections. Ideally, if classical search engines were providing standardised SERP, it should be usable on any engine. A more concrete way to integrate existing systems will be to provide tunable *wrappers* to adapt simply to any search engine. Extensions could integrate task oriented sequences of queries (with document display tracking) so that other features may be provided.

[7] https://lucene.apache.org

ACKNOWLEDGMENTS

The research presented in this article was partly funded by the Gelati Emergence project of the Grenoble Informatics Laboratory (UMR 5217).

REFERENCES

[1] Lucas Albarede, Francis Jambon, and Philippe Mulhem. 2019. Exploration de l'apport de l'analyse des perceptions oculaires : étude préliminaire pour le bouclage de pertinence. In *COnférence en Recherche d'Informations et Applications - CORIA 2019, 16th French Information Retrieval Conference. Lyon, France, May 25-29, 2019. Proceedings*. https://doi.org/doi:10.24348/coria.2019.CORIA_2019_paper_1

[2] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: a survey. *Information Processing & Management* 56, 5 (2019), 1698–1735.

[3] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[4] Lorena Leal Bando, Falk Scholer, and Andrew Turpin. 2010. Constructing query-biased summaries: a comparison of human and system generated snippets. In *IIiX*.

[5] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst. 2012. Attentive Documents: Eye Tracking As Implicit Feedback for Information Retrieval and Beyond. *ACM Trans. Interact. Intell. Syst.* 1, 2 (Jan. 2012), 9:1–9:30. https://doi.org/10.1145/2070719.2070722 http://gbuscher.com/publications/BuscherDengel12_AttentiveDocuments.pdf.

[6] Georg Buscher, Ludger Van Elst, and Andreas Dengel. 2009. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 67–74.

[7] Li Chen and Feng Wang. 2016. An Eye-Tracking Study: Implication to Implicit Critiquing Feedback Elicitation in Recommender Systems. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. Association for Computing Machinery, Halifax, Nova Scotia, Canada, 163–167. https://doi.org/10.1145/2930238.2930286

[8] Yongqiang Chen, Peng Zhang, Dawei Song, and Benyou Wang. 2015. A real-time eye tracking based query expansion approach via latent topic modeling. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1719–1722.

[9] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 13–22.

[10] Liana Ermakova and Josiane Mothe. 2016. Document re-ranking based on topic-comment structure. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 1–10.

[11] Francesco Bellotti, Riccardo Berta, Alessandro De Gloria, and Massimiliano Margarone. 2008. Widely Usable User Interfaces on Mobile Devices with RFID. In *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, Joanna Lumsden (Ed.). IGI Global, Hershey, PA, USA, 657–672. https://doi.org/10.4018/978-1-59904-871-0.ch039

[12] Donna Harman. 2010. Is the Cranfield Paradigm Outdated?. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. ACM, New York, NY, USA, 1–1.

[13] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

[14] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. 2012. From puppy to maturity: Experiences in developing Terrier. *Proc. of OSIR at SIGIR* (2012), 60–63.

[15] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American Society for Information Science* 48, 9 (1997), 810–832.

[16] IC Mogotsi, Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to information retrieval. *Information Retrieval* 13, 2 (2010), 192–195.

[17] Günther E Pfaff et al. 1985. *User interface management systems*. Vol. 1. Springer.

[18] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. 2–10. https://doi.org/10.1145/290941.290947

[19] ChengXiang Zhai and Sean Massung. 2016. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool.