Understanding Search Engines: Requirements for Explaining Search Results

Ntoulas Alexandros, Stamou Sofia, Tzagarakis Manolis, Christodoulakis Dimitris

Computer Engineering and Informatics Department
and Computer Technology Institute
University of Patras
Building B΄ GR 26500
Patras, Greece
{ntoulas, stamou, tzagara, dxri}@cti.gr

SUMMARY
There are three different groups that use Commercial Web Search Engines: the Developers, the Evaluators and the End-Users. Each group has different information needs and applies different criteria when examining the retrieved documents. Most Search Engines attempt to measure retrieval performance providing figures of recall and precision, indicative of the quantity of the obtained information but saying too little about the quality of this information. In this paper we present a survey of the requirements that each user group has and we propose a generic framework, independent of the details of the underlying Search Engine. Our aim is to provide users with explanation utilities regarding qualitative information of the returned documents. This study's motivation emerged from real life experience acquired during the development of a Web Search Engine for Greek and our purpose is to explain the most usual difficulties in user understanding of Search Engines' operation.

KEYWORDS : Information Retrieval, User Models, Evaluation Criteria, Search Engines.

INTRODUCTION
The rapid growth of information networks and especially the World Wide Web (WWW) has made accessible a huge quantity of both structured and unstructured information, which is available to a variety of users having different information needs. However, user's point of view is underestimated in the design as well as in the evaluation of Information Retrieval (IR) Systems, since in the IR domain, research studies (i.e. TREC experiments) mainly focus on models that support the retrieval of documents that better match the query [6]. Furthermore, the evaluation of an IR system is based on recall / precision figures that only provide quantitative information expressing the ability of the system to find the relevant documents. Despite the fact that the above-mentioned figures can be interpreted and well understood by IR designers and evaluators, the end users and their search interests are not taken into account in the evaluation phase. This point seems to be in conflict with the aim of an IR system, that is to meet users requirements [12].

Moreover, it must be clearly reported that there are three distinct user groups, which actually utilize web search engines, namely the developers, the evaluators and the end users. Although all of them have common expectations (retrieval of information items) from the system, each one applies different criteria prior to issuing a search request and uses different measures while evaluating the obtained results. In parallel, the usability of IR systems is a key concept in human interaction with IR systems and is concerned with making systems easy to learn and easy to use. With respect to usability, the information retrieved needs to be visualized in a non-uniform way for each user group in accordance with their different background and information needs.

In this paper we present a survey of the requirements set by the aforementioned user groups and propose a generic framework, independent of the underlying Search Engine, aiming at providing users with explanation facilities regarding qualitative information of the obtained results. The purpose of our study is to explain the most commonly observed problems in user understanding of Search Engines' operation in order to build a firmer foundation for improving users' interaction with them.

Thus, motivated by the experience we acquired during developing a web search engine for Greek and by focusing on the observation that commercial web search engines provide hardly any explanation concerning the presentation order (ranking) of the retrieved results, we examined the criteria applied by each of the three user groups while seeking for information on the web and we also studied how these requirements affect user's judgments on the performance of search engines.

In the following section (2) we discuss previous research conducted in this area and we continue with a detailed description of the profile of each user group and the role they have in IR tasks (3). In the same section we report on the criteria and requirements set by each user group

individually while performing a search request and we also examine the factors that are of importance for each group while judging the retrieved results. Following on from this, we propose a framework architecture model (4) and illustrate the applicability of this approach on a web search engine. Finally, we conclude with an overall assessment on how user requirements affect the evaluation of IR systems (5).

BACKGROUND

The goal of an Information Retrieval (IR) system is to locate relevant documents in response to a user's query. All of the methods currently used to evaluate performance of information retrieval systems have limitations in their ability to measure how well users are able to acquire information [7] and how they judge the accuracy of the obtained results. Many studies and experiments have been conducted trying to evaluate retrieval performance however, evaluation methodologies adopted tend to address a restricted formulation of the problem, often focusing solely on precision and recall figures and totally omitting users' information needs and how these are met [3]. A traditional level of retrieval evaluation has been to measure users' success at retrieving relevant documents using indices such as recall and precision. While these indices provide some information for determining the quantity of useful information obtained from IR system, they say little about the quality of that information [7].

The accuracy of IR systems is measured by Precision defined as the proportion of retrieved documents that are relevant. But relevance is inherently subjective [17] since relevance judgements are known to differ across judges and for the same judge at different times [14]. Meadow has argued that relevance is not fixed, but changes based on the users' past and current knowledge as well as over time [11]. Apart from the subjective nature of relevance judgements user satisfaction does not elucidate how users interact with or benefit from IR systems. The role of human actors in information retrieval is decisive since such systems are targeted towards satisfying users' information needs. Although, users of IR systems are human actors, the research aiming at developing such systems has often been system-oriented [8]. Moreover little attention is paid on the fact that there are three distinct user groups of IR systems, namely the developers, the evaluators and the end-users each of which applies different criteria while performing a search request and consequently adopts different measurements when judging the retrieved results. Each user group may interpret and handle the same information rather differently and use an IR system inconsistently [5]. In the design of information retrieval systems it would be very useful to know about the different working methods adopted by different human intermediaries. These differences most of the times are strongly related to the experience accumulated by human searchers since they select different search terms to describe a search request and they use different criteria when judging the retrieved results. One of the central issues for retrieval systems design is to support effective interaction between users and other components of the system [13].

To date most information retrieval research experimentation on relevance feedback systems has concentrated primarily on end users and it has been conducted by using static collections of textual information. Moreover, no clear distinction has been widely reported among the interaction each user group has with IR systems and no attempt has been made towards acquiring feedback from all user groups together while using or evaluating such a system. There is therefore an apparent need for carrying out real searching experiments in order to investigate the expectations of each user group while issuing search requests and how these requirements affect relevance judgements of the search results.

In the present paper we claim that understanding of different user groups' information seeking behaviour and expectations should influence information system design and thus should enable the tailoring of information systems better to each groups' needs. Our objective is to study the requirements applied by each user group when performing a search request on a commercial web search engine. The motivation behind selecting a web search engine for our study concentrates on the fact that we are looking for a real operational environment with real users and real search requests, as opposed to static text collections with artificial queries.

Taking into consideration the research conducted so far on how end users interact with IR systems and by focusing on the observation that there are actually three distinct user groups we wanted to examine the requirements issued by each group during performing search requests. In addition, it is of great interest to check and compare the criteria applied by each group while evaluating retrieval performance of web search engines. Our aim was to highlight the major inconsistencies between requirements applied by individual human actors so that future techniques incorporated into IR systems are user-centered thus providing emotive encouragement of the users to modify their queries and understand systems' performance.

In the remaining sections we describe in detail the different approaches adopted by each user group during information seeking and we examine in more detail how each user group interacts with search engines. Our suggestion is that in order to construct more effective retrieval systems and in order to help users understand performance of search engines there is a need for more

knowledge on how users communicate with such systems when they search for information. The assumption behind this study is that an associative relationship exists between the various types of criteria and requirements applied by each individual user group when performing a search request.

CRITERIA AND REQUIREMENTS ADOPTED BY EACH USER GROUP WHILE PERFORMING A SEARCH REQUEST

As pointed out by Belkin [1] information retrieval does not attempt to answer questions or solve problems but it is intended to help users find information that might be useful for those purposes. Information scientists have long acknowledged the fact that many factors contribute to human judgments of relevance in evaluating the effectiveness of IR systems [15]. Yet the field lacks a generally accepted technique for collecting data regarding the criteria or reasons that underline relevance judgments, particularly judgments by all user groups utilizing such systems. This is due to the fact that different parameters are taken into account by each user group when interpreting / evaluating the ranking order of the retrieved results. Lancaster and Fayen [10] listed 6 criteria for assessing the performance of IR systems, namely: coverage, recall, precision, response time, user effort and form of output. Although these criteria are still applicable, there are inconsistencies between those that each user group adopts.

Many experiments have been conducted trying to explain and evaluate the performance of web search engines, but they mainly concentrate on the analysis of traditional recall / precision scores as provided by the engines or by using algorithmic techniques, e.g. terms frequency (TF), inverse document frequency (IDF). In addition, users' relevance feedback is usually collected and analyzed from experiments conducted with static collections of data and with pre-specified query terms totally omitting users' feedback when seeking for information on the web.

Taking into account that there are three different user groups of IR systems we highlight the role each group plays in evaluating search engines and we also examine how the requirements applied by each group should influence the design and implementation of search engines. The main difference among the aforementioned user groups does not concern solely the evaluation but also the working methods adopted by each human intermediary in order to conduct a search request. In particular, since developers of IR systems are familiarized with search engines they count on each user group to search out the information they are looking for. On the other hand end users and evaluators often figure out how to use a search engine based on what they already know. Following on from these end users are mainly interested in getting a result regardless of the system they are working with and consequently they pay little attention to the system's services or help manuals provided.

In the following subsections the role each of the three groups, namely the developers, the evaluators and the end users, plays in IR applications is described aiming at a better understanding of the special characteristics and evaluation methodologies each group applies. The motivation for carrying out such a research came up from real life experience we acquired while developing a web search engine for Greek. More specifically, we observed that each user group adopts incompatible working methods while looking for information on the web and as a consequence different criteria are adopted in order to judge relevancy of the retrieved results. In particular, we noticed that end users are only interested in finding information paying no attention at all to the system's components or help facilities. The only factor they sometimes take into consideration is the form of the information output. On the other hand developers not only use a search engine to find information but are mostly interested in checking the engine's performance in terms of coverage, speed, recall and precision. Finally, evaluators are interested in both factors that is, the engine's performance and finding information and thus they share common characteristics with the other two groups.

The Role of Developers in Evaluating IR Systems

Developers of search engines form the user group, which has great indexing and information storage experience. Thus, selection of search terms is for them as if indexing search requests [8] and while conducting a search request they prefer both free-text terms and descriptors as search terms. Thus, they select carefully their search queries since they are aware of the engine's capabilities. While executing their queries they select both default and advanced modes of the engine (where available), they use Boolean logic operators and they adopt wildcards and field search capabilities where provided by the engine.

Once they have executed their search request and the engine returns a list of documents as relevant they evaluate the results according to the following measurements and criteria. Firstly, they take into serious consideration the recall and precision scores as given by the engine. From an engineering point of view they are interested in recall since they wish to investigate the engine's coverage and in precision since they are interested in the performance of the matching algorithms and techniques incorporated into the engine. In general, developers of IR systems examine the algorithmically ranked order which represents a list of decreasing degrees of objective relevance to a user's information need (query). In addition, they are also interested in the

engine's response time (i.e. between issuing a search command and displaying the first batch of search results on the screen) since it is the indicator of the engine's speed. Developers are not solely interested in finding the desired information but also in testing how the engine works and in evaluating its performance towards their initial design goals. The main goal for using a web search engine they developed concentrates on the fact that they wish to test the performance of its components and not simply satisfy their information needs. That is the main factor that differentiates developers from the other user groups while interacting with IR systems.

The Role of End-Users in Evaluating IR Systems

End users tend to make use of the functions they already know in order to acquire information from the web. Some characteristics of end users are that they are not familiarized with using Boolean operators or advanced search facilities (e.g. field search capabilities) supported by the engine and they almost never read the help manuals provided by the engine. As reported by Carroll [4] end users never read manuals but instead they start using the software immediately since they do not care about the system as such and do not want to spend any time just learning about it. Usually, they read the help manuals only after they have tried several unsuccessful search requests. Nevertheless even though manuals are targeted towards helping users and especially inexperienced ones, they cannot always be easily understood since they contain too specialized terminology restricted to the domain of computers or IR that end users are not familiarized with.

The first step taken by end users when seeking for information on the web is to form a query that defines their information needs. To be able to form a query the end user must initially select / specify some keywords considered by him as representatives of the topic of his search intention. It has been noticed that end users most of the times select single term queries to express their information needs and rarely use phrases or expressions instead. Once the user has executed his query the engine returns a list of retrieved documents that are supposed to be relevant to his search request. The next step requires the evaluation of the retrieved results by the user with respect to his search intention and goal. The way in which end users interpret the retrieved results strongly depends on their past experience and background in using IR systems and on how these results are presented / visualized by the system. After all end users' own request formulation is a representation of his current cognitive state concerned with an information need [9] and consequently they apply subjective evaluation measures based mainly on the notion of relevance accumulated by their experience and background.

Usually end users of search engines are interested in finding the desired information item(s) with the fewest possible clicks on the retrieved links and they pay little attention on the recall / precision scores accompanying each retrieved item. What they usually do is to examine briefly the first few retrieved documents and make their assessments on how well these meet their information needs at that time. They do not take into account how well the query term matches the retrieved documents and they hardly pay any attention to the relevance scores provided by the engine. In case the engine returns zero hits end users either reformulate their query since they assume that they selected wrong or misspelled keywords or quit searching on the particular system and try to satisfy their information needs through another search engine.

The Role of Evaluators in Evaluating IR Systems

Evaluators are experienced users of IR systems, by definition, since they have experience either on IR systems in general, or on evaluating search engines in particular. In both cases, they have established patterns of behaviour and understanding of the system's function.

Evaluators play a twofold role while judging performance capabilities of web search engines. In a cognitive sense they definitely act like end users since in a degree they are subjective to their own assessments especially when they have generated the search requests. Although the evaluator / assessor is supposed to act like an algorithmic entity in order to produce a strictly objective performance baseline, he or she will, to a degree, involuntarily become dependent of interpretations and subjectivity [2]. On the other hand assessors also act like developers since they are interested in the engine's search capabilities determined by recall / precision scores and in the effectiveness of the ranking algorithms. Thus, they express their search requests by using descriptors and text terms as well. Furthermore, they do not restrict their evaluation in the default mode of the engine but they also test advanced facilities where available and they are quite familiarized with Boolean operators, wildcards and field search capabilities supported by various web search engines. Once the results are retrieved they examine the relevance scores provided by the engine but they also make their own relevance judgments based on their experience. In case of zero hits they usually rephrase their query term(s) or select advanced mode(s) of the engine, where available.

One main difference among evaluators and end users is that the former do not terminate their search once they have found the information they are looking for but they keep on examining the retrieved results until coming to a conclusion on the engine's search capabilities. Even though there are no major differences between the

developers and the evaluators groups with respect to the duration of work experience in the field of information services, nevertheless evaluators have clearly shorter experience in using a particular search engine than the developers of the specific system. That is one of the main reason why evaluators while assessing the search capabilities of a search engine do not take into account solely the matching algorithms or precision figures in order to judge relevance of the obtained results. What they also examine is term frequency (TF), inverse documents frequency (IDF), whether the query terms appear on the title or abstract of the retrieved documents and they also make their own judgments on the ranking order of the retrieved documents. Thus, evaluators form a user group sharing common characteristics with both groups mentioned above since they have common requirements and relevance criteria.

PROPOSED FRAMEWORK

Communication among developers, evaluators and end users is often difficult, seldom studied and of paramount importance to design search engines [16]. The main difficulties arise from the fact that each user group adopts different searching behaviours and thus sets different requirements prior conducting a search request. The goal of this study is to improve the quality of a search engine by providing explanation facilities in order to assist all user groups issue their search requests and help them comprehend how the system works.

The framework we propose aims at narrowing the gap between theory and practice of a search engine's operation. It is well known that all search engines use some basic common criteria in order to estimate the relevance of a returned document with the specified query. These criteria normally include the number of occurrences of each query term in the document, its position in it (e.g. title, body, meta-tags etc.), the proportion of occurrences of each term to the total number of terms in the document etc. A sound framework should therefore include facilities that would enable the search engine's users not only to visualize such information but also to compare the various results.

A first attempt towards this direction has been made by Yahoo [19] and most recently by Google [18] where the user has at his disposal some extended facilities for searching or getting specific information about a web page. On the other hand, our aim is not to improve performance of the engine but instead to assist users understand how it works. Thus, some of the components included in the proposed framework are the following:

Query term highlighting. This feature is useful for all three user groups. The developers are interested in knowing the position of the search terms in the document in order to estimate the correctness or the performance of their retrieval and ranking algorithms. On the other hand the evaluators can use it to justify why a web page ranked higher than another. The end users can use it for the same reason as the evaluators and can also use it in order to track their areas of interest within the document. A feature similar to this is already implemented by Google [18].

Word lists/lemma lists/word frequency lists. Although the IDF algorithm is used to determine the representativeness of a query term within a document, such algorithmic measures are meaningful for a restricted and experienced group of users, namely the evaluators and the developers. Conversely such figures can be neither understood nor interpreted by end users since they lack technical background. Each of the proposed word lists functions as follows: word lists include all unique words of a document, lemma lists contain all unique words of a document induced to their first inflectional word form, while word frequency lists comprise of the aforementioned lists accompanied with frequency figures of each term, stemmed or not, in the document. Such lists, facilitate users to understand whether a specified document reports on a specialized subject by examining whether the desired terms exist in a word list, since looking up a term in a lemma list is easier and less time consuming than going through the whole document.

Visualization/graphs of query terms. This feature includes a visual representation of the distribution of the query terms within the returned document. Such a component would be a graphical representation of the document where the query terms existing in it would be represented by small dots, thus giving users an outlook of the distribution of these terms within the document. Another possible visualization can be a line graph of the number of occurrences of query terms versus the overall number of terms in the document.

Multiple algorithm independent ranking. This feature can be used as a benchmark facility in order to evaluate a newly developed retrieval and ranking algorithm. Developers will be able to see whether their implementation is in accordance with other ranking algorithms and can possibly track problems or misconceptions in their implementation. In addition an independent (uniformly calculated) score of relevance between a query and a returned document is not only a powerful tool in the hands of the evaluators, in order to facilitate their work, but also in the hands of the end users in order to assert the "reliability" of a search engine.

Apart from the above components a flexible and modular framework should be system and browser independent, in

which new services could be easily incorporated in cases where new or still untraced requirements arise.

OVERALL CONCLUSIONS

The work reported in this paper examines the working and evaluation methods adopted by each individual user group while using a web search engine. We have also shown that due to the fact that many inconsistencies exist among the criteria applied by each group there is a strong need for the establishment of a firm framework that would enable human intermediaries interact successfully with IR systems thus satisfying their information needs. Some components of the framework are briefly described in a previous section but a lot more needs still to be done towards this direction. Information retrieval and intermediary systems should be able to correspond to the different methods and criteria applied by users with different experiences and roles in IR tasks.

The combination of the requirements into a common framework is a real challenge for designers and developers in order to build a stable human-centered IR system targeted towards all users independently of their background or experience. Based on the observation that the easiest way to teach someone something (using effectively a search engine in our case) is after all to tell him directly we underline the need of establishing a common engine independent framework that would help users understand how the system works and thus be able to use it more effectively.

BIBLIOGRAPHY

1. Belkin, N. Interfaces for Information Retrieval Systems. User Modelling in Information Retrieval Systems. Lecture Material. *First European Summer School in IR* 1990, pp. 274-334.
2. Borlund, P. and Ingwersen, P. Measures of Relative Relevance and Ranked Half-Life: performance indicators for interactive IR. *In Proceedings of ACM SIGIR '98 Melbourne, Australia*, 1998.
3. Brajnik, G., Mizzaro, G., Tasso, C. Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support. 1996. *In Proceedings of ACM SIGIR '96 Conference , Zurich, Switzerland.*
4. Carroll, J.M, Ronson, M.B., The Paradox of the Active User. 1987. *In J.M Carroll (Ed.) Interfacing Though: Cognitive Aspects of Human-Computer Interaction. Cambridge, M.A:MIT Press.*
5. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T., The Vocabulary Problem in Human-System Communication. 1987. *In Communications of the ACM,* Nov. 1987, Vol.30, No. 11.
6. Harman, D., The TREC Conferences. 1995. *In HIM'95 Conference, Konstanz, Germany,* pp. 9-28.
7. Hersh, W.R., Molnar, A., Towards New Measures of Information Retrieval Evaluation. 1995 *In proceedings of the ACM SIGIR '95,* Seattle, WA, USA.
8. Iivonen, M., Searchers and Searchers: Differences Between the Most and Least Consistent Searchers. 1995. *In Proceedings* of the ACM SIGIR '95, Seatle, WA, USA.
9. Ingwersen, P. Polypresentation of Information Needs and Semantic Entities. Elements of a Cognitive Theory for Information Retrieval Interaction. 1994. *In Proceedings of the ACM SIGIR '94, Springer Verlag,* 1994, pp. 101-110.
10. Lancaster, F.W., Fayen, E.G., Information Retrieval On-Line, *Los Angeles, CA.* 1973, Melville Publishing Co. Chapter 6.
11. Meadow, c.T., Text Information Retrieval Systems. Academic Press: San Diego, 1992.
12. Mulhem, P., Nigay, L., Interactive Information Retrieval Systems: From a User Centered Interface Design to Software Design. 1996. *In proceedings of the ACM SIGIR '96,* Zurich, Switzerland.
13. Nordie, R., User Revealement – A Comparison of Initial Queries and Ensuing Question Development in Online Searching and in Human Reference Interactions. 1999. *In proceedings of the ACM SIGIR '99 Conference,* Berkeley, CA, USA.
14. Schamber, L. Relevance and Information Behaviour. Annual Review of Information Science and Technology, 1994, 29:3-48.
15. Scamber, L. Bateman. J., User Criteria In Relevance Evaluation: Toward Development of a Measurement Scale. 1996. *ASIS 1996 Annual Conference Proceedings.*
16. Sonnenwald, H.D., Developing A Theory to Guide the Process of Designing Information Retrieval Systems. 1992. *In Proceedings of the ACM SIGIR '92 Conference,* Denmark.
17. Voorhees, E.M., Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness. 1998. *In Proceedings of the ACM SIGIR '98 Conference,* Melbourne Australia.
18. The Google Toolbar. http://www.google.com/options/toolbar.html
19. Yahoo! Companion. http://edit.yahoo.com/config/download_compantion