Implicit Contextual Modelling for Information Seeking

Ryen W. White, Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow, Scotland. G12 8QQ
{whiter, jj}@dcs.gla.ac.uk

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland. G1 1XH
ir@cis.strath.ac.uk

ABSTRACT

In this position paper we present an overview of our current work in utilising contextual models to help web searchers find relevant web documents. The traditional approach to such information retrieval (IR) assumes that a user's information need is static and does not change as they peruse the documents the search system presents to them. However, this approach is simplistic, and does not consider the dynamic nature of an information need, something that has been well documented. In our approach we use implicit evidence, captured unobtrusively from searcher interaction to model such change. We are particularly interested in the *intentionality* behind this interaction, and which document representations (i.e. summaries, sentences) users view. This evidence is firstly used to enhance the user's query, then automatically update the display and if required, re-search the web.

1. CONTEXT IN INFORMATION SEEKING

Web search systems operate under a simple retrieval paradigm, where a user, with a need for information, motivated by some gap in their current knowledge, seeks the information required to close this gap and hence satisfy their need. Typically, users are expected to express this need via a set of query terms submitted to the search system. This query is then compared to each document in the collection, and a set of potentially relevant documents is returned.

However, web queries are typically short, ambiguous, and are often only an approximation to the searcher's *real* information need. The retrieval operation is a process of query-document inference, where the query infers relevant documents. In essence, an IR system operates on a 'quality-in, quality-out principle', where a query that closely represents the user's real need increases the likelihood that more of the documents suggested by the system will be relevant. Even a query that is a good approximation to a user's real need may still lack some of the terms necessary to adequately discern relevant documents from non-relevant documents. The influence of context on this is unquestionable.

Novice, or inexperienced, users can have difficulty in formulating queries that are sufficiently indicative of their real information need. The problem is amplified where the user's need is vague or 'ill-defined'. In such cases, the user's inability to even outline the nature of their need can lead to problems in both the retrieval and document assessment processes. For such users, perusing even some of the retrieved document set can lead to marked changes in their understanding and/or knowledge of the topic. Typically, the influence of the user's contextual predicament and/or awareness when interacting with IR systems is ignored. Instead, these systems model user need based only on the query terms they submit, relying heavily on the user's ability to formulate such queries. Current IR systems play only a passive role in the development of better defined information needs. In our work we develop systems and use approaches that help users more actively.

The 'interactive revolution' of the early 1980's pointed toward the need to address IR systems from an end-user perspective. Interactive Information Retrieval (IIR) systems are defined as those where the user dynamically conducts searching tasks and correspondingly reacts to system responses over session time.

Despite this shift of focus toward the user, and on increasing the quality of the IR system 'input', it is the traditional approach to IR that many web search systems adopt. The user represents their need in a one-off textual query, submitted to the system, which then processes the query and returns a list of results. This is more commonly known as the 'black-box' approach, shown in Figure 1.

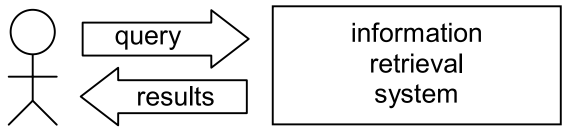

Fig. 1. The 'black-box' approach to information retrieval

Such systems are static, and not aware of the context in which they operate; an abstraction necessitated by the mathematical models that underlie them. The query (often combined with statistical information on the document collection) is the sole factor used in determining output, something that is problematic for those who have difficulty expressing their needs, either because they are ill-defined or they are unsure of what the system expects from them. Capturing implicit evidence from searcher interaction can provide additional information useful in devising a more complete query.

In IIR, context effectively constitutes all factors that influence both the user's and the system's role in the information seeking process, but are not explicitly defined. Contextual influence plays a vital role in relevance assessment, but is not considered by the systems intended to assess relevance on the user's behalf.

IR systems use 'algorithmic relevance' to determine whether a document pertains to a user's information need. Such algorithms assume the form of a query-document matching function, usually producing a ranked output of documents, ranked in descending order depending on their 'relevance'. It has been suggested that neither the system nor the query-document matching function are relevant to the context from which the user directs their query. We can therefore postulate that because their notion of relevance lacks some of the qualities intrinsic for users in determining relevance then the results presented by the system are only an approximation to relevance. Through developing IR systems that take account of the user's context, the *actual* relevance of the recommendations presented by these systems can be improved.

In general, context-aware computing systems are sensor-based, sometimes wearable devices, working through contextual feedback from the environment in which they operate. In IIR, context is traditionally represented through the content of active applications and their relation to user's long term goals and objectives. Typically, no attention is paid to what contextual information user interaction can yield and the use of current goals and search intentions. There is also no guarantee that the information currently being sought will coincide with long term aspirations. Users' behaviour when searching for information can yield much about the contextual influence (i.e. time constraints, user mood) being imposed upon them. Contextually-aware IIR systems take such influence into account, tailoring their responses to suit the current search context. It is important that such systems, as well as reacting to contextual change, are also proactive, recommending future search alternatives to users. Developing the model that underlies such a system is outlined in what follows in this paper.

2. IMPLICIT CONTEXTUAL MODELLING

In our current work we develop a contextual model for information seeking based around user interaction and information seeking behaviour. This model, once complete could be applied to operational retrieval environments, gathering implicitly, information on the context of the search, analysing and applying this to the benefit of the user.

2.1 BACKGROUND

Our work in this area is a continuation of our precursors in web document summarisation [1], implicit relevance feedback [2], and implicit context modelling [3]. The goal here is to find a better means by which we can generate an internal query that best represents contextual influence and a user's changing information need. Then, depending on the degree of change do one of two things, automatically update the results display or resubmit a new query to the search system. For explanatory purposes we combine a description of the model and the system that implements it.

2.2 MODEL and SYSTEM

Our system (currently under development) serves as an interface to a web search system result list. The query submitted is passed to the underlying search engine, the result list is parsed and threads are dispatched to each document in the list (up to a maximum of thirty at any one time). In a similar way to [1] the system then creates a real-time query-biased summary of these thirty documents in parallel. The entire process takes around 4 seconds.

We consider a user's search context to be constructed of three main components; the user's individual search strategies, their current emotional/affective state and any constraints imposed on them by the task they are performing and other situational factors. These components influence how users seek information, and the work presented in this paper outlines how we capture this search context by monitoring two observable behaviours: the use of a *relevance path* and *focused mouse movements*.

2.2.1 Observable Behaviours

In our model there are five different *representations* of the one document which combine to form the *relevance path* (Figure 2). Firstly we have a list of sentences from all documents retrieved (thirty at a time) scored in relation to the query, we call these sentences *top ranking sentences* (TRS). We then have the title of the document, a sentence in the summary of the document, that sentence in the context it occurs in the document and finally the document itself. Through their interaction, the user has control over which representations are shown on the display.

TRS ⇨	document title ⇨	summary sentence ⇨	sentence in context ⇨	document
①	②	③	④	⑤

Fig. 2. The relevance path

The user can enter and leave the path at any point, and the distance travelled along it can provide a clue as to how relevant the information in the path is. We assign a *path weight* depending on how far the user travels. If a user stops before the end of the path is reached (i.e. at a summary sentence), we do not count the terms in the summary sentence as this has dissuaded the user from further perusal. In contrast, since the user can leave the path at any time and go to a document, we increase the weight of the terms in the representation from where this occurs.

The journey between each step in the path (shown in Figure 2 by arrows) is a mouse movement. We consider this movement as *focused* if the user acts with intentionality from their current location towards the next step in the path. In our system, we recommend this next step to the user. For example, if they indicate an interest in a top ranking sentence we highlight the document title of the document where that sentence resides. If the user then goes *directly* to this document title (to view a summary of the document) then the interaction is focused. We use the time taken to go to this next step as a measure of just *how* focused the interaction is. This time is normalised for the length of the representation to avoid undue bias caused by the differing times involved in assessing the relevance of different sized representations.

2.2.2 Combining the Behaviours

We are currently testing how we can combine these factors to generate appropriate new queries based on what users have expressed an interest in. A representation is considered relevant if the user makes a focused movement towards it. At each step in the relevance path we remove all common stop words (i.e. 'the', 'of', 'a', etc.) from the representation's text and score each of the terms that remain based on representations viewed. We then rank all of the terms and compare the average score of the top six ranked terms across *all* relevant representations with the average score of the query terms across *all* relevant representations. We use this as a means of detecting change in the information need, and generating a new query comprised of the original query and the top six ranked terms. For a small change we use the new query to re-rank the TRS list, for a slightly larger change we use it to re-rank the document list and for large changes we submit this to the underlying web search system.

The system (with relevance path numbered) is shown in Figure 3.

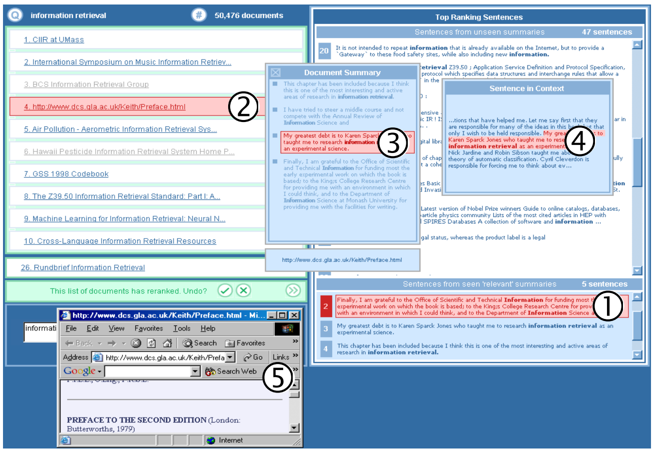

Fig. 3. Search interface

Our system uses query-biased summarisation and top ranking sentences, two components whose worth has been proven in previous experiments [1,2,3]. When a change in the information need is detected, the system proposes alternative courses of action and the user has the option of undoing any change the system makes. All such suggestions are made at the periphery of the interface, so not to interfere with the task at hand.

3. FUTURE INTENTIONS

We intend to formalise our contextual model and complete the system based upon it. We then have to experimentally test our hypotheses that our approach will (a) lead to more effective searching and (b) build a more accurate picture of search context and the user's real information need.

4. REFERENCES

[1] White, R.W., Jose, J.M. and Ruthven, I. 'The influencing effects of query biased summarisation in web searching'. *The International Journal of Information Processing and Management*. in press. 2002.

[2] White, R.W., Ruthven, I. and Jose, J.M. 'The use of implicit evidence for relevance feedback in web retrieval'. Proceedings of the 24th BCS-IRSG European Colloquium on IR Research. Lecture notes in Computer Science 2291. pp. 93-109. Glasgow, UK. 2002.

[3] White, R.W., Ruthven, I. and Jose, J.M. 'Finding Relevant Web Documents using Top Ranking Sentences: An Evaluation of Two Alternative Schemes'. Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 57-64. Tampere, Finland. 2002.