EFFORT-PRECISION AND GAIN-RECALL BASED ON A PROBABILISTIC NAVIGATION MODEL

Integrating Post-Query Navigation within a Measure of Retrieval Effectiveness

Keywords: Performance, Measurement

Abstract: Traditional evaluation of retrieval systems is based on implicit assumptions about the users' interaction with the system. It is assumed that the user is presented with the ranked results and examines the documents one after the other in the order they are listed. In this paper we argue that such a model is obsolete in the case of the Web and structured document retrieval, where navigation is an integral part of the user's search strategy. We advocate that post-query navigation needs to be reflected in the evaluation framework. We substantiate our proposal with the evidence of post-query navigation from user studies and discuss examples of systems that have been developed with the consideration of the user's browsing behaviour. In order to capture retrieval effectiveness for query and navigation based search, we introduce a measure of retrieval effectiveness that comprises a probabilistic model of the users' post-query navigation.

1 INTRODUCTION

Evaluation in IR has a long and rich history with work dating back to the development of the first IR systems in the 1950's, resulting in a wealth of evaluation studies and initiatives (Rijsbergen, 1979; Sparck Jones and Willett, 1997; Baeza-Yates and Ribeiro-Neto, 1999). Over the years, it has become common practice to evaluate retrieval systems' effectiveness using test collections consisting of a set of documents, user requests, and relevance assessments. This so-called *Cranfield tradition* of experimental evaluation has given rise to what is now known as 'standard IR evaluation practice'. It has become universal through the retrieval evaluations organised at the Text REtrieval Conference (TREC) (Voorhees and Harman, 2005).

These traditional evaluation experiments rely on implicit assumptions regarding the retrieval task and the user interaction model. A typical retrieval task, often referred to as flat document retrieval, is to return a ranked list of relevant documents in response to the user's query. The user of such a system is assumed to follow a simple model of interaction (Baeza-Yates and Ribeiro-Neto, 1999). First, the user poses a query, representing the information need, typically in the form of a bag of keywords. In response, the system returns a ranked list of documents estimated to satisfy the user's request. The user then examines the documents sequentially, starting from the top, inspecting each document until the end of the list or a predefined number of documents (e.g. 1000 at TREC) is reached.

In this scenario, established measures, such as precision and recall (Baeza-Yates and Ribeiro-Neto, 1999), provide a suitable and intuitive mechanism for evaluating the effectiveness of a retrieval system. The quality of the system's output is measured as a function of the retrieved and the relevant documents: Recall is defined as the ratio of retrieved relevant documents and the total number of relevant documents in the collection. Precision is computed as the ratio of relevant documents within the set of retrieved documents. Computation of precision at a given rank implicitly assumes that the user spends the same amount of time inspecting each document. Thus, the user's search effort per document is assumed constant. Based on the definition of search effectiveness as the ratio of output quality vs. user effort, quality is measured for a fixed amount of effort in this case.

With the emergence of new IR paradigms, such as Web, video and structured document retrieval (SDR), the validity of these assumptions has to be re-examined. For example, the simple model of user interaction described above does not reflect the typical browsing behaviour of Web users. Furthermore, the assumption of a constant effort per search result is questionable when, for example, variable size segments of videos are retrieved. Longer video segments require a longer review time and thus more effort from the user. In addition to these, the assumption of independent units of retrieval also does not apply to SDR, where multiple parts of the same document can be retrieved. SDR is in fact a key example as it combines querying and browsing, and generally provides results that vary in size. Although SDR predates the development of the eXtensible Markup Language (XML) by a decade (Bray et al., 1998), it is the popularity and widespread use of the XML standard on the Web and in digital libraries that made SDR into a vibrant research area. The aim of SDR (and XML IR) is to exploit the structure of documents and return relevant parts of the document (rather than render relevance for the whole document), thus reducing the user's effort in locating relevant content within the document.

Evaluation of systems that return parts of the document structure has to take into account a more complex user interaction. Depending on the result presentation, users may access other components that are structurally related to inspected results. Users may then locate additional relevant information by browsing along the structure of the document or by simply scrolling up and down the content display. This is particularly true when the user is presented with starting points for browsing. Furthermore, the system may return a paragraph or a section in proximity of the desired part of the document. When the content display enables the user to see both the retrieved and the relevant paragraphs, such a 'near miss' may be considered satisfactory since the cost of the user's effort is only slightly increased. Similar arguments hold for the Web and video case, where a web page or video clip is part of the Web structure and video, respectively.

These examples illustrate the need to consider a more comprehensive user interaction model and include post-query navigation within the evaluation framework. This requires new measures of retrieval effectiveness which are not limited by the assumptions of the Cranfield model. Performance measures that capture structural dependencies in the document are effort-precision and gain-recall (ep/gr), proposed in (Kazai and Lalmas, 2006). They are used for the evaluation of content-oriented XML IR within the INitiative for the Evaluation of XML IR (INEX)[1], a TREC-like evaluation forum for XML IR, launched in 2002. The measure of ep/gr takes into account the dependency among retrieved XML elements and, for example, facilitates rewards for near-misses and punishment for overlap (e.g., when a paragraph and its container section are both retrieved). The main shortcoming of this measure is the absence of a formal user model and, thus, reliance on heuristics when calculating the basic parameter of gain.

Another measure is PRUM proposed in (Piwowarski et al., 2007), which is an extension of Raghavan et al's probabilistic definition of recall-precision (Raghavan et al., 1989). The measure of EPRUM in (Piwowarski et al., 2007) investigates another way to extend traditional precision and recall, and is close to (Kazai and Lalmas, 2006) as it defines precision at a given recall level as the ratio of minimal search length over two different lists. However, EPRUM has a complex way of dealing with graded relevance assessments.

In this paper, we redefine effort-precision and gain-recall by incorporating the probabilistic model of users' post-query navigation developed in (Piwowarski et al., 2007). By grounding the measure on a formally derived user model, we expect to arrive at a theoretically well-founded evaluation framework that takes into account structural dependencies and allows for graded relevance.

The paper is structured as follows. Section 2 reviews selected user studies that motivate the explicit use of user models in the evaluation framework. Section 3 details the user tasks and relevance judgements at INEX. In Section 4 we introduce the evaluation measures and present the probabilistic user model in Section 5. The two are combined into the measure of effort-precision as discussed in Section 6. We close with a review of related works in Section 7 and conclusions and future work in Section 8.

2 POST QUERY NAVIGATION

Post-query navigation describes users interaction with the results of a search system, which is typically presented in the form of a ranked list of documents. In the context of the Web, navigation (colloquially known as surfing) is whereby users follow links and browse the destination web pages. This form of navigation can be considered as inter-document navigation. We can also talk about within-document navigation. For example, users of SDR systems can browse

[1] http://inex.is.informatik.uni-duisburg.de/2007/

from a result component to other document parts inside the container document. In this section, we detail studies of user behaviour in both Web and SDR contexts, and give examples of systems that build on this model to provide explicit support for users' post-query navigation. Our aim is to motivate the need for evaluation frameworks where post-query navigation is integrated within the measure of retrieval effectiveness.

2.1 Navigating on the Web

Navigation is a major part of users' experience on the World-Wide Web (WWW). Recognizing its importance, researchers have been studying the behavioural characteristics of Web users for many years now. These studies are typically based on survey data and information extracted from client-side log file analysis and are conducted with the aim to supplement the understanding of Web users in order to yield design and usability guidelines for Web pages, sites and browsers. From the wealth of research that exist today, studies focusing on audience analysis, such as navigation strategies and interface usage include (Weinreich et al., 2006; Juvina and van Oostendorp, 2004; Sellen et al., 2002; Cockburn and McKenzie, 2001; Byrne et al., 1999; Catledge and Pitkow, 1995).

In general, the literature distinguishes two strategies of users' information seeking: searching and browsing. Searching is typically associated with the act of executing queries in a search system, while browsing is generally described as an activity of exploring and picking out bits and pieces (Cove and Walsh, 1988). Browsing and searching are not mutually exclusive activities, but users often move back and forth between the two strategies. Bates summarizes this in his "berrypicking" model of information seeking (Bates, 1989), where a user's search strategy is constantly evolving through browsing. Several strategies to browsing have also been published. Cove and Walsh (Cove and Walsh, 1988), for example, distinguishes three strategies: 1) Search browsing, which is a directed search, where the goal is known; 2) General purpose browsing, where the user consults sources that have a high likelihood of items of interest; and 3) Serendipitous browsing, which is purely random. This continuum allows to distinguish between browsing as a method of completing a task and open ended browsing with no particular goal in mind.

One of the largest web browsing studies to date was conducted by Harald Weinreich et al. (Weinreich et al., 2006). They analyzed over 135,000 page visits by 25 experienced volunteers over a mean period of 105 days. Their study confirmed Web navigation to be a rapidly interactive activity. They found that link following was the most common navigation activity, accounting for 43.5% of page transitions. Direct access through bookmarks, typing in urls, or home page buttons has accounted for 10%. Navigation using the back button represented 15% of all page transitions (corresponding to 50% drop found five years earlier). Users' habits of within-page navigation showed that the most selected hyperlinks are those located in the top left corner of the screen. In addition, nearly a quarter of all cases, people chose links that required scrolling.

A study of interaction behaviour for users engaged in Web search activities that originate with the submission of a query to a search engine was done in (White and Drucker, 2007). The study placed particular emphasis on post-query navigation trails (i.e., pages viewed on the click stream following the query being issued), which were collected through client-side logging of 2,527 users over a five-month period. A search trail was defined as one originating with a directed search (i.e., a query issued to a search engine), and proceeding until a point of termination where it was assumed that the user has completed their information-seeking activity. Trails contained multiple query iterations, and pages that were either: search engine homepages, search results, or were connected to a search result page via a hyperlink trail. They found that out of approximately 80 million Web pages, 12.5% were part of such a search trail, with an average trail length of around 17 steps.

Building on the prevalence of navigation as a search tactic, systems are increasingly being developed where post-query navigation is an explicit component of the retrieval paradigm. In this extended model, the documents in the ranked result list of a search query represent starting points from which the user can commence exploration. This is the approach taken in the hypertext retrieval system of navigation-aided retrieval (NAR) (Pandit and Olston, 2007). They define good starting points as hypertext documents that, while they may not match the users query directly, permit easy navigation to many documents that do match the query, via one or more outgoing hyperlink paths. Best Trails (Wheeldon and Levene, 2003) is a similar retrieval system which selects starting points in response to queries.

Also related is the work in the area of topic distillation (Craswell et al., 2003), where retrieval systems aim to identify a small number of high-quality documents that are representative of a broad topic area. Much of this work is based on algorithms, such as HITS (Kleinberg, 1999), which identifies bipartite

subgraphs consisting of hubs (related to our notion of starting points) and authorities.

2.2 Navigating in SDR

One of the first studies to investigate searchers' information seeking behaviour in the context of SDR compared two variants of the same retrieval interface: One that highlighted relevant document parts and one that highlighted best entry points (BEPs; starting points for browsing). They found that users showed strong preference for the BEP interface, and in particular to BEPs deeper in the document's structure. Users browsing behaviour included actions such as jumping from one BEP to the next, and linear and hierarchical browsing supported by a table of contents. The study of (Reid et al., 2006a; Reid et al., 2006b) examined aspects influencing users' BEP selection strategies with the aim to support automatic BEP identification.

The largest user studies have been carried out as part of the interactive track at INEX 2004 and 2005 (Tombros et al., 2005a; Tombros et al., 2005b; Larsen et al., 2006). The aims of these studies were to study the behaviour of users when interacting with components of XML documents, and secondly to investigate approaches for XML retrieval which are effective in user-based environments. Their main findings include the general observation that overlapping components, i.e., components from the same document at different ranks in the ranked list, frustrated many users. This issue, however, could be levied when the display of results were clustered by the container documents. An important finding regarded the use of document structure as contextual information that users often consulted in order to decide on the usefulness of a document. The analysis of users' browsing behaviour indicated that they tend to browse to more specific information rather than to more exhaustive information. In addition, users found the table of contents and query term highlighting useful.

Approaches to SDR explicitly build on the navigational aspects of the retrieval model. One of the most influential works in SDR, is that of (Chiaramella and Kheirbek, 1996). The cornerstone of their work is the combination of the two modalities of information retrieval: searching and browsing. The proposed integrated model is based on the definition of so-called index units and the process of aggregation for calculating the index weight of terms contained at different levels of a document's hierarchy. Index units are defined as self-explaining units of information, which may be nested. Aggregation is defined as an indexing strategy, which recursively evaluates index expressions of index nodes in the document hierarchy, starting from the atomic index units moving up. Extending this work, a logical model is explored in (Fuhr and Grossjohann, 2001), where augmentation weights are introduced by means of probabilistic rules. The model proposed in (Lalmas, 1997) employs Dempster-Shafer's theory of evidence and implements an aggregation operator using Dempster's combination rule.

Since the development of XML, a wide range of XML IR systems have been developed, implementing the retrieval paradigm of SDR for XML document collections. A major catalyst of research in content-oriented XML IR came with the establishment of the INEX evaluation initiative in 2002 (Fuhr et al., 2003). Each year, INEX publishes an expanding volume of proceedings of its annual workshop, which provides an overview of the latest developments in the field. Summaries of the work described in the proceedings are also reported in ACM SIGIR Forum (e.g. (Fuhr and Lalmas, 2004; Tombros et al., 2005a; Lalmas and Kazai, 2006)). Another recent review of XML IR can be found in (Amer-Yahia and Lalmas, 2006).

3 THE INEX SETUP

3.1 User tasks

The main activity at INEX is the ad hoc retrieval task, where the collection consists of XML documents, composed of different granularity, nested XML elements, each of which represents a possible unit of retrieval. Within the umbrella of the ad hoc track, INEX 2007 defines three retrieval tasks: Focused, Relevant in Context, and Best in Context task.

The Focused task asks systems to return a ranked list of the "most focused" XML elements that satisfy the user's information need, without returning overlapping elements (e.g. a paragraph and its container section element). Here systems are required not only to estimate the relevance of elements, but also to decide which element(s), from a tree of relevant elements, are the most focused non-overlapping one(s).

The Relevant in Context task is much like the Focused task, but here the ranked list consists of groups of the most focused elements, clustered by the unit of the document. This task assumes a fixed result presentation format to the user: Systems are expected to return, for each relevant document, a set of elements that contains the relevant information within the document. This can be likened to asking systems to return a ranked list of documents and then inside each, highlight the relevant information for the user. The rel-

<article><name>Ali Baba></name>
...
<section><title>Story Summary</title>
Ali Baba, a poor woodcutter, ... forty thieves...
Ali Baba's rich brother, <link>Kasim</link>, finds out ...
The thieves, finding the body gone, realise that ... The first several times they are foiled by <link>Morgiana</link>, ...
The lead thief pretends ... invited to dinner at Ali Baba's house. He is recognised by Morgiana, who ...
</section>
...
</article>

Figure 1: Relevance assessments are collected as highlighted text fragments

evant text fragments inside the document are treated equally (they are not ranked).

Finally, in the Best in Context task, systems are required to return a ranked list of best entry points (one per document) to the user, representing the point in the document where users should start reading.

For all three tasks, it is reasonable to expect that users will navigate from a returned result element to other components within the document. This is in fact made explicit for the Best in Context task. In the case of the Focused task, users may browse the local context of the result. In the case of the Relevant in Context task, users may browse the whole document to locate the bits of texts highlighted by the system.

3.2 Relevance judgements

Apart from the Best in Context task, for which separate best entry point judgements are obtained, the evaluation of the first three tasks relies on the (same) set of relevance assessments collected from human judges. The relevance assessment procedure is based on a yellow-marker design and involves the highlighting of relevant text fragments in the document collection. The collected relevance assessments are, hence, in the form of arbitrary sized text passages, which are not constrained by XML element boundaries. This is illustrated in Figure 1.

The conversion of highlighted passages into assessments on XML elements is a simple process, whereby for each XML element, the length of the element (`@size`) and the number of highlighted characters (`@rsize`) is recorded (see Figure 2). From these, a specificity score can be automatically calculated for each XML element, reflecting the extent to which the document component focuses on the topic of request. The score is calculated as the ratio of the number of highlighted characters contained within the component c and the length of the component:

$$spec(c) = rsize(c)/size(c) \quad (1)$$

Specificity hence can take any value in $[0,1]$.

4 EFFORT-PRECISION AND GAIN-RECALL

4.1 Effort-precision based on ranks

Effort-precision (ep) and gain-recall (gr) are part of the eXtended Cumulated Gain (XCG) measures proposed in (Kazai and Lalmas, 2006) and employed as the official measures at INEX 2005 and 2006. They are extensions of the cumulated gain based measures of (Järvelin and Kekäläinen, 2002), developed for multi-graded relevance values, allowing to credit IR systems according to the retrieved documents' degree of relevance.

The underlying notion of the measure is that relevant information is associated with some level of gain. The meaning of the gain value within the evaluation may be compared to the notion of utility, reflecting the worth that a retrieved component represents to the user. The user obtains the gain that is associated with the accessed relevant information. As the user is presented with more relevant information, the gain is accumulated. In order to access relevant information and hence obtain the associated gain, the user has to invest some effort, e.g. click on a result in the ranking. The quality of a retrieval system's performance is then measured in terms of gain vs. effort.

Effort precision in (Kazai and Lalmas, 2006) is defined as a measure of the amount of relative effort (in terms of number of visited ranks) required of a user to reach a given level of cumulated gain when scanning a given ranking compared to an ideal ranking. Performance is hence reported in relation to an ideal ranking. This is illustrated in Figure 3: The horizontal line represents the cumulated gain of r, which is reached by the ideal curve at rank i_{ideal} and by the

```
<file collection="wikipedia" name="2267781"> <passage
start="/article[1]/body[1]/section[1]/p[5]/text()[1].89"
end="/article[1]/body[1]/section[1]/p[5]/text()[1].199"
size="111"/> <element path="/article[1]/body[1]/section[1]/p[5]"
exhaustivity="2" size="569" rsize="111"/> <element
path="/article[1]/body[1]/section[1]" exhaustivity="2" size="8470"
rsize="111"/> <element path="/article[1]/body[1]" exhaustivity="2"
size="20356" rsize="111"/> <element path="/article[1]"
exhaustivity="2" size="20376" rsize="111"/> </file>
```

Figure 2: Excerpt from an INEX'06 topic's relevance assessments file. The `@size` attribute stores the XML element's length (in characters), while the `@rsize` attribute corresponds to the number of highlighted characters within the element.

system at rank i_{run}. The ratio of the two ranks indicates the performance of the system. The closer the system can match the ideal curve, the closer to 1 the *ep* score. The ideal ranking is derived by sorting the components of the collection by decreasing gain value. For example, if the collection consists of three components c_1, c_2, and c_3, where the associated individual gains are 3, 6, and 2, respectively, then the ideal ranking would be c_2, c_1, and c_3.

4.2 Generalized effort-precision

In this section, we generalise the measure of effort-precision by removing the implicit assumption in (Kazai and Lalmas, 2006) to measure effort in units of rank. Calculating effort as the number of rank positions accessed by the user means that each result in the ranking represents equal cost to the user. This has the disadvantage that additional navigation costs incurred through browsing from a result to other parts of the document are ignored. In order to allow for variable effort per result, alternative solutions can calculate effort as a function of time, user clicks, or number of read characters. By shifting the unit of effort to, e.g., number of read characters, we are effectively re-scaling the x axis in Figure 3. We still take measurements at rank positions, but the distance between measurement points can now vary depending on the amount of effort the user accumulates when accessing a result and its structurally related components.

We denote the gain associated with a document component c as g_c and the effort as e_c. The cumulated gain at rank i is calculated by summing up the individual gains along the ranking up to and including rank i:

$$cg[i] := \sum_{j=1}^{i} g[j] \tag{2}$$

where $g[j]$ is the gain obtained at rank j. For example, the ranking c_1, c_2, and c_3, with associated gains of 3, 6, and 2, respectively, yields $cg[3] = 11$.

The total effort for a given rank can be defined as:

$$ce[i] := \sum_{1 \le j \le i} e[j] \tag{3}$$

where $e[j]$ is the effort associated to consulting rank j.

We can also define the minimum number of ranks a user has to consult in order to reach a given gain-recall level gr:

$$m[gr] := \min(i \text{ s.t. } cg[i] \ge gr) \tag{4}$$

The total effort for a given gain-recall level can be defined as:

$$ce[gr] := ce[m[gr]] \tag{5}$$

where $e[j]$ is the effort at rank j. Note that the total effort can be infinite if the gain-recall cannot be achieved. Note also that for the ideal list we assume that any gain-recall level can be achieved.

We will use the subscript *ideal* to denote the curve of the ideal ranking and *run* to refer to a system ranking. As mentioned above, the ideal ranking is given as the ranked list of components in the collection in decreasing order of gain value.

We define effort-precision (*ep*) as a measure of the amount of relative effort required of a user to reach a given level of cumulated gain when scanning a given ranking compared to an ideal ranking:

$$ep[gr] := \frac{ce_{ideal}[gr]}{ce_{run}[gr]} \tag{6}$$

where $ce_{ideal}[i]$ is the total accumulated effort at which the cumulated gain gr is reached by the ideal curve and $ce_{run}[gr]$ is the total effort at which the cumulated gain of $cg[i]$ is reached by the system under evaluation. Note that the latter value can be infinite if the gain cannot be reached, in which case the ep value is simply 0. A score of 1 reflects ideal performance, where the user needs to spend the minimum necessary effort to reach the given level of gain.

We define gain-recall at rank i as the cumulated gain value divided by the total cumulated gain

(Kekäläinen and Järvelin, 2002):

$$gr[i] := \frac{cg[i]}{cg[n]} = \frac{\sum_{j=1}^{i} g[j]}{\sum_{j=1}^{n} g[j]} \quad (7)$$

where n is the total number of relevant components.

Instead of taking measurements at absolute cumulated gain values, we can calculate effort-precision at arbitrary gain-recall points, $x \cdot cg[n]$.

The meaning of effort-precision at a given gain-recall value is the amount of relative effort that the user is required to spend when scanning a system's result ranking compared to the effort an ideal ranking would take in order to reach the given level of gain relative to the total gain that can be obtained.

This method follows the same viewpoint as standard precision/recall, where recall is the control variable and precision the dependent variable. In our case, gain-recall is the control variable and effort-precision the dependent variable. As with precision/recall, interpolation techniques are necessary to estimate effort-precision values at non-natural gain-recall points, e.g. when calculating effort-precision at standard recall points of $[0.1, 1]$.

As with standard precision/recall, a (non-interpolated)[2] mean average effort-precision, denoted as $MAep$, can be calculated by averaging the effort-precision at every point where there is an increase in the cumulated gain, and then averaging these over the set of test queries. For not retrieved relevant components a precision score of 0 is assigned. Analogue to recall/precision graphs, we may also plot effort-precision against gain-recall and obtain a detailed summary of a system's overall performance. The shape of the ep/gr graph is similar to that drawn for precision/recall.

So far, we have defined a general evaluation framework with two key parameters: gain and effort. The calculation of the user's gain and effort for a given retrieval result is where we require a model of user's browsing behaviour, which is described in the next section.

5 A MODEL OF PROBABILISTIC POST-QUERY NAVIGATION

5.1 User navigation model

We base our model on the generic navigational model proposed in (Piwowarski et al., 2007; Piwowarski and Dupret, 2006).

[2]Interpolation of the ideal curve is necessary here.

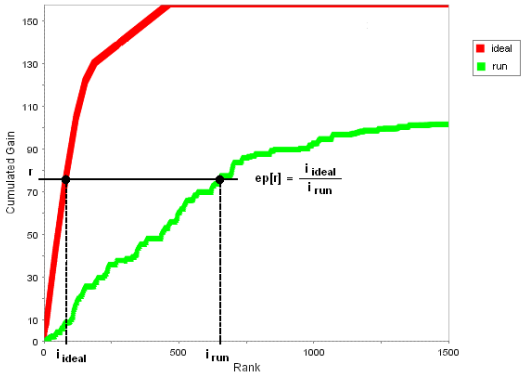

Figure 3: Illustration of effort-precision (ep), where effort is measured in units of rank

We assume that in response to a query, the user is returned a ranked list of document components by the search system. From one result component, the user can navigate to other, structurally related components by following hyperlinks or simply by scrolling in the container document. The user navigates from one component to another with a certain probability that is dependent on the user's navigational history. For example, the user may be more likely to explore around a relevant component than to browse from a non-relevant component. On the other hand, users may be less likely to follow long navigational paths that require, e.g., many clicks. After exploring the context of a result component, the user is assumed to return to the ranked list and proceed to the next result.

As described in the previous section, each component is associated with a level of gain that the user obtains when viewing (reading) its content. In order to obtain the gain, the user invests a certain amount of effort, e.g. clicking on the result and reading its content. The user picks up gain and spends effort for each visited component, whether the component was found in the result ranking or browsed to from a result.

More formally, we denote the ranked list of results as L of size $|L|$. The i-th component of the ranking is given as $c_i \in L$. L_i denotes the set of components in L retrieved up to and including rank i, thus $L_i \subseteq L$ and $c_i \in L_i$ also hold.

We define the context of a component c as the set of components that can be reached from c through navigation (by following links or scrolling), and denote it by C_c. From this definition, it is clear that the context of a component is dependent on the structure of the document collection.

We denote the set of components accessed (seen) by the user – either directly from the result list or via navigation – up to and including a given rank i as S_i.

We assume that the i-th component in the result ranking is always accessed by the user who has scanned the ranking up to rank i, thus $L_i \subseteq S_i$ and $|S_i| \geq |L_i|$.

Each accessed component is associated with a value of gain $g[c]$ and effort e_c. Both gain and effort are cumulative, that is, accessing more relevant content increases the user's total gain but at the same time requires more effort. Accessing non-relevant content increases the effort, but the gain remains unchanged. In general, effort is a monotonically increasing function while gain is monotonically non-decreasing.

In terms of probabilistic events, we denote $P(i \rightsquigarrow c)$ the probability that the user navigates to a component c from rank i. The component may be a result or a component in the context of a result. The probability takes into account all possible routes to c. For example, if the user can reach c'' directly from c (i.e. $c \rightsquigarrow c''$) and also through c' (i.e. $c \rightsquigarrow c' \rightsquigarrow c''$), then this is reflected in the probability $P(c \rightsquigarrow c'')$. For the set of components in the context of c, given as $c' \in C_c$, $P(c \rightsquigarrow c') > 0$ holds. In general, the probability $P(i \rightsquigarrow c)$ will depend on the user's navigation history, the structure of the collection, as well as the content of the component c.

We denote $P(c \in S_i)$ to mean the probability that the component c has been seen by a user who traversed the result ranking up to and including rank i. We denote the context associated to rank i as C_j. A component can be seen at rank i only if it belongs to one context C_j for some $j \leq i$.

5.2 Estimating the probabilities

In order to be able to tractably compute the necessary probabilities, we follow the approach of (Piwowarski et al., 2007) and adopt the same simplifying assumption to treat the set of events $i \rightsquigarrow c$ mutually independent. This means that a user is assumed to navigate the context of a result component returned at rank i independently from previous navigations.

The above assumption then simplifies the computation of the probability of the event that the user sees a document component after examining i ranks in the ranked list:

$$P(c \in S_i) = 1 - \prod_{j=1}^{i} (1 - P(j \rightsquigarrow c))$$

This means that all causes leading to a component being seen by the user are treated independently of each other.

As an example, consider a ranking of components, where the first rank is c_1, and c_2. Assume that both are linked to the document component c with navigational probabilities $P(1 \rightsquigarrow c) = 0.3$ and $P(2 \rightsquigarrow c) = 0.9$. At rank 1, the probability of the user seeing c is $P(c \in S_1) = 1 - (1 - 0.3) = 0.3$. At rank 2, the probability increases to $P(c \in S_2) = 1 - (1 - 0.3)(1 - 0.9) = 0.93$.

To obtain the actual navigational probabilities, some parametric user model, such as that of $P(c \rightsquigarrow c') = (1 + e^{\theta d(c,c')})^{-1}$, where $d(c,c')$ is the distance in number of characters between the components c and c', could be employed (Piwowarski and Dupret, 2006).

The sequence of links that a user may follow can be modelled as a stochastic process in terms of an absorbing Markov chain (Levene et al., 2001). In this stochastic process, the user accessing c is faced with a choice of continuing the navigation and following one of the available links from c or terminating the navigation session. Based on this the overall trail probability can be calculated. using web data and a random walk model, Huberman et al. (Huberman et al., 1998) found that the probability of a trail of length t is approximately proportional to $t^{-3/2}$.

6 CALCULATING GAIN AND EFFORT

Since the user model is now stochastic, the calculation of an effectiveness score requires the estimation of an expected value of the different measures. In this section, we focus on the expected cumulated gain and effort precision.

The following derivation closely follows (Piwowarski et al., 2007), with the difference that we need to deal with efforts associated to consulting a rank.

Let I denote the set of ideal elements. The only components that have an associated gain are components in I. The construction of I is task dependent and can be for example done as in XXXcite your notes on INEX metricsXXX.

Extending cumulated-gain $cg[i]$ Accepting the assumptions of the user model presented in the previous section, we can compute the expected cumulated gain value as:

$$\mathbb{E}[cg[i]] = \sum_{c \in I} g[c] P(c \in S_i) \tag{8}$$

We can see that the reward associated to an ideal element is bounded by $g[c]$, and that this gain is reached only when the ideal element is completely seen according to the user model.

Extending the minimum search length ($m[gr]$) We first note that the probability that the minimum number of ranks a user has to consult is

$$\begin{aligned} & P(m[gr]=i) \\ = \;& P(cg[i-1]<gr \wedge cg[i]\geq gr) \\ = \;& P(cg[i-1]<gr)-P(cg[i]<gr) \end{aligned}$$

where by definition the cumulated gain at $i=0$ is 0. We show latter how $P(cg[i-1]<gr)$ can be computed.

Extending effort-precision ($ep[gr]$) It is necessary to extend formula (6). As the user behaviour is now stochastic, so is the cumulated gain at a given rank. A solution is to compute the expectation of the ratio (6). We can compute the expectation of the effort-precision at the gain-recall value gr as:

$$\mathbb{E}[ep[gr]] \quad = \quad \mathbb{E}[ce_{ideal}[gr]] \times \mathbb{E}\left[\frac{1}{ce_{run}[gr]}\right] \quad (9)$$

We need to evaluate $\mathbb{E}[ce[gr]]$ and $\mathbb{E}\left[ce[gr]^{-1}\right]$, respectively for the ideal and for the evaluated list.

Then, we can compute the two expectations of fornula (9):

$$\begin{aligned} \mathbb{E}[ce[gr]] &= \mathbb{E}[ce[m[gr]]] \\ &= e[1]+\sum_{i\geq 1} e[i+1]P(cg[i]<gr) \end{aligned}$$

and

$$\mathbb{E}\left[\frac{1}{ce[gr]}\right] = \sum_{k\geq 1}\frac{1}{ce[k]}\left[P(cg[k+1]<gr)-P(cg[k]<gr)\right]$$

This latest formula assumes that the search length is infinite when the gain gr cannot be reached.

Computing $P(cg[i]<gr)$ The cumulated distribution $P(cg[i]<gr)$ is computed in the following manner. Let $\mathcal{T}_i$ be the set of elements seen with a probability 1, and $\mathcal{P}_i$ the set of partially seen (i.e. seen by a subset of users only):

$$\begin{aligned} \mathcal{T}_i &:= \{c\in I \; s.t. \; P(c\in S_i)=1\} \\ \mathcal{P}_i &:= \{c\in I \; s.t. \; P(c\in S_i)\in(0,1)\} \end{aligned}$$

The gain at rank i is then a random variable defined as

$$cg[i]=\sum_{c\in\mathcal{T}_i} g[c]+\sum_{c\in\mathcal{P}_i} g[c]\times \mathbf{1}_{c\in S_i}$$

where $\mathbf{1}_{c\in S_i}$ is a random variable equals to 1 if the user has seen the ideal component c and 0 otherwise. As the first part of the sum is deterministic, the problem is reduced to the computation of the second term $cg^{(P)}[i]$. It is possible to compute

$$P\left(\sum_{c\in\mathcal{P}_i} g[c]\times\mathbf{1}_{c\in S_i} < gr-\sum_{c\in\mathcal{T}_i} g[c]\right)$$

in a time quadratic in the number of componenents in $\mathcal{P}_i$. When $\mathcal{P}_i$ has a sufficiently large number of elements (experiments have shown that values over 10 were enough), the random variable $cg^{(\mathcal{P})}[i]$ can be approximated by a normal variable of mean

$$\sum_{c\in\mathcal{P}_i} g[c]\times P(c\in S_i)$$

and of variance

$$\sum_{c\in\mathcal{P}_i} g[c]^2 P(c\in S_i)\,(1-P(c\in S_i))$$

The ideal list In order to calculate effort-precision, we need to derive the ideal ranking. The ideal ranking in general will depend on the user model, and hence could involve a complex optimization process to maximize the user's gain for minimum effort for arbitrary user models. It is also possible that multiple ideal lists can be constructed for some user models, or that the ideal list is different depending on the rank i. In general, we can define the ideal list $\mathcal{L}$ for a given gain-recall value gr as:

$$ideal[gr]=\underset{\text{list } L}{\operatorname{argmax}}\,\mathbb{E}[ep_L[gr]]$$

In the case of most of the INEX user models, the ideal ranking can be easily obtained by ranking the document components or the container documents by their respective $g[c]$ value. A general method for computing easily the ideal list (or at least the $ce_{ideal}[gr]$) is yet to be found.

6.1 Example

Assume that we have three documents, one of them, d_3 is ideal (in the case of standard IR we could say simply relevant). The list is composed of the three documents (d_1,d_2,d_3). We assume that hyperlinks between documents induce the following behaviour on users:

- 40% will browse from d_1 to d_3.
- 30% will browse from d_2 to d_3.

Notice that we do not need to specify the behaviour to non ideal elements.

We assume that the efforts associated to inspecting each rank is a constant 1.

We assume that the gain of retrieving d_3 is 1. After the first rank, the probability that the user has seen d_3 is

$$1-(1-0.4)=0.4$$

. After rank 2, the probability is

$$1-(1-0.4)(1-0.3)=5/8=0.58$$

After rank 3, the probability is simply 1 since all the users will see document 3 at rank 3.

In this special case, it is easy to see that the cumulated gain is inferior to 1 if and only if document 3 is not seen. Thus, $P(cg[1]<1)=0.5$, $P(cg[2]<1)=0.35$ and $P(cg[1]<1)=0$. We then have

$$\mathbb{E}[ce[1]]=1-\frac{1}{3}\times 0-\frac{1}{2}\times 0.65-\frac{1}{6}\times 0.35\approx 0.63$$

We can also consider the classical user model, where the user does not browse between documents. In this case $P(cg[1]<1)=P(cg[2]<1)=1$ and $P(cg[3]<1)=0$. The expectation of the inverse search length is in this case equals to $1-1/2-1/6=1/3$. Since the ideal list is still the same, ep for a gain one is 0.3 in this case.

7 RELATED WORK

Only a limited number of measures have been proposed in the literature that allow to take into account post-query user navigation. We present a brief review of these here, and invite the reader to refer to (Pehcevski and Piwowarski,) for a more detailed review.

The work that is the most closely related to ours is that of (Piwowarski et al., 2007; Piwowarski and Dupret, 2006). We have in fact based our model of user navigation on the generic navigation models proposed by Piwowarski et al.

Precision-Recall with User Modelling (PRUM) (Piwowarski et al., 2007) is an extension of the probabilistic PRecall measure proposed by Raghavan et al., which allows to take into account users' browsing behaviour. It extends the original interpretation from the probability of a viewed document being relevant to the probability that the user sees a newly relevant XML element when he/she consults the context of a retrieved element, given that the user wants to see a given amount of relevant units:

$$PRUM[l]=P(Lur|Retr,L=l,Q=q)$$

where l is the Recall level wanted by the user, q is the query, *Retr* is the probability that the user consults the element, and *Lur* is the probability the the element leads to a relevant element that has not yet been seen by the user.

PRUM employs probability estimations for a user's browsing behaviour, and updates the probability of a node being seen by the user depending on its structural relationship to the currently visited node and assumptions about the user's interaction. The more structurally related elements that have been returned to the user, the more chances the user had to access the current result element, and hence, the more its score is reduced.

Based on the same user model, the measure of Expected Precision-Recall with User Modeling (EPRUM) (Piwowarski and Dupret, 2006) calculates the expectation over the ratio of two minimum values: The minimum rank that achieves a given level of recall l over all possible rankings and over the system's ranked list.

$$Precision@l=\mathbb{E}(\frac{min\_rank_i}{min\_rank_s})$$

min_rank_i is the minimum number of ranks the user has to consult to achieve the recall level of l from all possible rankings, and min_rank_s is the minimum number of ranks the user has to consult to achieve the recall level of l based on the given system ranking (it can also be infinite as in the case of the new ep/gr metric).

This measure is also related conceptually to our own as it is based on minimum lengths. An advantage of our measure is that varying costs of navigation per rank can be reflected directly in the score. The effort-precision presented in this paper is an alternative definition of the precision used in EPRUM; the difference is that we deal with gains and that the definition is simpler thus yielding a measure which is easier to compute.

Other related measures include the measures based on the concept of a user's tolerance to irrelevance (T_2I) (de Vries et al., 2004). The main idea is that a user merely needs an entry-point into the document that is 'close' to relevant information. Taking this view, a retrieval system produces a ranked list of entry points. The user starts reading the retrieved document from the suggested entry point, giving up when no relevant information is found before his or her tolerance to irrelevance limit is reached, at which point the user proceeds to the next system result.

The measure of *Expected Ratio of Relevant Elements* (ERR), proposed in (Piwowarski and Gallinari, 2004), is the expectation of the number of relevant XML elements a user sees when consulting the list of the first k returned results divided by the expectation of the number of relevant XML elements a user sees whilst exploring the whole collection:

$$ERR=\frac{\mathbb{E}(N_R|N=k)}{\mathbb{E}(N_R|N=|E|)}$$

where $N_R|N = k$ is the number of relevant XML elements in the first k results, and $N_R|N = |E|$ is the number of relevant XML elements in the collection.

8 CONCLUSIONS

In this paper, we consider extensions of the effort and gain model for document retrieval with an explicit user interaction model. The motivation to extend the evaluation in this direction stems from evidence presented by studies of user behaviour on the Web and in the realm of SDR at INEX. We chose the evaluation framework of the effort-precision and gain-recall measures as the basis for our current work as it was shown to reflect well on the goals of SDR approaches. In (Kazai and Lalmas, 2006) it has also been shown to perform reliably. The measure was also seen as particularly suited for our purpose as it divorces the user model from the actual calculation of effectiveness scores. However, one of its main criticms has been its use of various heuristics due to the lack of a formal user model. In this paper, we aimed to address this by marrying up our measure with a formal user model.

In order to include the aspects of the user interaction we followed the work by (Piwowarski et al., 2007) and incorporated the probabilistic model of user navigation from search results. We included the probability estimates into the calculation of the effort and gain measures and thus captured the effect of the specific user behaviour onto the effectiveness of the system.

This is a step towards a comprehensive information retrieval model that takes into account properties of the search engine (e.g., scoring functions), characteristics of the result displays (e.g., the viewing and scrolling properties), and explicit models of user interaction with the system (e.g., switching from navigation to search list examination and vice versa).

In our future work we will consider more sophisticated models for the user's achieved gain and expended effort. We shall look at alternative models for expended user effort that clearly differentiate between the cost associated with navigation and the cost of accessing the components in the ranked result list. That will lead to a natural switching mechanism between two search modes. It is natural to expect that the transition happens when the cost and benefit ratio from a navigation step is larger than the one for accessing the next component on the ranked list. By using more realistic models we will be able to compare the simulations for a given data set with observed user behaviour. Furthermore, we will introduce more detailed characterization of content structure and navigation properties that are exposed through the user interface. This will enable diversification of models for different content types, from search over structured content such as books to hyperlinked environments such as the Web, individual Web sites, or hypertext documents.

REFERENCES

Amer-Yahia, S. and Lalmas, M. (2006). Xml search: Languages, inex and scoring. *SIGMOD Record*.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

Bray, T., Paoli, J., and Sperberg-McQueen, C. M. (1998). Extensible Markup Language (XML) 1.0. http://www.w3.org/TR/1998/REC-xml-19980210, W3C Recommendation. Technical report, W3C (World Wide Web Consortium).

Byrne, M. D., John, B. E., Wehrle, N. S., and Crow, D. C. (1999). The tangled web we wove: a taskonomy of www use. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 544–551, New York, NY, USA. ACM Press.

Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the world-wide web. In *Proceedings of the Third International World-Wide Web conference on Technology, tools and applications*, pages 1065–1073, New York, NY, USA. Elsevier North-Holland, Inc.

Chiaramella, Y. and Kheirbek, A. (1996). An integrated model for hypermedia and information retrieval. In Agosti, M. and Smeaton, A. F., editors, *Information Retrieval and Hypertext*. Springer.

Cockburn, A. and McKenzie, B. (2001). What do web users do? an empirical analysis of web use. *Int. J. Hum.-Comput. Stud.*, 54(6):903–922.

Cove, J. F. and Walsh, B. C. (1988). Online text retrieval via browsing. *Inf. Process. Manage.*, 24(1):31–37.

Craswell, N., Hawking, D., Wilkinson, R., and Wu, M. (2003). Overview of the TREC 2003 Web Track. *Proceedings of TREC*.

de Vries, A., Kazai, G., and Lalmas, M. (2004). Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference Proceedings*, pages 463–473.

Fuhr, N., Gövert, N., Kazai, G., and Lalmas, M., editors (2003). *Proceedings of the 1st Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl, Germany, 8-11 December, 2002*. ERCIM, Sophia Antipolis, France.

Fuhr, N. and Grossjohann, K. (2001). Xirql: a query language for information retrieval in xml documents. In

SIGIR'01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 172–180, New York, NY, USA. ACM Press.

Fuhr, N. and Lalmas, M. (2004). Report on the inex 2003 workshop. *SIGIR Forum*, 38(1):46–51.

Huberman, B., Pirolli, P., Pitkow, J., and Lukose, R. (1998). Strong Regularities in World Wide Web Surfing. *Science*, 280(5360):95–97.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (ACM TOIS)*, 20(4):422–446.

Juvina, I. and van Oostendorp, H. (2004). Predicting user preferences: from semantic to pragmatic metrics of web navigation behavior. In *Proceedings of the conference on Dutch directions in HCI*, page 10, New York, NY, USA. ACM Press.

Kazai, G. and Lalmas, M. (2006). eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *ACM Transactions on Information Systems (To Appear)*, 24(4):503–542.

Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Lalmas, M. (1997). Dempster-shafer's theory of evidence applied to structured documents: Modelling uncertainty. In Belkin, N. J., Narasimhalu, A. D., and Willet, P., editors, *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 110–118, New York. ACM.

Lalmas, M. and Kazai, G. (2006). Report on the ad-hoc track of the inex 2005 workshop. *SIGIR Forum*, 40(1):49–57.

Larsen, B., Malik, S., and Tombros, A. (2006). The interactive track at INEX 2005. In Fuhr, N., Lalmas, M., Malik, S., and Kazai, G., editors, *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, 28-30 November 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 404–417. Springer-Verlag.

Levene, M., Borges, J., and Loizou, G. (2001). Zipf's Law for Web Surfers. *Knowledge and Information Systems*, 3(1):120–129.

Pandit, S. and Olston, C. (2007). Navigation-aided retrieval. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 391–400, New York, NY, USA. ACM Press.

Pehcevski, J. and Piwowarski, B. Evaluation metrics. to be published in the Encyclopedia of Database Systems, Springer.

Piwowarski, B. and Dupret, G. (2006). Evaluation in (xml) information retrieval: expected precision-recall with user modelling (eprum). In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 260–267, New York, NY, USA. ACM Press.

Piwowarski, B. and Gallinari, P. (2004). Expected ratio of relevant units: A measure for structured document information retrieval. In Fuhr, N., Lalmas, M., and Malik, S., editors, *Proceedings of the 2nd Workshop of the INitiative for the Evaluation of XML retrieval (INEX), Dagstuhl, Germany, December 2003*, pages 158–166.

Piwowarski, B., Gallinari, P., and Dupret, G. (2007). Precision recall with user modeling (prum): Application to structured information retrieval. *ACM Trans. Inf. Syst.*, 25(1):1.

Raghavan, V., Bollmann, P., and Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229.

Reid, J., Lalmas, M., Finesilver, K., and Hertzum, M. (2006a). Best entry points for structured document retrieval - part i: Characteristics. *Inf. Process. Manage.*, 42(1):74–88.

Reid, J., Lalmas, M., Finesilver, K., and Hertzum, M. (2006b). Best entry points for structured document retrieval - part ii: Types, usage and effectiveness. *Inf. Process. Manage.*, 42(1):89–105.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA. Out of print, available online from http://www.dcs.glasgow.ac.uk/Keith/Preface.html.

Sellen, A. J., Murphy, R., and Shaw, K. L. (2002). How knowledge workers use the web. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–234, New York, NY, USA. ACM Press.

Sparck Jones, K. and Willett, P., editors (1997). *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Tombros, A., Malik, S., and Larsen, B. (2005a). Report on the inex 2004 interactive track. *SIGIR Forum*, 39(1):43–49.

Tombros, T., Larsen, B., and Malik, S. (2005b). The interactive track at INEX 2004. In Fuhr, N., Lalmas, M., Malik, S., and Szlavik, Z., editors, *Proceedings of the 3rd Workshop of the INitiative for the Evaluation of XML retrieval (INEX), Dagstuhl, Germany, December 2004*.

Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

Weinreich, H., Obendorf, H., Herder, E., and Mayer, M. (2006). Off the beaten tracks: exploring three aspects of web navigation. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 133–142, New York, NY, USA. ACM Press.

Wheeldon, R. and Levene, M. (2003). The best trail algorithm for assisted navigation of Web sites. *Web*

Congress, 2003. Proceedings. First Latin American, pages 166–178.

White, R. W. and Drucker, S. M. (2007). Investigating behavioral variability in web search. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 21–30, New York, NY, USA. ACM Press.