Workshop on Novel Methodologies for Evaluation in Information Retrieval

European Conference on Information Retrieval – ECIR 2008, Glasgow, United Kingdom, 30 March 2008

Information retrieval is an empirical science; the field cannot move forward unless there are means of evaluating the innovations devised by researchers. However the methodologies conceived in the early years of IR and used in the campaigns of today are starting to show their age and new research is emerging to understand how to overcome the twin challenges of scale and diversity.

With such challenges in mind that we decided to hold the first Workshop on Novel Methodologies for Evaluation in Information Retrieval. The workshop is composed of long and short papers covering a range of important evaluation methods and tools. In addition, two invited talks from well known researchers in the evaluation field – Tetsuya Sakai of NewsWatch and Martin Braschler – provide a alternative perspective.

The workshop was chaired by Mark Sanderson; with co-organisation from Julio Gonzalo, Nicola Ferro and Martin Braschler. The papers were peer reviewed by members of our PC committee: they are...

- Paul Clough — University of Sheffield
- Franciska de Jong — University of Twente
- Thomas Deselaers — RWTH Aachen University
- Norbert Fuhr — University of Duisburg
- Frederic Gey — U.C. Berkeley
- Donna Harman — National Institute of Standards and Technology
- Gareth Jones — Dublin City University
- Noriko Kando — National Institute of Informatics
- Jussi Karlgren — Swedish Institute of Computer Science
- Bernardo Magnini — ITC-irst
- Paul McNamee — Johns Hopkins University
- Henning Müller — University & University Hospitals of Geneva
- Stephen Robertson — Microsoft Research
- Tetsuya Sakai — NewsWatch Inc.
- Diana Santos — Linguateca Sintef
- Peter Schäuble — Eurospider Information Technologies
- Ellen Voorhees — National Institute of Standards and Technology
- Christa Womser-Hacker — University of Hildesheim

The workshop was supported by the TrebleCLEF project: a coordinated action funded under ICT-1-4-1 Digital libraries and technology-enhanced learning; grant agreement 215231. More information about the TrebleCLEF can be found at http://www.trebleclef.eu/.

Table of Contents

Workshop on Novel Methodologies for Evaluation in Information Retrieval

30th European Conference on Information Retrieval (ECIR 2008)
March 30, 2008, Glasgow, United Kingdom

Workshop Program

9:00–9:05 *Welcome and Opening*
Mark Sanderson

9:05–9:50 Invited Talk
Evaluation: From Traditional Search to Exploratory Search
Tetsuya Sakai

9:50–10:30 Session 1: Evaluating the User Viewpoint
9:50–10:15 *Towards the Evaluation of Literature Based Discovery Systems*
Ulises Cerviño Beresi, Mark Baillie, and Ian Ruthven
10:15–10:30 *Changing the Subject — One Way of Measuring Trust in Information*
Jussi Karlgren

10:30:–11:00 BREAK

11:00–11:20 Session 2: Experimental Collections for Web Evaluation
11:00–11:10 *A Large Time-Aware Web Graph*
Paolo Boldi, Massimo Santini, and Sebastiano Vigna
11:10–11:20 *Compressed Collections for Simulated Crawling*
Alessio Orlandi and Sebastiano Vigna

11:20–12:10 Session 3a: Evaluation Challenges for New Tasks
11:20–11:45 *To Separate or Not to Separate: Reflections about Current GIR Practice*
Nuno Cardoso and Diana Santos
11:45–12:10 *Dynamic Focused Retrieval of XML Documents and Its Evaluation*
Toshiyuki Shimizu and Masatoshi Yoshikawa

12:10–12:30 Discussion

12:30–14:00 LUNCH

14:00–14:40 **Invited Talk**
Evaluation of the Search Functionality of Enterprise Web Portals
Martin Braschler

14:40–15:30 **Session 3b: Evaluation Challenges for New Tasks**
14:40–15:05 *Large-Scale Interactive Evaluation of Multilingual Information Access Systems – the iCLEF Flickr Challenge*
Javier Artiles, Emma Barker, Paul Clough, Julio Gonzalo, Jussi Karlgren, Victor Peinado
15:05–15:30 *How Many Experts?*
Gianluca Demartini

15:30–16:00 **BREAK**

16:00–17:15 **Session 4: Alternative Evaluation Methods**
16:00–16:25 *VisualVectora: An Interactive Visualization Tool for Cumulated Gain-based Retrieval Experiments*
Kalervo Järvelin, Illka Vähämöttönen, Heikki Keskustalo
16:25–16:50 *Document Accessibility: Evaluating the Access Afforded to a Document by the Retrieval System*
Leif Azzopardi, Vishwa Vinay
16:50–17:15 *Angle Seeking as a Scenario for Task-Based Evaluation of Information Access Technologies*
Emma Barker, Joe Polifroni, and Robert Gaizauskas

17:15–17:30 **Discussion and Future Directions**

Towards the Evaluation of Literature Based Discovery

Ulises Cervino Beresi
The Robert Gordon University
School of Computing
Aberdeen, UK
ucb@comp.rgu.ac.uk

Mark Baillie
University of Strathclyde
Department Computer and
Information Sciences
Glasgow, UK
mb@cis.strath.ac.uk

Ian Ruthven
University of Strathclyde
Department Computer and
Information Sciences
Glasgow, UK
ir@cis.strath.ac.uk

ABSTRACT

A consequence of the natural fragmentation of science into specialist fields is that disjoint but logically related literature exist. Literature Based Discovery (LBD) is the science of making these connections more evident. One of the challenges of LBD research is how to evaluate new approaches in a standard framework. For example, new techniques are often measured against a small set of examples. Also, the notion of relevance in the context of LBD goes beyond topicality which is often used as the assessment criterion for information retrieval. This paper reports on a pilot study motivated by the need to define a standard protocol for LBD evaluation. The aim of the study was to observe the various criteria users employed when assessing the relevance of retrieved documents. The main findings from this study indicated that users judge relevance at several levels of granularity and share a preference for *exemplary documents* which provide an easy entry point to unfamiliar, but logically related, research fields.

1. INTRODUCTION

As research fields become more specialised academics tend to interact more with researchers and literature from their chosen speciality and less with research outside of their own specific area of interest. Consequently, the interaction between fields, through cross-referencing across fields and shared use of common literature, becomes reduced and related fields *detach* from one another. The result is relatively isolated and highly specialised bodies of literature, a phenomenon that has recently accelerated due to the increased rate of new publications available online[15].

This detachment of research fields means that academics who share common interests and approaches but who work in different areas can miss important connections. It is becoming increasingly challenging for most researchers, especially in established fields, to keep up-to-date with important developments in their own chosen speciality[16]. However, keeping track of useful new developments in allied fields is even more demanding, relying far more heavily on the inefficient processes of manual literature searching and browsing, chance discoveries through personal communications or selective manual dissemination of information. More often, cross-disciplinary connections have to be engineered through dedicated, but often small-scale, initiatives.

The aim of research into Literature Based Discovery (LBD) is to help *discover* these connections between seemingly unrelated disciplines by mining publicly available academic literature. This area of research is motivated by the findings of Don Swanson, who in the mid-80s discovered two disjoint literature bodies that were complementary, i.e. when put together, they suggested an answer to a question which was not previously published. Swanson saw the potential in this procedure – combining knowledge from both literature sets to form an answer – and started to systematically investigate it under the name of Undiscovered Public Knowledge, more recently known as LBD[15].

The most famous example of a successful LBD is that of the relation between dietary fish oil and Raynaud's disease[14]. Dietary fish oil has been shown to have several effects on the blood circulation of patients. These effects counter-balance those that are considered symptoms of Raynaud's syndrome, e.g. fish oil lowers blood viscosity and high blood viscosity is a symptom of Raynaud's syndrome. When put together, these two ideas suggested an answer to a question that was not previously published. As Swanson had illustrated, the creative use of online information seeking and retrieval applications could lead to the discovery and detection of potentially unintended logical connections by bridging the gaps between isolated literature, resulting ultimately in new breakthroughs in science. This is the problem of LBD.

Evaluation of LBD techniques typically involve identifying whether key concepts from the Swanson studies were promoted through these semantic representations. For example, the appearance of phrases such as "blood viscosity" and "platelet aggregation", which were important in connecting "Raynaud's disease" with "Fish oil". However, focusing on a small subset of exemplar examples limits the inferences that can be drawn. To begin to address this shortcoming, this paper reports on an initial investigation into the problem of evaluating systems for LBD within a more general framework.

To achieve this aim, we first expand on LBD and its underlying models in Section 2. Then in Section 2.2, we provide a review of the major approaches at evaluating LBD systems and in Section 3 we describe the pilot study. Section

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

4 reports on some preliminary findings whereas Section 5 finalises the article with a discussion and suggested future work.

2. LITERATURE BASED DISCOVERY

In its most basic form LBD could be presented using the following analogy. Suppose that a scientist is interested in finding a novel treatment for Raynaud's syndrome. He might start by searching for the literature on the syndrome and reading more about the symptoms. Once he is familiar with the attempts at treating the syndrome, he may then embark on searching for an alternative treatment. By searching for a particular symptom he may be made aware of other diseases with the same symptom, alternative treatments that will have an effect on the symptom, etc. So, given the literature on the symptom, he could analyse which are the potential treatments related to it. This search model, the search for unknown but related information, is what has been called the open model[20].

Once our scientist from the previous example has a potential treatment in mind (whether because he already had some evidence supporting it or he had found it using the open model) he probably would be more interested in finding out what are the aspects that both the syndrome and the treatment have in common. That is, through which variables could the treatment affect the symptoms of the syndrome. Searching for those common aspects between two topics is what has been referred to as the closed model[20].

2.1 An abstraction of the models

The open model is an exploratory model where users begin a literature search on a known topic A (the syndrome in our previous example). From the resulting documents, a list of B topics is extracted (the symptoms for instance). This list is usually long so a post-processing step is needed where filtering/ranking is typically performed. The same procedure is then applied to each B topic and a list of new C topics is extracted (the potential treatments). The user is then presented with a list of potential ABC combinations.

The closed model is where users assume that a relationship exists between two known topics A and C (syndrome and potential treatment). Using topics A and C as a starting point, both literature sets are retrieved and from them common B topics are extracted (the interacting variables). Again, a post-processing step is usually taken on the resulting list.

2.2 Examples of evaluation in LBD

The main purpose in evaluation, whether in IR or any other area, is to measure the sensitivity of the measured variables to changes in system parameters. The evaluation methodology in Laboratory IR is fairly established, using test collections and relevance judgements as artifacts to simulate and measure these variables[19]. This is useful to measure the performance at a system level, however this approach leaves several questions unanswered that, in the case of LBD, might be central to measuring the success of a system. Our description of the LBD models suggests that the search behaviour might be a complementary 2-step one where searchers would firstly search for potential relations through the use of the open model and then find evidence on the selected ones using the closed model (also suggested by Weeber[21]). Even though the distinction between the two models may be a theoretical one, authors have found this to be beneficial when evaluating their systems since it simplifies matters by eliminating the possible interaction between models and helps isolate measurable variables.

Swanson's initial discoveries[14, 17] were confirmed by clinical experiments that provided evidence of the connections and by several papers published afterwards. Swanson's procedure had a promising start. Subsequent researchers attempted to replicate Swanson's discoveries to evaluate the performance of their systems. For instance, Gordon and Lindsay[9] focus on the the open model and use standard IR metrics (precision and recall) on the linking topics where Swanson's original discoveries are taken to be the relevance judgements. Taking a more holistic approach, Weeber[20] focuses on both the open model of discovery as well as the closed model. To evaluate the system, they too try to replicate Swanson's original discoveries and observe if they linking topics are within their top ranked discovered topics.

Prior to Gordon et al.[8] authors had concentrated on a single domain[1], considered a connection only if it was completely new or overlooked by the majority of the researchers in the field and the type of discoveries were of the form $disease \rightarrow treatment$. Gordon et al. approach these issues in a different, although related, fashion. Firstly, they apply LBD to the World Wide Web (WWW) therefore breaking free from the medical domain. Secondly, their starting topic is not a disease but rather a technology/technique and by trying to find new problems in which to apply them, they are effectively suggesting that not only complementary information is of value. Thirdly, they consider that LBD could be applied not just to make new discoveries but also to rediscover previous knowledge or to aid users by easing the transition into a new field of study. To evaluate their system they consult an expert in the field (genetic algorithms) and ask him to identify topics (retrieved by their system) that might be interesting to explore in connection to the starting technologies.

Approaches have followed general guidelines when it comes to evaluating systems. Swanson's initial discovered links are used as ground truth (relevance judgements in IR speak) and researchers observe how high the links are ranked by their systems. Results obtained with this methodology must be interpreted with care as has always been the case. Being tied to a collection also implies that results might not be as generalisable as one would like them to be. Moreover, restricting the evaluation to the replication of so few discoveries may lead to overfitting (tailoring of systems to accomplish this goal only) which also may lead to poor generalisation. In other words, the findings and inferences made from this small set of studies may not extrapolate to the wider problem of LBD.

2.3 Approaching an evaluation framework

The evaluation of LBD systems presents a considerable challenge since several aspects have to be covered. Most of these aspects have already been considered by Swanson (although in a different context) resulting in nine postulates of impotence[16]; a set of truisms for Information Retrieval research. These postulates, interpreted in the context of LBD, serve as guidelines to the design of an evaluation methodology. We briefly review here the ones we considered imme-

[1]There is another example of these efforts to branch out the use of LBD techniques in different domains in [6]

diately relevant. In LBD, users are in search for unknown, although related, information. It is, therefore, practically impossible for them to state their requests with any degree of fidelity (first postulate of impotence). Secondly, LBD is about finding logically related literatures that, when put together, will manifest the relations between topics. Therefore documents cannot be assessed independently of each other (third postulate of impotence). Thirdly, "aboutness" (or topicality) is certainly not what the searcher is after. It is not enough that the retrieved information is "on topic", it must also be related and complementary. Although being "on topic" will provide a floor to even start considering relevance, the multidimensional and dynamic essence of relevance must be acknowledged (sixth postulate of impotence). Fourthly LBD systems are interactive by nature. It is the interaction between the user and the system and between models what will help users to find hidden relations and derive full relevance (seventh postulate of impotence). To accommodate these considerations, the evaluation methodology would have to be such that it allowed to be done in a controlled environment and still as realistic as possible.

In the following section we report on a pilot study. The study was designed keeping Swanson's postulates in mind and was based on the methodology proposed by Borlund in[5]. Borlund's methodology is heavily based on the use of simulated work task situations. The aim of the study was to observe which components were the most important for users when doing LBD type of searches.

3. MATERIALS AND METHODS

The key components to the design and implementation of the pilot study were the use of simulated work task situations to trigger the participants' information needs and the gathering of verbal data, i.e. the recording of talk-aloud protocols. Moreover, while designing the study we attempted to provide practical approximate solutions to some of Swanson's postulates of impotence[16]. In the following subsections we describe each of these components in more detail.

3.1 Experimental design

The study consisted of two sessions with a gap between them of no more than a week. This time gap was chosen so that results from the first session would still be present in the participant's mind when doing the second session.

In the first session participants were given the task of finding five documents that described their area of research or an aspect of it (the task description can be seen in figure 1). To do this a simple keyword search system was provided. The only restriction applied in this session was that they had to provide us with exactly five documents as stated in the task description. These initial documents initiated the open search acting as a representation of the user's starting topic, i.e. they served as a request of the form "here's a description of my area of research, please show me the related topics to it".

As it was described in section 2.1 there are two stages in the open model:

1. Finding related intermediate topics (B topics) for the starting topic (the A topic)

2. For each intermediate B topic, finding related topics (C topics)

Figure 1: Task given to participants in session one of this study

> Dear Participant,
> I'd like to ask you to search for documents that describe your area of research or an aspect of it. Whenever you think you have found one, please write down the document ID (located at the top of the viewing window) on the provided sheet. The purpose of this search is so that the system under test can then suggest topics that might be related to your area of research for you to further investigate.

To model topics we used Probabilistic Topic Models (PTM)[13]. PTM are statistical models in which topics are represented as a probability distribution over a vocabulary. Representing topics as probability distributions means that each topic z defines a probability $P(w|z)$ of generating the word w. For each of the topics extracted from the initial set of five documents, the top three words w, ranked by $P(w|z)$, were used as a query to retrieve 50 more documents. This is a form of relevance feedback[11] and the assumption behind this procedure is that the retrieved documents will also contain topics related to the original topic, i.e. intermediate B topics. The topics in these documents were extracted (the B topics) using the same strategy and the top three terms of each topic were used as a query to retrieve one hundred more documents. The set of final topics (the C topics) were then estimated on each of the new document sets.

The final C topics were ranked according to a simple ranking function $rank(C) \propto P(C|B)P(B|A)$ were $P(t_i|t_j) \propto \prod_{w_i \in t_i} P(w_i|t_j)$ and the top ten topics were presented to the user in the second session. The participants were asked to investigate three of these topics and their relation to their area of research (the initial A topic). To complete the search task a time limit of one hour was allocated.

3.2 Collection

The collection searched by the participants is a collection of articles from several volumes of the Communications of the ACM (CACM)[2]. The original documents were in the Portable Document Format. Indexing was performed on the extracted plain text and a standard list of terms (stopwords) was used to remove commonly occurring words such as *the, a, at,* etc. The collection contains 7028 general articles covering several areas of Computer Science. The average document length is 2676 terms with 85% of the document containing between 0 and 5000 terms (after stopword removal). The longest document contains 34184 terms and the shortest only 79 terms.

3.3 Users group

The user group consisted of Computer Science PhD students. The choice of the group was motivated not only by the availability and access to it, but also because PhD students are well versed on their research topic. We assumed that they were be able to judge documents on their research

[2] `http://www.acm.org/publications/cacm`

Figure 2: Simulated work task description

Simulated work task situation: You just got out of a supervisory meeting and got bad news. Even though the work you've been doing is very good, it is somewhat too constrained/specific. Your supervisor suggested you should look for potential areas that might benefit from the techniques/theories you've developed as part of your PhD. Your supervisor suggested you identify these potential areas as well as find the pertinent literature so you can discuss them with him/her in your next meeting.
Indicative request: Find, for instance, about a technique employed in another area that might share commonalities with your work, e.g. an algorithmic process that could be abstracted or refined.

and also on potentially related fields, i.e. judge the *relatedness* of the information in regards to their research topic. Moreover, we also assume that they are at least familiar with literature searches and traditional search systems. For this study the size of our user group was 4 participants.

3.4 Tasks

A simulated work task situation is a brief description of a *real life* situation which results in the user using the IR system. It also not only ensures realism in a controlled environment but serves two main purposes, from[5]:

1. it triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real life; and

2. it is the platform against which situational relevance is judged

The work task description used in our pilot study was carefully crafted before recruiting the participants. The description can be found in figure 2. The information needs triggered by this task are very specific: the need for related information outside the user's area of research. This meant that users had to reach out to possibly unknown fields to *broaden* their work.

3.5 Measurements

Once the first session finished, participants were interviewed using open ended questions. During the second session participants were asked to talk-aloud as it progressed. Talk aloud protocols allow the experimenter to observe first hand and in a realistic manner how participants interact with the task they are trying to complete. It requires participants to verbalise anything and everything that is going through their minds as they interact with the system. Performing protocol analysis usually involves a number of steps to follow (for a much more in depth description please refer to Ericcson[7]):

1. The participant talks aloud during the experiment and this is recorded either in an audio tape or a video tape.
2. The recording is transcribed.
3. Data is segmented (divided into "utterances").
4. The experimenter chooses/designs a coding scheme.
5. The segmented data is encoded.
6. The encoded data is analysed.

For the first step participants were instructed to verbalise all that came through their minds and prior training was provided in the form of a ten minutes session using the LBD system on an example search task. Only digital audio recordings were done using a microphone connected to a PC. These recordings were then transcribed (step 2 of the protocol).

3.5.1 Analysis

In this study we gathered four types of information:

1. Interaction information,
2. Information about the user's intentions regarding the documents,
3. Relevance criteria information and
4. Information about relations between topics.

Step 4 of the talk aloud protocol requires that a coding scheme is used to tag the utterances obtained in step 3. Coding schemes that suited our purposes weren't readily available to analyse the data gathered in the sessions therefore we resorted to creating our own scheme. The encoding of the utterances was made according to the following criteria:

- Interaction: any utterance that indicates the participant is performing an operation on/with the system or interacting with it, e.g. reading, clicking on a document's surrogate, etc.
- Intent: any mention of the participant's intentions regarding the obtained information, e.g. using it to impress his/her supervisor.
- Relevance Criteria (RC): any mention of factors that affect the participant's choices regarding whether they are to keep or not a document, e.g. if the user picks the document because it is a survey.
- Trigger/Link: any mention of previously seen documents (selected or not) that affects the participant's behaviour (selection or consideration mainly). This denotes that the mentioned document is somehow related to the document being operated on.

3.5.2 Relevance Criteria

A second level of encoding was performed on the utterances coded as *Relevance Criteria (RC)*. Our intentions behind the use of this second level of encoding were to gain a better understanding of which of all the mentioned criteria is most important for users when performing this type of tasks. The coding scheme used was the one presented by Barry and Schamber in [3]. This coding scheme is the result of the comparison of the relevance criteria observed in two of their other studies (see Barry[1, 2] and Schamber[12] for more information). There is a total of ten criteria listed in this coding scheme. Here we briefly review them:

- Depth/Scope/Specificity: whether the information is in depth or focused, has enough detail or is specific to the user's needs. Also whether it provides a summary or overview or a sufficient variety or volume.
- Accuracy/Validity: whether the information found is accurate or valid.
- Clarity: whether the information is presented in a clear fashion. This includes well written documents and well as the presence of visual cues such as images.
- Currency: whether the information is current or is up to date.
- Tangibility: whether the information relates to tangible issues, hard data/facts are included or information provided was proven.
- Quality of Sources: whether the quality of the information can be derived from the quality of the sources of it. This includes authors as well as publications.
- Accessibility: whether there is some cost involved in obtaining the information.
- Availability of Information/Sources of Information: whether the information is available at that point in time.
- Verification: whether other information in the field, or the user, agrees with the presented information.
- Affectiveness: whether the user shows an affective or emotional response when presented the information.

According to the authors, *accessibility* refers to the cost or effort involved in obtaining the information. Effort, according to their interpretation, refers to physical and not mental effort. For instance, if a document is available only through an interlibrary loan, then it would require physical effort from the user to obtain it. Cost refers to possible fees involved in obtaining such documents. In our study documents were readily available and no fees were involved in obtaining them. Since the mental effort necessary to process the information is not interpreted to be a type of "effort", we didn't expect this criterion to be observed. By this we didn't mean to deny the existence of this type of effort. We merely stated that utterances expressing this would not be tagged with *accessibility*. However, as we will discuss in sections 4.3 and 4.4, mental processing effort might have played a crucial role.

In this study we decided to interpret *availability* as defined by Schamber[12]. This means that the code refers to physical availability of the document itself and not to personal availability nor environmental availability (Barry's interpretation of the code). Since documents were available at all times in our system we didn't expect this criterion to be observed.

Verifying the information coming from an unknown field is very hard to do for a newcomer to the field. The task given to the participants required them to branch out to unknown (to them) areas of science placing them in the spot as newcomers. Considering this, we didn't expect this criterion to be observed very often. We did expect, though, *information novelty* to play an important role when users judged documents. We included a code that would account for this. In the study done by Barry[2] three types of information novelty are mentioned:

- Content novelty: whether the information is new to the user
- Source novelty: whether the source of the document is new to the user, e.g. an unknown author
- Document novelty: whether the document is new to the author

We grouped all three of them under the label *seen before*. We used this code to tag utterances that expressed when a document had been seen before (in the current session or not), if the document was known to the participant or if the information was already known. We expected to observe more occurrences of document and information novelty however source novelty could have well appeared, for instance, in the form of a known author writing in a different field.

4. RESULTS

Different levels of analysis were performed on the data. The following sections present the findings. Firstly we describe the nature of the data captured. Next, we consider the observed relevance criteria; which aspects were the most important for the participants when doing an LBD type of task. Lastly, we analyse a particular criterion more in depth.

4.1 Nature of the data

A total of 868 codes were assigned to utterances across participants, with some utterances being assigned more than one code. *Interactions* were the most observed with a total of 359 (41.36%) followed by *relevance criteria* with 351 (40.44%), *intent* with 137 (15.78%) and *trigger/link* with 21 (2.42%) occurrences. We asked a second researcher to code a random sample of 50 utterances from the transcriptions. The second researcher achieved an 87% agreement.

4.2 Relevance criteria

The distribution of relevance criteria observed is shown in table 1. Overall, criteria dealing with the *tangibility* and with the *depth/scope/specificity* of the information were the most common. Recalling the explanation of the code *tangibility* from section 3.5.1 it is not surprising that this was the most common criterion used by the participants (it appeared 137 times representing a 39% of all the relevance criteria mentioned). *Tangibility* refers to the contents of the document, the actual information contained within it. Out of 137 occurrences of this criterion, 37 (27%) are mentions of what is usually described as "aboutness" or "topicality". This suggests that even when referring to the contents of a document, users found that "being on topic" was of limited importance. Some examples of the utterances coded with *tangibility* are:

- "[the document is] talking about"
- "it does illustrate"
- "an initial application"

The second most important criterion observed, with a total of 71 occurrences (20%), was *depth/scope/specificity*. This criterion is more ambiguous since it deals not only

with scope, but also with specificity, volume, detail and even genre of the document that contains the information. Reasonably so, participants were quite interested in these properties of documents. The observation of this criterion suggests, as many authors anticipated, that relevance as a whole doesn't depend solely on topicality (which in this study is included in *tangibility*). We delay the analysis of this criterion until we reach section 4.3. Examples of the utterances coded as *depth/scope/specificity* include:

- "general summary"
- "detailed enough"
- "lots of information"

As it was expected, utterances regarding the novelty of the information were the third most common. Encompassing document novelty, content novelty and sources novelty the code *seen before* was used to tag utterances 57 times (16.5%). According to almost all authors novelty, in the sense of something being "new", in LBD is an important factor. This is observable in that of all the mentions coded with *seen before* 26 (45%) were negative, i.e. the user decided not to pick the document when it wasn't "new" to him. This might seem contradictory with the belief that only "new" information is of value in knowledge discovery. A much more intuitive result would have been if all the mentions had been negative, i.e. the user had always rejected a document when it wasn't "new". However this was not the case.

Negative mentions of this criteria followed the expected pattern of "I've seen this before therefore I'm not interested in it". Examples of these negative utterances are:

- "I've seen already"
- "it is that damn article again!"
- "our old friend"

Positive mentions, on the other hand, followed a similar but reversed pattern: "*because* I've seen this before I'm interested in it". Some examples of the positive mentions include:

- "always getting that article" (the document is retrieved for different intermediate topics)
- "again here we have" (the document is retrieved for different intermediate topics)
- "it was this *TOPIC* again" (a new document on the same topic is retrieved for a different intermediate topic)

Some participants seemed to interpret the reoccurrence of a document as a positive reinforcing sign rather than a negative sign. Documents that were retrieved for different intermediate topics were deemed very relevant because they kept "cropping up" everywhere. Perhaps a document seen under a different light (a different intermediate topic) gains a new interpretation.

Affectiveness and quality of sources were the fourth and fifth most common codes observed with 44 (14%) and 31 (9%) occurrences each. When no other indicators of the quality of the information are present (which might be the case when investigating unknown fields of research) resorting to the reputation of the authors, their affiliation and/or the reputation of the publications seem to be a sensible approach. Participants expressed this in ways such as:

- "I see the name of"
- "never heard of him"
- "he's guest editor"

The code *clarity* was observed only 4 times (1%) whereas *currency* 2 times (0.5%). Finally, as we had expected, the codes *accessibility* and *availability* were not present. This reflects the fact that documents were always available and that there were no costs (physical or monetary) involved in searching/retrieving when using our system. A plausible explanation for why we didn't observe utterances referring to the verifiability of the data found is that, as it was briefly discussed in section 3.5.2, the participants were exploring new territories making the verification of the information found hard to do.

Table 1: Relevance Criteria and their occurrence across participants

Relevance Criterion Label	Global occurrences	Percentage
Depth/Scope/Specificity	71	20%
Accuracy/Validity	-	-
Clarity	4	1%
Currency	2	0.5%
Tangibility	137	39%
Quality of Sources	31	9%
Accessibility	-	-
Availability of Information/Sources of Information	-	-
Verification	-	-
Affectiveness	49	14%
Seen before	57	16.5%

4.3 Depth, Scope and Specificity

The most mentioned criterion, *tangibility*, deals with the contents of documents, specifically with what is referred to as "hard data", e.g. concrete examples. The second most commonly mentioned criterion deals with several aspects such as the volume and/or variety of information presented, the level of detail and even the genre of the document that contains it. Its code is *Depth/Scope/Specificity*. Out of 71 occurrences, 34 (47%) were references to *exemplary documents*. According to Blair and Kimbrough, "exemplary documents are those documents that describe or exhibit the intellectual structure of a particular field of interest"[4]. Vocabulary varies significantly across research fields in science. One function these exemplary documents perform is to provide a definition of the words included in these vocabularies. Moreover, by mapping the structure of the field they also provide a context in which the vocabulary is to be interpreted. An example of such documents in the scientific community is the *survey article*. Survey articles summarise

up to a point in time the most important advances and issues to be treated in a field, include a list of references to follow up and possibly a list of important academics and institutions. We could argue, up to a certain point, that an exemplary document of this type may, for instance, ease the entrance of a newcomer to the world of research in that field. It would ease this entrance by not only providing an overview of the field itself but also of the pertinent vocabulary and major players in it. It's reasonable to observe users preferring a survey article to the latest article on a specific topic when getting acquainted with the field being investigated. Amongst the references to this preference we found:

- "general summary"
- "gentle introduction"
- "good overview"

A possible answer for why this type of documents are preferred by users when entering a new field is because these documents may have a high ratio of information obtained vs. processing effort (both concepts introduced in Harter's theory of psychological relevance[10]). Perhaps by providing this roadmap to the field, together with the associated technical vocabulary, survey articles offer plenty information in exchange for little mental processing effort. This would afford users a quick judgement to whether or not it would be beneficial to go deeper into the field and search for possible connections.

4.4 Unclassified utterances

Several utterances were left unclassified due to their ambiguous nature. They might also require an extension to the coding scheme devised and presented in section 3.5.1. Some examples of these utterances are:

- "not sure whether I see the connection"
- "much more connected"
- "not sure how these things connect"

The utterances refer to connections either between articles or between areas of research. In the case of the positive examples one could argue that these utterances should be coded with *trigger/link*. However, in the case of these unclassified utterances there wasn't enough context to decide whether to code them with this tag or not. This was aggravated by the fact that *trigger/link* refers only to links between documents and not topics, making the need for a disambiguating context even greater.

In the case of the negative examples there could be several explanations for why the user is not being able to infer the connection such as:

- The documents/topics are indeed not connected,
- The documents/topics are connected but the user is not being able to process the information satisfactorily to infer the connection or
- The documents/topics are connected but the user has deemed the connection unimportant or irrelevant

This is an issue to investigate further. One possible implication would be that the information presented in these documents is complex requiring a big processing effort on the users behalf. This means that the ratio between the the information obtained vs. processing effort is small resulting in the document being judged irrelevant (or psychologically irrelevant[10]).

5. DISCUSSION AND FUTURE WORK

The aims of the reported study were to observe in a realistic but controlled fashion the manifestation of relevance criteria used by users when judging the relevance of documents in LBD. The study was largely based on the methodology for evaluation proposed by Borlund in [5] and its design was done keeping Swanson's postulates of impotence[16] in mind. Borlund's methodology is heavily based on simulated work task situations. These are descriptions of real life situations that develop in the users requiring the use of an IR system. To make use of this methodology we had to carefully craft the work task situations so that the appropriate information needs were triggered in our users: the need for related information outside the user's area of research. To complement the simulated work task situations we used talk aloud protocols. Talk aloud protocols permitted us to gather in a realistic and natural way data on the users' behaviour as they worked on the assigned task. One obstacle when doing protocol analysis is the choice of the coding scheme for tagging the utterances extracted from the transcriptions. We resorted to creating our own since no scheme that fitted our requirements was readily available.

From the data we were able to extract four types of information: interaction information, intent information, relevance criteria information and information about the relations between topics. When analysing the relevance criteria mentioned we observed that the two most common criteria referred to the contents of documents and to the properties of documents such as format, volume/diversity of information and genre. Users seemed to have a preference for what Blair and Kimbrough have referred to as exemplary documents[4]; documents that are representative of the intellectual structure of the collection of documents from that research field. Exemplary documents include survey articles, opinion pieces or editorials by leading academics, text books and seminal papers. Exemplary documents provide a topography of the intellectual landscape, as they discuss the issues of the field, demonstrations of how the field refers to, and also about these issues and topics. This type of document is important for searchers new to the area, as demonstration of how to solve existing problems can illustrate how solutions can be applied to new unresolved problems in their own research field. It would appear that exemplary documents may help by providing an entry point for a researcher to a new field of investigation. The popularity of this type of documents amongst the participants of this study may be influenced by the level of cognitive processing effort required to obtain the information from the documents. Harter suggested that the relevance of a document is guided by the ratio of this the amount of information obtained vs. the effort needed to process the information[10]. In the case of survey articles we suggest that this ratio is high. Harter's theory of relevance provides a plausible explanation not only to the observed preference of users towards exemplary documents but also to the observation of failed inferences. Several ut-

terances expressing the failure in establishing a connection between documents (or topics) were recorded and left unclassified due to the insufficient context around them. We have suggested that one possible explanation to this failed inferences is that the information found was complex and required to much cognitive processing effort, resulting in the user judging the (failed) connection irrelevant in the Harter sense. This is certainly an issue to investigate further.

The use of simulated work tasks to trigger information needs on users combined with the use of talk aloud protocols seemed to be an appropriate methodology to gather empirical data in this scenario. However a proper evaluation methodology should provide not only a mean of collecting data but also objective metrics with to measure the performance (success) of our systems. Swanson has recently taken a step in this direction by proposing modified versions of the traditional precision and recall metrics[18].

We have used Swanson's postulates of impotence[16] to guide the design of this study. By using example documents of a researcher's area of expertise as a starting point we tried to overcome what is stated in postulate number one: that the statement of a query may be practically impossible to come up with. This is of particular interest in LBD since the requested information is, a priori, unknown. We also observed that there is some evidence that supports postulate number three. Users explicitly mentioned the relations and links between documents. This provides evidence, even if little, that documents cannot be treated as isolated from each other and the appropriate metrics should acknowledge this. Should such metrics already exists, using our study design and coding scheme, researchers would be able to capture the appropriate data to conduct the evaluation of their systems.

The multidimensionality of relevance (sixth postulate) was explicitly acknowledged in the observation of different criteria used by users when judging the relevance of documents. Some criteria were more common than others and, although it was observed, "topicality" or "aboutness" represented only a small fraction of all the criteria observed.

We still believe that evaluating LBD systems is not a trivial task and we've observed that there are plenty factors affecting it. The appropriate mechanisms to capture the right data together with the right metrics must be made available before we can evaluate the performance of our systems. Small careful steps towards a satisfactory evaluation methodology are needed and we believe that this study might be one of them.

6. REFERENCES

[1] C.L. Barry. *The identification of user criteria of relevance and document characteristics: Beyond the topical approach to information retrieval*. PhD thesis, 1993.

[2] C.L. Barry. User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.

[3] C.L. Barry and L. Schamber. Users'criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2-3):219–236, 1998.

[4] D.C. Blair and S.O. Kimbrough. Exemplary documents: a foundation for information retrieval design. *Information Processing and Management*, 38(3):363–379, 2002.

[5] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):8–3, 2003.

[6] K.A. Cory. Discovering Hidden Analogies in an Online Humanities Database. *Computers and the Humanities*, 31(1):1–12, 1997.

[7] K.A. Ericsson and H.A. Simon. *Protocol analysis: verbal reports as data*. MIT Press Cambridge, MA:, 1993.

[8] M. Gordon, R.K. Lindsay, and W. Fan. Literature-based discovery on the world wide web. *ACM Trans. Inter. Tech.*, 2(4):261–275, 2002.

[9] M.D. Gordon and R.K. Lindsay. Toward discovery support systems: a replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil. *Journal of the American Society for Information Science and Technology*, 47(2):116–128, 1996.

[10] S.P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science and Technology*, 43(9):602–615, 1992.

[11] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145, 2003.

[12] L. Schamber. *Users'Criteria for Evaluation in Multimedia Information Seeking and Use Situations*. PhD thesis, Syracuse University, 1991.

[13] M. Steyvers and T.L. Griffiths. Probabilistic topic models. *Latent Semantic Analysis: A road to meaning*, 2005.

[14] D.R. Swanson. Fish oil, Raynaud's syndrome and undiscovered public knowlege. *Perspectives in Biology and Medicine*, 30:7–18, 1986.

[15] D.R. Swanson. Undiscovered public knowledge. *The Library quarterly(Chicago, IL)*, 56(2):103–118, 1986.

[16] D.R. Swanson. Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39(2):92–98, 1988.

[17] D.R. Swanson. Migraine and Magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31:526–557, 1988.

[18] D.R. Swanson, N.R. Smalheiser, and V.I. Torvik. Ranking Indirect Connections in Literature-Based Discovery: The Role of Medical Subject Headings. *Journal of the American Society for Information Science and Technology*, 57(11):1427–1439, 2006.

[19] E.M. Voorhees. The philosophy of information retrieval evaluation. *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF*, 2406:9–26, 2001.

[20] M. Weeber, H. Klein, L.TW de Jong-van den Berg, and R. Vos. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud–Fish Oil and Migraine–Magnesium Discoveries. *JASIST*, 52(7):548–557, 2001.

[21] Marc Weeber. Advances in literature-based discovery. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.

Changing the subject — One way of measuring trust in information

Jussi Karlgren
Swedish Institute of Computer Science
Kista, Sweden
jussi@sics.se

1. TARGET NOTION: TRUST

For the purposes of two recent student projects hosted at SICS, we defined a target notion based on *trust* in lieu of topical relevance. Specifically, the studies in question examined the effects of using annotation software viz Annozilla [2] and cooperation, contrasting paired test subjects to subjects working singly[1].

The studies were empirical, and based on a laboratory-style setting, where subjects recruited by notices posted in university halls were invited to work on a set number of tasks designed to be somewhat realistic in style but completely unrelated to any previous interest or activity on the part of the subjects. This research note discusses the target notion of *trust* defined for the student projects.

Most evaluations of information retrieval or information seeking presume the existence of some topic-related measure related to topical relevance. The everyday notion of topical relevance has been operationalised and formalised to the quantifiable relevance of TREC-like studies. This formal target notion of "relevance" is an effective tool for focused research. Much of the success of information retrieval as a research field is owed to this formalization. But relevance, in the form it is operationalised, has drawbacks.

Trying to extend the scope of an information retrieval system so that it is more task-effective, more personalized, or more enjoyable will practically always carry an attendant cost in terms of lowered formal precision and recall as measured by relevance judgments. This cost is not necessarily one that will be noticed, and most likely does not even mirror a deterioration in real terms – it may quite often be an artefact of the measurement metric itself. Instead of being the intellectually satisfying measure which ties together the disparate and vague notions of user satisfaction, pertinence to task, and system performance, it gets in the way of delivering all three.

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

In the studies referred to here, while the focus of the studies was investigating annotations and cooperative behaviour, respectively, they shared the common target notion of *trust*.

2. WHAT TO TRUST OR NOT?

How might one be able to establish whether the subjects trusted the information? While the measurement of trust is a research field in itself, and well approached with caution, these two student projects took the simple approach of asking the subjects themselves.

The subjects were presented with a web-based questionnaire which gave them two topics in turn each with a sequence of questions on the topic. The subjects were asked to find materials on the internet that pertained to the questions and to indicate whether they trusted the results or not. The topics were purposely chosen to be somewhat contentious and of current interest – both reflected recent discussions in the mainstream news media. This was to ensure that the topic itself would transcend the obvious – a vapid topic would not hold the intellectual tension necessary for trust or distrust.

3. MEASURING THE EFFECT OF TRUST

In order to measure trust, after completing each of the two topics, the subjects were given a paper questionnaire to fill out. A simple breakdown of answers is given in Table 1. While most users were somewhat careful about assuming they had found all information on the topic, and not entirely trusting as to its various qualities, a non-insignificant number users indicated that they had modified their opinion on the topic for both queries[1] and a somewhat larger number of subjects reported learning more about a topic.

Results from the student projects were mainly qualitative, but included the findings that subjects working in pairs were more likely to report learning more about a topic and reported higher level of trust in the found sources, while they retrieved fewer documents – which presumably reflects the benefit of cooperation and the attendand overhead effort associated with cooperative discourse. If the target measure had been topical relevance, the results would likely have shown a lowered recall for the cooperative condition. That specific data point would not significantly have improved the understanding of cooperation.

[1]This included, for the Aspartame question, unreported in the table, the test leader.

Aspartame					
	1	2	3	4	5
know	4	15	2	1	0
interest	0	3	6	6	7
learn	0	2	7	7	6
change	9	6	7	0	0
facets	2	6	9	4	1
trust	1	6	10	4	1
Echelon					
	1	2	3	4	5
know	14	5	2	1	0
interest	0	6	7	3	6
learn	1	4	8	6	3
change	12	6	3	1	0
facet	5	9	7	0	1
trust	0	7	11	3	1

Table 1: Self-reported aspects of trust in web sources for information

Crosstabulation was inconclusive, given the relatively small number of respondents, but showed e.g. that the user with the greatest previous knowledge did still change opinions for one of the topics.

4. CONCLUSIONS

Given controversial questions that interested them, subjects performed experiments with enthusiasm and reported that the experiment had influenced their state of mind. This forms an implicit test of trust in the retrieved material. While the respondents reported a medium, to low-medium range of trust in the materials, and did not believe they had found all pertinent facets of opinion pertaining to the topic, they still adjusted their opinions on the matter to some extent and reported having learned about the topic.

This attempt at evaluating trust both by explicit question and by indirect effect on the responents' state of mind gave rise to a number of questions. Setting ethical questions aside, the methodological issues are non-trivial. Firstly, editorial: how might one find questions that are suitably interesting (in this case, the students spent several days on formulating and testing questions, until the settled on the suitably provocative ones). Secondly, technical: how could this type of test be distributed to a larger number of respondents, and how can the results be calibrated to provide a stable and generalisable conclusion?

5. REFERENCES

[1] Djuna Franzén. *"Det är klart det är lättare när man är flera!" : en undersökning av samarbete inom informationssökning och tilltro till dokument på Internet*. Number 256 in Uppsatser inom biblioteks- och informationsvetenskap. Institutionen för ABM, Uppsala universitet, Uppsala, 2006.

[2] Åsa Johnson. *Tillit på webben – Annozilla som förtroendeingivande verktyg*. Number 2006:21 in Magisteruppsats. Institutionen Biblioteks- och informationsvetenskap, Högskolan i Borås, Borås, 2006.

Table 2: Post-topic questionnaire *(Translated from Swedish.)*

Did you have any previous knowledge of the topic?	
None	Know this topic very well
Did you find the topic interesting?	
Not at all	Very interesting
Did you learn more about this topic by completing this task?	
Nothing	A lot
Did you change your opinion on the topic after completing this task?	
Not at all	Completely new opinion
Did you find most facets and most different points of view for this topic during your session?	
No, one perspective only	Yes, all points of view
Do you trust the information you found?	
Not at all	Yes, fully

Table 3: Topic 1: The artificial sweetener Aspartame *(Translated from Swedish.)*

1.	What is Aspartame made of, and under what other names has been used for the same product?
2.	How many times sweeter than regular sugar is Aspartame?
3.	In what types of product is Aspartame used in Sweden?
4.	What company had latest the sole rights to manufacture Aspartame?
5.	Is using Aspartame products a good method to attain weight loss?
6.	Is Aspartame safe to ingest?
7.	Is Aspartame approved for human use all around the world?
8.	When was Aspartame first approved as a food sweetener?
9.	How high ADI-value does Aspartame have?
10.	Does ingesting Aspartame cause side effects?
11.	Are there categories of people who should not use Aspartame?

Table 4: Topic 2: Personal integrity on the internet *(Translated from Swedish.)*

1.	What two international treaties protect international communication?
2.	What are the five intelligence agencies that have signed the UKUSA agreement?
3.	What is TIA, total information awareness?
4.	Echelon is a globalt, digital communication tapping system based in the US. How does it work?
5.	How has the EU acted with respect to Echelon?
6.	To which e-mail program does the NSA have the encryption keys?
7.	What automobile corporation claims to have lost a major order to General Motors due to NSA communications intercepts?
8.	What did Hans Buehles do in Iran in 1992?
9.	What did Kjell Ove Widman do at Crypto AG?
10.	Does Sweden participate in Echelon in any way?
11.	Can a private individual avoid being tapped by Echelon?

A Large Time-Aware Web Graph

Paolo Boldi
Dipartimento di Scienze dell'Informazione
Università di Milano, Italy
boldi@dsi.unimi.it

Massimo Santini
Dipartimento di Scienze dell'Informazione
Università di Milano, Italy
santini@dsi.unimi.it

Sebastiano Vigna
Dipartimento di Scienze dell'Informazione
Università di Milano, Italy
vigna@dsi.unimi.it

ABSTRACT

We describe the techniques developed to gather and distribute in a highly compressed, yet accessible, form a series of twelve snapshot of the `.uk` domain. *Ad hoc* compression techniques that made it possible to store the twelve snapshots using just 1.9 bits per link, with constant-time access to temporal information. Our collection makes it possible to study the temporal evolution link-based scores (e.g., PageRank), the growth of online communities, and in general time-dependent phenomena related to the link structure.

1. INTRODUCTION

By now, several sources provide accessible snapshots of web data. The Stanford WebBase project, for instance, provides hundreds of Terabytes of such data. In the last years, the LAW (Laboratory for Web Algorithmics) focused its effort on web graphs instead.

Inside the DELIS project, an interest rose about *temporal* link analysis, that is, studying how the web-graph nodes and arcs evolves in time. Since without proper bounds the amount of information required by this activity is staggering, we decided to concentrate our efforts on gathering twelve monthly 100 Mpages snapshots of the `.uk` domain and store them in a format that would make temporal information accessible on a standard workstation.

The basis of our work is WebGraph [2], a framework for web graph compression that currently provides the best compression available (in terms of bits per link) and whose data is readily accessible using free Java or C++ code [6]. One of the main challenge of this work was to extend WebGraph so that it would compress efficiently labels representing temporal information. Moreover, we wanted to extend the standard WebGraph flexible approach, that makes it always possible to load data into main memory or access it in offline form (with a performance drop, of course).

2. GATHERING THE SNAPSHOTS

The snapshots have been taken at the start of each month, during a period of $7 - 10$ days, using the bandwidth provided by the Università degli Studi di Milano and a cluster of PCs that has been funded by the DELIS project. Some basic information about the snapshot is shown in Table 1.

Crawling parameters. As in any limited-size crawl, it is essential to define the crawl parameters (the stopping criterion is clearly that of reaching about 100 Mpages, without counting duplicates). We highlight the main features:

Crawl policy. We use UbiCrawler's [1] built-in per-host breadth-first visit. A number of threads scan in parallel distinct hosts, and newly discovered URLs are added to a queue. When a thread completes its visit, it extracts from the queue the first URL whose host has an IP address that is not currently visited, and starts visiting that host in breadth-first fashion.

Seed. The seed is a large (190 000 elements) set of URLs obtained from the Open Directory Project. The reason for such a large seed is that of making the crawl more stable and repeatable, and reduce the amount of spam (as links in the Open Directory Project are judged by humans).

Maximum number of pages per host. We limited each host to a maximum of 50 000 pages. This guarantees that we shall crawl at least 2 000 hosts, and limits the impact of web traps and database-driven sites.

Maximum inter-host depth. We do not delve more than 16 levels in a host. The main reason for a limit in depth is avoiding traps and also badly configured 404 pages, which sometimes generate an infinite number of links by prefix buildup.

URL normalisation. URLs are normalised following the strategy explained in the `BURL`[1] Java class. We apply all safe normalisations, escape all illegal characters, and treat in a special way square brackets as they are ubiquitously (although erroneously) used in an unescaped form.

Duplicate detection. Many pages are duplicates, and to detect their presence we maintain a set of 64-bit fingerprints obtained after stripping attributes (of HTML elements) and other non-relevant parts of the page. When a duplicate is detected we just store a pointer to the original page. About 25% of the overall pages happen to be duplicates.

3. ALIGNING THE SNAPSHOTS

The web is constantly changing, and network errors can affect the presence of a site or page during a given crawl, Nonetheless, Table 2 shows that we have a significant host overlap in the chosen twelve-months time span.

[1] BUbiNG URL.

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

	Pages	Size (GB)	GZip'd Size (GB)
June	112 386 763	1 893	402
July	136 956 559	2 287	477
August	141 395 895	2 424	507
September	148 965 298	2 756	546
October	129 558 491	2 336	478
November	150 146 132	2 637	546
December	144 489 446	2 552	525
January	151 578 113	2 651	553
February	153 966 540	2 692	564
March	151 427 461	2 568	545
April	150 606 689	2 700	559
May	150 054 551	2 658	556

	Nodes	Arcs	Size (GB)	bit/arc
June	80 644 902	2 481 281 617	0.89	3.07
July	96 395 298	3 030 665 444	1.16	3.30
August	100 751 978	3 250 153 746	1.23	3.25
September	106 288 541	3 871 625 613	1.32	2.93
October	93 463 772	3 130 910 405	1.03	2.83
November	106 783 458	3 479 400 938	1.16	2.86
December	103 098 631	3 768 836 665	1.34	2.77
January	108 563 230	3 929 837 236	1.38	2.72
February	110 123 614	3 944 932 566	1.39	2.74
March	107 565 084	3 642 701 825	1.34	2.84
April	106 867 191	3 790 305 474	1.36	2.79
May	105 896 555	3 738 733 648	1.30	2.69

Table 1: Per-snapshot full-text and web-graph stats.

	June	July	August	September	October	November	December	January	February	March	April	May
June	94 967	73 304	71 686	69 899	65 516	64 501	59 478	62 459	62 447	58 953	57 671	57 747
July		130 778	102 250	99 489	89 951	90 909	81 491	86 741	88 143	84 082	82 731	82 138
August			128 505	102 873	84 999	90 378	81 023	86 489	86 762	81 908	80 066	79 637
September				136 605	88 006	94 655	84 335	90 887	89 620	84 993	82 097	81 156
October					109 918	86 175	75 831	81 130	81 614	76 616	75 660	75 128
November						121 208	86 714	91 461	91 549	84 125	82 322	81 664
December							113 471	88 852	84 335	79 298	76 254	75 850
January								125 134	94 259	86 402	84 474	83 127
February									122 956	91 094	87 864	86 708
March										122 506	84 971	83 839
April											113 157	91 636
May												114 529

Table 2: Host overlap stats.

The first important step in getting a temporally labelled collection is *alignment*: identifying URLs in different snapshots that correspond to the same web page. Alignment is a non-trivial issue because if a URL is not static it might contain session-generated data (e.g., a session ID) that makes *de facto* identical URL appear to be distinct.

For the present collection, a radical choice was made: the only allowed URLs are static URLs (i.e., URLs that do not contain a question mark[2]). We plan to develop some reasonably sound alignment technique for dynamic URLs in the future.

4. TEMPORAL LABELLING A GRAPH

Once URLs are aligned, it is possible to build a *global* graph G that includes all static pages (and related links) appearing in each snapshot. The graph G must be labelled so that, for each node and each link, we can detect whether it was retrieved in a given snapshot. Essentially, we need to store twelve bits of information per node and per arc.

We decided to use the labelling facility implemented in WebGraph to store a twelve-bit label for each node and arc. To reduce significantly the space occupied, we generated 2^{12} canonical Huffman coders [5], one for each possible node label. Each coder contains an optimal, canonical code for the distribution of the labels of the arcs going out of nodes with a fixed label. The distribution on the outgoing arc is strongly dependent on the label of the source, and we exploit this fact to increase the compression ratio.

5. MEMORY AND DISK FOOTPRINT

Memory and disk occupation depend on which components are loaded in memory (as opposed to being accessed directly on disk). The underlying graph (representing each node and arc ever met during the twelve crawls) has 133.6 millions of nodes and 5.5 billions of arcs. WebGraph uses in this case ≈ 2.6 bits per link, resulting in a bitstream with a memory footprint of 1.7 GiB.

As we discussed previously, we use 2^{12} canonical Huffman decoders, which require around 120 MiB of core memory. Node and arc labels occupy around 1.2 GiB, implying a cost of just 2.16 bits per label. Since the overall graph represents 13.8 billion links, the cost per link of the overall representation is just 1.9 bits per link.

We still have to consider the about 200 million pointers that are necessary to access the graph bitstream and the label bit stream. WebGraph uses a broadword implementation [7] of the Elias–Fano representation of monotone functions [3, 4], which provides constant-time pointer access with minimal space occupancy. As a result, we use 9 bits per graph bitstream pointer and 11.6 bits per label pointer—a memory footprint of about to 500 MB.

Finally, the access to a labelled arc requires about 250 ns on an Opteron at 2.4Ghz, making it possible to apply standard algorithmic techniques requiring random access to the graph (for instance, whole visits) using a commodity workstation.

[2]This choice, unfortunately, cannot prevent *opaque* session-dependent URLs from generating noise in the collection.

6. REFERENCES

[1] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711–726, 2004.

[2] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[3] P. Elias. Efficient storage and retrieval by content and address of static files. *J. Assoc. Comput. Mach.*, 21(2):246–260, 1974.

[4] R. M. Fano. On the number of bits required to implement an associative memory. Memorandum 61, Computer Structures Group, Project MAC, MIT, Cambridge, Mass., n.d., 1971.

[5] D. S. Hirschberg and D. A. Lelewer. Efficient decoding of prefix codes. *Comm. ACM*, 33(4):449–459, 1990.

[6] J. Ratkiewicz. WebGraph++, 2006. `http://homer.informatics.indiana.edu/~nan/webgraph/`.

[7] S. Vigna. Broadword implementation of rank/select queries. In *Proc. WEA 2008*, Lecture Notes in Computer Science. Springer–Verlag, 2008.

Compressed Collections for Simulated Crawling

Alessio Orlandi
Dipartimento di Informatica
Università di Pisa, Italy
aorlandi@di.unipi.it

Sebastiano Vigna
Dipartimento di Scienze dell'Informazione
Università di Milano, Italy
vigna@dsi.unimi.it

ABSTRACT

Collections are a fundamental tool for reproducible evaluation of information retrieval techniques. We describe a new method for distributing the document lengths and term counts (a.k.a. within-document frequencies) of a web snapshot in a highly compressed and nonetheless quickly accessible form. Our main application is reproducibility of the behaviour of focused crawlers: by coupling our collection with the corresponding web graph compressed with WebGraph [3] we make it possible to apply text-based machine learning tools to the collection, while keeping the data set footprint small. Finally, we describe a collection based on a crawl of 100 Mpages of the `.uk` domain, publicly available in bundle with a Java open-source implementation of our techniques.

1. INTRODUCTION

Focused crawling is a term originally given by Chakrabarti *et al.* [6] to denote a crawling activity that gathers *relevant* pages (as opposed to all pages). The notion of relevance is dependent on the particular application: for instance, a user might provide a set of interesting pages as an example.

The main issue in designing a focused crawling policy is the prioritization of the queue containing the frontier. Since the pages in the frontier are known, but not crawled, to maximise the *harvest rate* (the number of relevant pages gathered averaged over time) it is essential to choose from the frontier either relevant pages, or non-relevant pages that will quickly lead to relevant ones.

Performances of crawling policies can be studied either on the real Internet or via visit simulation on stored graphs. The latter provide a scientifically reproducible playground suitable for comparative studies among strategies and for parameter testing. In our case, simulation requires both the graph structure and the page content. This work has been motivated by the absence of a standardized reference collection for focused crawling, and more in general for the application of text-feature based machine learning techniques to the web. Such a collection should be:

- *Large*: representing a sizeable portion of the World Wide Web.
- *Heterogeneous*: containing different topics and types of sites. For instance, a crawl containing pages only from a few domain names is not.
- *Open*: we require the collection to be small enough for distribution via network or optical discs and to be available for other researchers.
- *Easy*: provided with a portable software architecture for fast decompression and access.
- *Shrinkable*: since we expect the collection to be large, we want to include the possibility of sub-collections of different smaller size for incremental testing.

To accommodate the above goals and avoid copyright problems, we assume that page file format and structure are irrelevant toward classification; more precisely, we will restrict the data derived from document content to the *counts* (i.e., the number of occurrences, a.k.a. within-document frequencies) of the terms appearing in the document, and to the document lengths in words. This information is sufficient to rebuild all weighting schemes we are aware of (TF/IDF, BM25, etc.) and it is powerful enough to support many known machine learning tools (e.g. SVMs or Bayesian classifiers). On the other hand, it is clear that our choice has significant limitations, as it discards pieces of information that some focused crawlers could need (e.g. anchor text delimiters or positional information).

There are two issues in replicating a web crawl: describing the graph structure, and describing the page content. For the graph structure, we rely on the WebGraph framework [3]. We will provide instead tools to access quickly the information about the page contents.

2. ARCHIVES OF SUMMARIES

We now present our main contribution—the design and implementation of what we call a *bitstream archive of summaries*. More precisely, a summary is a tuple

$$\langle \ell, s, L = \langle t_0, f_0 \rangle, \langle t_1, f_1 \rangle, \ldots, \langle t_{s-1}, f_{s-1} \rangle \rangle,$$

where ℓ is the document length in words, s (the *size* of the summary) is the number of distinct terms appearing in the document and L is a list of s term/frequency pairs. We assume that our documents are numbered so that the node number in the graph matches its document identifier (albeit content might be missing for some pages due to HTTP server errors).

We want to use *gap encoding* techniques that are common in the compressed storage of inverted indices in this setting. To this

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

purpose, we make the key observation that *terms should be renumbered by frequency rank*. More explicitly, term 0 is the term of highest frequency, and so on.[1] As a result, the term lists contain numbers which are often smaller than original ones: this phenomenon improves significantly the results of gap encoding.

For coding counts we use the following observation: terms that have a high frequency (e.g., stopwords) have usually also a high count. Thus, we list counts in *inverse* order (e.g., starting from the less frequent term) and gap-code them. We have of course no guarantee that the counts will be increasing, so we use the mapping $\nu : \mathbf{Z} \to \mathbf{N}$ thus defined:

$$\nu(x) = \begin{cases} 2x & \text{if } x \geq 0 \\ 2|x| - 1 & \text{if } x < 0. \end{cases}$$

The difference between consecutive counts (which can be negative) is mapped through ν and the resulting (small) integer coded using an instantaneous code.

Summaries in the archive are stored contiguously in the format described above.

Choosing the codes. Once the strategy for storing our data is laid out, we just have to discuss which instantaneous codes we should use. As often happens on the web, several of the distributions involved are power laws or similar distributions. Thus, beside the classical γ and δ codes, we also experimented with ζ codes [4], which were designed for power laws. Our suggestion is γ coding for the gaps of counts, δ coding for document lengths, ζ_3 coding for the sizes of the summaries and ζ_2 coding for the gaps of the term list.

Random access. The bit stream containing all documents clearly provides sequential access to the entire collection, as each summary is self-delimiting. On the other hand, simulating a crawler requires random access to the entire archive. To avoid keeping in memory a very expensive array of pointers explicitly, we resort to *rank/select* data structures. Given a set of integers S, ranking an integer x gives the number of elements of S smaller than x, whereas selecting a rank r gives the r-th element in S (counting from 0). In particular, we use a broadword implementation [9] of Elias-Fano dictionaries [7, 8], which provide (almost) constant-time rank and select operation occupying a space close to the informational-theoretical lower bound.

To obtain efficient random access, we sort the summaries in increasing document identifier order and build two sets. The first set contains the identifiers of pages that are present in the graph but missing from the archive. By ranking in this set we can, given a document identifier d, find its ordinal position in the bit stream (by subtracting the number of missing documents that precede it). Since missing document are a small fraction, the dictionary for this set occupies little space.

The second set contains the bit pointers to each summary. In our case, for instance, the Elias-Fano dictionary provides direct access using less than 15 bits per summary. Subarchives can be easily layered over real archives using just additional rank/select structures.

Of course, the summaries contain now data based on a renumbered version of the terms, so we decorate our archive with an auxiliary file containing the permutation inverting the frequency-rank order. As a result, once a summary is read we can map the term indices to their original values. In our implementation this process is transparent to the user.

An open-source Java implementation of the techniques described will be soon made publicly available at the Laboratory of Web Algorithmics web site (`http://law.dsi.unimi.it/`) as part of the LAW library. Archives can be easily built from MG4J document collections (MG4J is a search engine developed at the DSI).

[1] The same idea has been used in [1] to store compactly the result of disjunctive queries.

3. A SAMPLE COLLECTION

As a sample of the techniques we described, we present a snapshot of about 100 million pages obtained from an original generic crawl of the `.uk` domain performed by the Laboratory of Web Algorithmics in May 2007 using UbiCrawler [2] (crawls made by UbiCrawler have already been used to build collections for evaluation [5]).

The crawler started from a seed of about 190000 different URLs. For each host, the crawler downloaded up to 50 Kpages with maximum depth 8. Only pages containing text were stored. The crawl was stopped at about 100 Mpages, resulting in about 500 GB of data stored in WARC/0.9 gzipped format.

There are of course a number of choices that have to be made to go from a snapshot to a stream of summaries. We parse HTML pages and apply the Porter stemmer; then, we remove English stopwords and all terms appearing in less than 20 documents. These choices are of course fairly arbitrary, and dictated by the need of eliminating *hapax legomena* and typos, but they are just parameters of our tools, and they can be set differently. We remark, moreover, that standardizing the preprocessing phase has also the effect of making the classification process further reproducible, by unifying the text preprocessing phase that most algorithms perform.

The bitstream archive storing the resulting summaries occupies about 26 GiB for data. Satellite data that must be loaded into main memory include 150 MiB for pointer data, 100 MiB for the term permutation and frequencies and about 1 GiB for the associated graph. This level of compression makes it possible to use the collection to simulate a crawl of 100 million pages on a standard PC in few hours.

4. REFERENCES

[1] H. Bast and I. Weber. Type less, find more: fast autocompletion search with a succinct index. *Proc. of the 29th annual international ACM SIGIR conference*, pp. 364–371. ACM Press, 2006.

[2] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience* 34(8):711–726, 2004.

[3] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. *Proc. of the 13th Intl. World Wide Web Conference*, pp. 595–601, 2004.

[4] P. Boldi and S. Vigna. Codes for the world wide web. *Internet mathematics* 2(4):405–427, 2005.

[5] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *ACM Sigir Forum* 40(2):11–24, 2006.

[6] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks* 31(11–16):1623–1640, 1999.

[7] P. Elias. Efficient storage and retrieval by content and address of static files. *J. Assoc. Comput. Mach.* 21(2):246–260, 1974.

[8] R. M. Fano. On the number of bits required to implement an associative memory. Memorandum 61, Computer Structures Group, Project MAC, MIT, Cambridge, Mass., n.d., 1971.

[9] S. Vigna. Broadword implementation of rank/select queries. *Proc. WEA 2008*. Springer–Verlag, Lecture Notes in Computer Science, 2008.

To separate or not to separate: reflections about current GIR practice

Nuno Cardoso
Faculty of Sciences
University of Lisbon
LASIGE group
ncardoso@xldb.di.fc.ul.pt

Diana Santos
Linguateca
SINTEF ICT
Oslo, Norway
diana.santos@sintef.no

ABSTRACT

Most geographical information retrieval (GIR) systems separate the treatment of the geographical and the non-geographical part, often called "thematic". In this paper, we provide an overview of this practice, and we advance arguments for and against. We also show some experimental results that apparently substantiate the non-separation argument. We conclude with the recommendation that this practice should receive more attention by the GIR community.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Evaluation

Keywords

Geographical IR, Geographical Query, Geographical Indexing, Evaluation

1. INTRODUCTION

The interest in geographical information systems and focused geographical search as a subarea of information retrieval is no longer new, with a regular workshop since 2004, GIR [18], and an annual evaluation contest in a cross-lingual setting, GeoCLEF, since 2005 [7, 8, 14]. However, we believe that there has not yet emerged a best practice approach, and we want to discuss a possible reason for this, namely the separation of the geographical terms from the rest of the terms.

Ever since its beginning as a new discipline, geographical information retrieval (GIR) has been thought as adding geographical dimension and processing to an already existing state-of-the-art IR. Cai's paper on geo-libraries [3], although primarily concerned with merging map and text approaches, has been influential in distinguishing among two subspaces in GIR: the thematic and the geographical. The thematic space concerns the subjects or themes that are relevant to the user, while the geographical subspace deals with the scope of the documents found. The thematic space is the usual domain of information retrieval, so, in order to advance the field, geographical information retrieval should concentrate on the geographical part, properly separated from the classic thematic part.

While this may appear a sensible inference, it soon faces the difficulties of dealing with text and textual queries, and the several properties of location in text (surveyed, for example, in Santos and Chaves[26]). In fact, GIR systems to date – possibly due to GeoCLEF – have been mainly trying to solve the problem of finding place names and information in text, which is a natural language processing task. And, to come right to the point, it is hard to separate geographical from non-geographical information in text. (For example, words do not come with a flag meaning "I convey geographic meaning, and only that meaning"...)

This paper addresses this issue in more detail: we start with a survey on the dividing strategies in GIR, to clarify the different approaches taken and eventually compare them, in Section 2. Then, we discuss possible reasons or arguments why these strategies may not work, from a natural language perspective, in section 3. Section 4 adduces some empirical data in favour of the non-dividing camp, while Section 5 concludes with the suggestion that the matter be further looked into by the GIR community.

2. SEPARATING THE LOCATION PART

The most straight-forward way to develop a GIR system is to adapt an off-the-shelf, standard IR engine, and augment it with geographical information and processing modules such as named entity recognizers and gazetteers, and then evaluate how this improves the overall results of the system, for geographical queries. This is the typical GIR approach used by participants along the three editions of GeoCLEF. Yet, no significant improvements over a pure IR approach were shown, which should perhaps ring a bell for the community.

2.1 Query parsing

A very common approach is to consider that a geographical query is a concatenation of two parts: i) the thematic part, and ii) the location part. The thematic part is handled by the classical text retrieval, while the geographical part is funneled to the newly developed geographical approaches [4, 16]. This approach assumes that most geographical queries are represented on a simple "what in where" format, that can therefore be easily divided into the two parts.

The query parsing pilot task in GeoCLEF 2007 [12] illustrates this assumption: it required that participants analysed 800,000 search engine queries, splitting the geographical queries into <*what, spatial relationship, where*> triplets.

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

Also, the first GeoCLEF pilot in 2005 provided an additional topic description in a similar form [8]. This was criticized in [22, 23] for lack of an adequate semantics for the relations, as well as for cross-lingual inappropriateness of the relations themselves, and was not used in later editions, although this might be reflect a lack of consensus among different organizers and not a shared position of GeoCLEF.

2.2 Document geo-indexing

Another frequently employed technique in GIR is the detection of location names in documents, and the creation of a separated geographical index, to store the extracted information.

For example, Leveling et al. [11] use an index of location indicators, that gather into a single index entry all location names and other derivative mentions such as adjective forms, acronyms or postal codes.

The SPIRIT project associates geometric footprints for each location in a separate index, and then used the calculation of polygon overlapping for inferring geographic similarity [9]. Kornai's approach is similar, assigning bounding boxes for each location present in the query, and using MetaCarta's local search engine [10].

With two distinct indexes serving the geographical retrieval module (a term index and a geographical index), the complexity of the GIR approach increases: with two indexes, and hence two independent ranking measures, what is the best way to combine these two relevance measures?

Although Overell et al. avoid this two-index merging problem, by converting the captured locations into unique identifiers that are also indexed along with the text, as terms [16], they are aware that they may simply be adding redundancy.

2.3 Geographic resources

Most GIR researchers rely in some way on geographical ontologies or gazetteers, that provide minimally, geographic names, classification, and coordinates. These can be accurately described as modelling separately the location relations such as inclusion, overlap, proximity and bordering.

This is, from our point of view, a natural and important addition. One has to have geographical knowledge encoded in a way that allows reasoning, and using such repositories will not be argued against, in the scope of this paper. But it is interesting to point out that, in fact, there have been also researchers who used WordNet, and Wikipedia, for getting geographical information from general resources [2, 16]. So this means that, for the sake of completeness, one could also discuss whether general ontologies (or specific ones) deal better with understanding the meaning of places in natural language (and for GIR).

One of the most common uses of such resources is for reasoning about the level of detail of a query (for example, in topic #54, "northern Europe", in an ontology, is likely to have countries such as Norway and Sweden with a "part-of" relationship). Another is to perform disambiguation, since most place names are not unique to a geographic place.

3. NOT SEPARATING THE LOCATION PART

There are nevertheless a set of arguments for not separating the location part, that we will now detail in the next subsections.

3.1 Geographical themes: a contradiction in terms?

Geographical terms are sometimes the theme of a query. To want to know something about Honolulu is as honorable and acceptable as to want to know something about judo. The difference is that the first information need has a strong geographic connotation, while the second has not. It is hard to defend that they should be treated separately a priori. (Nonetheless, it is also true that one might want to know where Honolulu is located, whereas "where is judo" does not make sense. We are not saying that geographic locations do not have different or specific properties, but this subject is not within this paper's discussion.)

3.2 Often the geographic part is contextual

Most geographically-implicit queries should not (and possibly don't) describe where the user is or comes from. This is a contextual datum which is or should be recovered by the query context and not by the query text.

In fact, this is done by major search engines that personalize or localize based on similar users, and one of the similarities may be the geographical origin.

This is the opposite of the case discussed in the previous section; here, the location is possibly extremely relevant but not necessarily expressed (if one is not already addict of search engine tricks).

3.3 Is separation at all possible?

Geographical queries (in the sense of having need for some geographical reasoning or awareness) come in several flavors. According to the typology initially suggested in [25] and then in [8], there are at least eight different kinds of queries that involve geography in some way. Just by considering those kinds of queries it becomes apparent that a separation between the geographical and non-geographical part becomes problematic.

Geographical queries like topic #40, "Cities near volcanos" or topic #56, "Lakes with monsters", just to mention two topic titles of last year's GeoCLEF, are hard to divide that way: the first because there apparently would be no non-geographical part left, the second because it is not exactly the same as the query "monsters in lakes" and therefore this query reformulation (allowing subsequent partitioning of the thematic part "monsters" and the geographical part "restricted to lakes") would miss the point. See [17, 20] for the importance of small words.

In fact, all concrete things occur in space, and the same is true for events. So, most words in natural language refer to more than one feature of an object or concept: its location and many other properties. Often, one needs to understand the text (and the user need) to understand which facet of a particular object or location is at stake. Although this is apparently similar to the ambiguity between *Washington* as a person or as district capital, it is more complex, because we are here pointing to the very **same** concept/object which can be seen from many angles [19, 21]. So, *Brussels* can denote the city, but most often than not by metonymy it describes the EU administration; the *Vienna circle* can denote a group of philosophers or a place in Vienna; while *Lisbon youth* can denote the young people living in Lisbon or the youth of a person spent in Lisbon. In all these cases, Brussels or Vienna or Lisbon are the **same** place with all their connotations, and the co-text selects what is being put in focus/referred to.

Another way of showing the problem with the a priori separation is applying the topic/focus distinction in linguistics, and see that sometimes geo and non-geo information swaps roles: For the type of topics only with scope, such as topic #73, "Events at St. Paul's Cathedral", the focus in on the geographical part: one is interested

in whatever is happening at some place, or at whatever objects or buildings exist at a certain location. For the type of topics that are restricted to a scope, such as "Dogs in Pittsburgh" [29], the focus is on the theme: it is the inverse of the previous case. One is interested in some topic, provided it occurs (or exists) in a certain part of the world. While this may be a useful distinction to understand that it is not trivial to assign geo and no-geo roles to topics, in practice the above topic/focus distinction does not take us far. Even if it is possible, in artificial venues, to produce clear-cut topics of the two above kinds, in most real cases it is not even clear what the user focus is: if one asks for "economy in the Bosphorus region" (topic #66), is one primarily interested in economy, or in the Bosphorus area? Does it really make sense to decide?

3.4 The search argument

Keeping the example of the Bosphorus area open, a typical informed person would also search for names of companies that they knew were operating on that area, or names of economical treaties, or related products. Eventually, names of factories (or factory locations) or ship names that had been in the news. (This is a remark that is relevant as far as log analysis is concerned. Expert searchers might be looking for "economy at the Bosphorus" with other keywords which would fail to be recognized as geographically related search in the first place. See Aires and Aluisio for a pertinent discussion of user intentions versus user activities [1]).

This tells against the current practice of defining geographical queries by those mentioning a geographical term of some sort. A more informed analysis of query logs might yield that a particular set of queries had a strong geographical glue even though no places had been mentioned.

3.5 Is separation useful?

Going back to the assumption that it is possible, in most cases, to separate geographical from non-geographical terms, separate processing misses the following relevant observation: Thematic keywords are often indirectly related to geographic knowledge. For instance, shipwrecks are often found near islands, or coast of oceans, and not on top of mountains or in the Sahara desert. To dismiss all this geographic knowledge (and its implicit co-occurrences for relevance) does not seem to be wise.

3.6 Is separation technically feasible?

Another argument, of a quite different nature, can also be advanced: there is not enough maturity in NLP to be able to really separate and identify all and only geographic terms and interpretations in text. There are still a lot of mistakes (failure to identify locations) and spurious hits (names or words that are considered locations when they shouldn't).

In view of this, a careful study of the importance of such deficiencies into the processing chain might be advisable. For example, Martins et al developed CaGE, a text mining module to capture locations from Web pages, based on a geographic ontology and basic context rules, in order to compute geographic signatures for Web pages to be used in GIR[15]. However, CaGE did not manage to capture most of the geographic evidence in the text collection used in HAREM, a NER evaluation contest for Portuguese [24, 27].

In HAREM we addressed seriously the issue of finding named entities which represented locations in context (and not simply names of places out of context). We therefore produced an evaluation resource which is unique and allows one to assess the difference between a gazetteer-based (or lexically based) and the real use of names for describing locations.

3.7 Summing up

In a nutshell, the problem of identifying something as purely geographical is not an easy task, if possible at all, as will also appear conspicuously when discussing geo-topics in the next section.

All the arguments just listed seem to show that the separation of geographical information from "the rest" may not have been well enough thought of in the first place. We proceed to show that actual practice in GIR systems and their evaluation also backs us in our warnings.

4. EXPERIMENTAL RESULTS

We start by reminding readers that, after three years of GeoCLEF, there is not a single GIR approach that clearly outperforms pure IR systems for the same GeoCLEF tasks [13]. This is indeed negative evidence of some strength for the need for a separate GIR strategy.

In this section, we will present a particular system developed by the first author and colleagues, and the results of the analysis of its performance in GeoCLEF. Although we are perfectly clear that there might be other design flaws in this system, the fact that explicitly investigating the issue of separation showed that it did not work for the particular architectures seems to be yet another valid counter-argument for it.

4.1 A case-study of a GIR system

XLDB's GIR system, co-developed by the first author, participated in all three editions of GeoCLEF, as part of a research project to give geographic capabilities for a Portuguese web search engine [28]. The architecture of the GIR system is shown in Figure 1, and described in detail in [4].

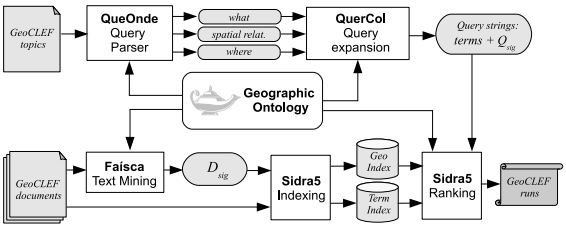


Figure 1: The architecture of the XLDB's GIR system.

In the 2007 edition, the GIR system embraced a purely segregational approach: the QueOnde query parser module divided the GeoCLEF topic titles into <*what, spatial relationship, where*> triplets; the QuerCol query expansion module had different strategies – blind relevance feedback for the thematic part, and an ontology-driven expansion for the geographic part – in order to generate a final query string; finally, the Sidra5 indexing module generates separated term and geographic indexes.

4.2 General analysis of its results

From a preliminary analysis of XLDB's GIR system, we came across the following practical results or doubts:

- The term query expansion (QE) approach adopted is based on blind relevance feedback set, using the top-5 documents and adding the top-8 expanded terms that were weighted higher by the $w_t(p_t\text{-}q_t)$ algorithm [6]. For the 2007 GeoCLEF topics, the QE step re-introduced geographic terms that were later injected in the thematic part.

	Portuguese topic title	English topic title
51	Extracção de petróleo e gás entre o Reino Unido e o continente europeu	Oil and gas extraction found between the UK and the European Continent
52	Crime perto de Santo André	Crime near St Andrews
53	Investigação científica em universidades da costa leste da Escócia	Scientific research at east coast Scottish Universities
54	Prejuízos causados por chuvas ácidas no Norte da Europa	Damage from acid rain in northern Europe
55	Mortes causadas por avalanches na Europa excluindo os Alpes	Deaths caused by avalanches occurring in Europe, but not in the Alps
56	Lagos com monstros	Lakes with monsters
57	Uísque de ilhas escocesas	Whisky making in the Scottish Islands
58	Problemas em aeroportos londrinos	Travel problems at major airports near to London
59	Cidades em que houve reuniões da comunidade dos países andinos	Meetings of the Andean Community of Nations (CAN)
60	Baixas em Nagorno-Karabakh	Casualties in fights in Nagorno-Karabakh
61	Acidentes de avião perto de cidades russas	Airplane crashes close to Russian cities
62	Reuniões da OSCE na Europa de Leste	OSCE meetings in Eastern Europe
63	Qualidade da água na costa mediterrânica	Water quality along coastlines of the Mediterranean Sea
64	Acontecimentos desportivos na Suíça francesa	Sport events in the french speaking part of Switzerland
65	Eleições livres em áfrica	Free elections in Africa
66	Economia no Bósforo	Economy at the Bosphorus
67	Pistas em que Ayrton Senna correu em 1994	F1 circuits where Ayrton Senna competed in 1994
68	Rios com cheias	Rivers with floods
69	Morte nos Himalaias	Death on the Himalaya
70	Turismo no Norte da Itália	Tourist attractions in Northern Italy
71	Problemas sociais na Grande Lisboa	Social problems in greater Lisbon
72	Costas com tubarões	Beaches with sharks
73	Ocorrências na catedral de São Paulo	Events at St. Paul's Cathedral
74	Tráfego marítimo nas ilhas portuguesas	Ship traffic around the Portuguese islands
75	Violações dos direitos humanos na antiga Birmânia	Violation of human rights in Burma

Table 1: Portuguese and English topic titles of GeoCLEF 2007.

- several geographical clues came in the form of landmarks (whose location is known), but which were missed because they were not in the geographic ontology.
- most geographical terms in our geographic signatures did not concern the geographic scope of the document: they could be case of metonymies or simply different facets of that term.

More specifically, a detailed analysis topic by topic, showed the following major sources of problems:

- local conveying property or association: Russian planes are not necessarily in Russia, Scottish research is not necessarily presented only in Scotland, France Press is not only read in France... in other words, the location association is hardly ever a restriction on geographical scope.
- as already referred, many query expansion terms are geographic, but not necessarily relevant for that either... it might be that the most significant expansion for football were Rio de Janeiro, but the topic one was interested in was "Italian football". Then, adding geographical terms outside Italy would probably only diminish performance.
- mention of theme and location in a document may not mean they were related in it: in fact, there was talk about acid rain in one context, and a location in Sweden in another context, and the document was returned as relevant. This is of course a general problem in IR – and thus not specific of GIR – but it tells against providing **one** geographical scope to a document based on the locations discussed in it.

4.3 Query expansion

As mentioned above, by analysing the behavior of the XLDB's GIR system on the GeoCLEF evaluation task revealed that the QE step re-introduced geographical terms in the thematic part, even considering that the initial query was stripped from all geographic names.

We have done an in-depth analysis of the results of this step for the 25 GeoCLEF topics of 2007. Table 1 list them both in English and in Portuguese, for convenience of the reader, but the results and the analysis was done for the Portuguese subtask.

Table 2 presents the top-8 terms re-introduced by the QE module, during the blind relevance feedback step. In bold stand the terms that are considered geographical by the GIR system: it is significant that, out of 192 terms, 71 (37%) are of clear geographic nature.

5. CONCLUDING REMARKS

We believe to have amassed enough data to raise doubts about whether an *a priori* separation between geographic and non-geographic information is appropriate for GIR, a separation we already theoretically attacked in 2006 [26].

Although we are aware that there are several different applications and contexts of use for GIR, and that we are speaking mainly from a GeoCLEF perspective, that is, one of querying geo-topics in newspaper text (and not Web pages or GIS papers), we believe that this reflection can be useful to the whole community, and we make a plea for people to test the particular separation flavour(s) they use in their systems with an open mind.

In particular, we believe that many further empirical studies – especially from the other architectures based on this separation – are required, as well as empirical studies of more general nature, both on

- linguistic issues: how geographical information is encoded in natural language(s) and which other clues may be relevant. For this, the recent trend of relation identification in information extraction may be an important one, see [5, 30].
- user studies: how do location matters really matter for users (of different IR systems). Most probably, different issues will be required for different kinds of task and different kinds of text. Maybe the new pilots at GeoCLEF this year will shed some light on this latter issue (one on Wikipedia and one on image search).

51	[**mar**, empresa, **unido**, **norte**, **reino**, gas, natural, mil]
52	[**santo**, luiz, oswaldo, silva, criminal, delegado, cruz, clodovil]
53	[**edimburgo**, efeito, gases, aquecimento, temperatura, **irlanda**, lugar, cientistas]
54	[cento, **dinamarca**, novo, reduzir, 2005, **oslo**, gases, **florestas**]
55	[**alemanha**, neve, **rios**, chuva, **holanda**, mau, assolar, continuam]
56	[**loch**, **ness**, **lago**, famoso, **ilha**, **mar**, volumoso, passada]
57	[bebida, **ilha_islay**, turfa, scotch, bourbon, single_malt, maltes, casa]
58	[**aeroporto**, **londres**, sido, **heathrow**, voo, passageiros, nomeadamente, contra]
59	[**tomarense**, igat, pedro, marques, autarquia, **tomar**, assistirem, praticava]
60	[**nagorno_karabakh**, **crimeia**, contra, itar_tass, presidente, kremlin, guerra, boris_ieltsin]
61	[**siberiana**, tupolev, 154, **irkutsk**, passageiros, **russo**, companhia, tripulantes]
62	[**hungria**, pacto, estabilidade, **europeia**, **checa**, nato, **leste**, apresentar]
63	[**mar**, objectivo, marinhas, **efluentes**, nascem, ecologistas, reivindicam, cento]
64	[**saas**, **valais**, final, esquiadores, esquiar, slalom, mil, **lausanne**]
65	[**senegal**, **marfim**, **costa**, **ruanda**, saraui, milhares, **ruandesa**, **ruandeses**]
66	[**capital**, sob, acordo, petroleiro_cargueiro, medidas, **turca**, **turcos**, **estreito**]
67	[**silverstone**, gp, pilotos, **pistas**, piloto, **pista**, lehto, grande]
68	[chuvas, problemas, **rio**, urbanos, abastecimento, parque, lercas_adjudicadas, suficiente]
69	[**himalaias**, evereste, alpinistas, gokyo, encontrados, monte, corpos, tinha]
70	[**veneza**, turistas, san, **veneziano**, piazza, comparados, guias, turista]
71	[**oeiras**, **xira**, **loures**, **amadora**, **cascais**, **sintra**, **franca**, **vila**]
72	[steven, brancos, entrar, atacando, comem, alimentam, spielberg, recife]
73	[]
74	[**ilhas**, **ilha**, **miguel**, **faial**, **graciosa**, **jorge**, milhas, **horta**]
75	[suu, nobel, **myanma**, aung, kyi, paz, san, anistia]

Table 2: Top 8 expanded terms for the GeoCLEF 2007 Portuguese subtask.

In fact, for each particular system following the separation architecture, one can always blame the lack of coverage of the ontology or the low recall level of the NER system employed, but this may only be masking a design flaw, which we bring to the consideration of the reader: that of trying to separating what cannot be separated.

Acknowledgements

This work was jointly funded by the European Union (FEDER and FSE) and the Portuguese government, under contracts ISFL/13/408 (FIRMS-FCT), 339/1.3/C/NAC (Linguateca)and PTDC/EIA/73614/2006 (GREASE II). The first author acknowledges FCT grant SFRH/BD/29817/2006.

6. REFERENCES

[1] R. V. X. Aires and S. M. Aluísio. Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em Português. *Revista Ciência da Informação*, 32(1):5–16, 2003. in Portuguese.

[2] D. Buscaldi, P. Rosso, and E. Sanchis. A WordNet-Based Indexing Technique for Geographical Information Retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *Lecture Notes on Computer Science*, pages 954–957. Springer-Verlag, 2007.

[3] G. Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In *Proceedings of the Second International Conference on Geographic Information Science, GIScience'02*, pages 65–79, London, UK, 2002. Springer-Verlag.

[4] N. Cardoso, D. Cruz, M. Chaves, and M. J. Silva. The University of Lisbon at GeoCLEF 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[5] J. Chu-Carroll and J. Prager. An Experimental Study of the Impact of Information Extraction Accuracy on Semantic Search Performance. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07*, Lisbon, Portugal, 6–8 November 2007.

[6] E. N. Efthimiadis. A user-centered evaluation of ranking algorithms for interactive query expansion. In *Proceedings of the 16th Conference on Research and Development in Information Retrieval, SIGIR'93*, pages 146–159, 1993.

[7] F. Gey, R. Larson, M. Sanderson, K. Bishoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, G. D. Nunzio, and N. Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *Lecture Notes on Computer Science*, pages 852–876. Springer-Verlag, 2007.

[8] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. Revised Selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 908–919. Springer, 2006.

[9] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the Third International Conference on Geographic Information Science, GIScience'2004*, pages 125–139, Adelphi, MD, USA, 20-23 October 2004.

[10] A. Kornai. Evaluating Geographic Information Retrieval. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 928–938. Springer-Verlag, 2006.

[11] J. Leveling and S. Hartrumpf. University of Hagen at GeoCLEF 2007: Exploring Location Indicators for Geographic Information Retrieval. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[12] Z. Li, C. Wang, X. Xie, and W.-Y. Ma. Query Parsing Task for GeoCLEF 2007 Report. In A. Nardi and C. Peters,

editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[13] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. *GeoCLEF 2007: the CLEF 2007 Cross Language Geographic Information Retrieval Track Overview.* Presentation held at CLEF 2007, Budapest, Hungary, 20 September, 2007.

[14] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[15] B. Martins, M. J. Silva, and M. S. Chaves. O Sistema CaGE no HAREM - Reconhecimento de Entidades Geográficas em Textos da Língua Portuguesa. In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, chapter 11*, pages 199–215. Linguateca, 2007. In Portuguese.

[16] S. Overell, J. Magalhães, and S. Rüger. GIR experiements with Forostar at GeoCLEF 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September 2007.

[17] K. Pastra, H. Saggion, and Y. Wilks. Extracting relational facts for indexing and retrieval of crime-scene photographs. *Knowledge-Based Systems*, 16(5-6):313–320, 2003.

[18] R. Purves and C. Jones. Workshop on Geographic Information Retrieval. *Computers, Environment and Urban Systems*, 30(4):375–377, 2006.

[19] J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, USA, 1995.

[20] E. Riloff. Little Words Can Make a Big Difference for Text Classification. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130–136, 1995.

[21] D. Santos. *What is natural language? Differences compared to artificial languages, and consequences for natural language processing*. Invited lecture at SBLP'2006 and PROPOR'2006, Itatiaia, RJ, Brazil. 15 May, 2006.

[22] D. Santos and N. Cardoso. Portuguese at CLEF 2005: Reflections and Challenges. In C. Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF Workshop, CLEF'2005*, Vienna, Austria, 21–23 September 2005.

[23] D. Santos and N. Cardoso. Portuguese at CLEF. In C. Peters, F. Gey, J. Gonzalo, H. Müeller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised selected papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 1007–1010. Springer-Verlag, 2006.

[24] D. Santos and N. Cardoso, editors. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, November 2007.

[25] D. Santos and M. S. Chaves. *The place of place in geographical IR*. Presentation held at the Geographic Information Retrieval workshop, held at SIGIR'2006. `http://www.linguateca.pt/Diana/download/acetSantosChavesGIR2006.pdf`.

[26] D. Santos and M. S. Chaves. The place of place in geographical IR. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval, GIR'2006 (held at SIGIR'2006)*, pages 5–8, Seattle, WA, USA, 10 August 2006.

[27] D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, and D. Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, pages 1986–1991, Genoa, Italy, 22-28 May 2006.

[28] M. J. Silva, B. Martins, M. S. Chaves, A. P. Afonso, and N. Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

[29] B. Yu and G. Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR'07 (held at CIKM'07)*, pages 49–54, Lisbon, Portugal, 2007. ACM.

[30] S. Zhao and R. Grishman. Extracting Relations with Integrated Information Using Kernel Methods. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05*, pages 419–426, Morristown, NJ, USA, 2005. ACL.

Dynamic Focused Retrieval of XML Documents and Its Evaluation

Toshiyuki Shimizu
Graduate School of Informatics
Kyoto University
shimizu@soc.i.kyoto-u.ac.jp

Masatoshi Yoshikawa
Graduate School of Informatics
Kyoto University
yoshikawa@i.kyoto-u.ac.jp

ABSTRACT

XML information retrieval (XML-IR) systems search for relevant elements in XML documents for given queries. Though XML-IR systems must handle nesting elements, the output of existing systems remains single ranked list. Ranked lists of all relevant elements may contain redundant contents by nestings, whereas single list of focused elements may lose possible benefit of XML-IR. We introduce the concepts of benefit and effort and propose to retrieve focused elements dynamically. The system flexibly retrieves non-overlapping elements which have larger benefit within the effort acceptable to users. To evaluate XML-IR systems with dynamic focused retrieval, we decided to use an upper bound of the benefit that is obtained by the system because we found no unique algorithm can be the optimal and practical solution for the problem. The performance of the system can be observed compared to the upper bound.

1. INTRODUCTION

To retrieve information about a topic from a large quantity of XML documents, a keyword search can be used to retrieve the document fragments that are relevant to the topic. XML information retrieval (XML-IR) systems generally use elements as basic search units. For example, in the case of academic articles marked up in XML, XML-IR systems use elements corresponding to sections, subsections, and paragraphs to construct search results.

One of the advantages of XML-IR systems is that users need to read only small portions of result documents, and XML-IR systems should retrieve higher relevant portions before they retrieve lower relevant portions. A critical focus of XML-IR is how to handle nesting elements. Some result elements may nest other results; for example, a paragraph element may be highly relevant to a query, while the section element that includes the paragraph is moderately relevant to the query.

INEX 2005 [5] defines three element retrieval strategies for evaluating the effectiveness of XML-IR systems. System using the Thorough strategy simply retrieve relevant elements from all elements and rank them in order of relevance. Elements retrieved using the Thorough strategy may overlap due to nestings. System that use the Focused strategy retrieve only focused elements by selecting the element with the highest score in a path and removing overlapping elements. Though the use of the Focused strategy avoids redundancy, it excludes non-focused elements from the results, which means that some of the possible benefits of XML-IR are lost [1] because it retrieves fixed ranked list of focused elements. Systems using the Fetch and browse strategy first identify relevant documents (fetching phase) and then identify relevant elements within a fetched document (browsing phase).

In INEX 2006 [6], Relevant in context task and Best in context task were also introduced. Relevant in context is a similar strategy to Fetch and browse except that it retrieves only focused elements according to their original document order for each article. Best in context searches for the Best Entry Point (BEP) for starting to read an article and retrieves only one best starting point per article. Retrieving elements grouped per article as in Fetch and browse or Relevant in context is another important point to consider, however it is not focus of this paper.

The retrieval strategies in INEX assume that XML-IR systems return a fixed ranked list of elements. Such retrieval strategies are important for system evaluation, however we considered that the style of single ranked list suffer from handling nesting elements. As we mentioned, the results of Thorough may contain redundant contents by nesting elements, or the Focused never retrieves ancestor elements of elements that have already been retrieved and lacks flexibility.

To handle nesting in XML-IR search results, Clarke [1] proposed controlling overlapping by re-ranking the descendant and ancestor elements of the reported element. However the retrieval style is yet fixed ranked list of elements.

To overcome this problem, we propose dynamic focused retrieval. We introduced the concepts of benefit, which is the amount of gain obtained by reading the element, and effort, which is the cost for browsing search results [10]. We supposed that the user interactively decide the amount of effort that can be spent to browse search results, and the system dynamically retrieve relevant elements within the specified effort. The system retrieves elements that provide larger benefit and to obtain more benefit from retrieved elements, nesting elements are removed from the search results since it is considered that there is no increase in benefit from reading

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

the repeated content. By using this dynamic retrieval style, the systems can retrieve non-overlapping elements flexibly.

The following situation can be considered as one of the actual usage scenarios of our XML-IR systems. A user first specifies a threshold of effort and the system retrieve focused elements for the specified effort. Then, in response to the retrieval result from the system, user interactively adjust the threshold of effort. If s/he thinks the returned set of subdocuments is too much (or too small), s/he will increase (or decrease) the threshold. For any threshold value of effort, the system always retrieves focused non-overlapping elements depending on the value.

We considered that effort consists of reading effort and switching effort. The reading effort is the effort for reading the contents of the result elements, and the switching effort is the effort for switch result items.

By using reading effort, we can directly handle element size. The elements retrieved by XML-IR systems vary greatly in size. A root element that corresponds to the whole document will be large, while other elements may be much smaller. Therefore, the cost of reading the content of a retrieved element is not known beforehand. In general, users of XML-IR systems browse search results from top ranked element to lower ranked elements. Therefore, fast retrieval of high ranked elements is important and there are some researches on top-k search of XML-IR [12, 3]. However, total output size of top-k elements is uncontrollable by simply giving an integer k. We considered using reading effort instead of an integer k is better alternative to control the total output size.

The systems can avoid long list of short elements by considering switching effort. In this paper, the discussions mainly based on only reading effort excluding switching effort, however we can extend the discussions by considering that we need extra effort, that is switching effort, for browsing each element in result lists.

We formalized the problem of maximizing benefit for a given effort when the system is given benefit and reading effort for each element, and defined the concept of search result continuity, which we consider a critical property for a practical system. Because systems based on our scheme retrieve elements flexibly according to the amount of effort, the contents of the elements retrieved as the optimal solution for the small amount of effort may not be included in the contents of the elements retrieved as the optimal solution for the larger effort. We believe a practical system should avoid this type of situation.

We found that the problem of finding an optimal solution for the formalized problem was an variant of the knapsack problem and was NP-hard. We also found that this optimal solution violated search result continuity, and therefore proposed greedy algorithms for the formalized problem. Furthermore, we improved the algorithm that we proposed in [10], and propose a recursive greedy retrieval algorithm.

We need to devise evaluation measure that is not based on the number of relevant results such as traditional Precision and Recall to evaluate XML-IR systems with dynamic focused retrieval, because such systems are based on benefit and effort and disregard the number of retrieved elements.

We assumed that the reading effort could be easily calculated using the length of the corresponding element, and the switching effort is a constant value in this paper. Thus, the performance of a system based on our scheme depends on the benefit calculated by the system. To evaluate a flexible retrieval system in which the result elements change depending on the specified effort, we assumed we could use the actual benefit of each element. Using actual benefit means that the optimal solution for the formalized problem gives a theoretical upper bound of benefit that the system can provide. We considered that the proportion of the total amount of actual benefit that the system provided compared to the upper bound changing the amount of specified effort could be used as a base for evaluation measures. However, as the problem of finding the optimal solution is NP-hard, we decided to use the upper bound of the benefit provided by the optimal solution as the target for comparison. We confirmed the effectiveness of the upper bound and the quality of the upper bound was sufficient for most queries of INEX 2005.

2. BENEFIT AND EFFORT

Our proposals on retrieval and evaluation for XML-IR are based on the concepts of benefit and effort. We basically assumed the following three points. 1) XML-IR systems retrieve elements, that is, partial text fragments in elements or aggregations of multiple elements can not be answers, and 2) users read full contents of retrieved elements, that is, when a element is retrieved and browsed by a user, the user read all descendant contents together. 3) Users do not read only a partial content of retrieved elements, and users read the content of the retrieved element to the end once s/he start reading. We believe these assumptions are also adopted in element retrieval of INEX [2].

2.1 Benefit

We considered users obtain benefit from relevant content to the query the user input. For a given query, the benefit of an element is the amount of gain obtained by reading the element. In this paper, we follow the assumption of traditional IR and XML-IR methods for the sake of simplicity. The target document set describe unique content, that is they do not repeat the same or similar content more than once. This assumption is adopted by traditional IR and XML-IR methods, which may retrieve redundant but different documents that describe the same or similar content. We can also observe the Precision / Recall measure does not consider such document similarity.

Under this assumption, basically, the benefit of an element can be considered to be the sum of the benefit of the child elements. However, when all the child elements are read together, the contents of the child elements may complement each other; hence the benefit of the parent element may be larger than sum of the benefit of the child elements.

In this paper, we assumed that the benefit of an element is greater than or equal to the sum of the benefit of its child elements.

2.2 Effort

We introduced two types of effort to model the actual user behavior. One is the reading effort which users spend in browsing the content of the search result, and the other is the switching effort which users spend in browsing the result items.

2.2.1 Reading Effort

We considered users spend reading effort when they browse the content of search results. The reading effort of an ele-

ment is the cost spent in reading the content of the element. Note that reading effort does not depend on the query and can be calculated based on the element itself. Basically, the reading effort of an element can be considered as the sum of the reading effort of the child elements. However, when all the elements are read together, users may be continuously reading within the same context. Hence, the reading effort of the parent element may be smaller than the sum of the reading effort of the child elements.

In this paper, we assumed that the reading effort of an element is less than or equal to the sum of the reading effort of its child elements.

2.2.2 *Switching Effort*

We considered users spend switching effort when they finished reading one result item and continue to read the next result item.

Intuitively, users need more effort for a result items that consist of many short paragraph elements than a result items that consist of only one section element which is the parent element of the paragraph elements and describes the same content in fact.

If we consider strictly, the switching effort is the value that depend on the relationship between the switched two result items. However, we handle the switching effort as a fixed constant value in this paper for the sake of simplicity.

2.3 b/e Graph

The XML-IR systems that we propose calculate the benefit of elements in the XML document set for a given query, and retrieve result elements. We assumed that users would specify the threshold amount c for effort, and that the system would retrieve elements that had larger benefit within the specified effort.

We assumed that reading the same content repeatedly would not increase the benefit. If a search result contains nesting elements, the system keeps only the eldest ancestor element of the nesting elements, and removes all the other elements of the nesting elements, because the amount of benefit provided remains the same. Therefore, the search results do not contain nesting elements.

Figure 1 shows an example of benefit and reading effort calculated by a system for a query. In Figure 1, the tree structure of an XML document is represented, and benefit and reading effort are shown in the form of benefit/reading effort adjacent to the element. For the sake of simplicity, we did not assume any concrete formula for calculating benefit and reading effort in Figure 1.

We assumed that we need to read the whole content of the element to obtain benefit from the element. That is, we can not obtain partial benefit for partial reading effort. When a system calculates benefit and reading effort for a query as in Figure 1 and does not take switching effort into consideration, if a user specifies a threshold of effort to 15, the element set that maximizes benefit is $\{e_3, e_2\}$, whereas if 20 is specified, the element set that maximizes benefit is $\{e_3, e_7\}$.

We considered system behaviors can be expressed in the form of a graph, which we call a benefit/effort graph (b/e graph, for short). The behavior of a retrieval algorithm is expressed by plotting the total amount of benefit obtained when the threshold c for effort is changed. Figure 3 shows system behaviors of three algorithms in Figure 2. Algorithms A1 and A2 only consider reading effort as effort, whereas A1' consider both reading effort and switching effort. The value of switching effort in this example is 5. For example, algorithm A1 retrieves $\{\phi\}$ when $0 \leq c < 10$, $\{e_3\}$ when $10 \leq c < 20$, and so on.

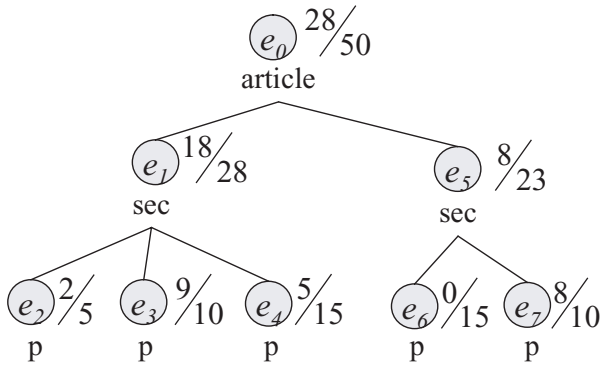


Figure 1: An example of calculated benefit and reading effort.

A1		A2		A1'	
$\{\phi\}$	$[0, 10)$	$\{\phi\}$	$[0, 5)$	$\{\phi\}$	$[0, 10)$
$\{e_3\}$	$[10, 20)$	$\{e_2\}$	$[5, 15)$	$\{e_3\}$	$[10, 25)$
$\{e_3, e_7\}$	$[20, 38)$	$\{e_2, e_3\}$	$[15, 25)$	$\{e_3, e_7\}$	$[25, 43)$
$\{e_1, e_7\}$	$[38, 50)$	$\{e_2, e_3, e_7\}$	$[25, 38)$	$\{e_1, e_7\}$	$[43, 50)$
$\{e_0\}$	$[50, \infty)$	$\{e_1, e_7\}$	$[38, 50)$	$\{e_0\}$	$[50, \infty)$
		$\{e_0\}$	$[50, \infty)$		

Figure 2: Examples of system behaviors.

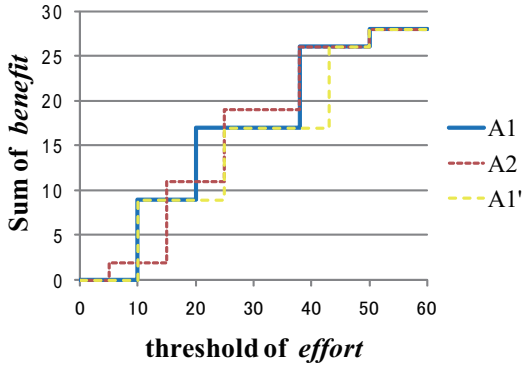


Figure 3: A1, A2 and A1' on b/e graph.

3. RETRIEVAL METHOD

We describe retrieval algorithms based on benefit and effort. First, we handle only reading effort as effort, and then introduce switching effort in Section 3.4. We formalize the problem in Section 3.1, and describe retrieval algorithms in Section 3.2 and 3.3. Finally, we extend the discussion and describe retrieval algorithms considering switching effort in Section 3.4.

3.1 Formalization of the Problem

The problem of maximizing benefit for a given effort is a variant of a knapsack problem [7] that has restriction of nestings. This problem (P) is formalized as follows.

$$
\text{(P)} \quad \left|
\begin{array}{ll}
\text{maximize} & z(x) = \sum_{i=1}^{n} b_i x_i \qquad (1) \\
\text{subject to} & \sum_{i=1}^{n} r_i x_i \leq c \qquad (2) \\
& x_i \in \{0,1\} \qquad (3) \\
& x_{j_1} + x_{j_2} + ... + x_{j_m} \leq 1 \qquad (4) \\
& \quad \text{for any } e_{j_1}, e_{j_2}, ..., , e_{j_m} \\
& \quad \text{which are elements} \\
& \quad \text{on a path from root to leaf}
\end{array}
\right.
$$

where b_i is the benefit of the element e_i, r_i is the reading effort of e_i, and c is the threshold value of effort input by the user. We can state that e_i is (not) contained in the search result by setting $x_i = 1$ ($x_i = 0$, respectively). The condition (4) shows that the search results do not contain nesting elements.

This problem (P) is considered to be an extension of the normal knapsack problem as it can be reduced to a normal knapsack problem if we handle XML documents that each contains only one element. We can say that the problem (P) is NP-hard since the normal knapsack problem is already NP-hard.

The system that maximizes benefit in the situation shown in Figure 1 retrieves $\{e_3, e_2\}$ when $c = 15$, and $\{e_3, e_7\}$ when $c = 20$. This means that the system does not output the content of e_2 when $c = 20$, though the content of e_2 is output when $c = 15$. Therefore, when a user specifies a threshold of effort to 20, s/he can not obtain information from e_2 though s/he pays more effort than the case when s/he specifies the threshold of effort to 15 and obtains information from e_2. To avoid such situations, we consider it important that systems have the property of search result continuity. Search result continuity is defined as follows.

Definition 1. Search result continuity

When we describe the result element set for threshold c of effort as E^c=$\{e_1^c, e_2^c, ..., e_n^c\}$, and the result element set for the threshold c' as $E^{c'}$=$\{e_1^{c'}, e_2^{c'}, ..., e_m^{c'}\}$, the algorithm has the property of search result continuity if the following holds for any c and c'. The function ancestor-or-self (e) returns an element set that consists of ancestor elements of e and e itself.

If $c \leq c'$, then $\forall\ e \in E^c$, $\exists\ e' \in E^{c'}$ s.t. $e' \in$ ancestor-or-self (e) □

In other words, the content of an element set for effort c must be contained in the content of the element set for effort c' if we increase the threshold value for effort from c to c'. Note that the element e' is unique to all e from Equation 3 and 4 in the context of problem (P). We consider that practical systems must provide search result continuity, as do algorithms A1, A2 and A1' in Section 2.3.

When we consider retrieval algorithms that provide search result continuity, there is no single algorithm that can provide benefit greater than or equal to all the other algorithms for any threshold value of effort in general. In the case of the situation in Figure 1, an algorithm such as A2, which can provide maximum benefit by retrieving $\{e_3$,$e_2\}$ when $c = 15$, can not retrieve $\{e_3, e_7\}$ and provide maximum benefit when $c = 20$.

3.2 Simple Greedy Retrieval Algorithm

As the problem (P) is NP-hard and the optimal solution violates search result continuity, we considered to solve the problem by greedy retrieving elements that we can obtain benefit efficiently.

We considered the Algorithm 1 to provide a greedy solution for (P). The inputs for the algorithm are $list_{in}$ and c. $list_{in}$ is the element list that holds $b_i/r_i \geq b_{i+1}/r_{i+1}$ for all i, and if b_i/r_i is equal to b_{i+1}/r_{i+1}, the list holds $r_i \leq r_{i+1}$. c is the threshold for effort. The outputs of the algorithm are the sum of obtained benefit z and $\boldsymbol{x}$= $\{x_1, x_2, ..., x_n\}$ for z.

When a element is retrieved, by adjusting the benefit and reading effort values of the ancestor elements, and retrieving the element that has the greatest b_i/r_i in that time, the system can retrieve elements efficiently. We call this algorithm a simple greedy retrieval algorithm. The output of this algorithm holds the conditions of (P) and search result continuity.

As an example, we describe a case in which a user specifies a threshold of reading effort c to 40, and benefit and reading effort are calculated as in Figure 1. In this case, $list_{in}$ is $\{e_3(b_3/r_3 = 9/10 = 0.9)$, $e_7(0.8)$, $e_1(0.64)$, $e_0(0.56)$, $e_2(0.4)$, $e_5(0.35)$, $e_4(0.33)\}$. First, e_3 is processed and $\boldsymbol{x}$=$\{e_0, e_1, e_2$, $..., e_7\}$ becomes $\{0,0,0,1,0,0,0,0\}$ and z becomes 9. At the same time, the benefit and reading effort of the ancestor elements e_1 and e_0 are adjusted. For e_1, b_1 is decreased to 9 and r_1 is decreased to 18. For e_0, b_0 is decreased to 19 and r_0 is decreased to 40. The benefit (reading effort) for the ancestor elements after these adjustments is obtained (is required), when the algorithm had a chance to subsequently process the ancestor elements later. The system reflects these adjustments and re-ranks the elements in $list_{in}$. In this case, $list_{in}$ becomes $\{e_7(0.8)$, $e_1(0.5)$, $e_0(0.48)$, $e_2(0.4)$, $e_5(0.35)$, $e_4(0.33)\}$. In addition, c becomes $40 - 10 = 30$. Then, e_7 is processed and $\boldsymbol{x}$ becomes $\{0,0,0,1,0,0,0,1\}$, z becomes $9+8 = 17$, $list_{in}$ becomes $\{e_1(9/18 = 0.5)$, $e_2(0.4)$, $e_0(0.37)$, $e_4(0.33)$, $e_5(0)\}$, and c becomes $30 - 10 = 20$. Next, e_1 is processed and the system sets $x_3 = 0$ because e_3 is the descendant of e_1. In addition, e_2 and e_4 are removed from $list_{in}$. $\boldsymbol{x}$ becomes $\{0,1,0,0,0,0,0,1\}$, z becomes $17 + 9 = 26$, $list_{in}$ becomes $\{e_0(2/12 \simeq 0.17)$, $e_5(0)\}$, and c becomes $20 - 18 = 2$. Next, e_0 is processed, however we can not retrieve e_0 within the specified effort, and the processing terminates . The outputs are $z = 26$ and $\boldsymbol{x}$=$\{0,1,0,0,0,0,0,1\}$.

The line labeled "simple" in Figure 4 shows the system behavior of a system using the simple greedy retrieval algorithm when the system calculates benefit for a query as in Figure 1.

3.3 Recursive Greedy Retrieval Algorithm

When the sum of the reading effort of elements retrieved using the simple greedy retrieval algorithm is less than the threshold value c, it is considered that the amount of benefit is increased by retrieving elements that have not yet been obtained using the remainder of effort.

However, it is important that systems have the property

Algorithm 1 Simple greedy retrieval algorithm

Input: $list_{in}$, c
Output: z, $\boldsymbol{x}$

$z = 0$
while $((e_i = top(list_{in}))! = null)$ do
  remove e_i from $list_{in}$
  if $(r_i > c)$ then
    break
  end if
  $adjust(e_i)$
  $x_i = 1$, $z+ = b_i$, $c- = r_i$
end while
return z, $\boldsymbol{x}$

function $\boldsymbol{adjust(e_i)}$ {
  for $(\boldsymbol{e_d \in e_i.descendants})$ do
    $\boldsymbol{x_d = 0}$
    $\boldsymbol{remove}$ $\boldsymbol{e_d}$ from $\boldsymbol{list_{in}}$
  end for
  for $(\boldsymbol{e_a \in e_i.ancestors})$ do
    $\boldsymbol{b_a - = b_i}$, $\boldsymbol{r_a - = r_i}$
    $\boldsymbol{rerank}$ $\boldsymbol{e_a}$ in $\boldsymbol{list_{in}}$
  end for
}

Algorithm 2 Recursive greedy retrieval algorithm

Input: $list_{in}$, c
Output: z, $\boldsymbol{x}$

$z = 0$
while $((e_i = top(list_{in}))! = null)$ do
  if $(!retrieve(e_i))$ then
    break
  end if
end while
return z, $\boldsymbol{x}$

function $\boldsymbol{retrieve(e_i)}$ {
  remove $\boldsymbol{e_i}$ from $\boldsymbol{list_{in}}$
  if $\boldsymbol{(r_i > c)}$ then
    $\boldsymbol{retrieveDescendants(e_i)}$
    return false
  end if
  $\boldsymbol{adjust(e_i)}$
  $\boldsymbol{x_i = 1}$, $\boldsymbol{z+ = b_i}$, $\boldsymbol{c- = r_i}$
  return true
}

function $\boldsymbol{retrieveDescendants(e_i)}$ {
  $\boldsymbol{e_m}$ = top ranked element that is descendant of $\boldsymbol{e_i}$ and $\boldsymbol{x_m = 0}$
  if $\boldsymbol{(e_m == null)}$ then
    return
  end if
  if $\boldsymbol{(retrieve(e_m))}$ then
    $\boldsymbol{retrieveDescendants(e_i)}$
  end if
}

of search result continuity. Therefore, it is not appropriate to obtain any elements for the remainder of effort. To satisfy search result continuity, the system needs to retrieve descendant elements of the element that is to be retrieved next. An algorithm that improves on the simple greedy retrieval algorithm by retrieving more elements for the remainder of effort is shown in Algorithm 2. We call this algorithm a recursive greedy retrieval algorithm.

In the case of the running example, when a user sets $c = 30$, the outputs of the simple greedy retrieval algorithm are $z = 17$ and $\boldsymbol{x}$=$\{0, 0, 0, 1, 0, 0, 0, 1\}$. However, a system using the recursive greedy retrieval algorithm can retrieve e_2 which is a descendant of e_1 after e_7; the outputs of the recursive greedy retrieval algorithm are $z = 19$ and $\boldsymbol{x}$=$\{0, 0, 1, 1, 0, 0, 0, 1\}$.

The line labeled "recursive" in Figure 4 shows the system behavior of a system using the recursive greedy retrieval algorithm when the system calculates benefit for a query as in Figure 1.

Theorem 1. Superiority of recursive greedy retrieval algorithm

The sum of benefit from elements retrieved by the recursive greedy retrieval algorithm is greater than or equal to that provided by the simple greedy retrieval algorithm for any threshold value of effort.

Proof omitted as trivial. We should therefore use the recursive greedy retrieval algorithm in the implementation of systems.

3.4 Consideration for Switching Effort

If we consider switching effort, we need to change the condition 2 in (P) to the following.

$$\sum_{i=1}^{n}(r_i + s)x_i - s \leq c \qquad (5)$$

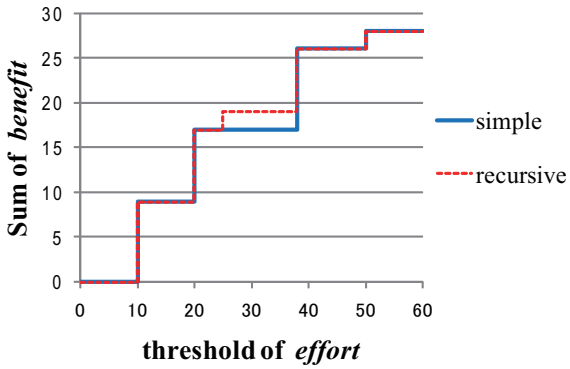


Figure 4: System behaviors of each algorithm on b/e graph.

where s is the switching effort.

We can extend retrieval algorithms in Section 3.2 and Section 3.3 by considering that we must spend additional switching effort to retrieve each element and c is also incremented by s.

We show the recursive greedy retrieval algorithm considering switching effort in Algorithm 3. The effort of each element and c is incremented by s. If we drop the $retrieveDescendants(e_i)$ function, the algorithm can be consider as simple greedy retrieval algorithm.

Algorithm 3 Retrieval algorithm considering switching effort

Input: $list_{in}$, c, s
Output: z, $\boldsymbol{x}$
1: $z = 0$
2: $c+ = s$
3: for ($e_i \in list_{in}$) do
4: $r_i+ = s$
5: end for
6: $e_{top} = top(list_{in})$
7: if ($!retrieve(e_{top})$) then
8: return z, $\boldsymbol{x}$
9: end if
10: $rerank\ list_{in}$
11: while $((e_i = top(list_{in}))! = null)$ do
12: if ($!retrieve(e_i)$) then
13: break
14: end if
15: end while
16: return z, $\boldsymbol{x}$

4. EVALUATION METRICS

We propose evaluation metrics based on the system behavior shown in the b/e graph. In order to evaluate systems based on benefit and effort, we supposed that the actual benefit for each element is available. Though systems calculate benefit for each element and retrieve results using recursive greedy retrieval algorithm, we must use the actual benefit by retrieved results for system evaluation not the benefit value expected by the system. Note that as Figure 3 or Figure 4 is drawn using the benefit values calculated by a system, the actual benefit obtained by the systems may differ from the values in these figures.

We decided to use an upper bound of the actual benefit that is obtained by systems. The performance of the system can be observed compared to the upper bound. We do not consider switching effort distinctly in this section, as we can consider that we need extra switching effort for browsing each element when the we want to take the switching effort into consideration. In this section, first we discuss how to calculate actual benefit and effort in Section 4.1, then describe why and how to calculate and use the upper bound in Section 4.2 and Section 4.3. Finally we compare our metrics with INEX metrics in Section 4.4.

4.1 How to Calculate Actual Benefit and Effort

INEX are manually developing relevance assessments for XML-IR. The relevance assessments of INEX 2005 consists of two parts, Exhaustivity (ex) and Specificity (sp) [1]. Exhaustivity is the extent to which the element discusses the topic of request, and it has three levels; Highly exhaustive (HE), Partially exhaustive (PE), and Not exhaustive (NE) [2]. We can convert HE, PE, and NE to numeric as 1, 0.5, 0, respectively. Specificity is the extent to which the element focuses on the topic of request, and it is calculated by dividing $rsize$, which is the length of the content relevant to the topic, by $size$, which is the whole length of the element.

We describe ex, sp, $rsize$, and $size$ of element e_i as $ex(e_i)$, $sp(e_i)$, $rsize(e_i)$, $size(e_i)$. We considered calculating actual benefit and reading effort from assessments of INEX 2005. For example, we can use following equations.

$$b_i = ex(e_i) * rsize(e_i) \quad (6)$$

$$r_i = size(e_i) \quad (7)$$

We can use only $rsize$ for calculating benefit when the assessments of INEX 2006 or later which include only Specificity are used.

$$b_i = rsize(e_i) \quad (8)$$

We assumed the switching effort is the fixed value in this paper. We need to use reasonable value to integrate with reading effort. Though the switching effort should be carefully investigated through user studies, we supposed that the reading effort corresponding to small but meaningful element size such as average size of paragraph elements is likely to be used as a starting value of switching effort.

4.2 Upper Bound of Benefit for Given Effort

When we consider the problem (P) in Section 3.1 and use actual benefit, we can calculate the maximum of actual benefit value for a certain c value by solving (P). If an algorithm could provide maximum benefit for all of the c values, we could use the benefit values of the algorithm as a basis for system evaluation. However, for this problem, there is no such algorithm because we must also consider search result continuity. We therefore decided to calculate the upper bound of the amount of benefit and use it as the basis for evaluating systems. The optimal solution of (P), which does not consider the search result continuity, can be an upper bound. However, this problem (P) is NP-hard and difficult to solve. Therefore, we decided to calculate an upper bound of (P). The upper bound of (P) is greater than or equal to that of the problem considering search result continuity.

Theorem 2. Upper bound of (P)
The optimal value of the continuous problem (P') of (P) that relaxes the condition (3) to $0 \leq x_i \leq 1$ provides the upper bound of (P).

Proof. The value that can be obtained in (P) can also be obtained in (P') because the problem (P') relaxes the condition in (P). The range of values in (P) is included in the range of values in (P'). Therefore, the optimal value of (P') provides an upper bound of (P). □

We considered the Algorithm 4 based on the Algorithm 1 to provide the optimal solution for (P'). If the system can not retrieve whole e_i (Line 4), it achieves optimal value by retrieving the partial amount that can be retrieved. As an example of calculating optimal value of (P'), we consider the case when a user specifies the threshold of effort c to 40, and benefit and reading effort are calculated like Figure 1, which is the same situation when we consider simple greedy retrieval algorithm. The processing basically follows the same steps except that the last step concerning e_0 is skipped. As the final step, system retrieves partial amount of e_0 using the remaining effort 2. x_0 is set to $2/12 \simeq 0.17$ and x_1 and x_7 are set to $1 - 2/12 \simeq 0.83$. $\boldsymbol{x}$ becomes $\{0.17, 0.83, 0, 0, 0, 0, 0, 0.83\}$, z' becomes $26 + 2 * 0.17 \simeq 26.3$, and then breaks.

When we pick up a non-overlapping element set $\{e_{k_1}, e_{k_2}, ..., e_{k_m}\}$, each of which is the descendant of the element e_r,

[1] The assessments of INEX 2006 or later only use Specificity.
[2] Too Small (TS) is introduced for small elements, however we regard TS is equal to NE.

Algorithm 4 Optimal solution for (P')

Input: $list_{in}$, c
Output: z', $\boldsymbol{x}$

```
1: z' = 0
2: while ((e_i = top(list_in))! = null) do
3:    remove e_i from list_in
4:    if (r_i > c) then
5:       x_i = c/r_i
6:       for (e_d ∈ e_i.descendants) do
7:          if (x_d == 1) then
8:             x_d = 1 − x_i
9:          end if
10:      end for
11:      z' += b_i * x_i
12:      break
13:   end if
14:   adjust(e_i)
15:   x_i = 1, z' += b_i, c− = r_i
16: end while
17: return z', x
```

the conditions in (P') hold if we set $x_r = \alpha$ and $x_{k_i} = 1 - \alpha$ $(1 \leq i \leq m)$. In this situation, the sum of benefit is calculated as follows.

$$\begin{aligned} & b_r * \alpha + \sum_{i=1}^{m} (b_{k_i} * (1-\alpha)) \\ = \; & b_r * \alpha + \sum_{i=1}^{m} b_{k_i} - \sum_{i=1}^{m} (b_{k_i} * \alpha) \\ = \; & (b_r - \sum_{i=1}^{m} b_{k_i}) * \alpha + \sum_{i=1}^{m} b_{k_i} \end{aligned} \quad (9)$$

Similarly, the sum of reading effort is calculated by replacing b_i by r_i in Equation 9. That is, setting $x_r = \alpha$ and $x_{k_i} = 1 - \alpha (1 \leq i \leq m)$ is the same as setting $x_v = \alpha$ and $x_{k_i} = 1$ $(1 \leq i \leq m)$ if we assume a virtual element e_v with benefit of $b_r - \sum_{i=1}^{m} b_{k_i}$ and reading effort of $r_r - \sum_{i=1}^{m} r_{k_i}$.

If the system can retrieve whole e_i, that is, it can set $x_i = 1$, setting $x_d = \alpha > 0$ for descendant element e_d of e_i is contrary to the optimal solution because the benefit obtained becomes $b_i*(1-\alpha)+b_d*\alpha$ and $b_i*(1-\alpha)+b_d*\alpha \leq b_i$ holds as b_i is greater than or equal to b_d. When the system retrieves descendant elements of e_i before retrieving e_i, the situation is considered to be $x_v = 1$ and $x_{k_i} = 1$ when we assume the virtual element e_v, and therefore $x_r = 1$ and $x_{k_i} = 0$, in fact.

4.3 Comparison with Upper Bound

The upper bound is represented in the b/e graph as a linear interpolation line of the plots when elements are retrieved by a simple greedy retrieval algorithm. A system using the simple greedy retrieval algorithm can obtain maximum benefit at such points.

For system evaluation, we need the actual benefit for each element. Note that this actual benefit can not be used in system implementation. Implementers of XML-IR systems can develop better systems by guessing the benefit of each element as closely as possible to the actual benefit. We assumed that we could use a common reading effort value for various systems because this value does not depend on the query.

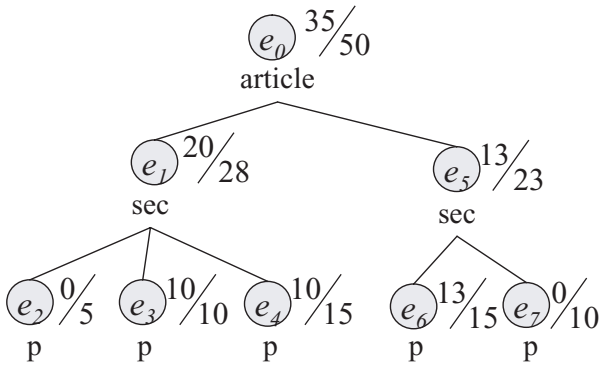


Figure 5: Actual benefit and reading effort.

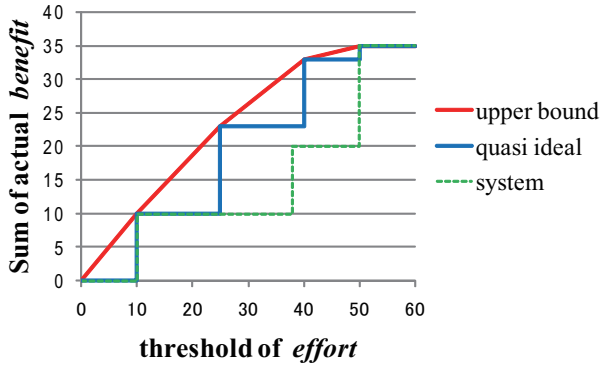


Figure 6: b/e graph for evaluation.

To evaluate systems based on benefit and effort, we can compare the behavior of an implemented system with the behavior of the upper bound on a b/e graph. If we implement a system that can calculate benefit values for each element that are the same value as the actual benefit, a system using the recursive greedy retrieval algorithm can provide benefit that is very close to the upper bound. We call such a system a quasi ideal system.

As an example, we explain the case in which the system calculates benefit and reading effort as in Figure 1, however the actual benefit and reading effort are those shown in Figure 5. In this case, using the recursive greedy ranking algorithm, with a threshold value of effort 40, the system can obtain 20 benefit by $\boldsymbol{x}$=$\{0, 0, 1, 1, 0, 0, 0, 1\}$, however the upper bound of benefit that can be obtained is 33 by $\boldsymbol{x}$=$\{0, 0, 0, 1, 1, 0, 1, 0\}$. Figure 6 shows a b/e graph for the running example. In Figure 6, 'upper bound' is for the upper bound and 'system' is for the system to be evaluated. If the system calculates benefit identical to the actual benefit, the behavior of the system is the line labeled 'quasi ideal'. The b/e graph enables us to intuitively understand the performance of the system in relation to the upper bound. By using the upper bound for comparison target, we can evaluate systems absolutely.

To evaluate XML-IR systems, the metrics based on XCG is used in INEX 2005 [4]. By considering the effectiveness

of the implemented systems in relation to the upper bound, we can apply this concept to our case. For example, we can introduce iMArep (interpolated Mean Average reading effort precision), which is calculated based on a b/e graph using the similar concept of iMAep (interpolated Mean Average effort precision) [4].

We calculated iArep values for quasi ideal system using INEX 2005 test collection. We used calculation formula for actual benefit and reading effort in Section 4.1 and disregarded switching effort. We obtained b/e graphs up to 50,000 effort, which means 50,000 characters that is corresponding to about the length of one article of INEX. Most of the iArep values for quasi ideal system were greater than 0.9, indicating that the quality of the upper bound was good. The iMArep value was 0.90 for all 29 Topics and 0.93 when we excluded Topics 209, 217, and 239 whose iArep value was relatively low because the good results tended to be large elements for such topics.

4.4 Metrics of INEX

Though the target of our metrics and that of INEX is differ, our evaluation metrics is motivated by that of INEX. Our metrics is for dynamic focused retrieval, whereas metrics of INEX is for single ranked list.

Though the evaluation metrics for XML-IR systems is still a task with several open issues, the metrics based on eXtended Cumulated Gain (XCG) is used in INEX 2005 [4]. System-oriented ep/gr (effort-precision/gain-recall) and user-oriented nxCG (normalized extended Cumulated Gain) measures are used by considering the relative effectiveness for the ideal system.

In the context of XCG based measures, effort is measured in terms of the number of visited ranks, whereas our reading effort is measured in terms of element length. In addition, such ideal system in XCG does not exist for the dynamic focused retrieval, and we decided to use the upper bound.

In INEX 2007, evaluation metrics called HiXEval based on relevant text length is used [8, 9]. The HiXEval metrics is based on the assumption that a system which retrieves elements that contain as much relevant text as possible, and as little irrelevant text as possible is preferred. In the HiXEval metrics, the Precision is measured by the length of retrieved relevant text compared to the total length of retrieved text, and the Recall is measured by the length of retrieved relevant text compared to the total length of relevant text.

If we regard element size as reading effort and relevant text size as benefit, the concept of HiXEval is similar to our metrics. In fact, element size can be seen as the factor of reading effort and relevant text size as that of benefit as in Section 4.1. However HiXEval does not consider ideal system as in XCG. This means that though HiXEval metrics can evaluate systems relatively, it can not evaluate systems absolutely. If the value obtained using HiXEval metrics is low, we can not say the system is ineffective, because the value may be low even for an ideal system. To evaluate systems absolutely, we need an ideal system for comparison. In our scheme, we used the upper bound of the optimal solution as the ideal value for comparison

5. CONCLUSIONS

We introduced the concepts of benefit and effort for XML-IR systems, and proposed a retrieval algorithm and evaluation metrics based on them. We examined situations in which users of XML-IR systems specify a threshold for effort and the system flexibly retrieves focused elements dynamically within the specified effort. We formalized the problem and calculated the upper bound of benefit for system evaluation.

In general, existing XML-IR systems calculate relevance score between each element and the input query using some scoring formula, and the length (in characters or in number of terms) of the element is included as a factor of the scoring formula [12], and scores are decreased for longer elements to avoid overestimation of longer elements. By handling element length separately as reading effort, we believe we can apply our scheme to other systems.

In future work, we will examine ways of distinguishing the portion of the retrieved element to be read. Furthermore, for XML documents created by marking up original PDF files, there is potential to show search result elements mapped on an image of a physical page [11]. We will look at ways of integrating our system with this type of user interface. A major drawback of our current scheme is that users must specify the threshold of effort. We believe that developing user interfaces that can smoothly retrieve result elements when users change the threshold value of effort is a promising solution. When we consider about the user interfaces, we can also imagine the opposite case in which users change the threshold value of benefit and the system retrieves elements which minimize effort.

6. REFERENCES

[1] C. L. A. Clarke. Controlling overlap in content-oriented XML retrieval. In SIGIR, pages 314–321, 2005.

[2] INEX. INitiative for the Evaluation of XML Retrieval. http://inex.is.informatik.uni-duisburg.de/.

[3] R. Kaushik, R. Krishnamurthy, J. F. Naughton, and R. Ramakrishnan. On the integration of structure indexes and inverted lists. In SIGMOD, pages 779–790, 2004.

[4] G. Kazai and M. Lalmas. INEX 2005 evaluation measures. In INEX, pages 16–29, 2005.

[5] S. Malik, G. Kazai, M. Lalmas, and N. Fuhr. Overview of INEX 2005. In INEX, pages 1–15, 2005.

[6] S. Malik, A. Trotman, M. Lalmas, and N. Fuhr. Overview of INEX 2006. In INEX, pages 1–11, 2006.

[7] S. Martello and P. Toth. Knapsack problems: algorithms and computer implementations. John Wiley & Sons Inc, New York, 1990.

[8] J. Pehcevski, J. Kamps, G. Kazai, M. Lalmas, P. Ogilvie, B. Piwowarski, and S. Robertson. INEX 2007 evaluation measures. In INEX 2007 Pre-Proceedings, 2007.

[9] J. Pehcevski and J. A. Thom. HiXEval: Highlighting XML retrieval evaluation. In INEX, pages 43–57, 2005.

[10] T. Shimizu and M. Yoshikawa. A ranking scheme for XML information retrieval based on benefit and reading effort. In ICADL, pages 230–240, 2007.

[11] T. Shimizu and M. Yoshikawa. XML information retrieval considering physical page layout of logical elements. In WebDB, 2007.

[12] M. Theobald, R. Schenkel, and G. Weikum. An efficient and versatile query engine for TopX search. In VLDB, pages 625–636, 2005.

Large-Scale Interactive Evaluation of Multilingual Information Access Systems – the iCLEF Flickr Challenge

[1]Paul Clough, [2]Julio Gonzalo, [3]Jussi Karlgren, [1]Emma Barker, [2]Javier Artiles, [2]Victor Peinado
[1]University of Sheffield, UK
[2]Universidad Nacional de Educación a Distancia, Spain
[3]Swedish Institute of Computer Science, Sweden

ABSTRACT

Participation in evaluation campaigns for interactive information retrieval systems has received variable success over the years. In this paper we discuss the large-scale interactive evaluation of multilingual information access systems, as part of the Cross-Language Evaluation Forum evaluation campaign. In particular, we describe the evaluation planned for 2008 which is based on interaction with content from Flickr, the popular online photo-sharing service. The proposed evaluation seeks to reduce entry costs, stimulate user evaluation and encourage greater participation in the interactive track of CLEF.

1. EVALUATION OF IR SYSTEMS

Evaluating the performance of Information Retrieval (IR) systems is an important part of the system development process from an engineering point of view, and a crucial part of the research process. It enables development of useful and effective technology, together with generalisable and sustainable knowledge for future development cycles. A systematic, transparent, and intuitively valid evaluation process has been a defining and unifying feature of the information access research field during the past decades ([14][15][16][6][3]), and has been instrumental in ensuring simultaneous commercial success and academic stringency. We should stay true to this tradition.

1.1. Traditional Evaluation Methodologies

The evaluation of retrieval systems tends to focus on either the system or the user. Saracevic [14] distinguishes six levels of evaluation for information systems that include information retrieval systems: (1) at the engineering level, (2) at the input level, (3) at the processing level, (4) at the output level, (5) at the use and user level and (6) at the social level. For many years information access evaluation has tended to focus on the first three levels, predominately through the use of standardized benchmarks (or reference collections) in a laboratory-style setting (also known as batch-mode evaluation).

The Cranfield experiments [5] were some of the first to develop and demonstrate the use of lab-based evaluation. However, information access systems are most commonly used interactively, within a task and social context, and this drives the need for user-centered evaluation to address performance at the latter three levels (output, use and user, and social). User-centered evaluation is important because it assesses the overall success of a retrieval system (as determined by end users of the systems) which takes into account other factors other than system performance, e.g. task context, cognitive influence, and the design of the user interface (see, e.g. [8]).

To enable reproducibility and comparison, standardized resources for evaluating document retrieval systems have been designed and used (a.k.a. test collections) for at least 30 years (first proposed in the Cranfield I and II projects [4]). Standardized resources have been used in major information access evaluation campaigns around the world such as TREC[1], CLEF[2] and NTCIR[3]. Researchers have recognized the value of testing retrieval systems within the large-scale setting through organized and managed campaigns, undoubtedly acting as a major influence in the design of information access systems over the past ten years or so. Not only have these events provided a testbed for evaluation, but also an interactive forum in which to exchange ideas and discuss techniques for successful system and algorithm design.

Although primarily a testbed for system-orientated evaluation, these campaigns (in particular TREC and CLEF) have also included user-oriented (or interactive) evaluation. However, evaluating interactive information access systems experimentally is challenging [2][7]. The high effort, cost, and overhead involved in recruiting test subjects, designing test systems, and formulating experimental scenarios risks both delivering unrealistic laboratory-based task formulations, and finding general results drowned in inter-user variation. The low reproducibility of experiments, failure to effectively generalize results, and the difficulty of comparison between different systems has limited the success of such initiatives (see, e.g. [7][12]).

[1] http://trec.nist.gov/ [accessed 11/03/08]
[2] http://www.clef-campaign.org/ [accessed 11/03/08]
[3] http://research.nii.ac.jp/ntcir/ [accessed 11/03/08]

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

1.2. The Challenge for Interactive Evaluation

However successful evaluation schemes have been in the past, new media pose challenges to content analysis and to established target notions of "relevance"; new modes of communication and contexts pose challenges to use cases and tasks underlying traditional ad-hoc evaluation schemes; multilingual materials, audience, and usage situations pose challenges to systems and processing resources. In addition, new interactive services are taken up by user communities, not by virtue of their engineering qualities or their ergonomics but by consumer evaluation based on social factors, marketing effectiveness, or even legal requirements: offering a well-built interface and providing solid content is no guarantee to commercial success. Evaluating interactive retrieval must make itself relevant to service providers by evaluating those aspects of interaction that are most crucial for the task a system is designed for: if the system has no underlying task model it must acquire one to be valuable. Traditional ad-hoc evaluation schemes have had an implicit use case and task model which does not necessarily carry over to new situations.

The next generation of evaluation methodologies must take into account not only changes in the underlying content, but the varying user base and societal and contextual factors surrounding the usage under study. How might we find a task that allows us to evaluate interactive retrieval, using multi-medial and multilingual data, possibly not in a standard collection, affording the potential to model new settings, new contexts, new tasks with large enough numbers of users to transcend inter-user noise, with a minimal amount of administrative overhead, and yet provide generalisable, intellectually appealing, and potentially interesting and useful results?

2. EVALUATING MULTILINGUAL IR

Multilingual information retrieval (MLIR) describes the situation in which a user searches for information in a language different from the query (see, e.g. [9]). Multilingual information retreival can be thought of as a combination of machine translation and traditional monolingual information retrieval. Most research has focused on locating and exploiting translation resources with which the user's search requests or target documents (or both) are translated into the same language. Campaigns such as the Cross Language Evaluation Forum (CLEF) [13)] and the Text REtrieval Conference (TREC) [2][17] multilingual track have helped encourage and promote international research, as well as create standardised resources for evaluating multi-lingual information access approaches.

2.1. Interactive CLEF (iCLEF)

The CLEF interactive track (iCLEF[4]) has been devoted, since 2001, to the study of Cross-Language Information Retrieval from a user-centered perspective. The aim has always been to investigate real-life cross-language searching problems in a realistic scenario, and to obtain indications on how best to aid users in solving them (see, e.g. [11]). Multilingual information retrieval is particularly interesting from an interactive point of view, because the need for search assistance is substantially higher than in monolingual information retrieval: normally, the user can quickly adapt to the system's modus operandi, but not to an unknown target language.

iCLEF experiments have investigated the problems of foreign-language text retrieval, question answering and image retrieval, including aspects such as query formulation, translation and refinement, document selection and document examination. The focus has always been on improving the outcome of the process in terms of a classic notion of relevance (documents meeting an information need that prompted a query), and the target collection (except for image search experiments) has always consisted of news texts in languages foreign to the user. Finally, the task has always involved the comparison of a reference system with a contrastive system, combining users, topics and systems with a Latin-Square design to detect system effects and filter out other effects (as used within the Interactive TREC track [7]).

Table 1: iCLEF task goals and participation (2001-2006).

Year	Task	Goal	Groups
2001	Ad-hoc	Document selection	3
2002	Ad-hoc	Document selection, query formulation & reformulation	5
2003	Ad-hoc	Full Cross-Language search	5
2004	QA	Full Cross-Language QA	5
2005	Image search/QA	Full Cross-Language QA / known-item image search	5 (2 image; 3 QA)
2006	Image search	open	3

Table 1 shows the progression of iCLEF since 2001. Overall, participation has always been low, with a high of 5 participating groups; a low of 3 groups. Although iCLEF in only a few years of activity has established the largest collected body of knowledge on the topic of interactive cross-language information retrieval, the experimental setup has proven limited in certain respects:

- The search task itself is unrealistic: news collections are comparable across languages, and most of the pertinent information tends to be available in the user's native language. Therefore, why would a user search for this information in an unknown language?
- The target notion of "relevance" does not cover all aspects that make an interactive search session successful (e.g. other factors could include satisfaction of results, usability of the interface itself, and the system's response time).
- The Latin-Square design imposes heavy constraints on the experiments, making them costly and with a limited validity (the number of users is necessarily limited, and statistically significant differences are hard to obtain).

2.2. Moving to Flickr

In order to overcome these limitations, the iCLEF track moved to a new pilot framework in 2006 [4] [10]: we decided to use the publicly available (and immensely popular) photo-sharing

[4] http://nlp.uned.es/iCLEF/ [accessed 11/03/08]

service Flickr[5] as the target collection. This is an inherently multi-lingual database through its lively tagging and commenting features, and it has the potential to offer a range of challenging and realistic multilingual search tasks for interactive experimentation. Although the database is in constant evolution – something which compromises reproducibility – the Flickr search API allows specifying timeframes (e.g. search images uploaded in the period 2004-2007), which permits defining a more stable dataset for experiments.

2.3. The Experience of iCLEF2006

Besides moving to Flickr as the target database, in 2006 we took the following additional decisions:

1. To lower the threshold of entry to the evaluation campaign, we offered a standard multi-lingual interface which various research sites can use to explore whatever features of interaction they are most interested in. The interface provides a (baseline) term translation service and a fine-grained log of user actions.
2. We designed three different search tasks: known-item search (find this image), topical search (find as many pictures as possible around this topic), and text illustration (find good images to illustrate this text). The illustration task naturally provides a search scenario where evaluation has to go beyond the traditional notion of topical relevance.
3. We did not impose any evaluation methodology on the participants. Being a novel evaluation scenario, we wanted to involve iCLEF participants in the exploration of novel evaluation methodologies as a key part of the campaign. This made the 2006 a collaborative exercise on how to study interactive issues in cross-lingual multi-medial information access.

Whilst we found enthusiastic support from the potential participants (fourteen groups signed up for the task), only three sites actually participated in the final evaluation (the three organizing groups themselves). We found that while the freedom of the task appeared to be attractive at first sight, the entry threshold was still too high: building an interface and designing an experiment proved too costly and the open design provided too little support for newcomers. In addition, we found that the submission schedule used in other CLEF tracks collapses with iCLEF due to the inherent time-consuming nature of implementing a user interface and running interactive experiments.

As in previous iCLEF editions, there was valuable knowledge acquired, but little participation from the research community. It can be concluded that, similar to Interactive TREC, the interactive CLEF task has not been as successful as the lab-based system-orientated tasks. Possible reasons for this include:

- Considering users is just not seen as important in information retrieval evaluation (compared to system-oriented evaluation).
- The large-scale setting of an evaluation campaign is simply ill-suited to interactive evaluation.
- Performing user experiments is time-consuming and difficult and little gain is seen for it (e.g. lack of generality and difficulty in comparing results).
- Developing efficient algorithms for information access is considered more important than user-orientated issues.
- System-orientated is well-understood; user-orientated evaluation is less clear and requires a deeper understanding (e.g. in the experimental design).

2.4. Remedies for iCLEF2008

One of the main limitations of iCLEF 2006 was that, although we moved into a realistic multilingual search setting, the experiment designed still did not facilitate having large-scale user logs. All three experiments employed less than 30 users that had to be recruited, trained, monitored and controlled. In 2008 we decided to concentrate on collecting user logs at a larger scale, and let participants concentrate on mining such logs to gain more knowledge about how users behave when they need to search in unfamiliar languages.

To be able to harvest a substantially larger set of search sessions, we decided to implement a single, basic multilingual search interface for Flickr, and make it available in the web for anyone. To attract – and specially to keep - potential users, we have made the search task a game. The basic task is simple: finding a given image (the user is shown a picture) in Flickr. Finding more images improves the user ranking in a "Hall of Fame". Note that this is a fully multilingual task: the image to be found can be annotated in any (or several) of the target languages, and the user does not know a priori which is the case.

This was modeled on the success of the ESP game for labeling images [1] and thought to increase interest in the task for both participating groups and their subjects. The entry costs of iCLEF2006 were clearly still too high, therefore for 2008 we provide groups with an experimental design, but still allow open extension for groups to adapt the design for their own investigations. As the evaluation has moved to Flickr/Web users, participants now have something in common with the subjects they recruit, therefore are more likely to be a captive set of subjects. Finally, to allow for the timing differences of running an interactive evaluation task, we have adjusted the deadlines of the standard iCLEF calendar, giving participants more time to run and analyse their experiments.

3. THE iCLEF2008 TRACK

We now describe the iCLEF track for 2008 in terms of what the organizers provide to participating groups, and what the groups must do.

3.1. Data and Resources

The organizers of iCLEF are providing the following to participants in 2008:

3.1.1. Task definition

The task for 2008 is known-item image retrieval based on photos from Flickr: the user is given an image, and the goal for them is to find the image again from Flickr. The advantage of this kind of search task is that it has clear goals for the user, it has a clearly defined measure of success (the image is either found or not) and whilst searching for the required image, users will invoke different (and potentially interesting) search

[5] http://www.flickr.com [accessed 11/03/08]

patterns. The user does not know in advance in which languages the image is annotated; therefore searching in multiple languages is essential to successfully find the images. The task is organised as a game: the more images found, the higher users (and user groups) will be ranked. Section 3.3 describes in more detail the selection of topics and example images.

Figure 1: The iCLEF2008 interface.

3.1.2. Default MLIR front-end to Flickr

We have designed and implemented a multilingual information retrieval interface to Flickr with the following functionalities (shown in Figure 1):

- Multilingual search: query in one language, get search results in up to six languages (English, Spanish, French, Italian, Dutch and German).
- Term-to-term translations between six languages (English, Spanish, German, French, Dutch and Italian) using freely available dictionaries (taken from http://xdxf.revdanica.com/down/).
- Selection of "best" target translations according to (i) presence in the Flickr *related terms* for the query, which often include target-language terms because they co-occur with the query terms in images annotated in multiple languages, something which is not unusual in the Flickr database; and (ii) string similarity between the source and target words. This was included because the free dictionaries used did not have information about the most frequent sense/translation.
- Enables user to pick/remove translations, and add their own translations (which go into a "personal dictionary"). We did not provide back-translations to support this process, in order to study correlations between target language abilities (active, passive, none) and selection of translations.
- Provision of search suggestions (Flickr related terms plus tags from displayed images).
- Control over the game-like features of the task: flow of images, users ranking, etc.

Note that we did not intend to provide the best possible cross-language assistance to search the Flickr collection. Our intention was to provide a rather standard, baseline interface where we can get information from users' behavior which is not too much dependent on a particular interface idiosyncrasy.

3.1.3. Experiment customization

In addition to harvesting search logs, we also offer this interface for groups interested in performing their own experiments with selected types of users, and we provide support for customization of the interface.

3.1.4. Generation of search logs

Search logs will be generated from the interface. We will focus on two user groups: (i) CLEF participants, which will be asked to play the Flickr game (the best team will receive an award at CLEF 2008), and (ii) Flickr/Web users at large. The game will be publicized in order to get a substantial amount of usage information. The idea of using CLEF researchers as a user group is not simply a matter of convenience: we believe that few cross-lingual information retrieval researchers have actually experienced cross-language search tasks as users, and the exercise we propose might broaden their vision of cross-lingual information retrieval research.

3.2. Participating in the Track

Participants in iCLEF2008 can essentially do two tasks: analyse log files based on all participating users (which is the default option) and perform their own interactive experiments with the interface provided by the organization. CLEF individuals will register in the interface as part of a team, so that a ranking of teams can be produced in addition to a ranking of individual users.

3.2.1 Generation of search logs

Participants can mine data from the search session logs, for example looking for differences in search behaviour according to language skills, or correlations between search success and search strategies.

3.2.2 Interactive experiments

Participants can recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc.

3.3. Topic Selection

In total, 180 example images will be available within the system for users to find (30 images in each language set: German, Spanish, English, French, Italian and Dutch). Classification of the language of an image is based on the "main" language of an image's text and tagset. Rather than select images randomly from Flickr, we wanted to maintain some element of experimental control and topic variation. The following points were considered during selection of the images:

- There should be sufficient text/tags accompanying an image to facilitate the task (i.e. we required "rich" text where possible).
- Ideally we wanted diverse topics in the test set and required roughly equivalent subject/topics in the different language groups, so the aim was to get at least one instance of a subject/topic group, for each of the language sets.

- When collecting images in different languages but with the same subject/topic, we aimed to find images with a similar visual perspective.
- The known item task must not be too hard: queries for finding images were manually recorded and an independent search carried out to check the images are not too hard to find.

Figure 2 shows example images from the current set of topics. As can be seen, these vary in aspects such as subject (topical content of an image), visual content, orientation, activity depicted in the image, and visual perspective (e.g. close-up, long distance).

Figure 2: Example topics for known-item search.

4. CONCLUSIONS

The iCLEF task has so far provided a substantial body of knowledge around the interactive aspects of Cross-Language Retrieval, but it has failed to engage the cross-lingual information retrieval research community, and it has always been restricted to experiments with a limited set of users, where statistically significant insights are hard to find. In the design of iCLEF 2008 we have made a significant change in our experiment design, focusing on acquiring a large set of search session logs and offering the data to iCLEF participants, so that the task focus is on mining search logs rather than designing interactive experiments. At the same time, we have decided to engage the CLEF research community as a user group for the experiment, hoping that this fully multilingual search exercise will broaden the scope of midstream cross-lingual information retrieval research into the essential – but hard to study systematically – interactive aspects of multilingual retrieval.

ACKNOWLEDGEMENTS

Work partially funded by the TrebleCLEF Coordination Action (FP7-ICT-2007-1).

5. REFERENCES

[1] von Ahn, L. Games with a Purpose, Computer, Vol. 39(6), pp. 92-94, June, 2006.

[2] Belkin, N.J., Dumais, S.T., Scholtz, J. & Wilkinson R. Evaluating interactive information retrieval systems: opportunities and challenges. CHI Extended Abstracts 2004: 1594-1595.

[3] Buckley, C. & E. M. Voorhees. Retrieval system Evaluation. In E. M. VOORHEES & HARMAN D. K. (Eds.), TREC: experiment and evaluation in information retrieval. London, England, MIT Press. 2005.

[4] Clough, P., Gonzalo, J. & Karlgren, J. Multilingual interactive experiments with Flickr. EACL 2006 Workshop on New Text - Wikis and blogs and other dynamic text sources. 2006.

[5] Harter, S. P. & Hert, C. A. Evaluation of information retrieval systems: Approaches, issues, and methods. Annual Review of Information Science and Technology (ARIST), 32, 3-94. 1997

[6] Hersh, W. Information Retrieval: A Health and Biomedical Perspective, 2nd edition, Springer-Verlag: New York, Berlin, Heidelberg, 2005.

[7] Hersh, W. & Over, P. Interactivity at the Text Retrieval Conference (TREC) Information Processing & Management, Volume 37, Issue 3, May 2001, Pages 365-367.

[8] Ingwersen, P. & Järvelin, K. The turn: integration of information seeking and retrieval in context, Springer. 2005.

[9] Jones, G.J.F., Beyond English Text: Multilingual and Multimedia Information Retrieval, in Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Honour of Karen Sparck Jones (ed. J. Tait), 2005, Kluwer. 2005.

[10] Karlgren, J., Gonzalo, J. & Clough, P. iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository. In CLEF 2006 Proceedings. 2007.

[11] Oard, D. & Gonzalo, J. The CLEF 2003 Interactive Track, Comparative Evaluation of Multilingual Information Access Systems. Results of the CLEF 2003 Evaluation Campaign. Springer-Verlag LNCS 3237, 2004.

[12] Over, P. The TREC interactive track: an annotated bibliography Information Processing & Management, Volume 37, Issue 3, May 2001, Pages 369-381.

[13] Peters, C. and Braschler, M.: Cross Language System Evaluation: The CLEF Campaigns. Journal of the American Soc. for Inf. Sci. and Tech. Vol. 52(12) (2001) 1067-1072.

[14] Saracevic, T. 1995. Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual international

ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

[15] Spark Jones, K. (Ed.). 1981. Information Retrieval Experiment. London: Butterworths.

[16] Spark Jones, K. & Willett, P. (Eds.). 1997. Readings in Information Retrieval, San Francisco, CA: Morgan. Kaufmann Publishers, Inc.

[17] Voorhees, E.M. & Harman, D.: Overview of TREC 2001, In NIST Special Publication 500-250: Proceedings of TREC2001, NIST, 2001.

How Many Experts?

A New Task for Enterprise Search Evaluation

Gianluca Demartini
L3S Research Center
Leibniz Universität Hannover
Appelstr. 9a D-30167 Hannover, Germany
demartini@l3s.de

ABSTRACT

Enterprise Search has attracted much attention from the Information Retrieval field because of the important economic outcome the search services can have in the context of enterprises. Consequently, an evaluation initiative such as TREC has provided, with the Enterprise Track, a standard approach to evaluate Enterprise Search System effectiveness.

In this paper we first discuss the evaluation approaches taken in the past. The main contribution is the proposal of a new search task to be performed and evaluated in the Enterprise Search context. We provide a motivation for the new task, a framework for its evaluation together with preliminary experiments, and some possible approaches that a system can adopt for performing the task.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Measurement, Experimentation

Keywords

Enterprise Search, Evaluation

1. INTRODUCTION

In the knowledge era, the amount of structured data that people in enterprises have to manage is increasing every day. People spend a big part of their working hours looking for information. For this reason information retrieval (IR) systems which are customized for finding items in the enterprise context have been developed and the need for comparing them fairly is growing as well.

Enterprise managers face the problem of selecting one Enterprise Search (ES) system which best satisfy their particular needs, and a comparison of available systems is usually done by third party companies. A standardized and reproducible evaluation process is needed in order to provide the managers with a fair comparison of ES systems helping them to take decisions. Moreover, the goal is to improve the ES systems effectiveness. In order to do this we need a proper way to evaluate the ES system effectiveness.

Enterprise Search Evaluation has been studied mainly in the context of the TREC initiative in the 2005, 2006, 2007 editions. The focus of this initiative was mainly on the evaluation of the Expert Search task, and some few other tasks, while several other ES tasks have not been considered. In this paper we propose a list of possible tasks to be evaluated in the context of ES together with some evaluation considerations (Section 3). We propose and describe in detail a new search task which is about finding the number of experts in the enterprise (Section 4). We first present the motivation for the development and evaluation of systems for answering this search task; we then redefine the concept of expertise in a way which is better suited for the task; finally, we discuss how the evaluation of this novel search task should be performed. In Section 5 we discuss the benefits and the disadvantages of different approaches for identifying the number of experts concluding that the most appropriate way is to harvest information from outside the enterprise in oder to compare the enterprise knowledge with the external environment. Finally, we conclude the paper outlining some future work.

2. DISCUSSION OF RELATED WORK

In this section we describe and discuss the previous work in the field of ES evaluation. We show how existing systems have been evaluated, which evaluation metrics and which definition of relevance have been used.

A proposal to evaluate the ES systems has been done in [14], but it does not address a standard testbed for evaluation. Instead, it mainly proposes a list of questions to consider when selecting one ES system and it does not really evaluate its effectiveness. In [13], the authors consider the choice of queries to be included in an hypothetical testbed for ES evaluation. The most successful initiative for evaluating ES systems in a standardized fashion is the TREC track for ES started in 2005 (TRECent[1]) with the aim of providing standard testbeds. The main problem, caused by obvious privacy issues, is that the test collections are based merely on public Web crawls (of W3C and CSIRO) thus do not reflect the whole available enterprise knowledge.

In the field of IR effectiveness evaluation, several measures to evaluate IR systems have been proposed [7], but they

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

[1] http://www.ins.cwi.nl/projects/trec-ent/

mainly target the evaluation of Web IR systems. In the context of ES, different from the Web where there are usually best answers (i.e., continuous relevance scores), there are only correct or wrong answers (i.e., binary relevance scores). In the task of Expert Search in TRECent, measures have been used with the assumption that not only one expert on the topic is needed by the user which is not always correct. In TRECent 2006 the main measure used was Mean Average Precision (MAP) over experts. The organizers also reported *bpref*, a metrics which is computed only on the explicitly assessed items, to control for incomplete judgments; and precision at 5. The reason for these measures was that this is a classical *ad hoc* task, but we argue that probably the number of retrieved items does matter (see Section 4), and this constraint is not taken into account by measures such as MAP where the quality of the entire ranking (of 1000 candidates) is evaluated.

Moreover, the conceptual framework surrounding relevance has coevolved with ES [12]. We argue that the evaluation of an ES system needs quite different concepts for relevance definition and, thus, also different measures given also that ES have been shown to be quite different from the standard Web search [9].

Privacy is the preeminent issue in the construction of an ES testbed. The problem is the same as in building a desktop search testbed [6]. Thus we propose to adapt solutions developed for desktop search in order to improve the quality of ES testbeds. Another related work in the context of evaluation testbeds is the "Mr.X" collection used in the TREC Spam Track [2] where the real data (i.e., emails) are not publicly available, but the systems are run by the TREC organizers who report the evaluation results. The same approach can be applied in the ES scenario where a "real" enterprise can offer the opportunity to test the ES systems on a real world dataset without disclosing private information.

3. TASKS AND METRICS

In this section we enumerate the list of the possible search tasks in the ES context also describing how these tasks should be correctly evaluated. A list of the most common ES users' needs is presented in [10], and it is compared with the one for the Web proposed by Broder [4]. We follow this categorization for presenting the existing ES tasks.

3.1 Navigational tasks

This group of tasks is about getting to a specific resource in the network: an Intranet page or a more generic entity.

3.1.1 Known Item Search.

The user wants to re-find an item (e.g., a document) that she remembers to exist but not its location within the Intranet. The search task is to find this item again. In this case there is only one relevant result, and, if the output of the system is a ranked list of retrieved items, the most appropriate evaluation measure is one depending on the position of the (first) relevant result. An example of a metrics which is best suited is Generalized Success@10 (GS10) [15].

3.1.2 Home Page Finding.

In this case the user wants to find the Intranet page of a department, group or unit of the enterprise. Also in this case the relevant result is unique and, having the assumption that finding the result after the first ten is the same as not finding it a measure like GS10 is appropriate. In the case where the user is supposed to browse the results after the first ten, a metrics like Mean Reciprocal Rank (MRR) [16] is the most suited.

[2] `http://plg.uwaterloo.ca/~gvcormac/spam/`

3.2 Informational tasks

In this group of tasks the user tries to fill an informational gap she has. That is, the user is trying to learn something new.

3.2.1 Document Search.

The user wants to find documents relevant to a given topic. Metrics like MAP are appropriate in the case where the ES system returns a ranked list of documents.

3.2.2 Email Search.

The user wants to find emails relevant to a given topic. In this case the structure of the document can be exploited in the query (e.g., "Find emails from John Brown about Web Services") and by the retrieval system. Metrics like MAP are appropriate in the case where the ES system returns a ranked list of emails.

3.2.3 Entity Search.

A newer task is the search for entities: for example, finding the phone number of an employee or a list of professors of a university department. Systems for solving this task have already been proposed (see for example [3, 5]), but it is not clear yet how to best evaluate and compare them. Given the fact that a ranked list of results is usually the most appropriate result format, and that the relevance is best defined as binary, a metrics like MAP should work well. In the previous works metrics like MRR and P@10 have been used.

3.3 Transactional tasks

There are many different but very similar tasks in this category. For example, both the task of downloading a software from the Intranet and resetting a password require finding one single relevant Intranet page. If there is only one correct answer for a query (as in the Question Answering problem), probably the best measure is Success@N, which has also been already used for ES evaluation in [8, 9].

3.4 People Search tasks

The last category of tasks is about finding people within the enterprise. This group of tasks might also be defined as a sub-class of Entity Search Tasks if we consider people as a specialization of entities.

3.4.1 Expert Search task.

This task is defined as "Find the experts on topic X". In the previous works (i.e., TRECent) metrics like MAP and *bpref* (in order to check for incomplete assessments) have been used. Given that the number of relevant experts does matter (see also Section 4) a good evaluation metrics would be R-Precision (P@R where R is the number of relevant experts). For the results of [1], which shows that R-Precision is highly correlated with MAP, the evaluation of this task has been performed correctly.

3.4.2 Number of experts.

This task is defined as finding the number of experts on a given topic in the enterprise, which is discussed in detail in Section 4. The evaluation can be based on a metrics which measures the distance from the correct number to the estimation made by the ES system (e.g., $|SNE - UNE|$, where SNE is the System estimation of the Number of Experts and UNE is its actual value).

4. HOW MANY EXPERTS ARE THERE?

In this section we describe in more detail the novel task of finding the number of experts in the enterprise. In this case the definition of "expert" is slightly different from those used for Expert Search (i.e., finding experts on a given topic) where the word "expert" is usually intended as "knowledgeable" and, therefore, we can say that "everyone is an expert on every topic", at least with a certain extent. In this case the need is defined as finding only the people who are *highly* expert on a given topic.

In this section we first present the motivation for the novel task we are defining. We then redefine the concept of expert in a better suited way for the new task of finding how many experts there are in an enterprise. Finally, we present and discuss a possible way to perform a standardized evaluation of the effectiveness of systems in answering to this task. In the next section, we present some possible approaches for finding the number of experts in the enterprise.

4.1 Motivation

There are several reasons why to find how many experts there are. For example, managers can better understand the knowledge power available: this information can be used while selecting the type of new project to acquire, or even to move the core business of the enterprise in the direction of what we can call the *enterprise knowledge* (that is, the aggregation of the employees' expertises) which is today the biggest competitive advantage a company can have. Moreover, it can also help managers in the identification of the need for new employees experts on certain topics.

One specific search task in the enterprise context is to find how many experts on a given topic there are in the enterprise. We can imagine that a general topic (e.g., Computer Science) should have a larger number of people with some expertise while a very specific topic (e.g., IR Evaluation) should have only few people who are *highly* expert.

In the context of Expert Search, the systems should be able to understand the *specificity* of the topic and retrieve a reasonable number of experts and not a fixed number for each topic. Moreover, in big enterprises it might not be always clear how many experts are there. For example, in a well known search engine company the researchers are free to work one day per week on a topic of their choice: this makes it difficult to monitor the personal grow of the employees. Another example is that employees who have particular interests in some topic (e.g., a certain technology) which they can not investigate during their working time, use to write, in their free time, technology related blogs where they explain and describe their discoveries and solution to problems. All these evidences might be difficult to collect for a human resources manager, and a retrieval system which integrates all such evidences in order to discover the quantity of experts present within the enterprise is useful.

4.2 A New Definition of Experts

While in the context of Expert Search the common agreement is to define experts as the people having the highest knowledge within the enterprise, for the task of finding how many experts there are this definition is not suitable. In order to answer correctly to this search task we need to consider not only experts within the enterprise but overall experts in the topic. We need to compare the most knowledgeable people working for the enterprise with the current state-of-the-art knowledge in the world on the given topic. Only in this way we can understand which is the need of the enterprise for new human resources who are experts on the topic or which is the placement of the enterprise among the competitors.

4.3 Evaluation of the task

For evaluating the task of finding how many experts there are in an enterprise, it is possible to reuse available test collections. Usually the Information Retrieval System (IRS) which performs expert search does not focus on retrieving the correct number of experts, but they rather output a ranked list of all possible candidate experts or of the top N. In the TRECent collections the number of relevant experts per query as well as the number of experts retrieved by the IRS is, anyway, available. We can then compute an evaluation metrics to assess the quality of the IRS in identifying the number of experts in the enterprise.

We performed experiments using the TRECent 2005, 2006, and 2007 collections. In 2005 and 2006 the W3C collection [17] was used. It had 30.18 and 28.4 average experts per topic respectively. In 2007 the CSIRO collection [2] was used: because the candidate experts were identified as the Science Communicators of the institution, the number of relevant experts was decreased to 3.04 on average per topic. We defined an evaluation metrics and computed it for each IRS and each topic of the 2006 collection in order to evaluate the ability of retrieving the correct number of experts. In order to be comparable with standard IR evaluation metrics, this metrics should have values between 0 and 1 where 1 is obtained with the ideal behavior. To this end, we compute the value of the H measure as

$$H := 1 - \frac{|SNE - UNE|}{|C|}, \tag{1}$$

where SNE is the System estimation of the Number of Experts, UNE is its actual value of the Number of Experts, and $|C|$ is the number of candidate experts[3], which assumes value 1 when the IRS retrieves the correct number of experts and 0 when the IRS output a ranked list of all the possible candidates and there are no relevant ones. The correlation values with standard IR evaluation metrics for the 2006 collection containing 91 runs are shown in Table 1.

From the absence of correlation (see [11] for a definition of the used correlation metrics) we can conclude that the current metrics do not take into account whether the IRSs retrieve a number of experts close to the real one or not. In particular, we can see, in the TRECent 2006 collection, that the best performing run in terms of H is 'basic', a baseline run which performed 78th out of 91 in terms of MAP showing that it might not be too hard for IR techniques to predict the amount of expertise available.

[3]For the 2006 collection 1092 candidates were present.

	Kendall τ	Spearman ρ	Kolmogorov-Smirnov D
MAP	-0.25	0.005	1
R-Prec	-0.22	0.008	1
GMAP	-0.27	0.0006	1
P5	-0.1	0.30	1
P10	-0.15	0.09	1

Table 1: Correlation values of the IRSs rankings done according standard IR metrics and H, for the TRECent 2006 collection.

5. IDENTIFYING HIGHLY EXPERT EMPLOYEES

There are several ways to identify *highly* expert employees (i.e., people highly knowledgeable on a given topic) in an enterprise. In this section we list some possibilities for computing the cardinality of such set of people.

A common approach in the context of the classic Expert Search is to return a fixed number of experts.

Thresholding the number of experts. The top k experts.

In the case of finding how many experts there are in an enterprise this approach would not help much. This metrics is not good for the new defined task because at each query the answer would be the same (i.e., k).

A more sophisticated approach is to put a threshold not on the number of experts but on the average expertise score so that the number of experts is different for different topics.

Thresholding on the average score. All the experts with score greater than the average score.

If the expertise scores are uniformly distributed among the employees, the number of experts will be close to half of the employees, while if there are few people with high scores and the rest has low scores, the number of experts will be given by the people with high score (see example in Table 2). In these examples we see that in one case (i.e., Ex1) only one *highly* expert employee is correctly identified, but in the other case (i.e., Ex2) two not so *highly* expert employees are identified.

	Ex1	Ex2
1	0.9	0.5
2	0.2	0.4
3	0.2	0.3
4	0.2	0.2

Table 2: Examples of thresholding on the average expertise score for finding how many experts there are.

A possible extension of the previously defined approach is to consider only some of the experts with a score greater than the average expertise score.

Top N% thresholding on the average score. Up to the best N% experts among who have a score greater than the average score.

If we define m as the number of experts above the average score, the drawback of this metrics is when there is only one (or a few, i.e. m is small) experts with a score greater than the average score (e.g. Ex1 in Table 2), and the rest has low scores. In this case, if $N \neq 100$, the result will be that there are no experts even if there is one *highly* expert employee. As another example, in the case where there are 5 people above the average score and $N = 20$ then the result is 1 independently from her score. This means that a careful choice of N is needed based on the value of m. That is, if m is high, then N can be decreased opportunely. If m is low then N should be increased. Moreover, if the average score is low, the result is anyway identifying a number of experts which represents people who are not strong experts (for example in the case where the scores are $\{0.3, 0.3, 0.2, 0.2, 0.2\}$, $m = 2$, $N = 50$, the result is 1 with score 0.3).

The most naive, but effective, approach is to compute an expertise score for each employee of the enterprise (i.e., the set of candidates), and consider as experts, only those with a score greater than a given threshold.

Thresholding the expertise score. All the experts with score greater than a given score (e.g., 0.9).

The weak point of this as well as of the previous approaches is that they consider only internal information about the enterprise knowledge. For example, an ES system might assign scores between 0 and 1 where 1 is the most expert employee on the topic. This does not help in identifying how many experts there are using the definition of expert presented in Section 4.2.

For correctly deciding how many experts there are in the enterprise, and for identifying the employees who are *highly* expert on the topic, we need to compare their knowledge with externally available expertise. In this case the Web can be used as a dataset for identifying experts because it is the best approximation of the external knowledge an ES system can process. We define two possible ways to use the Web in order to find *highly* expert employees in the enterprise:

1. Compare the internal candidates with external candidates. Find the strongest expert on the topic in the Web, assign her the maximum score (e.g., score 1.0), and compare the enterprise employees with her.

2. Compute the total knowledge available in the Web and compare it with the enterprise knowledge.

This brings a new need of finding experts on the Web: a topic which already started to be investigated in the research community [18].

6. DISCUSSION AND CONCLUSIONS

In this paper we first presented the current state of Enterprise Search evaluation. Some evaluation effort already started, but very recent search tasks (e.g., Entity Search) are coming together with systems aiming in solving new challenges. We have proposed for each search task the most appropriate evaluation metrics. Standard metrics such as Mean Average Precision are appropriate in cases when the ranking produced by the systems is important, but, in other cases, less popular metrics such as Generalized Success@N are more appropriate. Moreover, we have motivated and defined a new Enterprise Search task (Number of Experts). In this case the evaluation of the systems must focus on the quality of the estimation, where the possible errors are overestimation and underestimation. A similar scenario is the one of XML retrieval where the correct specificity must be

identified and the proper element size must be retrieved (not too generic and not too specific).

In the future we will investigate, evaluate, and compare possible ways of identifying *highly* expert employees in the enterprise setting also using external evidence of expertise for comparison. A two-steps approach can be followed where the first issue is to identify candidates on the Web for a given topic: one possible solution would be to use the Wikipedia corpus for identifying a list of candidates. The following step is to build an expert profile for each candidate using all the evidence available on the Web.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. Moreover, we thank Fan Deng and Peter Fankhauser for their help in producing the final manuscript, and for their valuable comments. This work is supported by the Okkam project funded by the European Commission under the 7th Framework Programme (IST Grant Agreement No. 215032).

8. REFERENCES

[1] J. A. Aslam, E. Yilmaz, and V. Pavlu. A geometric interpretation of R-precision and its correlation with average precision. In *28th SIGIR*, pages 573–574, 2005.

[2] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The csiro enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

[3] H. Bast, A. Chitea, F. M. Suchanek, and I. Weber. Ester: efficient search on text, entities, and relations. In *SIGIR*, pages 671–678, 2007.

[4] A. Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

[5] T. Cheng, X. Yan, and K. C.-C. Chang. Entityrank: Searching entities directly and holistically. In *VLDB*, pages 387–398, 2007.

[6] S. Chernov, P. Serdyukov, P. A. Chirita, G. Demartini, and W. Nejdl. Building a desktop search test-bed. In *ECIR*, pages 686–690, 2007.

[7] G. Demartini and S. Mizzaro. A classification of ir effectiveness metrics. In *ECIR*, pages 488–491, 2006.

[8] R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson. Searching the workplace web. *In WWW*, pages 366–375, 2003.

[9] D. Hawking. Challenges in enterprise search. *Proceedings of the Australasian Database Conference ADC2004, pages*, pages 15–26.

[10] H. Li, Y. Cao, J. Xu, Y. Hu, S. Li, and D. Meyerzon. A new approach to intranet search based on information extraction. *Proceedings of the 14th ACM CIKM*, pages 460–468, 2005.

[11] M. Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, 2007.

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

[13] T. Rowlands, D. Hawking, and R. Sankaranarayana. Workload sampling for enterprise search evaluation. *Proceedings of SIGIR 2007*, 2007.

[14] D. Stenmark. A Methodology for Intranet Search Engine Evaluation. *Proceedings of IRIS22, Department of CS/IS, University of Jyväskylä, Finland, August*, 1999.

[15] S. Tomlinson. Early precision measures: implications from the downside of blind feedback. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 705–706, 2006.

[16] E. Voorhees. The TREC-8 Question Answering Track Report. *Proceedings of TREC*, 8:77–82, 1999.

[17] W3C Text Collection, 2005. http://research.microsoft.com/users/nickcr/w3c-summary.html (Last visit: February 2008).

[18] A. V. Zhdanova, L. J. B. Nixon, M. Mochol, and J. G. Breslin, editors. *Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics, Busan, Korea, November 12, 2007*, volume 290 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

VisualVectora: An Interactive Visualization Tool for Cumulated Gain-based Retrieval Experiments

Kalervo Järvelin
Dept. of Information Studies
University of Tampere
Finland

Kalervo.Jarvelin@uta.fi

Ilkka Vähämöttönen
Dept. of Information Studies
University of Tampere
Finland

Ilkka.Vahamottonen@cs.uta.fi

Heikki Keskustalo
Dept. of Information Studies
University of Tampere
Finland

Heikki.Keskustalo@uta.fi

Jaana Kekäläinen
Dept. of Information Studies
University of Tampere
Finland

Jaana.Kekalainen@uta.fi

ABSTRACT

Cumulated Gain (CG) based evaluation of the results of IR experiments has gained in popularity. The metrics allow to test several user scenarios regarding the assumed user persistence in scanning the rankec result lists and user's preferences on document relevance. Testing multiple scenarios both aggregated across topics and by individual topics produces masses of evaluation data, which may require much persistence to analyze. We present a tool, VisualVectora, which allows one to visualize CG based evaluation results interactively on screen by topic, across topics and between experimental runs. The user may interactively change the number of ranks to consider, discounting, and relevance gain weighting. This contributes to IR evaluation methodology as the evaluator may easily test and see whether / which experimental runs behave interestingly under which scenarios. One may also easily identify topics which deviate from general trends, for example. The present paper describes VisualVectora and exemplifies several ways of using it in IR evaluation.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]

General Terms

Performance, Experimentation.

Keywords

Experiment result visualization, Discounted cumulated gain

1. INTRODUCTION

Modern IR evaluation focusing on user's viewpoint often takes into account relevance degrees of retrieved documents – assuming that all relevant documents are not equally valuable – and user's persistence in examining ranked retrieval results [3].

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

Cumulated Gain (CG) based evaluation of the results of IR experiments [1] is a popular means of performing such evaluation. The CG based metrics allow the experimenter to test several user scenarios regarding the assumed user persistence in scanning the result lists in ranked retrieval and user's preferences on document relevance. The metrics accommodate such scenarios through evaluation parameters which consist of (a) a discount factor, (b) the number of relevance levels employed in evaluation, and (c) the weighting of relevance levels.

Often experimenters employing CG based metrics need to test multiple scenarios both aggregated across topics and by individual topics. This produces masses of evaluation data, which may require much persistence to analyze. Consequently there is a demand for interactive visual analysis of experimental results providing both aggregated overall and topic-specific graphs representing the performance of various IR techniques under several experimental scenarios. Such an analysis tool should support:

- the analysis of user persistence in examining retrieval results
- relevance weighting of documents from liberal binary weighting to sharply graded weighting
- rapid identification of differences between runs or topics
- interactive analysis.

Such visual analysis allows rapid experimentation at the evaluation stage. The evaluation space is multidimensional due to the parameters the experimenter may want to work with. An interactive tool allows one to examine the stability of findings under different evaluation conditions and also to perform *what-if* – analysis. One may also easily identify topics which deviate from general trends, for example.

In the present paper we present a tool, VisualVectora, which allows one to visualize CG based evaluation results interactively on screen by topic, aggregated across topics, and between experimental runs. The user may interactively change the number of ranks to consider, discounting, and relevance gain weighting and then examine the effects through several CG based metrics – cumulated gain (CG), discounted cumulated gain (DCG), and their normalized variants (nCG, nDCG).

The present paper exemplifies several ways of using VisualVectora in IR evaluation. VisualVectora can be applied as soon as TREC-type of experimental run results (e.g., top thousand results of each query of each run, and the recall bases) are available.

2. VISUALVECTORA OVERVIEW

We shall first look at the system architecture, then the input interface and finally at the output interface and interaction.

2.1 System Architecture

The VisualVectora System architecture is illustrated in Figure 1. An IR experiment is conducted externally and the experimental results, such as TREC run results, are transformed to the VisualVectora format shown in Table 1. The Data file contains the ranked results from the experiment, each line reporting a topic number, document relevance score (e.g. score 0-3), document ID, and document rank (e.g. 1-1000). There may be 1 to 10 different data files representing the results of different runs. The Ideal file contains for each topic the recall base size for each relevance level. Typically the number of non-relevant documents is unlimited. These files are stored on the experimenter's computer and uploaded to the Vectora Server together with run parameters through the experimenter's browser.

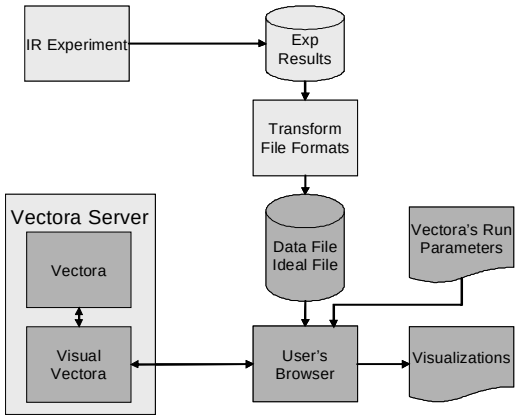


Figure 1. VisualVectora system architecture.

The Vectora Server is a Unix server hosting two programs, Vectora and VisualVectora. The former performs all (n)(D)CG calculations as defined in [1] and the latter visualizes the vector calculation results, giving graphs for all metrics both by individual topics and as averages across topics for each run given in a data file. Vectora is implemented in C and VisualVectora in java.

The experimenter starts VisualVectora through his/her browser and by identifying the data and ideal files and giving the parameters of evaluation (see below). The visualizations are shown in the browser window (also below).

2.2 The Interface - Input

The screen shot of Figure 2 shows the VisualVectora input interface. The main panel gives usage guidelines while the left panel allows the user to specify the evaluation to be carried out. The input parameters are:

- specification of the number of data files – opens equally many fields for data file names

Table 1(a). VisualVectora data file format

Topic #	Score	Document ID	Rank
001	3	FT922-15099	1
001	2	FT942-12805	2
001	1	FT941-9999	3
001	3	FT934-4848	4
001	2	FT921-6603	5
...	...	...	...
002	2	LA111690-0059	1
002	2	FT911-558	2
002	0	LA120389-0149	3
...	...	...	...

Table 1(b). VisualVectora ideal file format

Topic #	R=0 #	R=1 #	R=2 #	R=3 #
001	(0 unlimited)	(1 15)	(2 9)	(3 14)
002	(0 unlimited)	(1 33)	(2 19)	(3 4)
003	(0 unlimited)	(1 21)	(2 5)	(3 11)
004	(0 unlimited)	(1 0)	(2 8)	(3 41)
…	…	…	…	…

- fields for data file names – the names can be browsed in the user's directory
- field for the ideal file name – the name can be browsed in the user's directory
- the number of ranks to consider: the experimental data may contain to rank, e.g., 1000 but the user may want to focus on less, say, the top-100 ranks
- the log base for discounting
- the number of relevance levels used (2 to 10) – opens equally many fields for relevance level weighting
- the run button.

By clicking the 'Run Vectora' button the data are handed over to Vectora for checking and computation. If the data contain no syntax errors, the aggregated and by-topic results will be shown.

2.3 The Interface - Output - Interactive Use

Figure 3 shows some VisualVectora sample output. The main panel contains the visualization graphs. Each run (and thus data file) forms one row of graphs. The aggregate graphs are given first (nDCG, DCG, nCG and CG) followed by graphs for individual topics. Corresponding graphs (aggregations, topical) from different runs are arranged vertically so that they can be inspected together. The buttons below the graphs allow enlarging each graph, picking the data from several graphs and joining them into one graph, as well as saving the graph data for exporting into other systems, like a spreadsheet. The left panel provides buttons for focusing on specific metrics and topics as well as combining a range of visualizations into one.

The experimenter may return to the initial screen and change the evaluation parameters flexibly. However, all the data files are processed with the same parameters. In some situations the experimenter might want to examine one or more runs under different conditions – that is, using different parameters. In this case one opens two or more VisualVectora windows, aligns them vertically (in a overlayed fashion), and then runs the same data files under different parameters. This allows rapid testing of various user scenarios on the same data sets.

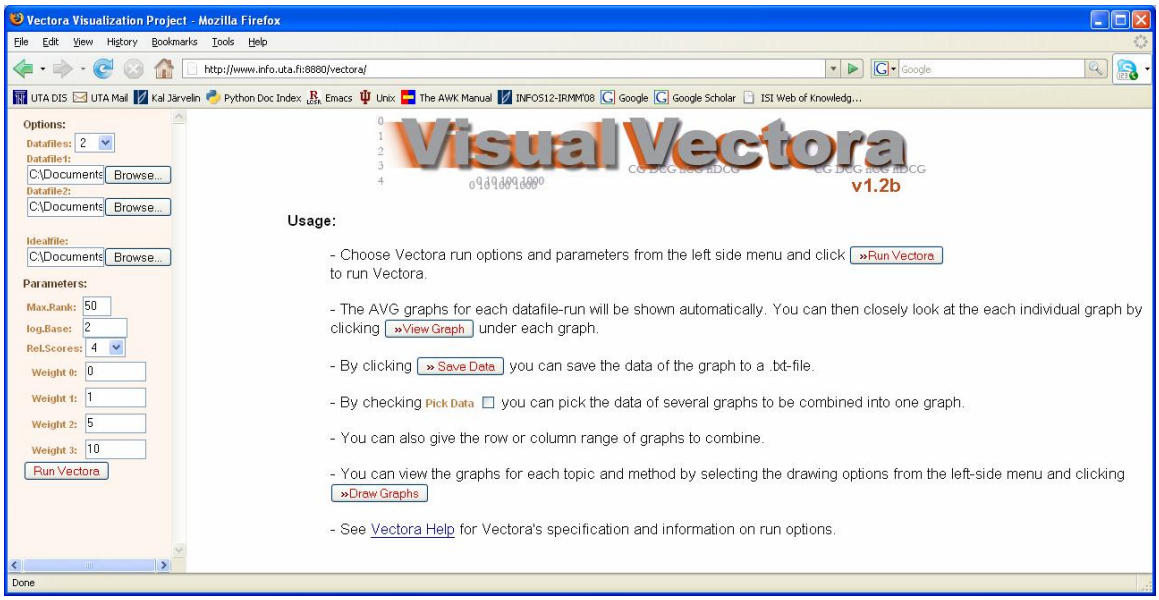


Figure 2. VisualVectora input: parameters on the left panel; instructions in the middle.

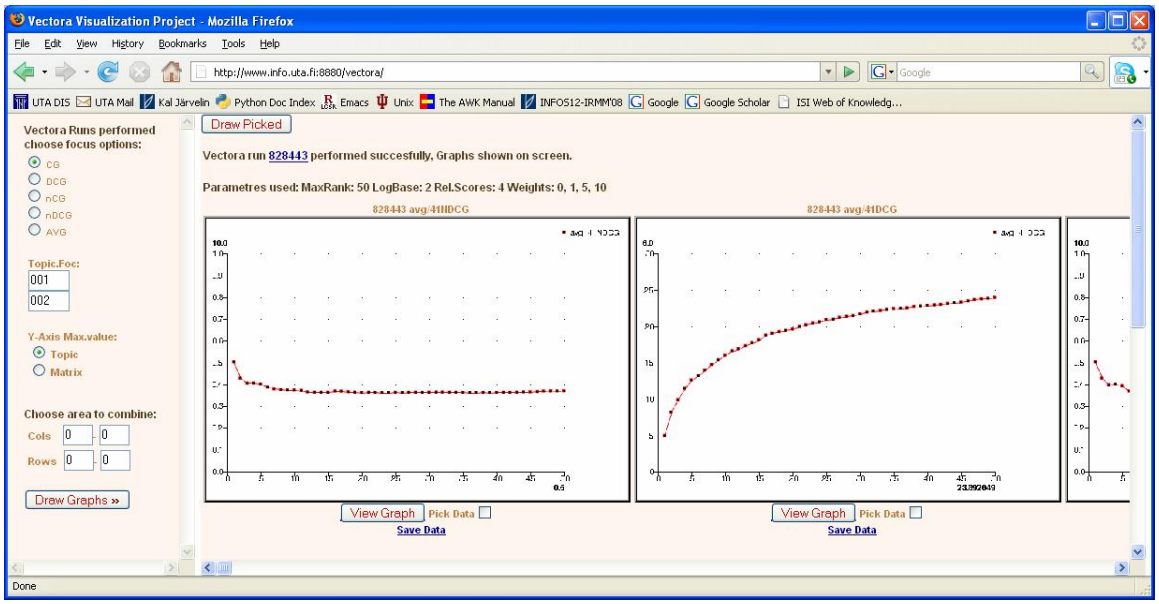


Figure 3. VisualVectora output: controls on the left panel, visualizations in the middle.

Below we shall use graphs generated by VisualVectora but we shall extract them from the VisualVectora output page in order to provide concise illustrations.

3. INTERACTIVE EVALUATION AND VISUALIZATION

We shall illustrate the use of the VisualVectora tool through the first five evaluation issues listed in the introduction.

3.1 The Data and the Runs

First, we use one of the best performing runs of the TREC 8 ad hoc track based on binary relevance judgments (50 topics). Second, we experiment with graded relevance test data (41 topics from TREC 7 and 8 having graded relevance assessments) – the documents are highly, fairly or marginally relevant, or non-relevant [9] in respect to a topic. We illustrate two runs: One consists of title-only queries and the other of queries based on titles, descriptions and narratives. The retrieval system used is

Lemur[1], and it was run with language modeling and two-stage smoothing options.

3.2 Several Metrics Automatically

Figure 4 shows four aggregate metrics for the TREC 50 topic run – nDCG, DCG, nCG and CG. The top row shows the nDCG and DCG curves, the bottom row the corresponding nCG and CG curves. One may quickly see that discounting bends the cumulated gain quickly toward a horizontal line and the total accumulated gain thus becomes much less (DCG max=12.8 vs. CG max=26.8). The normalized versions, however, do not have much difference.

Such a display provides a rapid overview of the behavior of the data set(s) under given evaluation conditions. By modifying the parameters one may test different assumptions, e.g., regarding searcher persistence and relevance criteria.

[1] Lemur is an open-source "toolkit designed to facilitate research in language modeling and information retrieval." (http://www.lemurproject.org/)

3.3 Aggregated vs. Topic-by-Topic Evaluation

While IR techniques are mainly tested for their average performance in a test collection, it is hardly sufficient to leave the analysis at the aggregate level. It often is necessary to analyze and present the topic-by-topic variation in the performance of a technique (e.g. [5], [8]). Figure 5 extracts the nDCG visualizations of four individual TREC topics (numbers #405 #407, #409, #410) from the preceding visualization run. In this way it is easy to see which topics are difficult (e.g. #409) and which are easy (e.g. #410) for a given run or across runs. Figure 8 at the end of the paper shows three query expansion runs (baseline and two expansions – details omitted) for 6 topics measured by nDCG and DCG – high topical variation becomes apparent.

VisualVectora may be used to identify query types, or performance variation across queries. Further, the researcher gains insight into the robustness of an IR technique. Corresponding visualizations for the same topics from different runs can be joined to illustrate the differences. It is also easy to export the data for the calculation of the curve averages (cf. the average precision of a precision-recall curve) for comparative by-topic histograms.

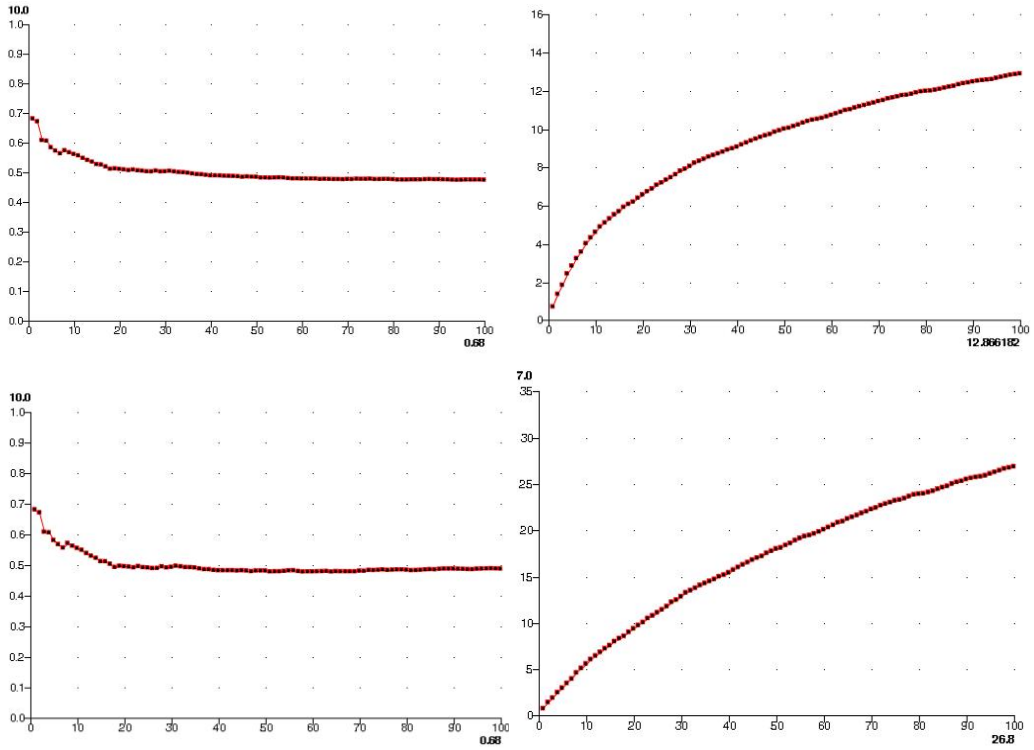


Figure 4. Four aggregate metrics for a TREC-8 ad hoc run (50 topics). Top row: nDCG and DCG; bottom row: nCG and CG. Binary weighting 0/1 and discount log=4.

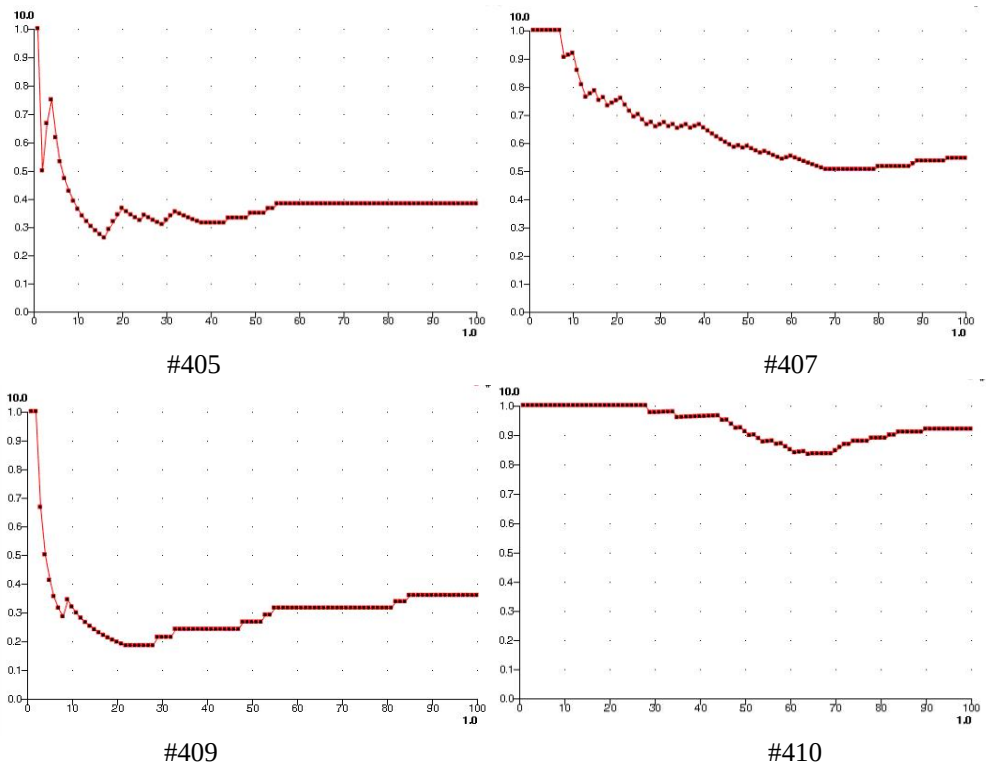


Figure 5. Performance curves (nDCG) for four individual topics (#405 #407, #409, #410) of a TREC-8 ad hoc run (50 topics). Binary weighting 0/1 and discount log=4.

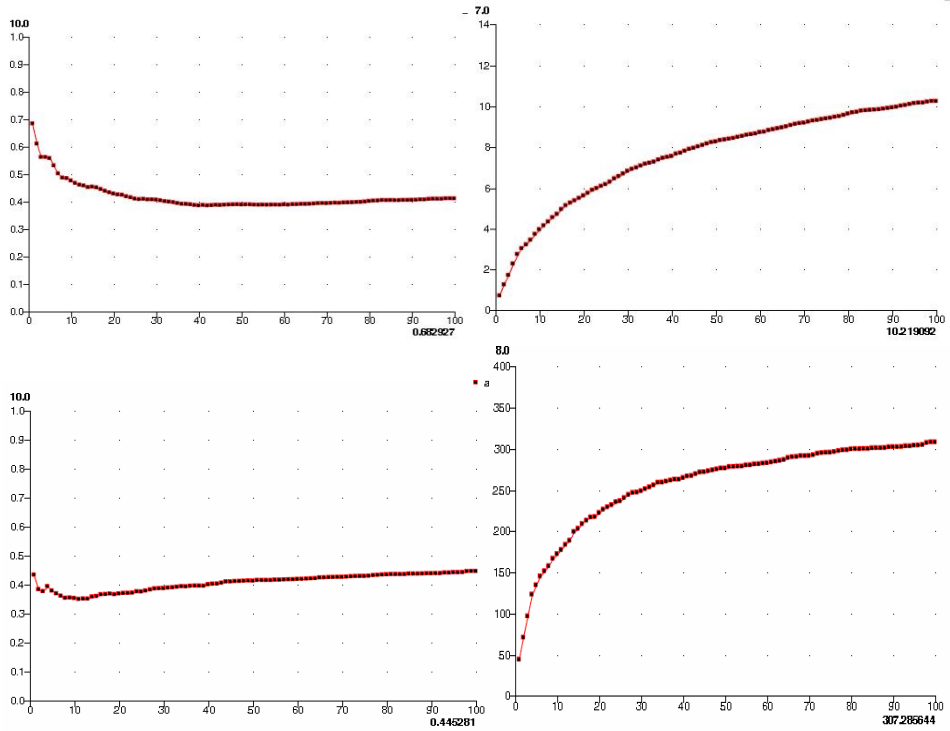


Figure 6. Aggregated performance curves (nDCG and DCG) for 41 topics, T+D queries and graded assessments. Top row: Binary weighting 0/1/1/1 and discount log=4; DCG top score 10.2. Bottom row: graded weighting 0/1/10/100 and discount log=4; DCG top score 307.3.

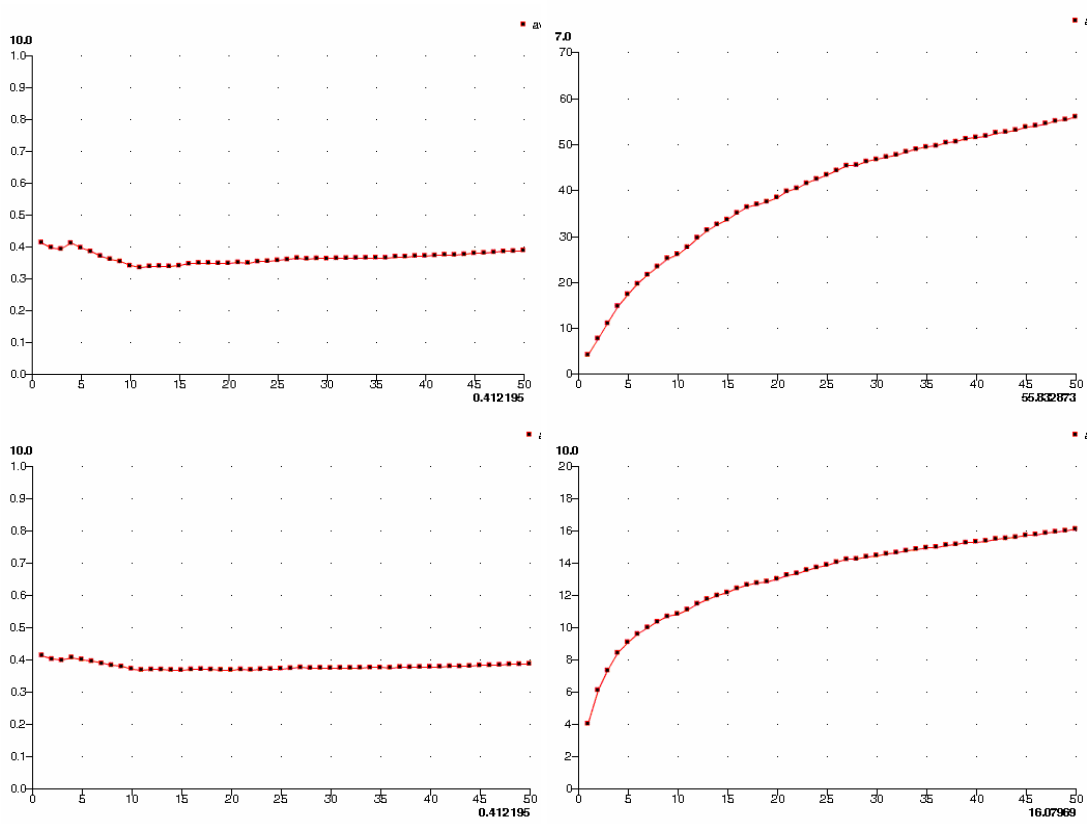


Figure 7. Performance curves (nDCG) and DCG for 41 topics, T+D+N queries and graded assessments. Top row: graded weighting 0-1-5-10 and discount log=10; nDCG @50 0.383 and DCG top score 55.8. Bottom row: graded weighting 0-1-5-10 and discount log=1.5; nDCG @50 0.386 and DCG top score 16.1.

3.4 The Effects of Relevance Weighting

In user oriented IR evaluation, one may need to take into account the varying relevance preferences searchers may have. Some users value any document contributing at least something to the topic, if not more than proving its existence in the collection. They may be represented by flat TREC-style relevance weights as in Figure 6, top row. Other users may ignore marginal documents and value highly relevant documents high as in Figure 6, bottom row. Different scenarios have been tested by e.g. [2],[11]. One may observe that the graph types, nDCG and DCG, yield quite different evaluation results for the same retrieval run. In this way, VisualVectora supports rapid testing of the effects of relevance weighting of retrieved documents.

3.5 The Effects of Discounting

In addition to relevance weighting, user oriented IR evaluation may require to take into account the rank positions of relevant documents. In cumulated gain based evaluation, this happens through the discounting factor [1]. Patient searchers may be willing to scan the search results at extended lengths while impatient ones hardly check more than 20 results. The patient searchers may be represented by a large log base (say, 10) for discounting as in Figure 7, top row. The impatient searchers may be represented by a small log base (say, 1.5) for discounting as in Figure 7, bottom row. One may observe that harder discounting (log=1.5) soon turns the curves insensitive to accumulating gain, making them horizontal, while the softer discount allows the curves climb even at rank 50. In this way, VisualVectora supports rapid testing of the effects of gain discounting of retrieved documents.

4. DISCUSSION

IR evaluation methodology is the study of IR evaluation methods and metrics. Therefore it analyzes, compares and discusses (a) the methods / metrics as algorithms, i.e., how to calculate, (b) the problems or issues on which the methods / metrics are applicable, and (c) the justifications for the application of the methods / metrics on the problems.[2] IR evaluation methodology tells that one should perform evaluation not only at the aggregate level, as cross-topic averages, but also at the level of individual topics in order to be able to assess the variance under the average evaluation figures. Moreover, the methodology advices that in user-oriented evaluation one should look at several evaluation scenarios to assess the stability of findings across the scenarios – after all, users exhibit quite a lot of variability regarding their persistence in examining retrieved results and in their relevance requirements / criteria. In the case of cumulated gain-based evalua-

[2] Methodology is the study of methods. The definition of method is the triple (domain, algorithm, justification) – see [7].

tion metrics, the evaluation scenarios vary across (a) the number of result ranks considered in evaluation, (b) the relevance weighting applied on documents of different degrees of relevance, and (c) the discount applied on the weights of late-ranked documents. Testing a range of evaluation scenarios across these three dimensions is fruitful but also produces a lot of data – as in TREC-style of evaluation, one thousand lines per run and topic, for example. Data visualization can represent such results in a concise form that supports intuition and rapid scanning of evaluation results.

The VisualVectora tool presented in the present paper is such a visualization tool that supports CG-based evaluation across the three dimensions: the number of result ranks, relevance weighting, and the discounts, both as overall average performance and by individual topics. It applies automatically four metrics, CG, nCG, DCG and nDCG. For each parameter configuration, it computes the visualizations for one or more retrieval runs, depending on how many the experimenter wants to work with. Each run produces a horizontal sequence of graphs and corresponding aggregate or topical graphs for different runs are aligned vertically. Moreover, the experimenter may open several VisualVectora windows for the same runs but different parameter configurations. On a large screen they may be partially overlaid to support rapid swapping and comparison.

There are some limitations in the current implementation of VisualVectora tool which call for further development. Like much of laboratory IR evaluation, VisualVectora is query-based, not session-based. There is no way to analyze multiple query sessions but as individual queries. In a similar fashion, evaluation of relevance feedback is not directly supported while the metrics have been applied to the issue [4]. The VisualVectora architecture allows one to swap the Vectora component for another one, as long as the inputs and outputs remain structurally the same. This opens the possibility of using other components for working with exotic document weighting, such as rewarding partially relevant documents [10] or penalizing for non-relevant documents [6]. Further development ideas contain including other metrics (e.g., MAP) and producing comparative topic-by-topic performance histograms automatically. The latter requires that, instead of a performance curve, a single-figure performance indicator, such as the mean of an nDCG curve, is first computed.

5. CONCLUSION

Cumulated Gain (CG) based evaluation in IR experiments is gaining ground. The metrics allow the experimenter to evaluate several user scenarios concerning the assumed user persistence in scanning the ranked result lists and user's relevance preferences. Testing several scenarios produces piles of data which is difficult to manage and analyze. We have described in this paper a tool, VisualVectora, which supports interactive visual analysis of retrieval results in IR evaluation. The experimenter may interactively change several parameters in the evaluation setting: the number of ranks to consider, discounting level, and relevance gain weighting. VisualVectora contributes to IR evaluation methodology as the experimenter may easily test and see how each experimental run behaves under each scenario. Detecting topics which deviate from general trends is also easy.

6. ACKNOWLEDGMENTS

This research was supported by the Academy of Finland Project Numbers 1209960, 1124131 and 1115480. VisualVectora is a password-protected Web application. Access can be granted for academic research – please contact the first author.

7. REFERENCES

[1] Järvelin, K. and Kekäläinen, J. 2002. Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems (ACM TOIS) 20(4), 422-446.

[2] Kekäläinen, J. 2005. Binary and graded relevance in IR evaluations – comparison of the effects on ranking of IR systems. Information Proc. & Management. 41 (5), 1019-1033.

[3] 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.

[4] Keskustalo, H., Järvelin, K. and Pirkola, A. 2008. Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value. Information Retrieval 11, in press.

[5] Monz, C. and de Rijke, M. 2002. Shallow morphological analysis in monolingual information retrieval or Dutch, German and Italian. In: C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (eds.), Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, LNCS vol. 2406, 262-277.

[6] Myaeng, S. H. & Korfhage, R. R. (1990). Integration of User Profiles: Models and Experiments in Information Retrieval. Information Proc. & Management 26(6), 719-738.

[7] Newell, A. 1969. Heuristic programming: Ill-structured problems. In: Aronofsky, J. (ed.) Progress in Operations Research, III. New York: John Wiley & Sons, 360-414.

[8] Qu, Y., Hull, D. A., Grefenstette, G., Evans, D. A., Ishikawa, M., Nara, S., Ueda, T., Noda, D., Arita, K., Funakoshi, Y., and Matsuda, H. 2005. Towards effective strategies for monolingual and bilingual information retrieval: Lessons learned from NTCIR-4. In: ACM Transactions on Asian Language Information Processing (TALIP) 4, 2 (Jun. 2005), 78-110. DOI= http://doi.acm.org/10.1145/1105696.1105698.

[9] Sormunen, E. 2002. Liberal relevance criteria of TREC – Counting on negligible documents? In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '02) (Tampere, Finland, August 2002) ACM Press, New York, NY, 324–330.

[10] Spink, A., Greisdorf, H. and Bateman, J. 1998. From highly relevant to not relevant: examining different regions of relevance. Information Proc. & Management 34 (5), 599-621.

[11] Voorhees, E. 2001. Evaluation by highly relevant documents. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '01) (New Orleans, LA, September 2001) ACM Press, New York, 74–82.

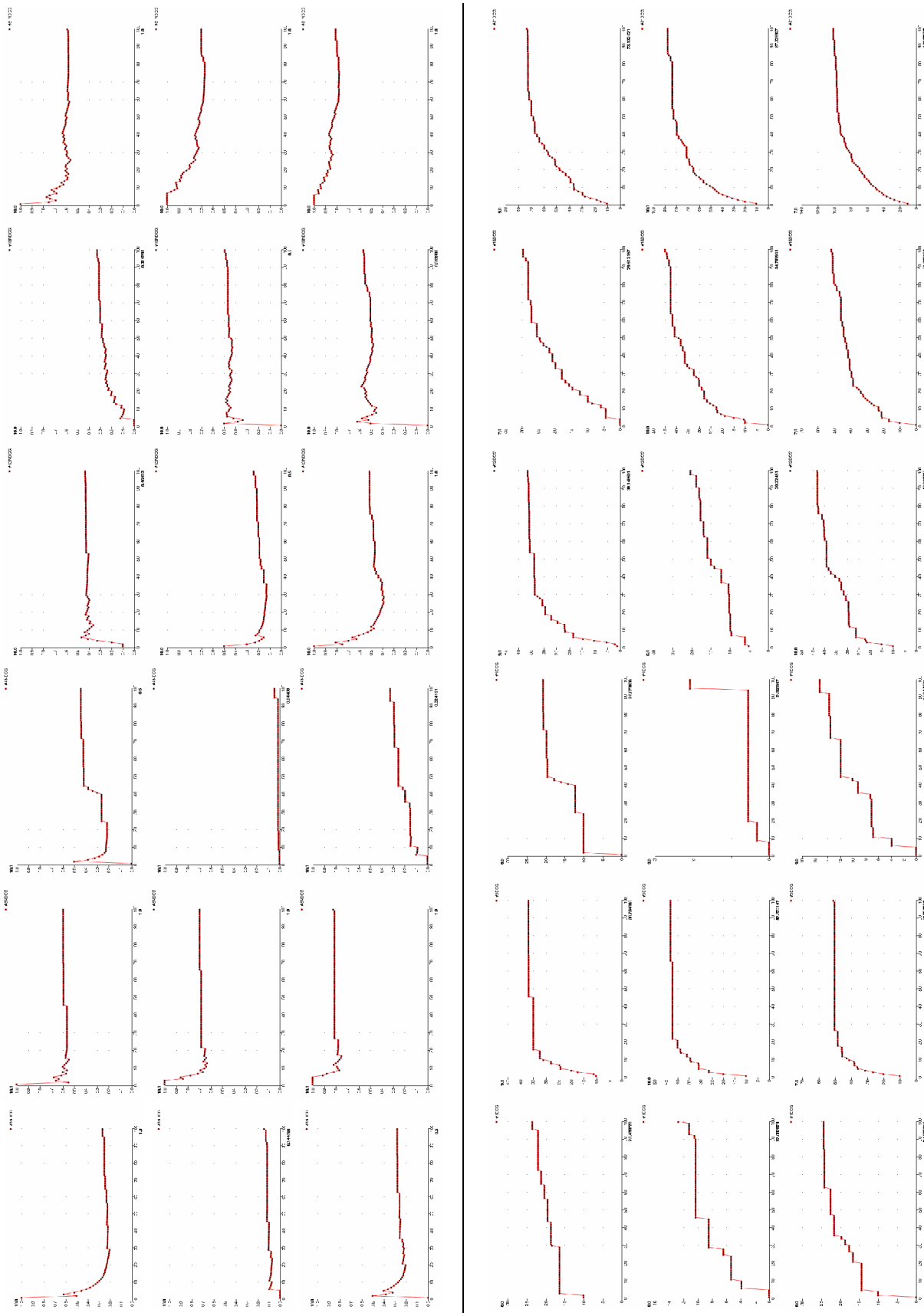


Figure 8. nDCG (top 3) and DCG (bottom 3) performance curves for 6 topics (one per column). Each vertical group of three graphs consists of three different runs (Baseline, Query Expansion A and Query Expansion B) for a topic and metric. Performance variations between runs at topical level explicated. In the DCG curves, the vertical axis height is not fixed. Finnish graded collection TUTK, 0/1/5/10 weighting, discount log=2.

Document Accessibility: Evaluating the access afforded to a document by the retrieval system

Leif Azzopardi
Department of Computing Science
University of Glasgow
Glasgow, UK
leif@dcs.gla.ac.uk

Vishwa Vinay
Microsoft Research Cambridge
7 J J Thomson Avenue
Cambridge, UK
vvinay@microsoft.com

ABSTRACT

How accessible a document is through an information retrieval system dictates how easily the document can be retrieved. This so called *document accessibility* provides a different way in which to evaluate an information retrieval system. This is because the focus is not on relevance but on retrieval; and retrieval is a precursor to relevance. In this workshop paper, we empirically explore the use of recently proposed measures of document accessibility in a pilot study on the TREC AQUAINT collection. Our experiments show how different retrieval models provide different levels of access to documents in the collection. This suggests that the measures could be useful for identifying bias towards certain parts of the collection, as a result of employing a particular retrieval model.

1. INTRODUCTION

Accessibility[1] is an abstract concept coined over fifty years ago in the field of transportation planning and land use [5]. In [1], this concept is adopted in the context of information retrieval. Instead of considering the accessibility of resources like employment within a city given the transportation system, the authors consider the accessibility of documents in a collection given an information retrieval system. In adopting this concept, they propose measures which aim to capture the ease with which a document can be retrieved given the retrieval system. In this paper, we shall refer to the access afforded to a document by the retrieval system as *document accessibility*. Measuring how accessible documents are provides new directions for the evaluation of an information retrieval system; ones that are not directly related to issues regarding effectiveness and efficiency but instead address the evaluation of issues regarding the access afforded to information.

With search providing a crucial role in the access to information, there are growing concerns over the "accessibility of information" through this technology [7, 11, 14, 8, 9]. This is because increasing amounts of information is being made available online, and using an IR system is becoming the primary means of accessing this information. One of the main concerns is over the manipulation of document rankings to favor certain groups of documents over others [14, 8]. Such bias is a very real concern on the web, and perceived bias in rankings has led to legal action being taken against a well known search engine company[2]. Measures of document accessibility could be used to determine in an objective way the presence of bias, if any, towards some documents over others. In the area of e-Government, ensuring that online content is accessible is very important because citizens of a democratic country have a right to the information. If the information is hidden from the public then this could jeopardize the integrity of the government. The importance of e-Government content being made accessible through search technology was highlighted in a recent report[3]. This resulted in changes to U.S. legislation[4] requiring that government websites be monitored and assessed in terms of how "searchable" they are so as to ensure that government information is accessible. Measures of document accessibility could be used to assess whether a sufficient amount of access is afforded to the information housed within e-Government websites. Currently, there are no quantitative measures and methodologies that can be employed or are recognized for this task.

In this workshop paper, we first outline the proposed document accessibility measures, and then report a number of experiments measuring the accessibility of documents given three different retrieval models. Our experiments show that measuring document accessibility provides interesting and new insights into the influence of the retrieval models on the access to documents within the collection. Our results show that some documents in the collection are substantially more accessible than others, and different retrieval models favor different parts of the collection tested. Our work suggests that measures of document accessibility are potentially useful in the evaluation of information retrieval systems for

[1]Accessibility, in the sense used in the current papers is unlike that under the Web Accessibility Initiative (WAI) which focuses on the usability and mobility issues concerning access to information on the web.

[2]see `http://www.searchenginewatch.com`

[3]Hiding in Plain Sight: Why Important Government Information Cannot be Found Through Commercial Search Engines, Center for Democracy and Technology, `http://www.ombwatch.org/info/searchability.pdf`

[4]U.S. Legislation: E-Government Act 2002, and the E-Government Reauthorization act 2007

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.
.

tasks like bias detection or ensuring sufficiency of access.

The structure of this paper is as follows. In the next section, we provide an overview of similar ideas that have been discussed within IR and briefly introduce the measures of document accessibility. Section 4 provides experiments that first calibrate the proposed measures and then explore how these measures could be used in an example on the TREC AQUAINT collection. In Section 5, we conclude with a summary of the empirical study conducted and details of some directions for future work.

2. RELATED WORK

The purpose of an information retrieval system is to deliver relevant content to the user and it should do this effectively and efficiently; evaluation in IR is the process of quantifying a system's ability to achieve this end. Effectiveness measurements attempt to capture *user effort* and are based on metrics that are derived from models of assumed user behavior (e.g. binary or graded relevance, top-heavy metrics that favor higher ranks, etc.). Efficiency metrics are designed to indicate *system effort* in terms of resources (e.g. average time to process a query, index size, etc.). The process of a user receiving access to a document however includes many more steps, often addressed by different components of the IR system.

Firstly, a necessary condition for a document to be a candidate result is that it should be present in the system's index. In the context of web search engines, knowledge of the existance of a page is dictated by how likely it is that the search engine's crawler will reach this page [13]. Some web-pages are essentially unreachable while going through a web search engine, unaccessible pages include dynamic webpages that a search engine cannot crawl and subject matter that is explicitly excluded from the crawl process. An interesting third kind are the undiscovered pages, crawlers start from a few seed pages and expand the pool of pages to be crawled by following links from pages already seen. If the initial set of seed pages are seen as transport routes into the entire web, accessible pages are only the ones that are part of the publicly linked web and connected (at varying distances) to the seed pages. Dasgupta et al [4] refer to this as the *discoverability* of pages on the web, and a related discussion of "dark matter" on the web can be found in [2].

Once a page is part of the search engine's index, how likely is it that a user will be presented with this page as part of a result set? The answer to this question consists of at least two components. Firstly, how likely is it that a query for which this page is a potential answer will be received by the system? Secondly, given such a query, what does the system see as the degree of match between the query and this page? The first criterion differentiates pages which cover a large number of relatively common queries from pages which are only ever returned for extremely rare and often highly specific queries. Retrieval algorithms differ in the scores they attribute to a query-document pair and therefore the likelihood of a document being in the result set of a query is clearly dependant on the specifics of the scoring function.

In hypertext environments, pages can be accessed during a browsing session by traversing hyperlinks, metrics like PageRank capture the possibility of a page being an intermediate step in such a path. In a world with search engines, pages can either be results themselves (i.e., links to them are displayed in response to a query) or are reached by navigation from a result page (known as post-query navigation [10]).

As the earlier paragraphs indicate, measures that focus on particular facets of the findability of content already exist. These include estimating how crawlable a site is (and how reachable the pages within it are), how easily a user can navigate around the site, etc. The one component of an IR system that has been ignored so far is the retrieval function. Once a document index has been built, the accessibility of a given document from the collection is dictated by the properties of this scoring function, the queries issued to the system and the document representations.

The rest of this paper concentrates on *document accessibility*, where the document representations and the set of queries are fixed. By choosing a particular retrieval function (including any parameters it might have), the question is what *a priori* bias has been imposed upon the access to information within the index? The next section outlines measures of document accessibility that capture how easily a document can be retrieved, given a particular retrieval function, as a way to determine this. The basis of these measures is inspired by work in the area of transportation planning and land use [5]. In this context, accessibility captures the potential to access opportunities (such as employment) at locations in a physical space (such as a city) given the transportation system (i.e., the road network and the bus, cycle path and a bicycle, etc). The accessibility is affected by the desirability of the opportunities and the willingness of the users of the system to travel in order to reach these opportunities.

Measuring accessibility in this context enables studies to be performed which consider the levels of accessibility to employment opportunities, schools, shops, etc., and how changes in the levels of accessibility affect the area (in terms of economic impact, social changes and so forth). The results of such studies provides valuable information to transportation planners and city designers in the development of land use, which is then used to inform the development of the transportation system. Before this, planners and designers would focus on measures which were based on the efficiency of the transportation system (for instance, the travel time between particular locations). However such approaches only provided very localized information about specific instances, whereas accessibility measures provided a global view of the quality of the system and its impact upon users.

An analogy of accessibility in information retrieval can be made as follows [1]. Instead of a physical space, in IR, we are concerned with accessing information within a collection (or information space), and instead of a transportation system, we have an IR system. Entering a query is like choosing a particular bus, where the ordered list of documents returned is like the order of destinations reached for that bus route. Opportunities to interact with resources while traveling along the route are reflected by going through the documents returned in the ranking by the retrieval system. The accessibility of the documents is dependant on the willingness of the user to travel a certain distance along the route (i.e., traverse down the ranked list) and all the queries that users are likely to travel along. In this way, the potential of a documents being retrieved can be measured, as a way to capture the document's accessibility; or the ease with which

a document can be retrieved.

3. DOCUMENT ACCESSIBILITY

More formally, given a collection $\mathbf{D}$, an IR system accepts a user query $\mathbf{q}$ and returns a ranking of documents $\mathbf{R_q}$. We can consider the accessibility of a document as a system dependent factor that measures how retrievable it is, with respect to the collection $\mathbf{D}$ and the ranking function used by the IR system. The general measure of the accessibility of a document is defined as ([1]):

$$A(\mathbf{d}) = \sum_{\mathbf{q} \in \mathbf{Q}} o_q \cdot f(c_{dq}, \theta) \quad (1)$$

where o_q denotes the likelihood of expressing query $\mathbf{q}$ from the set of queries $\mathbf{Q}$ or its *importance*. $f(c_{dq}, \theta)$ is a generalized utility/cost function where c_{dq} is the distance associated with accessing $\mathbf{d}$ through $\mathbf{q}$ which is defined by the rank of the document, and θ is a parameter or set of parameters given the specific type of measure.

A cumulative based measure can then be defined as follows: $\theta = c$, where c denotes the maximum rank that a user is willing to proceed down the ranked list. The function $f(c_{dq}, c)$ returns a value of 1 if $c_{dq} \leq c$ (with the top-most position considered as rank 1), and 0 otherwise. So, if returning a document in response to a given query has a distance greater than c associated with it, then it is considered unaccessible (for this query). For another query however, the document may be accessible because the cost of accessing it is within the distance c. Alternatively, the document could be considered accessible for the same query but to a user who has a higher cost threshold. Since all the documents within the cutoff defined by c are equally weighted, this type of measure emphasizes the number of times the document can be retrieved within that cutoff over the set $\mathbf{Q}$.

A gravity based measure can also be defined by setting the function to reflect the effort of going further down the ranked list, such that documents at lower ranks are considered less accessible. For instance, the accessibility of the document could be set inversely proportional to the rank of the document, such that:

$$f(c_{dq}, \beta) = \frac{1}{(c_{dq})^{\beta}} \quad (2)$$

where, the set of parameters θ includes β which is a dampening factor that adjusts how accessible the document is in the ranking. Preference for higher ranks has been observed in studies of user search behaviour (e.g. [6]) and the gravity-based measure provides a simple mechanism to incorporate such information.

Given either measure, $A(\mathbf{d})$ provides an indication of the opportunity of retrieving $\mathbf{d}$. This value can be obtained for each document $\mathbf{d} \in \mathbf{D}$ so that we can compare whether there is more opportunity to retrieve one document over another. Using this measure to compare groups of documents has potential to aid in the design, management and tuning of retrieval systems in a number of ways. Imagine that for a given collection of documents and a given IR system, the average $A(\mathbf{d})$ of a set of documents is extremely high, while for another set of documents the average $A(\mathbf{d})$ is very low. Perhaps, the first set of documents was a group of site entry pages, and our system has a prior towards such pages, thus we would expect these pages to have a higher $A(\mathbf{d})$. In this case, it is desirable that these documents are more accessible. On the other hand, if the set of highly accessible pages was composed of spam pages, because these pages have used "tricks" to artificially inflate the number of queries for which they are retrieved, then this is not desirable and the system needs to be adjusted. Alternatively, if there is a set of documents which are virtually inaccessible in the collection, either the documents' content needs to be altered or the retrieval system needs to be changed, or both.

At a higher level, the measure $A(\mathbf{d})$ motivates questions regarding how accessible documents in the collection should be, and whether we are interested in trying to "hide" or "promote" certain documents within the collection. Or whether we should adopt an approach that ensures access to the information is free from bias, so that *any document is as accessible as any other document* in the collection. We refer to the latter notion as "universal access"[5]. Measures of document accessibility provide a novel way in which these questions and issues can be considered objectively.

4. EXPERIMENTS

In this section, we report the results from a set of experiments designed to achieve two purposes: (a) illustrate the behavior of the proposed measures with respect to their parameters, and (b) evaluate the behavior of three standard retrieval models on a standard IR dataset as reflected by measures of $A(\mathbf{d})$.

Experimental Setup.

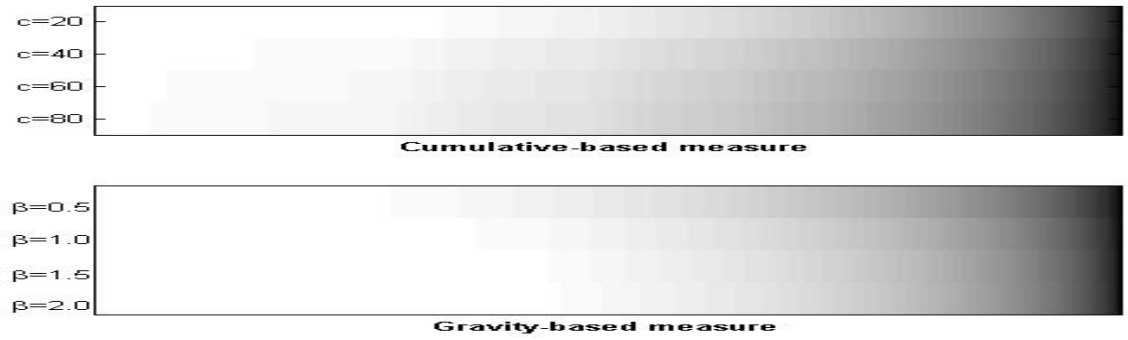
The data we used for our experiments was the AQUAINT collection which consists of three different news sources (APW, NYT and XIE) compiled over a number of years (1996-2000). The documents were indexed using the Lemur Toolkit[6] where Porter stemming was applied and stop words removed. For the purposes of this example, we considered three popular retrieval methods available in Lemur - TFIDF, BM25 and Language Model(LM). The default parameter settings were used for TFIDF and BM25. When a comparison across retrieval models is being made, LM with Bayes smoothing parameter $\mu = 1000$ was used and is referred to as LM1000. The effect of the setting of parameter μ is also investigated.

The set of queries, $\mathbf{Q}$ used plays an important role in the estimation of $A(\mathbf{d})$. In this paper, the set of queries was compiled using one query for every single term in the vocabulary (leading to $663,158$ single-term queries). $\mathbf{Q}$ defines the reference set of queries with respect to which the accessibility of the documents is calculated. It should be expected that with different definitions of $\mathbf{Q}$, alternative views of the accessibility of documents in the collection would be obtained. For instance, if $\mathbf{Q}$ is restricted to a particular topic set, then the accessibility of the documents given a particular topic could be evaluated. Here, we chose single word queries in order to try and obtain a topically unbiased sample of the collection, an approach similar to query based sampling [3].

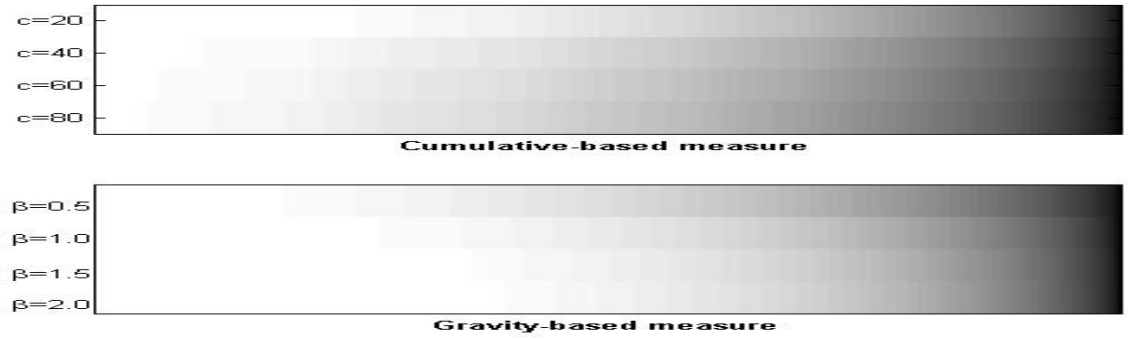
We also treat each query equally, such that $o_q = 1$ in Equation (1). Essentially, this approximation of $A(\mathbf{d})$ is formed on the basis that any route into the collection is equally possible. Investigation of how different weighting

[5]The disability rights movement advocates equal access and terms this notion as universal access. This differs from our use of the phrase.

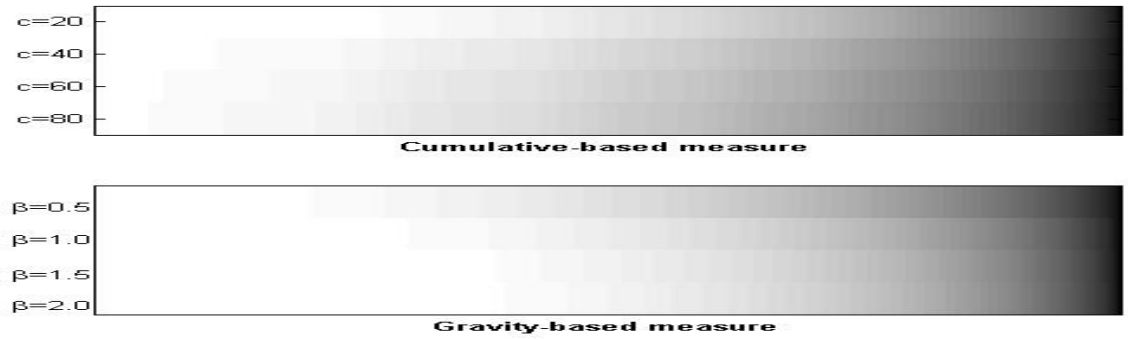
[6]`http://www.lemurproject.org`



(a) TFIDF



(b) BM25



(c) LM1000

Figure 1: Behavior of the three retrieval models with respect to settings of parameters of the two proposed measures. As c increases the coverage (i.e., the unique set of documents retrieved at least once) increases because more documents are retrieved as more documents are considered. While, as β decreases, more documents are deemed accessible, because documents low down in the ranking are penalised less.

		Cumulative-based Measure			Gravity-based Measure		
		c = 40	c = 60	c = 80	$\beta = 1.0$	$\beta = 1.5$	$\beta = 2.0$
TFIDF	c = 20 $\beta = 0.5$	0.980	0.962	0.945	0.955	0.897	0.858
BM25	c = 20 $\beta = 0.5$	0.966	0.947	0.928	0.947	0.884	0.845
LM1000	c = 20 $\beta = 0.5$	0.968	0.948	0.931	0.945	0.886	0.846

Table 1: Pearson's correlation coefficient between A(d) values calculated for different parameter settings of the Cumulative and Gravity based measures

schemes (e.g. "more common queries should be weighted higher") will affect the accessibility of documents is left for future work.

Parameter Settings.

For the first experiment, we consider how sensitive $A(\mathbf{d})$ is to the choice of parameters involved in its calculation (i.e., we calibrate the measure). The accessibility scores of each document were calculated for the two measures (Cumulative and Gravity). For each measure, a number of parameter values were tested: $c = 20, 40, 60, 80$ for Cumulative and $\beta = 0.5, 1, 1.5, 2$ for Gravity[7] on each of the retrieval algorithms.

Given the set of single-term queries, $\mathbf{Q}$, and a particular set of parameters, we calculate $A(\mathbf{d})$ for each document in the collection. Figure 1 plots the accessibility values for the documents, after sorting them in ascending order, in the form of a grayscale map. White regions correspond to *cold documents*, i.e., documents with low $A(\mathbf{d})$ values. Conversely, black indicates documents which have high $A(\mathbf{d})$ values.

The immediate observation from the plot is the large proportion of white, indicating that a substantial portion of the collection has very low accessibility, irrespective of the retrieval algorithm used. This is obviously a function of the query set being used and our particular method for estimating accessibility. We however think that it is an interesting observation to note that a large number of documents (over a third) have $A(\mathbf{d}) \approx \mathbf{0}$, which indicates that they do not get retrieved in the result set of any single-term query.

Our objective for this experiment was to monitor the behavior of the $A(\mathbf{d})$ measure with respect to changes in the parameters of the calculation. For the cumulative measure, increasing c corresponds to a user reading a larger result set, so a larger part of collection is accessed for each query. The reducing amount of white in the grayscale heatmaps for increasing values of c indicates that $A(\mathbf{d})$ captures this intuition. Similarly for the gravity-based measure, an increase in β indicates a larger importance to being ranked high, i.e., documents lower down in the list of results are less likely to be accessed. Thus, increasing β leads to reduced accessibility for some documents in the collection (i.e., a decrease in the proportion of black in the maps).

We also provide in Table 1 the linear correlation coefficient between the $A(\mathbf{d})$ values calculated for different settings of the parameters. The high values indicate that while the specific values representing the accessibility of documents is sensitive to parameter settings (as they should be), the general trend across all the documents in the collection is stable. We wish to again highlight the fact that the general definition of $A(\mathbf{d})$ is designed to provide flexibility to adapt the measure depending on the scenario; thereby allowing it to be tailored to specific applications.

We wish to point out that the values of $A(\mathbf{d})$ reported in the experiments here are not only dependant on the choice of parameters and measure (Gravity / Cumulative) being used, but also vary according to

- the choice of the query set $\mathbf{Q}$
- the weight factor o_q

[7]Due to storage and computational restrictions, we also employed a rank cutoff of 100, when computing the gravity-based measure.

both of which are described in Equation 1. Alternate choices for these two factors (e.g. use of bi-term queries, giving a higher weight to popular queries, etc.) will lead to different numerical values of the $A(\mathbf{d})$ for each document. A separate set of experiments that calibrate our measures with respect to these choices can be performed.

Equality Objective - Universal Access.

What would happen if all documents in the collection have equal values of $A(\mathbf{d})$? i.e., every document in the index is equally accessible. This may not necessarily be a desired objective for a retrieval function, e.g. a scoring function that picks documents randomly will lead to all documents having roughly the same $A(\mathbf{d})$ values across a set of queries, but the effectiveness (i.e., precision/recall) of such a system would most probably be poor. Similarly, a web search engine administrator may wish to promote/favor particular documents or sets of documents within the collection (e.g. homepages over non-homepages, sponsored over not sponsored) or match the accessibility of the collection to the usage of the collection. However, there are other scenarios where such an objective would be valid, for instance within a library which aims to ensure impartiality or within a patent database where it is important that documents are accessible. Here, we use the objective of universal access as a common reference point with respect to which we compare retrieval algorithms.

We first evaluate the three standard retrieval functions in terms of the access they provide to individual documents across the collection. Clues are available in Figure 1 - BM25 and LM1000 make a larger proportion of the collection accessible (i.e., lesser white). In the current experiment, we concentrate on the $A(\mathbf{d})$ values calculated using the Cumulative measure with $c = 20$.

First, the documents are arranged in increasing order of their $A(\mathbf{d})$ values. A normalised version of $A(\mathbf{d})$ is calculated as

$$A_n(\mathbf{d}) = \frac{\mathbf{A}(\mathbf{d})}{\sum_{\mathbf{d}'} \mathbf{A}(\mathbf{d}')} \tag{3}$$

so that the $A_n(\mathbf{d})$ values sum to 1 across the collection. We then plot the cumulative $A_n(\mathbf{d})$ on the Y-axis with the X-axis representing increasing document numbers. If all documents had equal access, the cumulative plot would be a 45° line. As can be seen from Figure 2, TFIDF is the furthest away from the equality objective and BM25 appears the *fairest*.

Tuning a retrieval model by changing parameter settings will also have an impact on $A(\mathbf{d})$. Here we compare the accessibility of documents in the collection as dictated by different values of the Bayes smoothing parameter(μ) in the language model. We find that increasing μ marginally increases the bias of the retrieval function, in terms of making some subset of the collection less accessible. It would be interesting to further investigate how the change in accessibility in the collection affects the effectiveness of the retrieval systems; but this is left for future work.

The skew of the plots in Figure 2 provides a global view of the distribution of accessibility imposed by a particular retrieval function. We next take a closer look at how the algorithms differ, in terms of the $A(\mathbf{d})$ values they attribute to individual documents. We provide pair-wise comparisons of retrieval functions in the form scatter plots. In each,

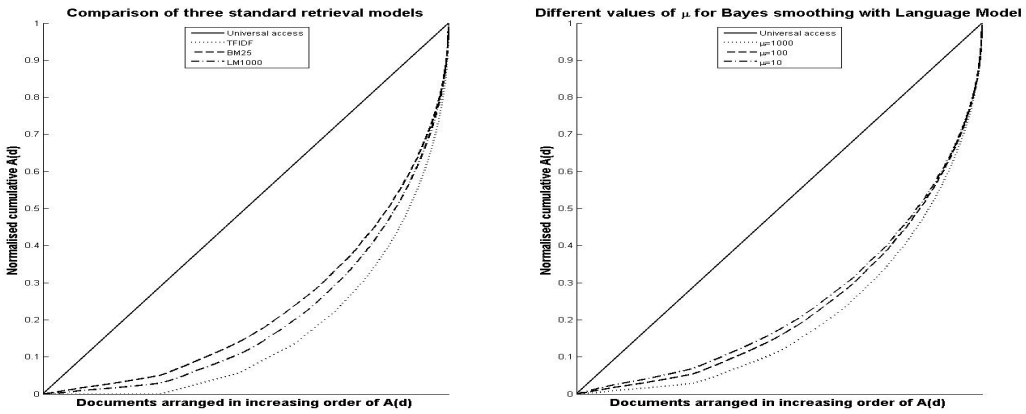


Figure 2: The cumulative normalised distribution of A(d). Left: Different retrieval models, TFIDF is most skewed indicating most accessibility bias. Right: Different levels of Bayes smoothing, higher values of μ lead to higher bias

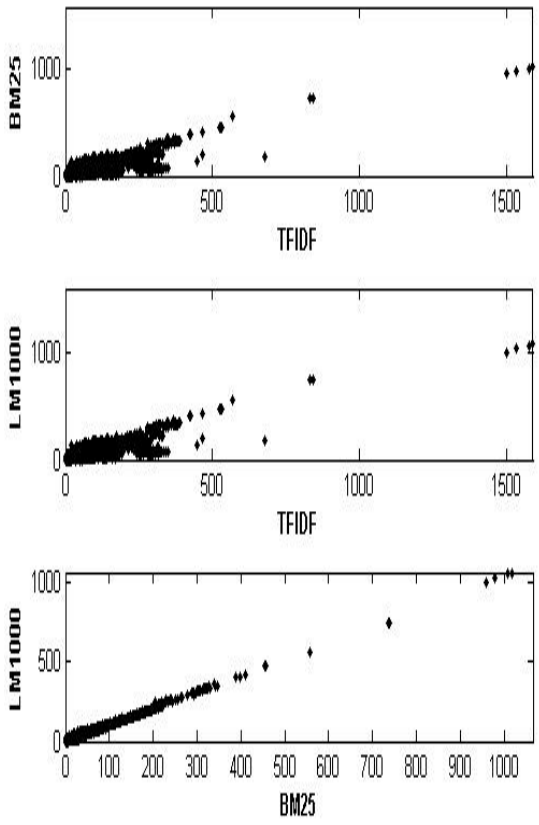


Figure 3: Pairwise comparisons between the three retrieval models (TFIDF, BM25, LM1000). LM1000 and BM25 provide a similar amount of access to each document, while compared to TFIDF this is not the case.

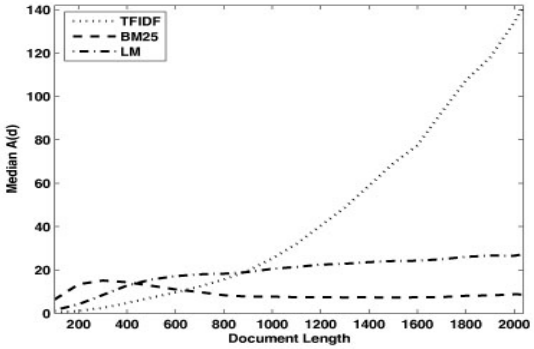


Figure 4: Median A(d) Vs Document Length. The Y-Axis shows how easily a document can be retrieved and the X-Axis denotes the length of the document. As document length increases, TFIDF disproportionately favors longer documents over shorter documents.

the X-axis corresponds to the (unnormalised) $A(\mathbf{d})$ value calculated using one retrieval function, and the Y-axis is the value given by an alternate retrieval function for the same document. If all points lie along the diagonal line, it indicates that the two retrieval systems agree on the level of accessibility they provide that specific document. We generate these plots for pair-wise comparisons of the three retrieval models (TFIDF, BM25, LM1000).

As can be seen from Figure 3, there is almost complete agreement between LM1000 and BM25, with TFIDF being somewhat different. All three algorithms agree on the most accessible documents. The scatter of points below the diagonal $y = x$ line when TFIDF is on the X-axis indicates that there are some documents that LM1000/BM25 think should be less accessible but these documents have larger $A(\mathbf{d})$ values when using TFIDF. Plots of pairwise comparisons between the three parameter settings for smoothing with LM are uninteresting, all three variations agree almost completely.

		APW	NYT	XIE	by Year
TFIDF	**1996**	-/-	-/-	8/1.8	8/1.8
	1997	-/-	-/-	8/1.7	8/1.7
	1998	33/6.4	81/15.0	7/1.5	24/4.8
	1999	37/6.6	72/13.1	7/1.5	28/5.4
	2000	42/7.8	68/12.4	7/1.4	25/4.9
	by Src	36/6.8	73/13.4	7/1.6	19/3.9
		APW	**NYT**	**XIE**	**by Year**
BM25	**1996**	-/-	-/-	63/12.3	63/11.1
	1997	-/-	-/-	63/12.1	63/12.3
	1998	75/13.9	84/9.6	65/12.2	64/12.1
	1999	72/13.2	80/9.1	64/12.1	58/12.0
	2000	74/13.7	77/8.93	63/12.0	58/12.0
	by Src	74/13.6	81/9.2	64/12.1	61/11.5
		APW	**NYT**	**XIE**	**by Year**
LM	1996	-/-	-/-	29/5.6	29/5.6
	1997	-/-	-/-	30/5.7	30/5.7
	1998	65/11.8	87/16.6	31/5.8	58/10.7
	1999	72/12.9	85/16.2	31/5.8	63/11.7
	2000	81/14.7	85/16.2	30/5.7	63/11.7
	by Src	71/12.8	85/16.3	30/5.7	54/10.1

Table 2: Median $A(\mathbf{d})$ of Cumulative $c = 100$ and Gravity $\beta = 0.5$ in each portion of the collection. For BM25, notice how the Cumulative scores for NYT are larger than the Cumulative scores for XIE, but the Gravity scores are smaller. That is, NYT documents are retrieved more often but XIE documents are ranked higher.

Document properties.

A known bias that the TFIDF algorithm suffers is with respect to length [12]. A sanity check for our $A(\mathbf{d})$ is to ensure that it reflects this known behaviour. Figure 4 shows the median of the cumulative measure for each algorithm given the length of the document (measured by the number of terms). For TFDIF (and to a substantially lesser extent, LM) as the length of the document increases the average number of times a document is retrieved also increases. For example, a document of length 2000 is on average 7 times more likely to be retrieved than a document of length 1000 (i.e., ≈ 140 divided by ≈ 20). On the other hand, BM25 tends to favor shorter documents, but overall is less biased with respect to document length. BM25 favors short documents initially, and then the length of the document begins to have a greater influence on the accessibility which slowly begins to increase. Consequently, we can see that BM25 is more robust to the problem of document length in comparison to the other algorithms considered.

Accessibility of documents in collection components.

Given the objective of universal access, we aim to determine whether the retrieval system provides such access to documents in the collection, or determine whether the systems have any inherent biases towards subsets of the collection. Here we look at divisions of the collection made in terms of date of publication (Year) and source of publication (Src).

In Table 2, the median $A(\mathbf{d})$ values are given for the Cumulative and Gravity measures for each algorithm by source and by year. If we consider each by year, then the accessibility BM25 provides is similar across years, whereas LM and TFIDF both favor documents from 1998-2000. If we consider each by source, then again the accessibility BM25 provides is similar across collections, but tends to favor NYT over APW and then XIE given the Cumulative measure, but using the Gravity measure XIE is favored over NYT. This suggests that while NYT documents are retrieved more often, XIE documents are ranked more highly. This result illustrates a key difference between the two types of measures. While TFIDF and LM both favor NYT documents, the extent of this preference is about 10 to 1 for TFIDF, whereas it is about 2.5 to 1 for LM (regardless of the type of measure).

This example illustrates how our measures can be used to quantify the accessibility of documents, and subsets of documents within the collection given a particular retrieval algorithm. By doing so, we can see the accessibility bias due to a retrieval algorithm given a collection. This provides system administrators with a tool to consider the influence of their algorithms on different populations of documents. For instance, it may be desired that more recent documents are favored in the collection, so the algorithms can be modified or tuned to reflect such objectives. Here, we can see that BM25 will provide more universal access to documents in the collection, whereas LM or TFIDF will tend to favor longer documents and NYT documents, whether this is desirable or not is entirely up to the administrators. However, other applications are possible, for instance, investigations into why particular subset of documents are being retrieved more/less often than other documents could be helpful in applications such as detecting spam (i.e., documents which are "too accessible" may be artificially inflating their $A(\mathbf{d})$) or hiding information (for example in patent databases to obtain "security through obscurity").

5. CONCLUSIONS

In this workshop paper, we have empirically explored the recently proposed *document accessibility* measures. These measures were designed to reflect the ease of a document be-

ing accessed through an information retrieval system. Where it was assumed that this property of the document is dependent on the retrieval function used and the relationship of document with the rest of the collection.

We have performed a number of experiments that compared three retrieval models (TFIDF, BM25, LM) on a standard IR dataset (TREC AQUAINT). We considered logical sub-divisions of the AQUAINT dataset, based on year and source, and measured the accessibility of documents in each subset. Experiments revealed that there is a substantial difference in the distribution of accessibility, with biases present towards some sources and years more than others. While these biases differed depending on which retrieval algorithm was being used, some general trends were observed. E.g. a preference for newer articles and that the XIE collection almost always consisted of the least accessible documents. We also considered the equality objective in order to provide a reference point with respect to which the bias in the accessibility across documents as imposed by the retrieval algorithm can be measured. In relation to this objective, we witnessed that TFIDF was the most biased while BM25 was the least. Subsequent analysis confirmed that TFIDF considerably favored the retrieval of longer documents, in comparison to BM25 or LM1000. However, building a retrieval function necessarily involves making some documents more accessible than others. This is because certain documents in the collection are more desirable than others by the users of the system. The accessibility of documents should ideally match how desirable the documents are (as opposed to making all documents equally accessible).

This paper has explored how document accessibility can be measured as a way to capture the access afforded to a document given a particular retrieval function. By measuring document accessibility, we have seen that it is possible to determine whether particular parts of the collection are favored over others. This suggests that it would be possible to devise methodologies which use these measures to investigate if there was any untoward bias within a retrieval system. But, while we have witnessed a disparity in accessibility among documents, the precise nature of the relationship between accessibility and effectiveness is as yet unknown, and provides an interesting direction for future work. Finally, while we have explored some operational aspects of using document accessibility measures, more work needs to performed in order to improve the estimation, approximation and calibration of the document accessibility measures.

6. ACKNOWLEDGMENTS

The first author would like to thank the Information Retrieval Facility (`www.ir-facility.org`) for the use of their computational resources.

7. REFERENCES

[1] L. Azzopardi and V. Vinay. Accessibility in information retrieval. In *To appear in the Proceedings of the European Conference in Information Retrieval (ECIR)*, 2008.

[2] P. Bailey, N. Craswell, and D. Hawking. Chart of darkness: Mapping a large intranet. Technical report, CSIRO Mathematical and Information Sciences, 2000.

[3] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, 2001.

[4] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins. The discoverability of the web. In *Proceedings of WWW '07*, pages 421–430, 2007.

[5] W. Hansen. How accessibility shape land use. *Journal of the American Institute of Planners*, 25(2):73–76, 1959.

[6] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the SIGIR '05*, pages 154–161, 2005.

[7] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400(6740):107–107, 1999.

[8] A. Mowshowitz and A. Kawaguchi. Assessing bias in search engines. *Inf. Process. Manage.*, 38(1):141–156, 2002.

[9] S. Pandey, K. Dhamdhere, and C. Olston. Wic: A general-purpose algorithm for monitoring web information sources. In *Proceedings of the 30th International Conference on Very Large Data Bases*, 2004.

[10] S. Pandit and C. Olston. Navigation-aided retrieval. In *Proceedings of WWW '07*, pages 391–400, 2007.

[11] V. Petricek, T. Escher, I. J. Cox, and H. Margetts. The web structure of e-government - developing a methodology for quantitative evaluation. In *Proceedings of WWW '06*, pages 669–678, 2006.

[12] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of SIGIR '96*, pages 21–29, 1996.

[13] T. Upstill, N. Craswell, and D. Hawking. Buying bestsellers online: A case study in search & searchability. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

[14] L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Inf. Process. Manage.*, 40(4):693–707, 2004.

Angle Seeking as a Scenario for Task-Based Evaluation of Information Access Technologies

E. Barker, J. Polifroni, M. Walker and R. Gaizauskas
Department of Computer Science
University of Sheffield
initial.surname@dcs.shef.ac.uk

ABSTRACT

In this paper we propose *angle seeking* as an appropriate task for the evaluation of information access technologies. We first describe angle seeking in the context of writing background to breaking news, analysing the types of information seeking activity it typically engenders, and then present a case study in which angle seeking forms the basis for a task-based evaluation in which a novel associative summary technology is compared with a conventional document retrieval engine. While neither technology is conclusively proved superior, this study both provides insights into these technologies and shows how a novel task-based evaluation can provide new information access technologies with a forum in which to establish themselves.

1. INTRODUCTION

Information seeking is typically not an end in itself, but rather occurs in some wider task setting. For example, information may be sought by someone writing a news report to deadline or carrying out a scientific research investigation or deciding what car to buy. The wider task may (1) require different sorts of information seeking activity (e.g. finding all relevant information, finding just one trustworthy source, developing a hypothesis, answering a factoid question) (2) impose production constraints on the information seeking (e.g. deadlines, form of output) and (3) typically be carried out by users with characteristic knowledge states (e.g. scientific investigations are carried out by those already expert in their field; news reports may be written on topics a reporter may know little about before beginning). Such diversity in tasks and in associated information seeking means "one-size-fits-all" information access tools, such as document retrieval engines, are unlikely to be optimal for every task with an information seeking component. It follows that designers of information access technologies should attend to the differing requirements that different task settings throw up for information seekers (as has long been recognised – [17, 7, 10]). One way to drive this process is to design evaluations of information access technologies that assess how well a tool assists a user in carrying out the wider task in whose service information seeking is undertaken. Doing so may help to liberate information access evaluation from the domination of a few standard evaluation measures, such as precision at rank 10, where relevance of retrieved documents to a query is all that is assessed, rather than the utility of a system for carrying out a task.

Various researchers have carried out evaluation of information access technologies in task settings, both simulated and real, e.g. [16, 8, 21] – see Section 3 below. In our view there is room for much more such work, until the implications of different task settings for information seeking are better understood.

One little explored task setting with significant requirements for information seeking is that of writing background to breaking news events, for example for a natural disaster, a political resignation, or company takeover. This task setting is one whose potential to inform the design of novel evaluations we have already explored [5, 12]. In brief the proposal in this earlier work was to assess the utility of different information access technologies by assessing the quality, as determined by task experts (professional journalists), of the written outputs of those using the technologies. That is, how good are the background pieces produced using information access technology A versus those produced using technology B? Experiments showed high intersubjective agreement between judges when they were asked to rank backgrounders written by different users on the same topic - i.e the task appears well-founded. However, there are various logistical difficulties in mounting an evaluation based on this task. In particular one needs a large pool of journalists prepared to write backgrounders on a range of topics, so that one can control for user and topic; one also needs sufficient qualified judges to assess the resulting background pieces. Since producing and assessing each background piece is a significant amount of work, mustering resources to carry out such an evaluation is not easy. Furthermore, since the resulting information artefacts are so rich (full texts) and the steps taken to produce one are so numerous (including, e.g. all the information seeking that may have contributed to the writer's understanding but did not yield any content that found its way into the final product), this task setting makes it difficult to gain understanding into which aspects of a system's behaviour may have contributed positively or negatively to the overall result.

To address these difficulties with the background writing task while retaining its advantages as an evaluation scenario

NMEIR Workshop at ECIR 2008 , March 30, 2008, Glasgow, UK.

– a real task setting with a strong information seeking component – we have focussed on one central but limited aspect of the task: *angle seeking*. Angles, discussed in further detail in the next section, are unifying ideas or overarching propositions which frame or position the information reported in the rest of a text. In news articles they are typically conveyed in the headline or lead sentence. Angle seeking is a key, early step in writing a news article, one which can require extensive information seeking but results in a concise output – usually a proposition expressed in a single sentence. As such, angle seeking is an appealing task for task-based evaluation of information access technologies.

To explore the utility and feasibility of angle seeking as a scenario for task-based evaluation of information access technologies we have made two contributions, which we report in this paper: (1) an analysis of the task, what the task is, the information seeking strategies that may be involved, and why is it an interesting challenge task for information access systems (section2); and (2), the design and execution of an evaluation using angle seeking as the task in order to assess two information access systems – a novel association-based approach and a conventional document retrieval engine. We report this work in the rest of the paper. In section 3 we discuss related work on task based evaluations for IA technologies. Section 4 describes the experiment we have carried out based on an angle seeking scenario, including details of the experimental design, the systems compared and the results of the evaluation. In the final section we draw conclusions about the utility of angle seeking as a scenario for evaluation of IA technologies.

2. ANGLES AND ANGLE SEEKING IN NEWS WRITING

The term "angle" may be used to describe both an information artefact and the activity or process that people carry out in producing such an artefact. The OED, reflecting these two uses, describes an angle as a noun: "a position from which something is viewed or along which it travels or acts", and as a verb: "to present information to reflect a particular view or have a particular focus". The term has currency in a number of domains, such as writing and politics, but it has particular significance for journalists researching and writing background for breaking news stories.

A news wire "backgrounder" is an extended prose piece, of around 500 words, sometimes referred to as a sidebar, which is produced when a news editor deems a particular story worthy of dedicated background material. The function of a backgrounder is not to continue to report details of new events, but rather to provide text that supports and contextualises these events. Speed is essential in the production of news wire content. Yet a backgrounder may appear some time after the early instalments of a story have been published on the wire, since the news room requires details of the breaking news to determine whether the story merits a background piece. Furthermore, research must be carried out, typically against a news archive, so that the journalist has a topic of interest to write about.

Developing a newsworthy "angle" is a key goal in the background research and writing scenario. While a precise definition is not something which is easily articulated, journalists have an intuitive understanding of what an angle is. Interviews with journalists and an analysis of a collection of background news wire texts suggest that we can see an angle as a unifying idea, an organizing construct, which links together information such that it might be used to frame the current event in a narrative text that is both coherent and compelling to an audience. We can find intuitive examples of angles expressed in the headline and the opening statements of a background piece, which journalists refer to as the "lead". Together, these lines provide a summary of what the backgrounder is about. Figure 1 shows two backgrounders for the same news event – Clare Short's resignation from the British Cabinet in 2003 – and illustrates how the angle taken in a backgrounder can profoundly affect the interpretation of a foreground event. In the first piece the angle taken is that the resignation is a consequence of Clare Short's character and the piece goes on to supply details of Short's colourful career. In the second, the angle is that Short's resignation is the continuation of a trend of resignations and sackings that have characterised Blair's government.

12/05/03: Clare Short resigns from Tony Blair's cabinet.

Background 1:

'SERIAL RESIGNER' WHO LED A CHARMED LIFE
The surprising thing about firebrand Clare Short's resignation is that her departure from the Cabinet did not happen much earlier.
Ms Short seems to have lived a charmed life as Secretary for International Development, first by describing the Prime Minister as "reckless" and then by missing a key vote last week on the contentious issue of foundation hospitals.
It looked as though she was almost begging to be sacked.
Those who have watched her progress are still astonished that such a volatile person ... has lasted for so long in the top echelons of Government.
Her reputation as what someone once described as "a serial resigner" was made when she served under Neil Kinnock as Leader of the Opposition ...

Background 2:

BLAIR'S CABINET CASUALTIES
Since sweeping to power in 1997, Tony Blair has had to deal with a string of high-profile resignations from his cabinet - and has felt obliged to remove several other senior ministers himself.
The first to quit following Labour's 1997 landslide triumph was Welsh Secretary Ron Davies, who stepped down after a "moment of madness" ...
Social Security Secretary Harriet Harman and her second-in-command Frank Field were both victims of Tony Blair's first major reshuffle - after apparently falling out ...
Peter Mandelson made history when he became the first Secretary of State to resign twice ...

Source: PA News Archive

Figure 1: Two Backgrounders for the same Event

Attfield and Dowell [3] present a model of journalistic information seeking in the context of the task of writing a news story. While not specifically concerned with the scenario of background news writing, their model provides some insights into how angles are sought and developed and the role that they play in the broader context of a news writing task. Given a news topic assignment by a news editor and a set of product and resource constraints, the three stages in the Attwood and Dowell model are:

1. *Initiation* A provisional angle is established and a deadline and word count constraints are determined. (This usually takes place during the initial assignment brief).

2. *Preparation* The angle is tested and either confirmed or refuted. Potential content is gathered, personal understanding is developed and a plan for the report is evolved. During this stage an assignment-specific collection of materials, paper or electronic, is assembled for later use.

3. *Production* The story is written, consulting the assignment collection, based on the understanding and plan developed so far. The writing process may provoke further information seeking and alteration of the plan.

The notion of an angle is central to their model. It is described by them elsewhere [2] as a "proposition, or central factual claim that is to be made by the report. Where the claim involves some speculation the angle takes the form of a working hypothesis or conjecture" and again as the "clearly focused perspective or guiding idea which determines both a solution's space and the writer's information requirements".

This is a compelling account. However, Attfield and Dowell stop short of pursuing in depth the process by which journalists iteratively gather potential content and refine their understanding of a topic. Based on observations of and interviews with journalists engaged in background seeking and a preliminary analysis of a corpus of information seeking dialogues between journalists where one was seeking background and the other providing it [6], we can elaborate on the processes described in the Attfield and Dowell model:

1. *Initiation* When journalists are seeking background information for a breaking news story, they may not always be provided with an angle. Often their job is to discover and establish angles for the story. They often begin the research process by formulating an idea of a topic, or perspective which they want to explore. This is typically derived from the details of the news story and their background knowledge. It may be as simple as a general topic area, e.g. "hurricanes", or more elaborate, e.g. "despite years of worsening weather, this is the worst storm since 1987".

2. *Preparation* The journalist tests and/or refines the provisional angle. Here the journalist is looking for patterns in the data, such as trends or interesting associations, which in his judgement will be sufficient to form the basis for a compelling background to the news story. Our research suggests that journalists have an expert understanding of the kind of information that needs to be examined in order to develop and support an angle and that they may engage in a number of strategies for finding patterns. We note that these are similar to the strategies Collins and Gentner [11] propose for developing and maniuplating ideas in their prescriptive model of the writing process:

   (a) *Collecting similar events* For example, finding other people who have left a Cabinet Office.

   (b) *Comparison* Comparing the current event with (1) a similar event or (2) a group of similar events (e.g. where does this fit on the scale of things?) – i.e. establishing differences or similarities.

   (c) *Viewing and sorting similar events by different attributes* E.g. arranging examples of protests at pay increases in chronological order; grouping earthquakes by their location; ordering hurricanes by windspeed, in the 5 categories of hurricane.

   (d) *Aggregating over similar events* E.g. numbers of caving accidents in a location; how many of these resulted in serious injuries or deaths.

   (e) *Aggregating over attributes* E.g. total numbers of fatalities in earthquakes in Asia in the past fifty years.

   (f) *Finding extreme similar instances* Based on different attributes, e.g. the earthquake to have killed the largest number of people; the most grisly kind of death etc.

   (g) *Newsworthy similar instances* Similar to (f), finding similar events with a newsworthy characteristic, for example "any funded science projects which have been associated with animal rights activity".

When the journalist is satisfied with the angle, he typically selects content from the materials he has examined in order to support and elaborate on the angle in the written background piece (stage 3 in the Attfield and Dowell model).

3. RELATED WORK: TASK-BASED EVALUATIONS FOR INFORMATION ACCESS

For more than a decade there has been growing interest in task-based user evaluations of information access systems.

One line of such work has concentrated on studying the effect that priming a subject with a task context has on the retrieval of relevant documents from a document collection, e.g. [8, 15]. Hansen and Karlgren [15], for example, consider the effect that a work-task scenario description may have on a reader's assessment of the relevance of documents retrieved in a non-native language they know well. While these sorts of study can yield insights into document retrieval technologies, they cannot, given their focus on document retrieval, give insights into the utility of other information access technologies for tasks that could potentially benefit from them

In contrast to this work, and perhaps less well explored, is work on evaluations in which the assessment has focused on measuring the outcomes of system use. Here the emphasis has been on evaluating information access systems indirectly, assessing how well systems have enabled the user to carry out some wider task, such as: answering a clinical question [16], writing a report [21], revealing the topic structure of an archive [22], etc. Apart from providing valuable insights into the benefits systems may bring to tasks, this approach is notable in that it allows for a comparison of systems which have different outputs, e.g. a list of document headlines vs. summaries of document clusters.

We note the task scenario used in McKeown's work [21] is in the same domain as the angle seeking task we describe in this paper. The authors asked subjects to help write reports for an issue in the news e.g. Hurricane Ivan's effects. Key differences are that they described this as a "fact gathering" scenario, where users answer three related questions about an issue in the news. So, a pre-specified topic guides information seeking and as such there is less emphasis on discovery and analysis for the written result, which is in contrast to what we have observed for the angle seeking task. Other task-based evaluations where the user task

shares some characteristics with those of angle seeking for background news task include Baldonado and Winograd [4] who used the wider task of writing a term paper for a graduate seminar (on either cryptography or neural networks) to focus a comparative evaluation for two variations of the Sensemaker information-exploration interface. They asked users to determine the specific topics and then to write down the titles of one or two promising references. However, the evaluation did not include a measure of the task outcome, focussing instead on the character of the interactions in the different conditions and on user satisfaction. There has been a notable line of work on developing IR applications to support the task of generating and testing hypotheses founded in literature collections, e.g. [24]. But to date, and to the best of our knowledge, evaluations have been restricted to demonstrating by critical example as opposed to more systematic evaluations involving multiple users carrying out multiple tasks in different system conditions.

4. AN ANGLE SEEKING EVALUATION

The information seeking activities typical of angle seeking, identified above in Section 2, suggest that a large range of possible information access systems could be applied to the angle seeking task. Document retrieval, similar event searching, topic tracking technology, overview technologies (e.g. scattergather), association mining techniques could all potentially be of help. Furthermore in current practice journalists are limited to document retrieval systems, but express considerable dissatisfaction with this technology for the task. Therefore there is a strong motivation to investigate the benefits which alternate approaches might bring to the task and for an evaluation which allows potential benefits to be assessed.

Since different information access technologies may differ in their objectives and outputs, in the role of the system in application setup, in the type user interactions, and so on, directly comparing the outputs of such technologies may not be feasible. This is one of the strong arguments mentioned above for devising an extrinsic evaluation.

To do this we proceed as follows: (1) identify a task output; (2) gather task outputs as produced by users who employ different information access technologies; (3) get experts to evaluate the "goodness" of the task outputs. This approach is based on the assumption that if two setups A and B, in which humans work with an information system to complete some task, differ only in their embedded information systems S_A and S_B, and A outperforms B according to some evaluation criteria, then S_A is more positively evaluated that S_B.

For the angle seeking task, we propose a setup consisting of a journalist together with an information access system and a text information source, or digital archive. Input to the setup is a breaking news story. The subject is asked to read this story and use the information resources to find as many angles for a background piece to the new event as possible within 15 minutes. The output is a list of angles and for each a list of documents which support the angle.

For this task we can identify various possible evaluation criteria: user satisfaction, effort, quality of output from the setup (the angle plus supporting content), and time to complete. To carry out an evaluation we must operationalise these criteria as measures. For example, user satisfaction could be measured by a post task questionnaire; effort by the number and type (productive or non-productive) of user interactions with the system, quality by experts' judgements on the angles plus supporting documents found by users. In the case study reported below we used two evaluation criteria only: (1) subjects' perception of the utility of each interface as a mechanism for searching for background information; and (2) the quality of the information provided by each interface.

In the rest of this section we describe a case study in using angle seeking as a scenario for evaluation two information access technologies. We first provide some details of the technologies, describe the design of the experimental setup in more detail and then present results.

4.1 Technologies Compared

A new technology that might be suitable for the task of seeking angles for breaking news events is what we refer to as "associative summaries", an approach that takes semantically annotated documents that are topically related to the breaking news event, looks for strong associations in the annotations, and then presents these associations as indexes to document clusters. The intuition here is that these summaries will give the user an idea of what content is available in the archive and of patterns in the data. Our hypothesis is that, given that angle seeking is a task that frequently requires a new event to be seen as the continuation of a pattern or trend, then a technology that actively discovers patterns in the data in areas topically related to the new event will be of more benefit than one which leaves the user, who may know little about either the topic or the archive content, to drive the information seeking process himself. In the evaluation below we compared associative summaries with a conventional document retrieval system, as a baseline, using the angle seeking task as an evaluation scenario.

4.1.1 Associative Summaries for Information Access

The associative summary technique may be summarised as follows (for full details see [23]). First it is assumed that an archive has been semantically annotated for entity types such as *person*, *location*, *date*, *organization* and so on and for *keyphrases* where the latter are single or multiwords terms that are indicative of document content (a variety of techniques exist for identifying these, such as [25]). For the experiment reported here a subset of these entity types was selected, consisting of just *person*, *location* and *keyphrase*.

The technique is applied to a topically coherent subset of documents from the archive. This subset, called the *topic set*, is assembled using a query to a search engine running over the archive (e.g. "China AND polution" – in the experiment one query was selected for each breaking news story for which subjects had to find angles). From the lead segment of each document in this topic set a fixed number of most frequently occurring instances of each of the nominated entity types is identified – in the current case the ten most frequent persons, locations and keyphrases. For each document in the topic set a binary vector representation of length 30 is then created, one position for each of the 30 frequently occurring entities, a 1 in any position in the vector indicating that there is a mention of this entity in this document.

The vector representations of the topic set are input to a clustering algorithm, in this case a modified version of Predictive Apriori with bottom-up agglomerative clustering [1]. The resulting clusters, representing potentially signif-

icant associations, are presented to users using one of two interfaces. The first interface (called the *Full Associations* interface below) shows associations grouped according to the entity types found in the associations. So, for example, all associations involving say keyphrases and locations – for instance "river Russia spill" in the "China AND pollution" topic set – are shown together, as are all associations involving just persons, and so on. Selecting any association takes the user to a page listing the titles of all documents in the archive (not just in the topic set) containing occurrences of the terms in the association (in our example, all documents containing occurrences of "river", "Russia" "spill" "China" and "pollution"). The second interface (*Combined Associations*) simply shows all associations, without grouping them by the types of entities found within them. Again selecting any association leads the user to page listing titles of all documents in which the association is instantiated and links to the full documents.

4.1.2 Baseline Document Retrieval System

The baseline system was the document retrieval system within Ontotext's KIM semantic annotation platform [19], itself built on the Lucence open source search engine library [1], an implemention of the vector space model. For the baseline interface, users constructed search terms themselves for the breaking news story and typed these directly into an interface to Ontotext's search facility. Rather than use the interface provided by Ontotext, a separate page was designed that preserves the look-and-feel of the other two interfaces.

4.1.3 Data Resources

Ontotext Corporation provides an interface to roughly 500,000 news articles from sources such as Reuters, the PA, ABC News, the BBC, and CNN. Each document has been automatically annotated for keyphrases and named entities using the KIM platform. For the experiments described here, Ontotext provided a Java applet that enabled us to query the archive by key term and receive a set of semantically annotated documents in XML format in return.

4.2 Experimental Design

We recruited a total of 18 subjects on the basis of their experience in news writing. Participants included sixteen MSc graduate students in the Department of Journalism Studies, University of Sheffield, and two professional journalists working for the Sheffield *Star*. We asked each participant to read a breaking news story and then, using one of the three interfaces to the Ontotext news archive described above, to find angles that might help in the preparation of the best possible background to the story. We set a 15 minute time limit for the task and asked subjects to find as many good angles as possible within the allotted time. When satisfied with an angle participants were to write down the angle (e.g., "Previous chemical spill in river in China") and to save any documents which supported the angle.

To help them carry out this task, we provided a short scenario which asked a participant to imagine him/herself as a reporter working for an international newswire agency and that the news editor had called for a 500 word background report for the wire to support a breaking story. Each subject carried out three tasks, each on one of three topics, real news stories chosen from AP newswire via Google, from within two weeks of the date of the start of the experiment. Of the three breaking news stories, one was about riots in France following the election of Nicolas Sarkozy as President, one was about a threatened lawsuit by the European Union against Microsoft, and one about new Chinese government measures to address pollution. These topics contain a range of event types/entities: one focussed on a person (e.g. Nicolas Sarkozy), one focussed on an organization and a political entity (e.g. Microsoft and the EU), and one focussed on a country and a keyword (e.g. China and pollution).

Each subject completed three tasks by interacting with each of three interfaces in turn, in a within-subject design.

We varied the interface order across subjects in order to assess the effects of the interface on user behavior and experimental judgment. Across the 18 subjects, each interface was used six times as the first, second, or third interface, respectively. To mitigate the confounding effects of story type on subjects' perception of the interface, we did not also vary story type. Each subject completed the Nicolas Sarkozy task first, the EU/Microsoft task second, and the China task third.

Subjects were given a sample breaking news story as a "warm-up". The three interfaces used in the experiment had been configured for the warm-up story, and subjects were given as much time as they wanted to work through the warm-up task while familiarizing themselves with the interfaces. Experimenters were present to answer questions at this point.

After completing the warmup, subjects returned to the main experimental page, where they were asked to indicate, in general, how familiar they were with each of the topics used in the experiment, rating their familiarity on a Likert scale from 1 to 5, with 5 being "Very familiar".

Subjects then carried out the experiment with a fifteen minute time constraint per task. After finishing each task, subjects were asked to answer two questions about each interface, using a 5-point Likert scale:

- How confident are you that you were able to fully explore the contents of the corpus? (with '1' indicating *Not confident* and '5' indicating *Very confident*)
- Would you use such a system again? (with '1' indicating *Not likely* and '5' indicating *Very likely*)

User input on the first of these is analyzed as the *confidence* metric in Section 4.3; the second as the *reuse* metric. After completing all three tasks and seeing all three interfaces, subjects ranked each interface by its usefulness, again on a Likert scale from 1 to 5 ('1' being *Not useful* and '5' being *Very useful*). This is called the *rank* metric in Section 4.3.

Finally, users were asked to tell us what they liked best and least about each interface, using a free-form text box. This last set of questions was optional, but all subjects except one provided feedback here.

4.3 Results and Analysis

4.3.1 User judgments/input

Overall, the two cluster-based interfaces were ranked as top-choice by our subjects 56% of the time and as either top or equivalent to the *Baseline* 67% of the time.

The average rank users assigned to each of the interfaces is shown in the second column of Table 1. Overall, the

[1] lucene.apache.org

Interface	Rank (average)	Confidence (average)	Use again (average)
Full	3.11	3.11	3.06
Combined	2.94	3.17	3.0
Baseline	3.28	3.33	3.7

Table 1: The scores users assigned to each interface, for overall rank, confidence, and reuse.

Topic/ order	Rank (avg)	Confidence (avg)	Reuse (avg)	Fam. (avg)
Sarkozy	3.06	2.94	3.22	2.722
Microsoft	2.83	3.06	2.83	2.22
China	3.44	3.61	3.67	2.33

Table 2: The scores subjects assigned each story for overall rank, confidence, reuse, and familiarity. The order in the table reflects the order in which the subjects saw each story topic.

highest ranking interface was the *Baseline* system. Preference for the *Baseline* was not significant, however, compared with the *Full Associations* interface, based on paired *t*-tests and a multivariate analysis of variance (MANOVA; Wilks' Lambda, $F(2,16) = .423$, $p = .662$). This lack of significant difference indicates that subjects had no strong preferences among the three interfaces.

Users' confidence in the usefulness of each interface for exploring the archive was also not significantly different in paired *t*-tests and a MANOVA (Wilks' Lambda, $F(2,16) = .242$, $p = .788$). The third column of Table 1 shows these scores. For the *reuse* metric, reflecting users' response to the question about using each particular interface again, averages are shown in the fourth column. As with rank and confidence, paired *t*-tests and a MANOVA showed no significant differences (Wilks' Lambda, $F(2,16)=1.8$, $p = .198$).

We next performed a series of ANOVAs using each of the subjective measures elicited from users as the dependent variable, and type (i.e., *Full Associations*, *Combined Associations*, or *Baseline*) and story topic (i.e., Sarkozy, Microsoft, or China) as independent variables. Table 2 shows these subjective measures as they correspond to story topic. We did not find significant effects or interactions with the independent variable *rank*. However, we found a marginal effect of story topic on confidence ($p = .098$, $F = 2.24$, $df = 2$), although no interaction effects. We also found a slightly stronger, though still marginal effect of story topic on reuse ($p = .066$, $F = 3.4$, $df = 2$), again with no interaction effects.

Recall that, because stories were always presented to users in the same order, story topic is a proxy for order in our analysis. Although there was a marginal effect of story topic on confidence, there was a significant correlation between users' confidence in the systems and story topic (i.e., order). Users' confidence increased monotonically over the course of the experiment, regardless of the order of the interface (see Table 2). The effect of order on user judgment has been seen elsewhere in search-based tasks [9], although with a much smaller subject population. For our subjects, confidence grew as they progressed through the experiment. This suggests that a longitudinal study using these interfaces might yield interesting results.

One possible explanation for the effects shown by story topic is the users familiarity with the story itself. The final column in Table 2 shows familiarity scores by topic, which were not significantly different. Familiarity had a marginal effect on confidence (ANOVA, $p = .06$, $F = 2.9$, $df = 3$), but no effect on rank or reuse. The topic users expressed the greatest familiarity with *a priori*, Nicolas Sarkozy, was also the one that had the lowest confidence scores. The interface to stories about China, about which users had expressed a lower degree of familiarity, had the highest confidence scores.

The lack of a significant difference in any of our subjective measures matches what has been found elsewhere in the literature [14, 18] when comparing interfaces that process data to a Google-like baseline. Simple keyword search interfaces are well-known and frequently used tools, and it is not easy in an hour-long experiment to show superior benefits from a new interface. The fact that two associative summary interfaces were preferred more than half the time is a positive indicator of the utility of associative summaries. Users' confidence grew as they progressed through the experiment, even when they were using the associative summaries in later stages. This indicates that all interfaces met subjects' information-seeking needs to some degree.

4.3.2 *Expert judgment*

In addition to the ratings we elicited from subjects, we also asked a Senior Lecturer in the Department of Journalism Studies at the University of Sheffield, who teaches on the the topic of angles in news stories, to serve as an expert judge on subject output. This expert judge was presented with 54 separate "packages" of documents, one for each of the stories (3) used by each of the subjects (18) to complete their tasks. Each package consisted of a set of angles, followed by the stories the subject found to support each angle [2].

The expert judge read the breaking news story for each topic/interface and answered three questions about each package of angles and background stories. Answers to the questions, listed below, were on a Likert scale of 1 to 5, with 1 indicating a negative opinion:

- How would you rank this package for its usefulness in building a background for the breaking news story?
- How would you rank this package for richness/comprehensiveness of background?
- How would you rank this package for originality/unexpectedness (i.e., does it contain something that is both novel and helps contextualize the event)?

Each package was examined blindly, i.e., the expert had no idea who created the package or what interface was used.

[2]A concern in this part of the experimental protocol was that it was not possible to elicit judgments from more than one expert, given the level and specificity of expertise needed to rate background material, and the time required to examine 54 sets of background angles and supporting documents. Because this is the first time, to our knowledge, that this technology has been both used and evaluated by experts in the same field, we felt that one set of judgments here would contribute to an understanding of the usefulness of the technology, while helping refine an evaluation protocol for use in follow-up experiments.

Interface	Usefulness	Richness	Originality
Full	2.67	2.22	2.06
Combined	2.72	2.61	2.28
Baseline	3.22	3.28	2.83

Table 3: Rankings from the expert judge for each interface, on usefulness, richness, and originality.

Topic	Usefulness	Richness	Originality
Sarkozy	2.83	2.22	2.11
Microsoft	2.78	2.78	2.5
China	3.0	3.11	2.56

Table 4: Rankings from the expert judge for each topic, on usefulness, richness, and originality.

High-achieving subjects			
Interface	Avg. rank	Conf.	Reuse
Full	3.50	3.50	3.50
Combined	3.00	2.75	3.00
Baseline	2.00	3.00	2.50
Low-achieving subjects			
Interface	Avg. rank	Conf.	Reuse
Full	3.25	3.00	3.25
Combined	2.75	3.50	3.50
Baseline	4.00	3.75	4.00

Table 5: Rankings from users, shown by groups, reflecting those whose angles were ranked highly by the expert judge and those that were ranked poorly.

The results of the expert judgments were used to associate measures of *usefulness*, *richness*, and *originality* (corresponding, respectively, to the questions above) with the other experimental variables. Table 3 shows the expert measures as a function of the interface used for the package. Table 4 shows the expert ratings for usefulness, richness, and originality, by topic.

Examined by interface, the baseline performs best along all dimensions. The differences are not significant, however, for usefulness or originality, although a MANOVA indicates significant differences in richness (Wilks' Lambda, $F(2,16) = 4.6$, $p < .05$). Paired t-tests showed a significant difference between the richness scores for *Full Associations* vs. *Baseline* interfaces ($p = .006$, $df = 17$) and for *Combined Associations* vs. *Baseline* interfaces ($p = .048$, $df = 17$).

Scores associated with topic also show significant differences only for richness measured by a MANOVA (Wilks' Lambda, $F(2,16) = 3.8$, $p < .05$). Paired t-tests showed significant differences in richness between Sarkozy angles and China angles ($p < .05$, $df = 17$).

The expert judge used in this experiment was able to provide insight into the interaction between topic type and richness, in an interview conducted after the judgments were elicited. He hypothesized that the topic with the highest scores along all dimensions, "China and pollution", lent itself naturally to the type of background information that he would score highly for richness. He further hypothesized that angles found for the other two topics would be, by their nature, not as interesting from his perspective.

After rating each of the packets, we asked the expert to go back through the angles found for each of the three topics and flag the angle+story combinations that he thought were most interesting. Not surprisingly, he found none that he felt were outstanding along this dimension for either the Sarkozy or the Microsoft story. However, he did find two for the China story, both from the same subject. These two stories were both found using the *Full Associations* interface, and, furthermore, were found by a subject that rated that interface the highest for usefulness. The fact that the only angles felt to be truly outstanding by the expert were found by the same subject, suggested that individual subject performance might be an interesting dimension to investigate.

4.3.3 Examining subjects by performance

Previous research has examined the effects of issues such as personality [13], and experience [20] on users' acceptance of a variety of technologies related to information presentation. Here, we investigate correlations between measures from the expert judge, which evaluate subjects' abilities to identify background for stories, and the preferences expressed by users.

Out of 18 subjects, four fell into a group that scored cumulatively highest on the measures of usefulness, richness, and originality. Those four subjects are classified as the *high achieving* set. Four subjects fell into a group that scored cumulatively lowest on these same three measures and those four subjects are classified as the *low achieving* set. Table 5 shows user-elicited scores for each interface, divided by the high-achieving and low-achieving subjects.

High-achieving subjects ranked the two associative clustering interfaces the highest. The differences here are significant between the rank of the *Full Associations* and the *Baseline* interfaces ($p < .05$, $df = 3$), with the baseline scoring significantly lower. Low-scoring subjects tended to prefer the baseline (although not significantly). Although the dataset is small, and we have only the opinion of one expert judge, this result seems to indicate that the associative summaries technology was able to be used effectively by high-achievers at the task.

5. CONCLUSION

We have introduced the task of angle seeking in the background news domain and shown it to be a rich task with much potential for focussing new applications and evaluations of information access technologies. We also presented an angle seeking evaluation which incorporates an expert's assessment of task outcome. Moreover, we have demonstrated this evaluation in an experiment which compared users' performance on the task in three different information access system setups, two using variants of a novel "associative summary" technology based on finding associations in semantically annotated text and a third using a conventional IR search engine. While the results were inconclusive, in so far as they were not able to establish whether the new technology was more effective in this task setting, they provide important insights into the relative strengths and weaknesses of the new and the conventional information access technologies. In particular by showing that the new technology was preferred by users who were good at the task, the evaluation has helped to establish the potential utility of a new technology, validating the observation we made at the outset that appropriate design of evaluations can help

advance technologies for information access.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the UK EPSRC grant GR/R91465/01. They would also like to thank sincerely members of the University of Sheffield Department of Journalism for their enthusiastic participation in many aspects of this work, most especially Jonathan Foster, Bob Bennett and David Holmes. Finally we would like to acknowledge the UK Press Association for access to their archive and for expert advice and comment and Ontotext for access to their system and news archive.

7. REFERENCES

[1] J. Alipio. Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proc., 2004 SIAM Int'l. Conference on Data Mining*, 2004.

[2] S. Attfield, A. Blandford, and J. Dowell. Information seeking in the context of writing: a design psychology interpretation of the 'problematic situation'. *Journal of Documentation*, 59(4):430–453, 2003.

[3] S. Attfield and J. Dowell. Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2):187–204, 2003.

[4] M. Q. W. Baldonado and T. Winograd. SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'97*, pages 11–18, Atlanta, Ga., 1997. ACM Press, New York.

[5] E. Barker and R. Gaizauskas. Evaluating Cub Reporter: proposals for extrinsic evaluation of journalists using language technologies to access a news archive in research. In A. Bailey, I. Ruthven, and L. Azzopardi, editors, *Proceedings of the Workshop on Evaluating User Studies in Information Access, 5th International Conference on Conceptions of Library and Information SCience,(COLIS 2005)*, 2005.

[6] E. Barker, R. Higashinaka, F. Mairesse, R. Gaizauskas, M. Walker, and J. Foster. Simulating Cub Reporter dialogues: The collection of naturalistic human-human dialogues for information access to text archives. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006.

[7] N. J. Belkin, P. G. Marchetti, and C. Cool. Braque: design of an interface to support user interaction in information retrieval. *Inf. Process. Manage.*, 29(3):325–344, 1993.

[8] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[9] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.

[10] K. Bystrom and K. Jarvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, March-April 1995.

[11] A. Collins and D. Gentner. A framework for a cognitive theory of writing. In L. W. Gregg and E. Steinberg, editors, *Cognitive processes in writing: An interdisciplinary approach*, pages 51–72. Lawrence Erlbaum Associates, 1980.

[12] R. Gaizauskas and E. J. Barker. Mice from a mountain: Reflections on current issues in evaluation of written language technology. In J.Tait, ed., *Charting a New Course: Natural Language Processing and Information Retrieval*, pages 195–238. Springer, 2005.

[13] D. Goren-Bar, I. Graziola, F. Pianesi, and M. Zancanaro. The influence of personality factors on visitor attitudes towards adaptivity dimensions for mobile museum guides. *User Modeling and User-Adapted Interaction*, 16(1):31–62, 2006.

[14] C. Gutwin, G. Paynter, I. Witten, C. NevillManning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. Technical report, Dept. of Computer Science, University of Saskatchewan, 1998.

[15] P. Hansen and J. Karlgren. Effects of foreign and task scenario on relevance assessment. *Journal of Documentation*, 61(3):623–639, 2005.

[16] W. Hersh, J. Pentecost, and D. Hickam. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, 47(1):50–56, 1996.

[17] P. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, 1992.

[18] H. Joho, M. Sanderson, and M. Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In *Proc., 26th European Conf. on Information Retrieval*, pages 42–56, 2004.

[19] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[20] S. K. Lippert and H. Forman. Utilization of information technology: Examining cognitive and experiential factors of post-adoption behavior. *IEEE Trans. on Engineering Management*, 52(3), 2005.

[21] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2005. ACM.

[22] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220, New York, NY, USA, 1996. ACM.

[23] J. Polifroni. *Enabling Browsing in Interactive Systems*. PhD thesis, University of Sheffield, Department of Computer Science, January 2008.

[24] M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, and R. Vos. Using concepts in literature-based discovery: simulating swanson's raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.*, 52(7):548–557, 2001.

[25] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. In *ACM DL*, pages 254–255, 1999.