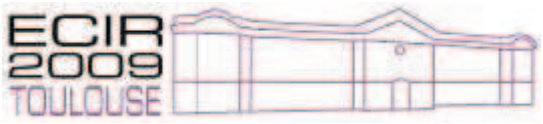

ECIR 2009 Workshop

Contextual Information Access, Seeking and Retrieval Evaluation

held in conjunction with the 31th European Conference on Information Retrieval International
in cooperation with the BCS-IRSG and is supported by ACM-SIGIR, ARIA and E-IRSG

6 April 2009, Toulouse

Organizers

Bich-Liên Doan
Supélec, France
Joemon Jose
University of Glasgow, UK
Massimo Melucci
University of Padua, Italy
Lynda Tamine-Lechani
University of Toulouse IRIT, France

Table of contents

Suspension of Disbelief in Interactive Information Retrieval Evaluation

Diane Kelly
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC USA
dianek@email.unc.edu

ABSTRACT

In literature and art, suspension of disbelief is the willingness of the audience to accept a premise as true even if it is not possible in the real world. Audience members must set-aside their beliefs about the world and accept one or more premises which form the foundation for the new world. Research also requires a suspension of disbelief, especially research conducted in the laboratory. In IIR evaluation, researchers suspend much disbelief. For example, most researchers understand (and believe) that measures developed in the context of system-centered evaluation (e.g., mean average precision) are inappropriate for evaluations involving users, but they still use them anyway. The design and development of IIR systems have been pursued diligently over the years, but less effort and care have been spent on developing IIR evaluation methods and measures. Instead, researchers often use methods and measures that are convenient, even when they are inappropriate and require extraordinary suspension of disbelief.

In this talk, I will first describe the IIR study landscape and standard IIR evaluation practices by presenting results of a systematic review of IIR evaluation studies that have been published in 31 sources over the last 40 years (1966-2006). The following questions will be addressed: How many IIR evaluation studies have been published during this period? How has this number changed over time? Where are IIR evaluations most likely to be published? What have been the most common subjects, tasks, collections and measures used in IIR evaluations?

In the second part of the talk, I will discuss problems and limitations of some of the standard ways of doing things and argue that more effort needs to be spent developing, evaluating and using research methods and measures that are tailored to IIR study situations, including IIR system evaluations and studies of IIR user behavior. A certain amount of suspension of disbelief is almost always required for research, but new research efforts can reduce the extent to which this is necessary in IIR.

Towards a Methodology for Contextual Information Retrieval

Emanuele Di Buccio
Department of Information Engineering
University of Padua, Italy
dibuccio@dei.unipd.it

Massimo Melucci
Department of Information Engineering
University of Padua, Italy
melo@dei.unipd.it

ABSTRACT

The crucial role of the user in the information seeking activities makes his behavior a useful evidence to seek and rank relevant objects. Considering the data about the user interaction in the ranking process can be interpreted as taking into account the user as an entity of the context where the information seeking activities are performed, and the behavior as an observable to represent the user. But the user is only one of the entities chosen to represent the context: the topic, the task are other involved entities. The complexity of modeling uniformly and simultaneously several entities through their observables can be faced by the adoption of a geometric framework. Starting from such geometric framework of the context, a methodology which aims at exploiting the contribution of different contextual entities is introduced in this paper.

1. INTRODUCTION

One challenge in Information Retrieval (IR) is to predict the objects relevant to the user's information needs. When the user is seeking information, the use of the evidence about his interests or behavior helps the system to predict the relevant information objects. However, there are many relationships between different observables, such as the user behaviour, the search task or the document content [1, 2] and discovering them is crucial for improving the retrieval effectiveness. For this reason, there is an increasing interest in these relationships, which can be viewed as the elements of the context affecting the relevance judgement. As a consequence, the IR system has to be designed to combine different observables of the context. In other words, the IR system has to be designed to be *context-aware*. Such a system should work as follows: When the user is seeking information, a set of features are collected for each entity involved; such features are the available evidence for the considered entity. Consider, for instance, the user as an entity. The features monitored during the user interaction are the evidence of the user behavior – examples are the display time and the amount of bookmarking or scrolling.

The difficulty of the design of a context-aware IR system increases in the event that several entities of the context are considered: besides the user, the topic, the task or the location are entities. The design and the implementation of distinct ranking algorithms, one for each type of feature can be an approach to use the features observed from several entities. Such an incremental approach to the combination of the features observed from different entities may fail to deal with the challenge posed by the context because the relationships between the entities are not explicitly modeled. Therefore, one issue is how to infer the relationships between the entities, e.g. between the task and the user, from the collected features.

The main goal of the methodology presented in this paper is to address the combination of the features observed from the different entities and then the representation of the relationships between them. The methodology will then be the basis upon which algorithms for ranking information object can be designed and evaluated by taking into account the context in a uniform way. In order to achieve this goal, the methodology has been structured in two steps. The first step is computing a representation of the context from the collected sets of features. To this end, a geometric framework proposed in [3, 4] and based on the vector space basis is used. However, in this paper, the problem of mapping the collected features to a vector space basis is addressed. The second step concerns how to rank information objects in the computed context.

2. MODELING CONTEXT

2.1 A Recycling Scenario

Let suppose Alice is looking for information about *possible benefit of recycling cans* because she wants to write a report for a university class. In order to perform this task, she submits the query *recycling can and why* to a search engine which returns a ranked list of results. Suppose she examines the results represented with title and snippet from the top of the result list, and she clicks on some of them, for instance the second and the fourth. Perhaps, she decides to inspect such results because they are promising. Alice's decision can help her to disambiguate the query. For instance, if the term "environmental" appears in the title or the snippet of the second documents, this information can help her to realize she is interested in one particular kind of benefit, that is, the one concerning the environment. During the examination of the pages, Alice performs several activities by interacting

with the browser. For instance, she can bookmark the page, print it and spend a certain amount of time on that page by performing scrolling activities and reading part or all of that. These features are evidence in support of the hypothesis that "environmental" is the meaning wanted by Alice. Suppose all the features about the mentioned behavior and the content of the displayed pages can be collected: how can we exploit the collected data to help Alice in the attainment of her information goal?

2.2 A Geometric Framework to Model Context

As stated above, many types of entity, e.g. the user, the task or the location are involved when seeking information. In the geometric framework adopted in this paper, each entity is characterized by a set of observables. In the scenario described in Section 2.1, whereas the user is an entity of interest, Alice's behavior is an observable of this entity – other observables of this entity may be, for example, her past queries or her cultural background.

Each observable can assume different *values*; for example, "ad-hoc task" or "home-page finding" are the values of the observable "type" of the entity "task". In this framework, however, the values are not scalars, but are vectors. This is necessary in order to employ the model illustrated in [4] upon which this paper is based. In particular, the set of values of an observable is a *basis* of the vector space in which the data are observed and the objects are represented.[1]

The rank, the display time and the occurrence of the term "environmental" are *features*; each of such features is associated to a (feature) vector and is depicted as a dashed line in Figure 1; for example, the rank is represented by $(1, 0, 0)$, the display time is represented by $(0, 1, 0)$ and the occurrence by $(0, 0, 1)$. When, for example, rank is 3, then $(1, 0, 0)$ is multiplied by 3 and $(3, 0, 0)$ is obtained.

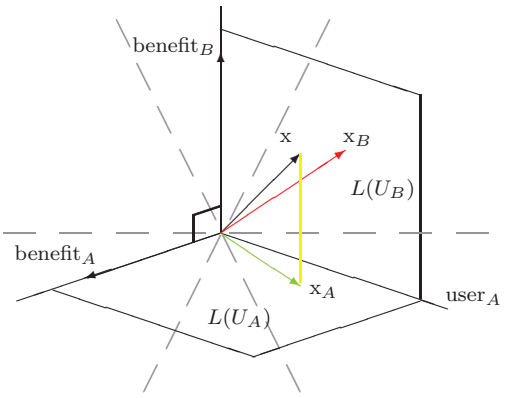

Figure 1: Example of different contexts.

In the previous scenario, for example, the vector behaviour$_A$ $= (3, 30, 0)$ is a value observed for Alice's behavior, the latter being an observable, where 3 is the rank and 30 is time spent on the clicked page. behaviour$_A$ can be represented by linearly combining the feature vectors with appropriate weights, for example, $(3, 30, 0) = 3(1, 0, 0) + 30(0, 1, 0) + 0(0, 0, 1)$. Moreover, the vector benefit$_A$ $= (3, 0, 1)$ is a value of the meaning of "benefit", where 3 is the rank and 1 means the occurrence of the term "environmental" in the snippet or the title of the document. Finally, benefit$_B$ is the meaning of "benefit" when the term "environmental" does not occur and the rank is 3, thus a different vector is obtained, i.e. $(3, 0, 0) = 3(1, 0, 0)+0(0, 1, 0)+0(0, 0, 1)$. Note that the display time is not involved in the latter observable.

The main problem is how to define all these vectors, that is the way for computing a basis which represents an observable observed from an entity. Indeed, once these vectors have been computed, an entity in a specific context can be described as a linear combination of the values of its observables. For instance, the information object x in Fig. 1 considered in the context where the user is Alice (i.e. user$_A$) and the environmental interpretation of benefit is considered (i.e. benefit$_A$), is obtained as a linear combination of the values user$_A$ and benefit$_A$ of the observables user and meaning of "benefit", thus obtaining the vector x$_A$.

One option is to define these vectors "manually" as did for the feature vectors. However, the manual definition is cumbersome, somehow arbitrary and prone to error. In the next section, a methodology for "automatically" defining the vectors which represent the values of the observables is illustrated.

3. METHODOLOGY

The previous section has been mainly focused on the geometric framework and on how this framework can be employed for representing the information objects and the context in which these objects are immersed. Although the previous work [4] gave a detailed account on the framework, it did not discuss an essential requirement, that is, the way for computing a basis which represents an observable (e.g. search type) and its values (e.g. ad-hoc resource finding) observed from an entity (e.g. task).

In contrast, this paper discusses how to compute the vector space basis. The basic idea reported in this paper is to build the basis which represents an observable by first of all considering the evidence collected through questionnaires or log files [5] when observing the information objects and the information seeking activities involving the entities of the context. Then, the evidence collected when observing the information objects and the information seeking activities is organized in well defined mathematical structures which are then "processed" to extract the bases.

However, an idea is not enough. What is needed is a methodology which precisely and completely describes how a well-defined geometric notion, e.g. the vector space basis, can be computed, perhaps, from a confused set of features observed during the information seeking activities and collected through noisy channels (e.g. questionnaires, HTTP log files). A methodology would not only help orderly collect features from the entities, but it also would help the IR researcher to design and then to evaluate innovative contextual retrieval systems as discussed in Section 4.

The methodology we consider in this paper is mainly structured in two parts, that is, *computing the context* of interest and *exploiting the computed context*. Here, "context" refers to the vector space basis and is used to stress the role played by the basis in representing the observables of an entity of the context – "context" is then, in a sense, a short form. These two parts will be discussed respectively in Section 3.1 and Section 3.2.

[1] A basis is a set of linearly independent vectors.

3.1 Computing the Context

The first part of the methodology consists in two steps, that is, (*i*) the features are collected to represent the context in which the information-seeking activities are performed by the user, (*ii*) the vector space basis which represents the context is computed by using the geometric framework described in Section 2.2.

As for the first step, in the scenario of Section 2.1, a tool (e.g. a browser extension) collects all the data concerning Alice's behaviour when she is examining the documents returned, for instance, by a search engine.[2] Examples of behaviour features are the data collected during the longitudinal user study of online information-seeking behavior described in [6]. Other features of interest can be extracted from the displayed results, that is, the title or the textual content of the snippet of the visited documents. The data are organized as a matrix whose rows refer to the objects observed (e.g. documents, clicks, events) and the columns refer to the features observed for each object. A set of points in a vector space (i.e. the rows) is thus obtained where the dimensionality of the vector space is the number of features.

The second step of the computing part pertains to the mapping from the features collected at the first step to the vector space basis. This point is crucial because it makes the geometric framework useable for IR purposes. While only the behavioral observable is taken into consideration in [3], in this paper, also the information about the meaning of the words is available and, as stated above, can be collected. Thus, two observables have been defined: an observable refers to the behaviour of the user, the other refers to the word meaning.

There is no reason not to exploit both the observables. For instance, the occurrence of the term "environmental" in the title or in the snippet of one of the displayed results when Alice is looking for information about "environmental benefit of recycling cans" can encourage her to select such result.

To this end, two distinct ranking algorithms, one for each observable, can be designed. In practice, two matrices are prepared – one matrix for each observable – and used as input to the ranking algorithms. The results of these two algorithms are then heuristically combined. This incremental approach misses the relationships, if any, between the two observables, thus failing to deal with the challenge posed by the context.

An alternative approach would be to "simultaneously" model the two observables. In practice, one single matrix is prepared in which the features of both the observables are recorded. However, modeling simultaneously the two observables implies some issues pertaining to the mapping from the features to the vector space basis as exemplified in Figure 2.

In the lower part of Figure 2 the use of two matrices is exemplified. The arrow labeled by B and the arrow labeled by K refer, respectively, to the user behaviour and to the meaning of the keywords. According to the geometric framework described in Section 2.2, two different vector space basis, respectively the matrix $\{b_{ij}\}$ and $\{k_{ih}\}$ are computed from these two matrices – one basis from each matrix. The basis computation can be performed by using the methods proposed in [3] and [4]. Note that a basis for the "entire" context can be obtained starting from the bases of the behavioral and the meaning context and combining them as shown by the gray arrows.

[2]Such a tool is under development at our laboratory.

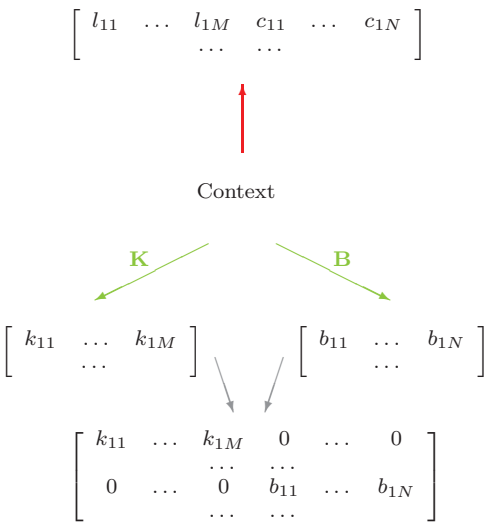

Figure 2: Issues pertaining the feature to vector-space basis mapping.

The limit of the use of two distinct matrices is that the two observables are considered as independent, although they refer to the same information seeking activity. In order to clarify this remark let come back to the scenario. Suppose Alice clicks on the document at rank 2 of the displayed list and spends 30 seconds on the selected page. However, the assumption that this behavior is independent from the occurrence of the term "environmental" does not necessarily hold. On the contrary, it may be well that the time spent by Alice to look at or perhaps investigate the clicked document depends on the occurrence of the term "environmental" in the textual content of the document.

Therefore, an alternative approach is to compute one single basis from all the features collected during an information seeking activity. As a result, one single matrix is prepared to collect all the features. The rationale underlying this approach is that the assumption that the features observed from an observable (e.g. the behaviour) of an entity (e.g. a user) are "entangled" with the features observed from an observable (e.g. the keywords meaning) of another entity (e.g. a document). This alternative approach is depicted in the upper part of Figure 2 where a basis simultaneously exploits all the features which are considered as representing the same information seeking activity.

3.2 Ranking in Context

The ranking process is affected by the approach adopted to compute the basis. Consider the first approach which produces two distinct matrices and then two vector space basis. In this case, an information object, e.g. a document, is represented as a vector of the space spanned by the basis which has been computed from a matrix. The document is then represented as another vector of the space spanned by the other basis which has been computed from the other matrix. The document is thus ranked against the two contexts separately since each document vector refers to a distinct space.

As an example, the vector x_A depicted in Figure 1 represents the information object x in the context $L(U_A)$ and the similarity with an object y in the same context $L(U_A)$ can be computed. When the meaning is considered, x_B represents the information object x in the context in the context $L(U_B)$ and the similarity with an object y in the same context $L(U_B)$ can be computed. The similarities computed in this way allow to rank the x's objects against the y's. Finally, the two computed rankings have to be fused in order to obtain a final list which considers simultaneously both the ranking information.

The alternative method is to work in the context where both the observables are considered simultaneously and rank the information objects in this context. This approach has the clear advantage that avoids the problems of combining different rankings produced by the observables of the context where the information seeking activities are performed. One advantage of this approach is that, by a suitable selection of the observables, also the relationships between the parts can be taken into consideration. This is for instance the rationale underlying the work reported in [7] where user and document are modeled as two subsystems interacting each other. The results reported show how the information provided by the system which considers both the subsystems simultaneously as a unique system can not always be derived from the information provided by the two subsystems considered as distinct.

4. FUTURE PERSPECTIVES

4.1 Methodology Issues

As regards the methodology described in Section 3, there are several issues. This first issue pertains the computation of the basis for the context concerning the user behavior. Indeed, some statistical procedures based on Principal Component Analysis (PCA) are unable to automatically discover the most effective vector space basis in terms of retrieval effectiveness [3] since they aim at maximizing the variance of the data, which is not the same as maximizing the effectiveness. A strategy to automatically find a solution to this issue is currently matter of investigation. Another issue concerns the possible strategy to combine rankings in the event of considering the two contexts distinctly, thus obtaining a unique final list. But not necessarily the two problems, that is mapping and ranking, have to be considered as distinct: the uniform way of modeling the different variables and the different observables involved, suggests that the two set of features can be exploited simultaneously to obtain a representation of the entire context. Exploiting both the observables might help provide the best direction in the data in terms of retrieval effectiveness, by affecting the "best variance direction", also in the case of the behavioral observable.

4.2 Evaluation

Evaluation is another crucial issue since a well established protocol which allows the variables of the context to be taken into consideration during the evaluation does not exist. The human-subject centered experiments take into consideration several observables, but refer to a specific situation and, as a consequence, are little generalizable. Generalization is required for several reasons. A first motivation is that each system is immersed in an environment; the environment itself can be modeled as a system constituted by a set of elements in interaction. As stated in [8], a crucial question is to understand which are the boundaries of the IR system, that is where it begins and where it ends. Is there a way to model the context, that is, the environment, so that also when the boundaries are modified the IR system can be evaluated? In our terminology, understanding which are the boundaries is equivalent to select the contextual entities and the observables for each entity. The problem is that the data used as a ground to evaluate the IR system changes when the boundaries are changed. The model described in Section 2, does not constitute only a ground to "build" ranking functions. The potential of such model is that, once the mapping between the features collected to model an observable and a basis is defined, we may produce "synthetic" contexts by tuning the features, thus obtaining a particular context of interest where the ranking algorithms can be tested. Finally, another issue regards the media. As stated in [4], the power of the geometric framework is also to be general and applicable to different media: the methodology proposed in [3] is not necessarily restricted to textual documents, but can be easily modified to non textual domains; the problem is that is not clear which are the suitable set of features to be used to represent other medium, e.g. images, thus affecting the possibility to test the effectiveness of the proposed model in non-textual domains.

5. ACKNOWLEDGEMENT

The work has been partially supported by the TELplus Targeted Project for digital libraries, as part of the eContent*plus* Program of the European Commission (Contract ECP-2006-DILI-510003).

6. REFERENCES

[1] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM '06*, pages 297–306, New York, NY, USA, 2006. ACM.

[2] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of SIGIR '08*, pages 163–170, New York, NY, USA, 2008. ACM.

[3] M. Melucci and R. W. White. Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM '07*, pages 273–282, New York, NY, USA, 2007. ACM.

[4] M. Melucci. A basis for information retrieval in context. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–41, 2008.

[5] M. Agosti, F. Crivellari, and G.M. Di Nunzio. A method for combining and analyzing implicit interaction data and explicit preferences of users. In *Proceedings of CIRSE 2009*, Toulouse, France, 2009. To Appear.

[6] D. Kelly. *Understanding implicit feedback and document preference: a naturalistic user study*. PhD thesis, New Brunswick, NJ, USA, 2004.

[7] M. Melucci. Towards modeling implicit feedback with quantum entanglement. In *Proceedings of QI 2008*, pages 154–159, Oxford, UK, 2008.

[8] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of SIGIR '95*, pages 138–146, New York, NY, USA, 1995. ACM.

Cognitive Effects in Information Seeking and Retrieval

Hideo Joho
Department of Computing Science
University of Glagsow
Sir Alwyn Williams Building
Lilybank Gardens
Glasgow G12 8QQ, UK.
hideo@dcs.gla.ac.uk

ABSTRACT

Lack of theoretical foundation in the design and development of interactive information retrieval (IIR) systems has been the major limitation in the field. Application of information seeking theories and models has also been limited in IIR. This paper is concerned with an approach to facilitate theoretical development in IIR based on cognitive effects, which can be considered in the evaluation of information seeking and retrieval (IS&R) research. In particular, I argue that looking at an "effect" in information seeking might facilitate our communication on IS&R phenomena and theory building in IIR.

1. INTRODUCTION

This paper is concerned with theoretical development in interactive information retrieval (IIR). By theoretical, I mean the use of scientific tools such as *framework*, *theory*, *law*, *principle*, *model*, and *hypothesis* [11]. By development, I mean to continuously build research upon existing research using the scientific tools. We are all aware that theory is the foundation of science. Shoemaker, et al. [17] state that a scientific theory is useful for summarising knowledge, developing practical applications, and guiding research. Nevertheless, we often focus on solving a problem at hand practically by developing systems, interfaces, or algorithms, rather than theoretically. Perhaps, the educational systems tend to an unbalanced weight on teaching practical solutions (e.g., how to build a program) and on teaching theoretical solutions (e.g., how to build and use a theory). Besides, theory building is a difficult job.

The design and development of interactive information retrieval systems have also been critisised as an ad-hoc process and lacking theoretical underpinning [6]. To address this issue, this paper explores ways to facilitate theoretical development in IIR in the context of Information Seeking and Retrieval (IS&R) research. The goals of IS&R research include 1) theoretical understanding of IS&R, 2) empirical explanation of IS&R, and 3) supporting the design and development of information systems [20, 16]. However, my survey of existing studies was originally motivated by the search for a theoretical framework for the design of search interfaces, and thus, the discussions in this paper are sometimes limited to this application, although often they have implications for a wide audience. The range of variables considered in IIR has expanded significantly due to the recent advance of IR in Context. Therefore, the role of theory has become essential for principled advance in this area.

The rest of paper is structured as follows. I first discuss the gap between information seeking models and information retrieval models, and how researchers have bridged it. Then I discuss cognitive effects often observed in Psychology as a potential building block of theoretical development. Finally, I discuss the implications of cognitive effects on the evaluation of search interfaces and other IIR techniques.

It should be noted that this paper does not provide detailed discussions on the definitions of all scientific tools for theoretical development. Interested readers might refer to an excellent paper by Järvelin [11] on conceptual tools in IS&R, and a book by Shoemaker, et al. [17] on theory building.

2. THE GAP

The designers and developers of interactive systems often seek theory to justify the design of (part of) a system and interface. One of the areas they look to is Information Seeking (IS), which studies the use and preference of information sources and channels, or more generally, information seeking behaviour. However, there are certain differences between IS and IIR which make this collaboration difficult. Järvelin and Vakkari noted that "by comparing the typical dependent variables in IR and IS studies it is easy to see that these two fields are interested in difference phenomena. The former is focused on the precision and recall of retrieval methods used by humans in interactive IR. ... The latter is focused on the preferences and use of channels" [20, p .124]. "If dependent and independent variables and their relationships do not overlap in the studies, it is understandable if representatives of both fields do not consider results and ideas from the other field useful" [20, p .124].

Another cause of the gap appears to be the characteristics of theoretical models used in the two fields. In IR, the models tend to be formal (e.g., Language Model), while they tend to be more conceptual in IS [24]. Ingwersen and Järvelin

To appear in the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation at ECIR 2009, 6th April 2009, Toulouse, France.
Website: http://www.irit.fr/CIRSE/

[10] also distinguish summary models and analytical models in IS&R. They noted that "summary models seek to summarize the central objects in an IS&R process - not necessarily all concrete - and their gross relationships without classifying and analyzing either. Analytical models, often narrower in scope, seek to classify the objects and relationships, and generate testable hypotheses. [10, P. 15]". An example of summary models is a nested model of information behaviour proposed by Wilson [24], while an example of analytical models is Byström and Järvelin's model on task complexity [3].

A model's ability to generate testable hypotheses is strongly related to *predictive power* of theory. Rutherford and Ahlgren [15] emphasised the importance of predictive power as follows. "The essence of science is validation by observation. But it is not enough for scientific theories to fit only the observations that are already known. Theories should also fit additional observations that were not used in formulating the theories in the first place; that is, theories should have predictive power. Demonstrating the predictive power of a theory does not necessarily require the prediction of events in the future. The predictions may be about evidence from the past that has not yet been found or studied." It is the predictive power that facilitates the design and development of practical applications based on a theory. However, in social science, the predictive power is often ignored and explanatory power is the focus of a model [17]. As a consequence, it is difficult to derive testable hypothesis from IS models and to base the design of search systems on them.

However, this trend does not seem to be unique to IS. Gibbs [7] noted that "throughout the field's history, sociologists have displayed an astonishing tolerance of untestable theories. Indeed, given the veneration in sociology of grand theories and the common indifference of sociologists to the few testable theories in their field, sociologies clearly pay little attention to testability when assessing theories." The limitation of summary models were echoed by Wilson [24] noting that "it does little more than provide a map of the area and draw attention to gaps in research: it provides no suggestion of causative factors in information behaviour and, consequently, it does not directly suggest hypotheses to be tested."

There is limited work which has attempted to overcome this gap between the two fields (e.g., [5, 12, 19]). For example, Fidel and Pejtersen demonstrated that Cognitive Work Analysis provided one approach to make studies in information seeking behaviour relevant to system design. A disadvantage of their approach is that it involved an extensive field study involved in the process which may not be available to us, and it is not easily repeatable. Vakkari [19] derived hypotheses from Kuhlthau's Information Search Process (ISP) model [13], and extended it to IIR context. For example, search tactics in ISP model was simply browsing or querying, Vakkari's extension contained more detailed query options. This shows that it is possible to bridge the gap between these two fields using scientific tools. However, as emphasised in Järvelin and Vakkari [20], this is partly due to the exceptionally good predictive power of the ISP model. Again, this reinforces the importance of predictive power in theoretical development[1].

3. COGNITIVE EFFECTS

How can we gain predictive power in the research? A way in which an understanding of human behaviour is represented in Psychology offers us one approach to solve this problem. This is often called an *effect*. An effect is a result of a cause (i.e., causality). A cognitive effect is a form of change in our cognitive state or process such as perception, learning, problem solving, memory, attention, and language caused by an event. An example of effects which is relevant to IIR research is the *novelty* effect [4]. The novelty effect refers to our tendency of temporally liking new technologies which diminishes as time goes by. Participants of user studies might prefer an experimental system over a baseline system due to this effect. A large number of effects has been observed and, to a different degree, studied by the researchers in Cognitive and Behavioural Science (See the `List of effects` section[2], `List of cognitive biases`[3], and `List of memory biases`[4] section in Wikipedia).

Effects have interesting traits as a conceptual tool of research. First, they can be a smaller unit than a model or theory. An effect might occur in a part of the model. This means that we can present and discuss effects more easily than models or theories. Second, they tend to have predictive power. Effects help us to predict what will happen based on a condition or function (e.g., when a new technology is introduced, people will initially like it). As we have discussed earlier, this allows us to formulate hypotheses to investigate this effect. Third, they do not always have a clearly understood cause, which is the principle of causality. This may sound contradictory to the previous point. However, the condition might represent a situation where an effect occurs, but it does not necessarily explain why it occurs. This means that we can present and discuss an effect without complete understanding. Fourth, effects are not necessarily about new ideas. It can be a conceptualisation of an effect we are empirically familiar with. An example is picture superiority effect, which essentially says concepts are more likely to be remembered if they are presented as pictures rather than as words. This simple effect, nevertheless, provides a theoretical justification for using a thumbnail of webpages as a form of bookmarking in search systems. Finally, some effects are a compound of other effects. For example, serial position effect shows that when people remember a sequence of items, they can recall the start and end items better than the middle items. This effect is a result of the primacy effect (for the first items) and the recency effect (for the last items).

An excellent example of cognitive effect in IS&R is the work on task complexity by Byström and Järvelin [3]. The overall effect of task complexity in their study can be defined as the changes of the needs of information caused the per-

[1]However, this is not to say explanatory power of theory can be ignored. If a theory is represented in a very abstract way, it will be difficult to use it in system or interface design.
[2]`http://en.wikipedia.org/wiki/List_of_effects`
[3]`http://en.wikipedia.org/wiki/List_of_cognitive_biases`
[4]`http://en.wikipedia.org/wiki/List_of_memory_biases`

ceived complexity of the task at hand. More specifically, it models that as task complexity increases 1) the complexity of information needed increases; 2) the needs of domain information and problem solving information increase; 3) the share of general-purpose sources increases, and that of problem and fact-oriented sources decreases; 4)the success of information seeking decreases; 5) the internality of channels decreases; and 6) the number of sources increases [3]. As can be seen, the effects help us predict a consequence for a given change. This helps us to design a responsive support for a given situation. The effects are also testable hypotheses [21]. Another task complexity effect was people's searching behaviour. For example, Machionini [14] observed that people are more likely to perform keyword search in low complexity tasks while they are more likely to perform browsing in high complexity tasks. Some of the accounts made by Byström and Järvelin help us understand the cause of this observed effect. Now we can see that a potential theory of task complexity in IS&R has been developing as a result of accumulated studies that look at effects and their causes.

Van Rijsbergen's clustering hypothesis [22] is another example. It hypothesises that similar documents are likely to be relevant for the same query. This hypothesis has been examined by numerous studies (e.g., [18]) and thus has become one of the most frequently used techniques in information systems. Arguably, it was the form of an effect, document similarity to topical relevancy, represented in the hypothesis that allowed researchers to use it as a building block of more complex applications.

Unfortunately, we do not have many cases like task complexity or the clustering hypothesis in IS&R. The next section discusses the implications of cognitive effects on IS&R research.

4. IMPLICATIONS ON IS&R EVALUATION

There are several directions of research based on cognitive effects. First, we can investigate existing cognitive effects found in Psychology in the IS&R context. Novelty effect, for example, is often used to describe a short-time positive effect on people's task performance or perception caused by a new technology. A new technology tends to increase a level of attention, which results in increased efforts or persistence, which in turn yields an achievement gain [4]. However, when they increase their familiarity with the technology this positive effect is known to disappear. Therefore, when participants of a user study are asked for their preference towards a baseline system (with low novelty) or an experimental system (with high novelty), a higher level of preference on the experimental system might be a result of its higher novelty rather than increased level of effectiveness, efficiency, or usability.

However, the increased level of effort and persistency caused by novelty appear to have more implications on IS&R. Recently, there have been several studies that tested the effectiveness of search interface that offer a form of grouping function that allows users to organise search results [23]. This type of interfaces is often designed to support exploratory tasks where information needs and task goals are ill-defined, and thus, iterative searching and browsing are required. One of the common results in such studies is that the number of interactions with system increased in experimental systems (with grouping function) when compared to baseline systems (without grouping). As discussed above, there is technological novelty in experimental systems. However, there also appears to be topical novelty. In other words, participants were finding new relevant information or new aspects/facets/instances in a topic with experimental systems, which could encourage them to make more effort and to be more persistent with artificial tasks. Therefore, continuously providing topical novelty might be a key factor to facilitate exploratory tasks, and thus, the design of exploratory search interfaces.

A second direction is to revisit existing IS&R findings to derive cognitive effects. The task complexity effect was such as example. Belkin, et al. [1] found that people's search queries can be lengthened when a short text "Information problem description (the more you say, the better the results are likely to be)" was added to the beneath of the query box. Since there is a higher degree of ambiguity in short queries, longer queries tend to perform better in search. This might be called the *query box* effect. Another example is click-through behaviour from query log analysis. It often shows a high click rate on the top ranked documents (e.g., 1st, 2nd, 3rd), but the URLs shown at the bottom of the result page (10th) are sometimes clicked more frequently than the middle (5th-9th). This is often explained by a scrolling action or paging action. When searchers try to go to the next page of search result, they saw the 10th result which was hidden before scrolling. This might be called the *page navigation* effect. How can we prevent this in terms of search interface design? Or is it worth placing a higher score result in the 10th place than 5-9th?

A third direction is to derive an effect by looking at the cases where proposed techniques did not work. This is a type of failure analysis. Although the designer of the system or search interface often assumes rational behaviour of the users, cognitive bias and irrational behaviour are common and persistent in human activities. By examining the cause of unsuccessful performance, one may hypothesise a potential effect that explains the situation. This can be a more constructive way of reporting our negative result rather than insignificant performance differences.

Readers may find the use of the label "effect" inappropriate and may prefer to use "hypothesis". As I discussed above, both have common traits, although a hypothesis is not necessarily about a casualty. The anomalous state of knowledge (ASK) hypothesis [2] is such an example. The ASK is about the nature of information needs. "Bias" and "fallacy" are also frequently used for a negative effect in Psychology. However, a more prominent assertion I would like to make in this paper was to encourage researchers to observe and formulate an effect as a part of evaluation and present it as a theoretical contribution of the research. The observed effect might be weak or sparse. You may not have a complete explanation, or later we might find that it was a false effect. However, at least this enables other researchers to build their work upon hypotheses derived from the effect. Gould [8] noted "Science is not a linear march to truth but a tortuous road with blind alleys and a rubbernecking delay every mile or two" (cited in [17]).

Measures used in laboratory-based experiments of IIR studies have been predominantly retrieval effectiveness (e.g., precision, recall, cumulative gain), interaction efforts or efficiency (e.g., task completion time, number of queries, frequency of browsing), and subjective assessments (e.g., task perceptions, system perceptions). This has resulted in a number of descriptive studies without much theoretical development. Shoemaker, et al. [17] noted "scholarly journals offer a wide variety of descriptive studies. ... These kinds of atheoretical research areas have their uses but they tend to produce isolated studies that do not move our knowledge forward on important questions in the field. ... such studies become like bricks lying around the brickyard rather than bricks that are used to build a wall" (p.168).

A strong tradition of measuring improvement in IR might have been preventing us from deriving potential effects from studies. Finding a cognitive effect in an evaluation can be a starting point of theoretical development as illustrated in this paper. The form of effects appears to be worth considering to represent our understanding of IS&R phenomena and to facilitate our communication about the phenomena.

5. SUMMARY

As discussed in the Introduction, we often focus on developing practical solutions to a problem, putting theoretical development at a lower priority in research. Also, theory building can be a difficult and abstract task. Exploring an effect in our studies might offer us a light-weight starting point for theoretical development which might eventually lead to an interesting theory. As Interactive IR is sometimes called Cognitive-oriented IR [10], the field has considered cognitive aspects in information seeking since its early days (e.g., [2, 9]). This paper discussed another aspect, a much smaller aspect, to contribute to the development of cognitive aspects in Information Seeking and Retrieval.

6. ACKNOWLEDGEMENT

The author thanks Robert Villa for useful discussions during the preparation of this paper. The author is also grateful for the constructive comments from anonymous reviewers.

7. REFERENCES

[1] N. J. Belkin, D. Kelly, G. Kim, J. Y. Kim, H. J. Lee, G. Muresan, M. C. Tang, X. J. Yuan, and C. Cool. Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 205–212, 2003.

[2] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval: Part 1. background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[3] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Inf. Process. Manage.*, 31(2):191–213, 1995.

[4] R. E. Clark. Reconsidering research on learning from media. *Review of Educational Research*, 53(4):445–59, 1983.

[5] R. Fidel and A. Pejtersen. From information behaviour research to the design of information systems: the Cognitive Work Analysis framework. *Information Research*, 10(1), 2004.

[6] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, June 2008.

[7] J. P. Gibbs. Resistance in sociology to formal theory construction. In J. Hage, editor, *Formal Theory in Sociology: Opportunity or Pitfall?*, pages 90–104. SUNY Press, 1994.

[8] S. J. Gould. Pretty pebbles. *Natural History*, 97:14–26, 1988.

[9] P. Ingwersen. Search procedures in the library - analysed from the cognitive point of view. *Journal of Documentation*, 38(3):165–191, 1982.

[10] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer-Verlag New York, Inc., 2005.

[11] K. Järvelin. An analysis of two approaches in information retrieval: From frameworks to study designs. *Journal of the American Society for Information Science and Technology*, 58(7):971–986, 2007.

[12] D. Johnstone, M. Bonner, and M. Tate. Bringing human information behaviour into information systems research: an application of systems modelling. *Information Research*, 9(4), 2004.

[13] C. Kuhlthau and Anonymous. *Seeking Meaning*, volume 2nd. Libraries Unlimited, 2003.

[14] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1997.

[15] J. F. Rutherford and A. Ahlgren. *Science for All Americans*. Oxford University Press, December 1990.

[16] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology*, 42, 2008.

[17] P. J. Shoemaker, J. W. Tankard, and D. L. Lasorsa. *How to Build Social Science Theories*. Sage Publications, Inc, December 2003.

[18] A. Tombros, R. Villa, and C. J. Van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Process. Manage.*, 38(4):559–582, July 2002.

[19] P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation*, 57(1):44–60, 2001.

[20] P. Vakkari and K. Järvelin. Explanation in information seeking and retrieval. In A. Spink and C. Cole, editors, *New Directions in Cognitive Information Retrieval*, pages 113–138. Springer, 2005.

[21] P. Vakkari and M. Kuokkanen. Theory growth in information science: applications of the theory of science to a theory of information seeking. *Journal of Documentation*, 53(5):497–519, 1997.

[22] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[23] R. W. White, B. Kules, and B. Bederson. Exploratory search interfaces: categorization, clustering and beyond: report on the xsi 2005 workshop. *SIGIR Forum*, 39(2):52–56, December 2005.

[24] T. D. Wilson. Models in information behaviour research. *Journal of Documentation*, pages 249–270, 1999.

A Method for Combining and Analyzing Implicit Interaction Data and Explicit Preferences of Users

Maristella Agosti
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
agosti@dei.unipd.it

Franco Crivellari
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
crive@dei.unipd.it

Giorgio Maria Di Nunzio
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
dinunzio@dei.unipd.it

ABSTRACT

A method of collecting and analyzing user data derived from the interaction with a Web searching service is introduced together with initial results. The selected and interleaved sources of data are both implicitly and explicitly collected during user interaction time. Preliminary results confirm that some insights can be gained by analyzing only data derived from the chosen and combined sources of data.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: User Issues; H.3.4 [**Systems and Software**]: User profiles and alert services

General Terms

Experimentation

Keywords

Log analysis, user studies, digital library evaluation

1. INTRODUCTION

User studies are a valuable method for understanding user behaviors in different situations. User studies require a significant amount of time and effort, so an accurate design of the process has to be carried out. Questionnaires are types of extensive methods that can be used to interview users, with the goal of learning how to better develop the service under investigation, and they explicitly collect information on the use of a specific service by its users [4]. Questionnaires can also provide simple feedback to build up an understanding of the different way users perceive the search tools provided by a service.

The interaction between the user and the system can be analyzed and studied in order to gather user preferences and "learn" what the user likes the most, and use this information to present the results in different ways. User preferences can be learned explicitly, for example asking the user to fill-in questionnaires, or implicitly, by studying the actions of the user which are recorded in the search log of a system. The second choice is certainly much less intrusive but requires more effort to reconstruct each search session a user made in order to learn his/her preferences.

Log is a concept commonly used in computer science; in fact, log data are collected by programs to make a permanent record of events during their usage. The log data can also be used to study the usage of a specific application, and to better adapt it to the objectives the users were expecting to reach. In the context of the Web, the storage and the analysis of Web log files are mainly used to gain knowledge on the users and improve the services offered by a Web portal, without the need to bother the users with explicit collection of information on the use of the portal.

In general, user studies and logs are used in a separate way, since they are adopted with different aims in mind. Ingwersen and Järvelin report in [3] that it seems more scientifically informative to combine logs together with observation in naturalistic settings. Pharo and Järvelin in [6] suggest systematic use of the triangulation of different data collection techniques as a general approach in order to get better knowledge of the Web information search process. Taking inspiration from this general approach, we have conceived a method of combining implicit and explicit user interaction data to gain information to be used also for personalization purposes. In the following we present this method together with initial experimental results that we have used to validate it.

The paper is organized as follows: Section 2 introduces the proposed method of combining different sources of user interaction data to extract knowledge on the user interaction process with a portal giving access to a two level sources of information: at the first level the user interacts with a union catalogue of bibliographic descriptions, at the second level the user interacts with national libraries catalogues that give access both to bibliographic descriptions and to digital objects. Section 3 reports on results that have been reached so far in the application of the proposed method in the real setting of The European Library. Finally, Section 4 gives insights on future developments of the application of the proposed method and possible presentation of it in a more structured and formal way.

2. A METHOD FOR DERIVING KNOWLEDGE ON WEB PORTAL USER INTERACTION

In the context of previous studies analyzing the interaction data of users who use on-line Web information services, we made the consideration that the information which can be extracted independently from log user interaction data and user questionnaire data could be combined together to understand how search process related factors are influenced by various personal and contextual user aspects [1]. From this starting point we have developed a method for collecting data derived from the user interaction log, "implicit" data, and data collected from user questionnaires, "explicit" data, for analyzing user Web searching and browsing. The combination of implicitly and explicitly collected data improves understanding of behavior with respect to the understanding that can be gained by analyzing the sets of data separately. In particular, the combined sets provide the opportunity of reaching insights towards user personalization of Web searching services, and also make possible results that can be generalized and applied not only to the users that have participated in filling out the questionnaires but to many other users that use the same service.

The method we propose was envisaged during the study of the Web portal of The European Library[1], which provides a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage.

The quality of the services and documents The European Library supplies are very important for all the different categories of users of a digital library system. Log data constitute a relevant aspect in the evaluation process of the quality of a digital library system and of the quality of interoperability of digital library services. Together with log data analysis, user studies can be performed on groups of users that freely crawl and navigate a Web portal, for example The European Library portal, and then fill in specifically designed questionnaires to report and describe their impressions. The conceived method is based on the combination and analysis of the following data sources: 1) HTTP log which contains the HTTP requests sent by the Web client to the Web server during a user browsing session, 2) search log which contains the actions performed by the user during a search, and 3) questionnaire data which are collected at the end of a user browsing and searching session.

If the previous sources of data were independently studied, only partial knowledge on the user would be gained. By analyzing the three data sources together the corresponding data can be interrelated, thus providing information not previously available. Such new information sheds light on previously unknown or hidden aspects about the user. The sources of data used and the interrelation among them are depicted in Figure 1 where an example of a user session is drawn.

The first problem of recreating the context of the actions performed by a user is identifying user sessions with certainty. In fact, the organization of the requests in a single session provides a better view of the actions performed by visitors. Authentication of users, which may provide the exact time of a session, is not always mandatory especially for free online services. For this reason a procedure named "session reconstruction" may be used in order to map the list of activities performed by every single user visiting the site. A possible approach to isolate a single session of a user is the use of the pair IP address and user agent[2] [5], which permits only a fixed gap of time between two consecutive requests [2].

3. CURRENT RESULTS AND INSIGHTS IN APPLYING THE PROPOSED METHOD

To validate the proposed method, a study was conducted in a controlled setting at the end of 2007 – beginning of 2008, in the computer laboratories of different faculties of the University of Padua, Italy, where students were requested to conduct a free navigation and search for information on The European Library portal and to fill in a questionnaire specifically designed to harvest the data that can be used to extract information on users satisfaction on the use of different parts of the portal. The students were mostly Italians, equally distributed between males and females, and with an age range typical of students of Bachelor and Master Degree (in most of the cases between 19 to 25 years old). Since the only way to recognize a particular user in the HTTP logs is the user agent, we asked students to modify the user agent string of the browser in order to uniquely identify each of them. A total of 216 students participated in the study, but a complete session reconstruction was possible only for 155 students, because 61 students made some mistakes in the procedure of the user agent string modification making it impossible to recognize them in the logs, so we decided to discard the data of this remaining 61 students to be sure of working on a quality data set.

The analysis of the results was done in the following order: the analysis of each stream of data - i.e. HTTP log, search log, questionnaires - was first conducted, while the analysis of possible interrelation among these sources was conducted later.

The analysis done on the HTTP logs regarded the provenance of the users and the selection of the collection of documents, i.e. if a user changes the default list of national libraries in which the search is performed. Not surprisingly, given the nature of The European Library portal, the users are spread throughout the world, and in particular Europe. However, users seem not to be using advanced search facilities and they prefer to leave default settings. Interestingly, for those users who intentionally changed the collections, it seems that there is not a correlation between the nationality of the user and the language of the collection (for example, Polish users who changed the collections prefer to search in the French Gallica collection instead of collections of Polish documents). A somewhat different result from the HTTP log is found on the correlation between the nationality of the user and the language of the interface. In fact, even though in general users tend to leave the default English interface, there is a correlation between the nationality and the language chosen for the interface. The same happens for the

[1] `http://www.theeuropeanlibrary.org/`

[2] A string which identifies the browser used by the user.

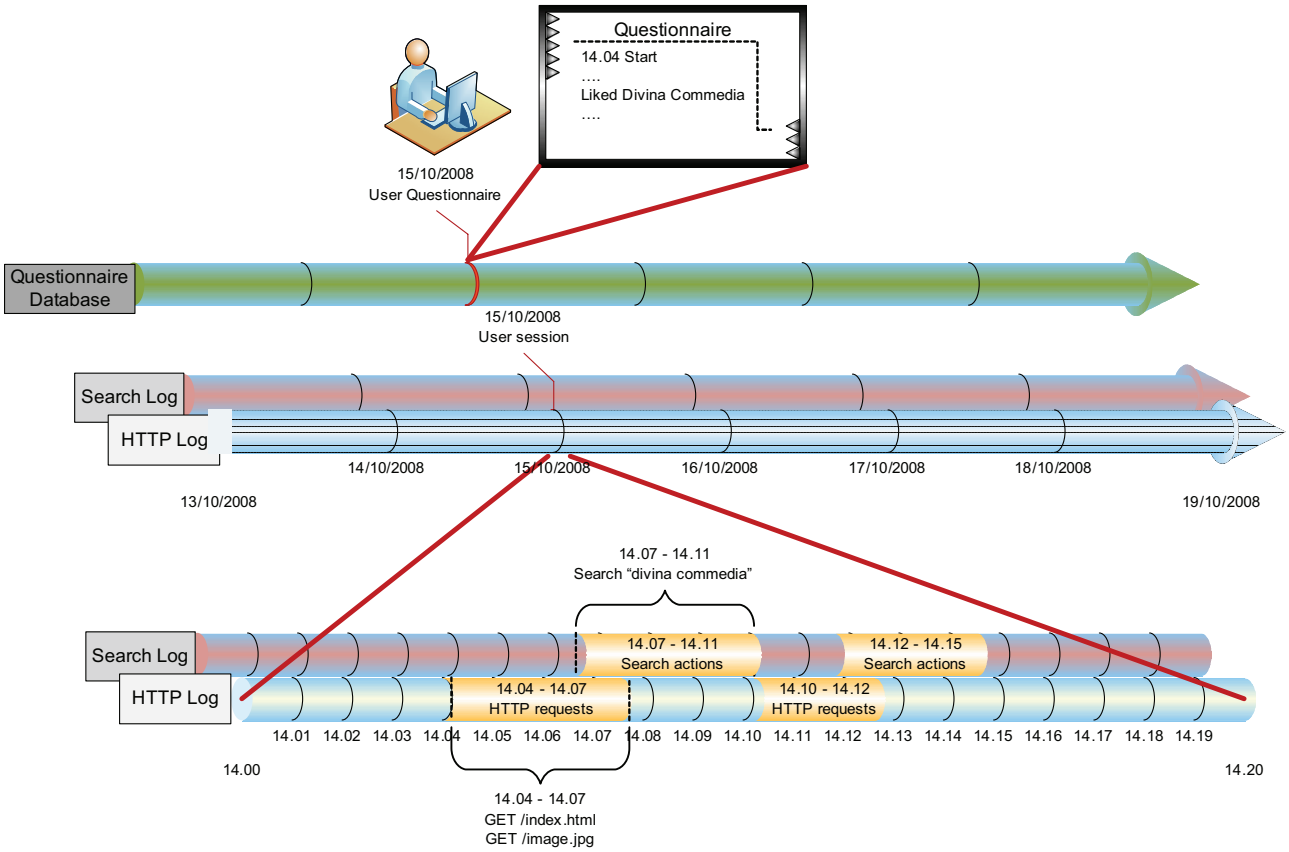

Figure 1: The three arrows represent the three sources of data that are collected and combined together by the proposed method. An example of a user session is shown to highlight the way the three sources are generated in an interleaved way during the user activities.

selection of the documents to be displayed, for example Italian users prefer on average to see documents of the National Libraries of Florence and Rome.

The analysis on search logs confirms some of the outcomes of the HTTP logs. For example, search sessions are on average short, around 2 minutes. We have to remember that the time of a session in the search log does not need to be exactly equal the session found in the HTTP logs since the information about the action and the HTTP requests may be different in time (a user may issue many queries without browsing the portal at all). It is interesting to note that there is a clear distinction between unregistered users, users who do not log in, and registered users, users who enter the portal and authenticate themselves: authenticated users spend more time on the portal compared to the users who do not authenticate themselves. However, the number of registered user sessions are less than 1% of the total number of sessions.

The analysis of the data collected from the questionnaires confirms the important question of languages. The analysis showed that, in general, students prefer to work in their own mother tongue instead of English. A total of 55% of users of the controlled study said that the first thing they did when they accessed the portal was change the language of the interface. However, the same people are willing to search, browse collections, and visualize documents in different languages. This is a first important point where the analysis of the interrelation between the three sources of data produces a more explicative situation than the one a single source could produce: users prefer to interact with the portal in their own language instead of English which is the default language (which means performing an intentional action of searching the Web page choosing to "click" the button which changes the language), nevertheless this does not exclude the fact that users search and browse collections of documents written in different languages. In particular, users do not choose the collection before the search (they leave the default list of collections) but they apply this sort of filtering at query result time when they show their preference for collections from their own country instead of others.

A first analysis of possible correlations between sessions calculated on different logs is shown in Table 1. The identification of sessions was based on: a TEL session identifier, the IP of the client who made the request, and the distance (in time) between two consecutive requests. For example, a session is a sequence of request made by the same client IP, with the same TEL session identifier, and whose distance between two consecutive request is always less than 30 minutes. The "Search log" column shows the statistics of the times, in minutes, of sessions found in the search logs, and between brackets the times of sessions of users who regis-

Table 1: Summary of statistics for the time of a user session in minutes calculated in the search logs (between brackets registered user only), HTTP logs (between brackets user who participated in the study), and the time for filling-in the questionnaire.

	Search log	HTTP log	Questionnaire
Median	2.0 (4.0)	1.3 (30.25)	31.0
Mean	6.0 (8.0)	4.7 (31.80)	33.0

tered to the portal. This shows that logging on is a clear intention of users who are willing to spend time in the portal and search more, compared to random users. The "HTTP log" column shows the times of sessions found in the HTTP logs computed in October 2007, and between brackets the times of the sessions of users who participated in the user study at the University of Padua. In this case, there is a strong bias of the students of the user study due to the time slot which was about 30/45 minutes. The times of random users are comparable to those found in the search logs. The last column shows the times of sessions for filling-in the questionnaires, which are obviously very similar to the times of HTTP sessions of the user study. There is one important aspect which emerges from the data: sessions are very short, browsing and searching activity lasts less than 2 minutes in 50% of the cases. This particular situation can be explained only by studying the answers of the users to the questionnaire where there are clear indications about some difficulties they found in understanding how to read the list of the results, and how to use some functions of the interface. These are also the reasons why they would have left the portal before if they were not asked to stay and fill in the questionnaire.

There is another important interrelation found among questionnaires and log data which may explain the short length of a user session. One of the outcomes of the questionnaire was the disorientation of the user upon entering The European Library portal for the first time, in particular it seems not to be clear what kind of information can be accessed through this portal. Users are in general ready to search in a Google-like fashion and obtain documents, in terms of links to pages or documents online, in the case of The European Library they are essentially in front of an online public access catalogue which retrieves bibliographic records. Obtaining library catalogue records after a search is a source of confusion which leaves the user unhappy and willing to leave the portal quickly.

Questionnaires also show that images in particular seem to be very appealing for users; both the "treasures" section, a section which shows high resolution images of ancient documents, and the "exhibition" section, a section which shows pictures of the national libraries buildings, were thoroughly browsed by users even before making any query in the portal. This is an important clue which may suggest that there should be more linking from the images to the catalogue records. The interrelation among the information about users which prefer images and the HTTP log and searches log is still under investigation. In fact, we would like to see if this willingness expressed in the questionnaire is also reflected in user actions: for example, a user who is interested in images clicks more frequently on images or search for documents like maps or paintings; or a user expresses this interest in images but actually does not perform any action in the portal which confirms this interest.

It is necessary to note that the sample of students cannot be considered as a significant sample of all the users of The European Library; however, the results of the analysis on this group of students are useful to understand the behavior of university students (specifically from Italy, and in particular from University of Padua). These are users who can be interested in using the portal and searching of bibliographic records and, for this reason, their judgements about the interface and the services have to be taken into consideration for personalization purposes.

4. FUTURE WORK

The continuation of the work foresees the identification of all the user sessions that are present in the logs and the matching of those sessions with the students ones, in this way we can assign a known profile to unknown users who can be automatically offered the same advanced services that at present can be offered only to known users. This can be considered a way of offering personalized services to a wider audience without explicitly asking for personal preferences.

Finally, a future development of the work is to provide a formal and structured model of the method.

5. ACKNOWLEDGEMENTS

The work has been partially supported by the TELplus Targeted Project for digital libraries, as part of the eContent*plus* Program of the EC, ad by the TrebleCLEF Coordination Action, as part of the 7FP of the EC.

6. REFERENCES

[1] M. Agosti, G. Angelaki, T. Coppotelli, and G. M. Di Nunzio. Analysing HTTP Logs of a European DL Initiative to Maximize Usage and Usability. In D. H.-L. Goh et al. eds, *Proc. 10th Int. Conf. on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam*, pages 35–44. LNCS 4822, Springer, 2007.

[2] B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. volume 2703 of *Lecture Notes in Computer Science*, pages 159–179. Springer, 2003.

[3] P. Ingwersen and K. Järvelin. *The Turn*. Springer, The Netherlands, 2005.

[4] D. Kelly, D. J. Harper, and B. Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information Processing & Management*, 44(1):122 – 141, 2008. Evaluation of Interactive Information Retrieval Systems.

[5] D. Nicholas, P. Huntington, and A. Watkinson. Scholarly Journal Usage: the Results of Deep Log Analysis. *Jour. of Documentation*, 61(2):248–280, 2005.

[6] N. Pharo and K. Järvelin. The SST method: a tool for analysing Web information search processes. *Information Processing & Management*, 40(4):633–654, July 2004.

A contextual evaluation protocol for a session-based personalized search

Mariam Daoud
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
daoud@irit.fr

Lynda Tamine-Lechani
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
lechani@irit.fr

Mohand Boughanem
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
bougha@irit.fr

ABSTRACT

Most existing evaluation protocols in IR are laboratory-based ones. They are based on a controlled evaluation methodology and lack generally of user evidences expressing his search context in the data set and the evaluation protocol as well. This paper proposes an evaluation protocol for a session-based personalized search. It is based on an enhanced TREC HARD collection with simulated user profiles issued from simulated search sessions. The experimental results show the effectiveness of our personalized search approach according to the proposed evaluation protocol.

Keywords

Evaluation protocol, personalized search, user profile, search session, simulation

1. INTRODUCTION

The main goal of an IR system evaluation framework is to measure the corresponding retrieval performance according to efficiency and/or effectiveness. Regarding effectiveness evaluation, most existing evaluation protocols are laboratory-based ones where the user need is generally represented by the user query. Within the emergence of contextual IR, these protocols are not sufficient for evaluating the system performance under the challenge of context. Attempts to extend the laboratory model to user centred evaluation have been achieved via the TREC Interactive track [6] and HARD track [1] by integrating the user context in the data set. Despite these extensions, the overall evaluation still is restricted due to the exploitation of only specific contextual features. To alleviate such limitations, contextual evaluation methodologies have been proposed to support simulated user profile through contextual simulations [9] or real evaluation scenarios through user studies [3].

In this paper, we present an evaluation protocol devoted for a session-based personalized search. The user profile is simulated using hypothetic user interactions on documents provided by TREC and the search session is simulated by aligning generated sub-queries of a query along a sequence. The paper is organized as follows. Section 2 presents a short overview of contextual IR evaluation. Section 3 presents our approach of search personalization and a contextual IR evaluation integrating simulated user profiles and search sessions. Section 4 presents a conclusion and our perspectives for future works.

2. CONTEXTUAL IR EVALUATION: A SHORT OVERVIEW

Contextual IR evaluation aims at measuring the system performance by integrating the user context in the evaluation scenario. There are two main types of contextual evaluation: evaluation by context simulations and evaluation by user studies.

The first kind of evaluation simulates users and interactions by means of well defined retrieval scenarios (hypothesis) [10]. A contextual simulation in [9] used a document collection issued from a predefined Web ontology and simulates the user profile by a concept of the ontology. For a specific simulated concept /user profile, queries are generated automatically by the top terms representing the concept. The user profile is built using a set of documents classified under this concept, called the profile set. Other personalized approaches carry out a contextual simulation by enhancing TREC collection with simulated user profile [11, 4] represented by a TREC domain and built using the relevant documents of the queries annotated by the domain. Evaluation measures are the precision and recall in the case of extending a laboratory-based collection (TREC) or the average rank of the documents returned by the system and that are classified under a simulated concept in [9]. This evaluation method is worthwhile since (a) it is less time consuming and costly than experiments with real users and (b) allows comparative evaluation with respect to the defined scenarios [12].

The evaluations by user studies are carried out with real users to test the system performance through real user interactions with the system. There are two types of user studies adopted in the domain. The first one [8] consists of using a search interface plugged within a TREC collection where the user is asked to reformulate queries related to a predefined topic by TREC in order to define a search session. The user profile is represented by the user search history in a search session. The second kind of contextual evaluation by user studies [3, 7] is carried out using an API search interface

(like the Google API) that allows the user to perform natural search. Evaluation measures are the average rank of the clicked documents by the user [3] or the average precision and recall over the top N returned documents by the system [7, 8]. Recall is computed using all judgements given by all the users for the query.

3. A CONTEXTUAL EVALUATION FOR A SESSION-BASED PERSONALIZED SEARCH

3.1 A session-based personalized search

Our approach of search personalization [5] aims at representing and personalizing the search using a graph-based user profile issued from a predefined ontology. The user profile is built over a search session by combining graph-based query profiles; personalization is achieved re-ranking the search results of queries related to the same search session.

A query profile G_q^s is built by exploiting the documents clicked D_r^s by the user and returned with respect to the query q^s submitted at time s. First a Keyword query context K^s is calculated as the centroid of documents in D_r^s:

$$K^s(t) = \frac{1}{|D_r^s|} \sum_{d \in D_r^s} w_{td} \qquad (1)$$

K^s is matched with each concept c_j of the ontology represented by single term vector $\vec{c_j}$ using the cosine similarity measure. The scores of the obtained concepts are propagated over the semantic links as explained in [5]. We select the most weighted graph of concepts to represent the query profile G_q^s at time s. The user profile G_u^0 is initialized by the profile of the first query submitted by the user. It is updated by combining it with the query profile G_q^{s+1} of a new related query submitted at time $s+1$.

A session boundary delimitation based on the Kendall rank correlation measure is used to compute the correlation degree between a new submitted query q^{s+1} and the user profile G_u^s. When the correlation value $\Delta I = q^{s+1} o G_u^s$ is above a predefined threshold value $\sigma*$, the search results returned by the system are re-ranked by combining their original score with their contextual score as follows:

$$S_f(d_k) = \gamma * S_i(q, d_k) + (1 - \gamma) * S_c(d_k, G_u^s) \qquad (2)$$

$$0 < \gamma < 1$$

The contextual score S_c is computed between the result d_k and the top h weighted concepts of the user profile G_u^s as follows:

$$S_c(d_k, G_u^s) = \frac{1}{h}. \sum_{j=1..h} score(c_j) * cos(\vec{d_k}, \vec{c_j}) \qquad (3)$$

3.2 Contextual evaluation protocol derived from HARD TREC

We present a contextual evaluation protocol which integrates the user profile into the evaluation process. Involved components are the query set consisting of related sub-topics of the HARD TREC topic set, the user profile built across a simulated search session defined by a sequence of related subtopics and the evaluation strategy that aims at training the session boundary delimitation and then testing the personalized search using the best system parameter.

3.2.1 Queries

Queries are the topics provided by TREC[1] 2003 HARD Track [2]. As the user profile is built over a search session (a sequence of related queries) and there is no information available in the collection concerning the correlation between the queries, we generate three subtopics per topic that define a simulated search session. The adopted strategy for generating the *sub-topics* of a topic consists of:

- extracting a document *profile set* that consists of the top r relevant documents returned by the system with respect to the topic.
- dividing the document *profile set* randomly into equally-sized three *profile subsets*.
- creating the centroid vector of each profile subset using formula (1) by representing each document as an ordered term vector using *tf*idf* weighting scheme.
- building each *sub-topic* by selecting the top three terms of the centroid vector.

In our experiments, we excluded topics that achieve null average precision and we set $r = 9$, which implies excluding topics that have less than 9 relevant documents returned by the system. We obtain a total of 8 topics in the main data query set.

In order to validate the reliability of the generated *subtopics*, we computed the percentage of average subtopic-topic relevant document overlap and the percentage of non-overlapping documents over the Top-n results (Top-20 and Top-40) returned by the system between the *subtopics* themselves. Results in figure 1 prove that the subtopics have more than (50%) of common relevant documents with the topic which confirms that they are related. Moreover, results in figure 2 prove that even though the *subtopics* were built from the same topic, they contains different terms which leads to do not return same documents (average non-overlapping estimated higher than 40% at Top-20). For the topic 48, we obtain a null percentage of non-overlapping documents at Top-40 because the generated subtopics contains two common terms over three. Subtopics still correlated as they are considered as different reformulations of the same topic.

3.2.2 Document collection

The main document collection used in the HARD track contains the newswire text from AQUAINT corpus and U.S. government documents.

3.2.3 User profile

The user profile is integrated in the evaluation strategy according to a simulation algorithm that generates it using hypothetic user interactions for each topic provided by TREC assessors. We consider that a topic holds an information need that represents a simulated user interest / user profile. The simulated user profile is created across related subtopics as follows:

1. Creating the ontological query profile for each *subtopic*. We notice that the query context K^s for a subtopic q^s is created using its appropriate relevant document *profile subset*.

[1]Text REtrieval Conference:http://trec.nist.gov

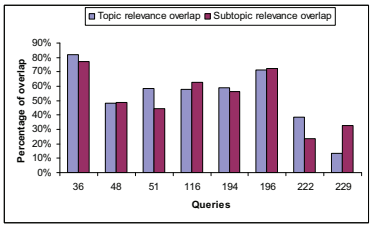

Figure 1: Percentage of relevant overlapping documents between the subtopics and the main topic

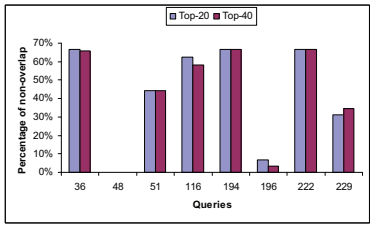

Figure 2: Percentage of non-overlapping documents between the subtopics themeselves

2. Initializing the user profile by the ontological profile of the first query of the session.

3. Applying the session boundary recognition mechanism using an appropriate threshold value (σ^*) when a new *subtopic* has to be processed along a query sequence. If the *subtopic* is correlated to the current user profile, this latter is then updated by combining it with the the profile of the new subtopic.

3.2.4 *Evaluation protocol*

The evaluation protocol is designed to tune the session boundary parameter in a training stage and evaluate the effectiveness of the personalized search in the testing stage. For this purpose, we divided the HARD topics into two topic sets: a *training topic set* and a *testing topic set*.

A. Training stage

In this stage, we tune the optimal correlation value of the session boundary delimitation that achieves the best performance. To do so, we set a training query sequence created by aligning successively subtopics of the training topic set given by the HARD track.

First we computed *subtopic-profile* correlations values according to the Kendall measure between each *subtopic* on the training sequence and the user profile built across previous and related *subtopics* (related to the same topic). Once the correlation values are computed, we proceed to tune an average correlation value that maximize the precision of detecting correct session boundaries $P_{intra}(\sigma)$ and correct correlated queries $P_{inter}(\sigma)$ defined as follows:

$$P_{intra}(\sigma) = \frac{|RQ|}{|TRQ|}, P_{inter}(\sigma) = \frac{|BQ|}{|TBQ|} \quad (4)$$

Where $|RQ|$ is the number of *subtopics* identified as correctly correlated according to the *subtopic sequence*, and $|TRQ|$ is the total number of *subtopics* that should be identified as correlated, $|BQ|$ is the number of *subtopics* indicating correct session boundaries, and $|TBQ|$ is the total number of session boundaries in the *subtopic sequence*. The optimal session boundary threshold value is identified when both measures ($P_{intra}(\sigma)$ and $P_{inter}(\sigma)$ have the maximum accuracy.

$$\sigma* = argmax_{\sigma}(P_{intra}(\sigma) * P_{inter}(\sigma)) \quad (5)$$

B. Testing stage

In this stage, we evaluate the search personalization along a testing query sequence (different from the training query sequence). The evaluation is based on comparing the typical search performed using only the query to the personalized search using the query and the correlated user profile.

This stage could be explained by the following steps:

- Creating the testing query sequence by aligning the subtopics of the testing topic set.

- Along this sequence, we used the optimal threshold value in the session boundary recognition mechanism to build the user profile across related testing subtopics. For each subtopic having a correlation value greater than the optimal threshold value $\sigma*$, we proceed to:

  – perform the personalized search on this subtopic by re-ranking its search results using the correlated user profile,

  – update the user profile by combining it with the query profile of the subtopic being processed.

3.3 Experimental Results

The evaluation protocol presented above is used to evaluate our search personalization approach. A total of 8 topics are divided equally into training topics and testing topics.

3.3.1 *Evaluating the session boundary delimitation*

According to the training stage, we create a sequence of 12 subtopics of 4 training topics by aligning the most correlated topics successively. So, we have 3 total session boundaries to be identified (TBQ=3) and 8 queries to be identified as correlated (TRQ=8) by considering two correlated subtopics per topic along the sequence. We computed the subtopic-profile correlation along the training sequence. The optimal session boundary threshold is identified at $\sigma^* = -0.41$ achieving significant precision of identifying correct session boundaries (66%) and correct correlated queries (75%).

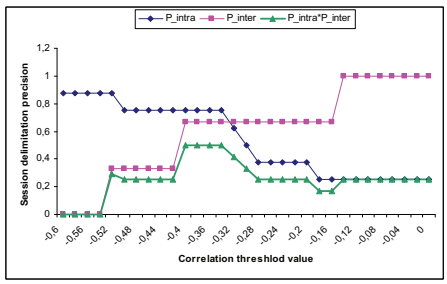

Figure 3: Kendall correlation values computed across the training *subtopic sequence*

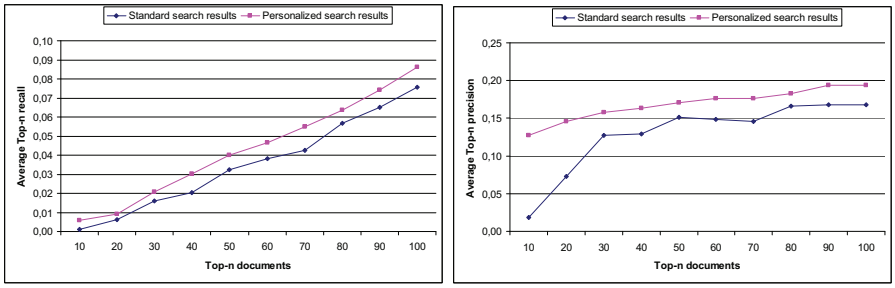

Figure 4: Average Top-n recall and Top-n precision comparison between the personalized search and the standard search on the subtopic sequence: profile built from the top ranked documents of the topic

3.3.2 Evaluating the personalized search performance

According to the testing stage, we evaluate the effectiveness of the personalized search by comparing it to the typical search. A testing sequence of 12 subtopics is created by aligning 4 topics according to their number in the HARD trec. We used the optimal session boundary identification ($\sigma^* = -0.41$) to detect related subtopics.
Figure 4 shows the average Top-n precision and Top-n recall achieved by personalized search comparatively to the standard one on the *subtopic* sequence using $\gamma = 0.3$ (tuned as the best value) of equation (2). Results prove that personalized search achieves higher retrieval precision and recall comparatively to the standard search. The best performance is achieved by the personalized search in terms of top-10 precision (12.73) and top-10 recall $(0, 57)$ comparatively to the standard search having lower top-10 precision (1.82) and lower top-10 recall $(0, 10)$. We have conducted experiments in previous work [5] where the profile set of each topic was defined by the first 30 relevant documents listed in Qrels [5]. We notice that the improvement is much more higher when the user profile is built using the top ranked documents returned by the system with respect to the topic. This confirms that our method achieves an effective personalization as in real world search engines.

4. CONCLUSION

In this paper, we have presented an evaluation protocol devoted for a session-based personalized search. More precisely, this protocol is suitable for the evaluation of a personalized search that requires a session boundary delimitation to build the user profile. It is based on the TREC HARD collection where the user profile is simulated for each topic using a set of relevant returned documents by the system and that are previously judged by TREC assessors. We defined also a session-based evaluation scenario that integrates the search session as a sequence of subtopics generated for a specific topic. We evaluated our approach according the proposed evaluation protocol and show that our approach is effective. In future work, we plan to extend this protocol by using real user data provided from a search engine log file. Extending the protocol aims at testing the effectiveness of the personalized search based on relevance judgements given by the user who submits the query. It consists of defining a query set provided from real users and the associated relevance judgements provided from the clickthrough data available in the log file.

5. REFERENCES

[1] J. Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. In *Proceedings of the 12^{th} text retrieval conference (TREC-12)*, pages 24–37, 2003.

[2] J. Allan. Hard track overview in trec 2003: High accuracy retrieval from documents. In *TREC*, pages 24–37, 2003.

[3] V. Challam, S. Gauch, and A. Chandramouli. Contextual search using ontology-based user profiles. In *Proceedings of RIAO 2007, Pittsburgh USA*, 2007.

[4] M. Daoud, L. Tamine, and M. Boughanem. Learning user interests for session-based personalized search. In *ACM Information Interaction in context (IIiX), London*, pages 57–64, octobre 2008.

[5] M. Daoud, L. Tamine, M. Boughanem, and B. Chebaro. A Session Based Personalized Search Using An Ontological User Profile. In *ACM Symposium on Applied Computing (SAC), Haiwai (USA)*, pages 1031–1035, march 2009.

[6] D. Harman. Overview of the the 4^{th} text retrieval conference (trec-4). pages 1–24, 1995.

[7] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.

[8] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 43–50, 2005.

[9] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of CIKM'07*, pages 525–534, 2007.

[10] L. Tamine, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval: overview of issues and research. *Knowledge and Information Systems (Kais)*, 2009.

[11] L. Tamine, M. Boughanem, and W. Zemirli. Exploiting multi-evidence from multiple user's interests to personalizing information retrieval. *IEEE International Conference on Digital Information Management(ICDIM 2007)*, 2007.

[12] R. White, I. Ruthven, J. Jose, and C. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.

Evaluating Information Access Tasks for Personal Lifelogs

Gareth J. F. Jones
Centre for Digital Video Processing &
Centre for Next Generation Localisation
Dublin City University, Ireland
Gareth.Jones@computing.dcu.ie

ABSTRACT

Emerging personal lifelog (PL) collections contain permanent digital records of information associated with individuals' daily lives. These archives can include materials such as emails, web content, personal documents, photographs, videos and music, logs of phone calls and text messages, and also personal and contextual data such as location (e.g. via GPS sensors), persons and objects present (e.g. via Bluetooth) and physiological state (e.g. via biometric sensors). PL archives have many potential applications including helping individuals recover partial forgotten information, sharing experiences with friends or family, telling the story of one's life, clinical applications for the memory impaired, and fundamental psychological investigations of memory. Realising these potential applications requires the effective combination of content and context information within the information access process. PLs can be collected by individuals over very extended periods, potentially running to many years.

At DCU we are currently engaged in the collection and exploration of applications of large PLs. We are collecting rich archives of daily life including textual and visual materials, and a range of related context data. An important part of this work is to consider how the effectiveness of our ideas can be evaluated. For example, what metrics should be used, and how can meaningful experiments be conducted where the only person who can judge the accuracy of the application is the owner of their PL? While these studies might have considerable similarity with traditional evaluation activities in areas such as information retrieval and summarization, the characteristics of PLs mean that new challenges and questions emerge. We are currently exploring the issues through a series of pilot studies and questionnaires. Our initial results indicate that there are many research questions to be explored and that the relationships between personal memory, context and content for these tasks is complex and fascinating.

Evaluation of a Personal Information Agent derived from the Context Modelling of Evolving Information Needs

Desmond Elliott
Department of Computing Science
University of Glasgow
Sir Alwyn Williams Building
17 Lilybank Gardens
Glasgow
G12 8QQ
delliott@dcs.gla.ac.uk

Joemon Jose
Department of Computing Science
University of Glasgow
Sir Alwyn Williams Building
17 Lilybank Gardens
Glasgow
G12 8QQ
jj@dcs.gla.ac.uk

ABSTRACT
This paper presents a pilot study of a novel context modelling method for capturing an evolving model of information needs during search tasks to assist users in satisfying their long-term needs. Short-term sessions are clustered using terms extracted from explicitly marked search results to create and update aspects of a user profile. The terms that most significantly define the aspects of a user profile are determined using the Ostensive Model and these aspects are used to create a Personal Information Agent to help users satisfy their long-term needs. A novel evaluation technique is introduced comprising four participants performing two simulated work situation tasks over three search sessions where user questionnaires and log file analysis measure the performance of the context modelling method.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Systems and Software-*Relevance feedback*

General Terms
Experimentation, Human Factors, Measurement

Keywords
Ostensive model, context modelling, personalised retrieval, evaluation techniques.

1. INTRODUCTION
In information retrieval, long-term information needs are typically characterised by users seeking resources on the same topic over multiple search sessions. They differ from short-term needs because the user usually has a personal or professional interest driving their behaviour. The majority of long-term information needs have multiple aspects; for example, somebody following developments in U.S. politics may be interested in the legacy George W. Bush leaves behind and the challenges facing Barack Obama during his presidency.

In traditional information retrieval techniques [1] there is an assumption that users will be able to satisfy their information needs with a single query during single session. Relevance feedback techniques [2] emerged to assist users in satisfying their needs with multiple queries on the same topic in a single session based on result list interactions. The emerging area of contextual information retrieval seeks to develop techniques to capture and represent information needs that evolve over time.

There are several examples of previous research on contextual information retrieval. Harper and Kelly [3] present an interface that allows users to organise their needs into piles and provides contextual feedback through these piles. Their evaluation combines click-through data and a user questionnaire. Martin and Jose [4] present an interface where users bundle documents and receive contextual feedback based on those bundles. Their evaluation also combines click-through data and user questionnaire responses.

Our motivation is help users satisfy their long-term information needs by modelling their interactions with search results and retrieving new web pages on their behalf using these models. In comparison with the approaches in [3][4], our approach combines the data extracted from explicit relevance feedback across multiple sessions to capture the evolving interests of users. We adapt the evaluation techniques used in [3][4] and perform a small user study with two tasks designed to simulate a long-term information need in participants.

We investigate the following questions in this paper:

- How useful is a retrieval system that helps users construct a profile of their evolving information needs?
- How useful is a personal information agent based on aspects of user profiles captured from explicit interactions?

To evaluate our novel context-modelling technique we will change the method of producing terms the system uses and suggests terms that define aspects of user profiles. The *baseline* system will construct user profiles using terms extracted from search queries issued during previous sessions. The *experimental* system will construct user profiles using terms extracted from explicitly marked documents. The evolving significance these terms contribute to the profile will be calculated using the Ostensive Model.

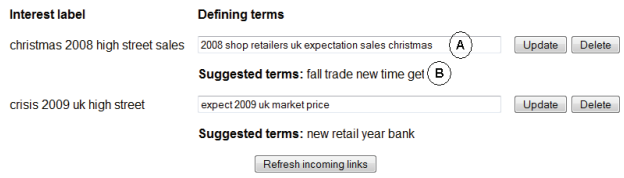

Figure 1: User profile management interface.

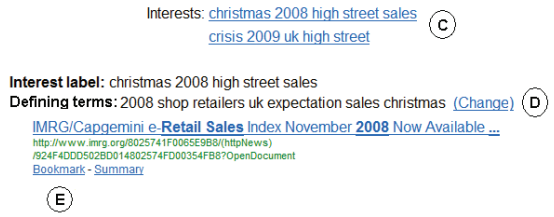

Figure 2: Personal information agent results.

The remainder of this paper is structured as follows. Section 2 describes the retrieval system used in our study. Section 3 describes our experimental and baseline context modelling techniques. Section 4 focuses on our evaluation technique and experimental methodology. Section 5 offers some preliminary results and discussion based on an experiment with four participants.

2. SYSTEM DESCRIPTION

Our retrieval system is based on the system originally presented in [5]**.** It is a web application comprising a separate backend and frontend. The backend consists of components for ad-hoc indexing, session clustering, context modelling, and logging. The frontend consists of interfaces for user profile management, bookmark management, search results, and personal retrieval agent results.

In the backend, a new session is created when a user issues a query and terminated when the user issues a different query. The Google SOAP Search API[1] is used to retrieve results, which are indexed ad-hoc using *tf-idf weighting* on the extracted text content of each result. The current search session is added to the user profile if one or more results are marked as explicitly relevant during the session and clustered with other search sessions by the terms extracted from the results that define the session.

In the frontend, users can search the Web using a typical search box. Users can click-through to inspect result content and mark search results as explicitly relevant. The different aspects of a user profile can be managed from the profile management interface, shown in Figure 1. For each aspect of their profile, users can view the defining terms (A), view a list of suggested terms derived from the context-modelling component in the backend (B), delete the aspect from their profile, and change the defining terms of the aspect. Search results retrieved by the personal information agent can be viewed and interacted with as shown in Figure 2. For each aspect of the user profile, the defining terms used to retrieve documents on behalf of the user are shown (D) and users can click-through to view the result content and explicitly mark the result as relevant (E).

3. CONTEXT MODELLING METHODS

We use the Ostensive Model [6] to determine the evolving nature of aspects of a user profile in long-term search tasks. The Ostensive Model is a model of developing information needs where the user's needs are described as a set of evolving information objects that exemplify the need. Campbell uses the Ostensive Model in a browse-based system to allow users define their needs through implicit feedback [7]. Upon selecting an object that defines their current need, the system immediately updates the set of objects available to the user.

[1] http://code.google.com/apis/soapsearch/

In a query-based system, we believe that immediately changing the search results would frustrate users. We propose that each set of interactions with respect to a query represent a change in the knowledge state of the user and this change is reflected when a search session terminates. Search query and results will be updated based on this change when the search session resumes.

3.1 Experimental Method

Suppose a user issues a *query* to a retrieval system. This action marks the start of a *search session* from which an aspect of the user's evolving information needs can be captured. Search sessions are represented as date-ordered collections of explicitly relevant search results, from either personal information agent results or from search query results, alongside the terms extracted from the text content of those results using a *tf-idf weighting* on the set of search results returned for that query.

Clusters of search sessions represent an *aspect* of the user's *interests*. An interest represents the history of an aspect of the user's evolving information needs. The relevance of the terms that define interests is calculated using an increasing-uncertainty-with-age discounting function [7]. This type of discounting function promotes the significance of the terms that define interactions in the most recent search sessions.

The five terms with the highest weight extracted from the clustered search sessions are the defining terms of an interest in the user profile. These defining terms represent the evolving needs of the user based on their explicit interactions with search results.

3.2 Baseline Method

The baseline context modelling method differs from the experimental method in how the defining terms for interests are obtained. The defining terms of each interest in the user profile are the five most frequently occurring terms used in queries that define the clustered search sessions. The baseline defining query represents the evolving information need of the user but it does not consider the explicit interactions performed by the user after they have issued a search query.

4. EVALUATION TECHNIQUE

Evaluating a contextual retrieval system presents challenges which are not appropriately covered using traditional measures such as precision and recall [1]. In Web search tasks there is rarely an available set of relevance assessments for a task and traditional search tasks are unsuitable to evaluate a system that attempts to model evolving information needs. We propose to use two simulated work situation tasks [8] to evaluate a system that captures a multiple-aspect user profile because we believe this type of task will allow the user more freedom in their searching.

ID	Question
A	The incoming results were useful
B	The system successfully modeled my search sessions
C	The initial defining terms were accurate
D	The suggested defining terms were useful
E	I regularly changed the defining terms of my profile

Table 1: User questionnaire differentials

4.1 Hypotheses

We have the following hypotheses from our research questions:

1. Modelling the interactions of users through explicit feedback will assist in the construction of user profiles that represent multiple-aspect of evolving information needs.
2. A Personal Information Agent based on a multiple-aspect user profile will retrieve useful documents on behalf of users.

4.2 Measurements

To test the first hypothesis, we will measure the number of manual changes performed by each user to their profile. We hypothesise that the experimental system will show a trend of fewer changes to the user profile if the profile accurately represents their needs. From the user questionnaire differentials shown in Table 1, we will use differentials B, C, D, and E. We believe the responses to these questions will help us evaluate participants' impressions of the context-modelling component of the systems.

To test the second hypothesis, we will measure click-through and explicit marking rates recorded from the interface. We hypothesise that the experimental system will show a trend of higher click-through and higher explicit marking than the baseline system. From the user questionnaire in Table 1, we will use differential A. We believe a trend of support will be present for this question if participants find the results useful.

4.3 Tasks

The tasks associated with traditional test collections such a TREC[2] and Cranfield [1] are too narrow in scope to evaluate an adaptive retrieval system. These tasks assume users have static and narrow information needs, for example TREC-9 Web Topics 451[3]: "Provide information on the Bengal cat breed." To address this problem, we have devised two simulated work tasks, of which Task B is shown in Table 2. The purpose of these tasks is to simulate a reasonably complex information need that can be divided into several subtasks and addressed in an order that suits the user's searching style.

4.4 Methodology

Four participants were recruited from within the Computing Science Department, their ages between 18 and 45. They were required to complete two simulated work situation tasks, over three 15 minute sessions for each work task, on three different days and were paid £25 upon completion of the experiment.

Each participant used both the baseline and experimental system during the experiment. The allocation of systems and tasks were rotated in a Latin square to attempt to minimise learning effects. Participants had no knowledge of whether they were using the experimental or baseline system for either task; the interface for the systems provided labels of System A and System B. With the exception of the labelling of each system, the baseline and experimental system interfaces presented to users were identical.

Imagine you are journalist writing an article on the sales performance of shops on the High Street in the United Kingdom over the Christmas 2008 period. Your article will be published in a major newspaper and read by hundreds of thousands of people. You will want to include a short-guide to the current financial crisis to set the context for your article, some well-sourced material on the affect of the crisis on Christmas sales, and what expectations are for the 1st quarter of 2009 on the High Street.

Table 2: Task description for simulated work situation task B.

In response to an entry questionnaire, all participants were educated to at least Master's degree level, were extremely confident using computers to complete everyday tasks, used Web search engines every day, and mostly found what they wanted using Web search engines.

Before commencing the experiment, participants were given a 10-minute training task to familiarise them with the features of the retrieval system and were allowed to ask questions. After completing the 1st session for both tasks, participants were asked to complete an interim questionnaire. Participants were able to complete the second and third sessions unsupervised from their own computer. After the 3rd session for both tasks, participants completed a post-task questionnaire; and finally, the participants completed an exit questionnaire.

5. RESULTS AND DISCUSSION

In this section we present preliminary results from our user study. Figure 3 shows the mean intermediate and post-task responses to the user questionnaire described in Table 1. The responses range from 1-5 on a Likert scale, where 1 represents a negative response and 5 represents a positive response, with the exception of differential E which has a reversed scale.

5.1 Context-modelling Method

The result of the log file analysis, shown in Figure 4, appears to contradict our initial hypothesis. In the experimental system each user made additions to the terms defining aspects of their user profile, however, insufficient data was captured to determine if the additions made by users to their user profile in the experimental system were a result of the term suggestions. Responses to questionnaire differentials B, C, and D in Figure 3 present little difference between the baseline and experimental systems.

5.2 Personal Information Agent

The results of the log file analysis, shown in Figure 5, show that three out of four participants marked more personal information agent results as relevant in the experimental system. For questionnaire differential A, in Figure 3, there is little different in responses at between the baseline and experimental systems.

[2] http://trec.nist.gov/

[3] http://trec.nist.gov/data/topics_eng/index.html

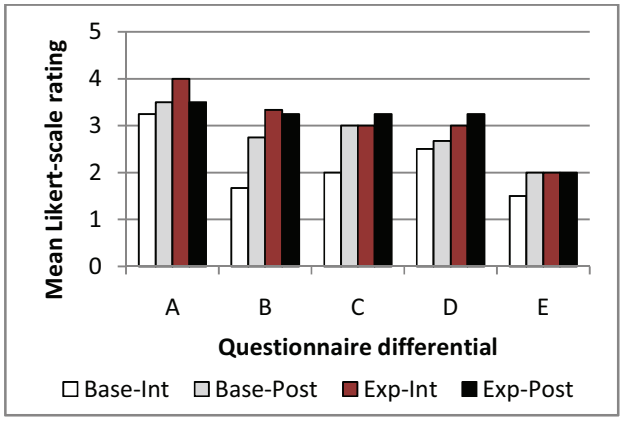

Figure 3: Graph of questionnaire responses and interim stage and post-task stage on differentials A –E from Table 1.

6. CONCLUSION AND FUTURE WORK

The preliminary results presented in this paper support the utility of a contextual retrieval system that assists users with their long-term information needs. The support is weak, however, and further studies with a refined evaluation design and more users are required to draw conclusions.

We believe the difficulty of designing an evaluation technique for a multi-aspect personalised retrieval system attributed to these results. In the future, we plan to redesign the evaluation technique presented in this paper. Evaluating a retrieval system across multiple search sessions poses several problems, including how to ensure the tasks assigned to users are complex enough to warrant multiple sessions, whether to allow users the freedom to complete the tasks unsupervised, and which measures will appropriately evaluate the performance of such a retrieval system.

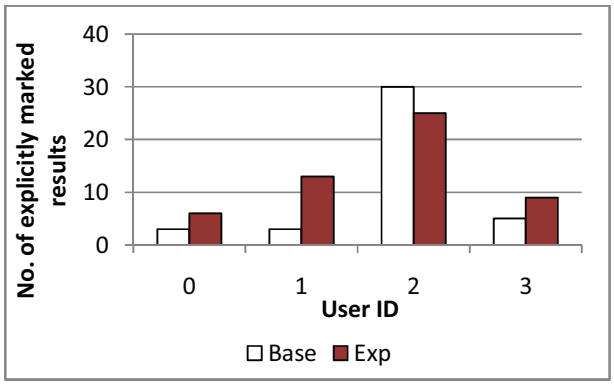

Figure 4: Graph of number of results retrieved by each user's PIA marked as explicitly relevant over the experiment.

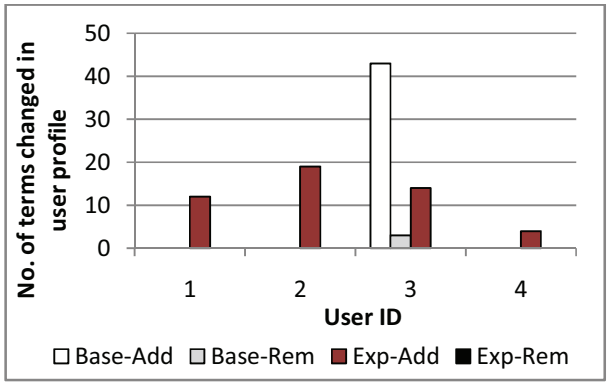

Figure 5: Graph of the number of additions and removals made to the defining terms of aspects of each user's profile.

7. REFERENCES

[1] Cleverdon, C.W., and Keen, M. 1968. Factors determining the performance of indexing systems, Vol. 1: Design, Vol. 2: Test Results. Aslib Cranfield Research Project, Cranfield, Bedford, England.

[2] Rocchio, J.J. 1971. Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton, Eds. Prentice-Hall, Englewood Cliffs, NJ, USA, 313-323.

[3] Harper, D.J., and Kelly, D. 2006. Contextual Relevance Feedback. In IIiX: Proceedings of the 1st International Conference on Information Interaction in Context, Copenhagen, Denmark, 129-137.

[4] Martin, I., and Jose, J. 2003. A Personalised Information Retrieval Tool. In SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 423-424.

[5] Psarras, I., and Jose, J. 2006. A System for Adaptive Information Retrieval. In Proceedings of 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland, 313-317.

[6] Campbell, I., and van Rijsbergen, K. 1996. The Ostensive Model of Developing Information Needs. In Proceedings of CoLIS-2, Copenhagen, Denmark, 251-263.

[7] Campbell, I. 2000. Interactive Evaluation of the Ostensive Model using a New Test-Collection of Images with Multiple Relevance Assessments. Journal of Information Retrieval 2, 89-114.

[8] Borlund, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56, 71-90.

Topic template queries to enhance document retrieval

Antonio Jimeno-Yepes
yepes@ebi.ac.uk
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, UK

Rafael Berlanga-Llavori
berlanga@lsi.uji.es
Dept. of Computer Systems and Languages
Universitat Jaume I, Spain

Dietrich Rebholz-Schuhmann
rebholz@ebi.ac.uk
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, UK

ABSTRACT

Topic template queries are focused on a facet of a user information need. The study of explicit feedback may provide an improved retrieval on new queries. We have analyzed two solutions that integrate the analysis of existing results based on query reformulation and the boosting of documents based on text categorization. Preliminary results show that both approaches produce interesting results when enough example queries are provided.

1. INTRODUCTION

In our work we are interested in topic template queries (TTQ). These queries are defined by a theme or subject which denotes a specific facet of related types of entities (e.g. the role of gene X in disease Y). Several instantiations of the templates are possible (e.g. the role of the *APC gene* in *colon cancer*). This setup might be useful for researchers in the biomedical domain which are in charge of curating a database.

In information retrieval (IR) we identify several tasks like ad-hoc information retrieval and text categorization. In ad-hoc information retrieval (AIR) the information need is variable and the document collection is fixed; even though the document collection content is updated over time. In text categorization (TC), the information need is fixed but there is a document stream instead of a document collection.

On the one hand, ad-hoc information retrieval is not restricted to a specific topic template. On the other hand, multiple instances of the same template can be generated, thus different from text categorization. In our work we present two approaches to deal with topic template queries that combine existing AIR and TC approaches.

In the next section we present the related work. Thereafter we introduce the methods we propose. Finally, we present the results and conclusions.

2. RELATED WORK

Our work is related to several approaches in IR that we briefly present in this section. We can find techniques like scatter/gather where the retrieved documents are organized according to a clustering of the documents. This means a post-processing of the search results which allows the user to identify interesting groups of documents [6]. However, the groups identified by this technique may not satisfy the interest of the user.

Techniques like query reformulation might provide a better representation of the original user query. Relevance feedback is a well known technique in which the user runs a query and selects some of the documents that considers relevant and a new query is produced based on this selection [2].

Text classifiers build models given example documents for predefined categories (which represent a static information need). Even though the former text categorizers were built manually, automatic techniques such ass query expansion and machine learning approaches [12] are commonly used nowadays.

Despite the variety of techniques for improving query results, there is no method that analyses the set of queries from a topic template to identify commonalities that might improve the performance of unseen queries for a topic template.

3. METHODS FOR TOPIC FEATURE SELECTION

In this paper we compare two approaches than analyze explicit feedback in retrieval tasks for a set of queries and produce a model that improves the performance on unseen queries with the same topic template. The first one is based on a combination of text categorization and ad-hoc information retrieval. Retrieved documents are boosted according to a text categorizer that determines the relevance of the document to the topic template. The second one is a query reformulation approach. The query is complemented with terms related to the topic template obtained after the analysis of relevant and irrelevant documents.

3.1 Text categorization

A text classifier is applied on the retrieved document list for a given query and is used to boost documents that are deemed relevant. These documents are boosted to the top of the retrieved list keeping the original rank among them. This is similar to the work of Ruch and Geissbühler [11] which combines a traditional vector space model result and a rule based system.

The text categorizers are trained using provided feedback; e.g. explicit feedback provided by users. Several categorizers based on machine learning algorithms [5] with different learning bias have been compared: decision trees (J48), naïve bayes (NB), support vector machines (SVM) and k-nearest neighbors (K-NN). A cross-validation analysis is used to select the most adequate classifier for the task.

3.2 Query reformulation

The query reformulation used in the experiments is based on the query reformulation we proposed in [8]. We have modified our Ontology Query Model (OQM) to integrate the terms denoting the topic template.

3.2.1 Ontology Query Model

The main aim of the ontology query model (OQM) [8] is to produce an IR query from a set of concepts $\mathcal{C}$ selected by a user browsing the ontology. We start from the set of words provided by the lexicon of the ontology, denoted by $LexW(T)$; where T is the set of terms. This means that the term *breast cancer* in T will be represented as the words *breast* and *cancer* in $LexW$. The terms are grouped in synsets $LexT$; e.g. *breast cancer* is placed in the same synset as *mammary carcinoma*. A given synset is linked to a concept in the ontology.

In the OQM, we need to estimate $P(w_i|\mathcal{C})$, that is, the probability of generating the word w_i given a set of concepts $\mathcal{C}$. In other words, we want to estimate the probability of choosing the word w_i when expressing the concepts in $\mathcal{C}$ in written documents. The model of $\mathcal{C}$ is then compared to each document model D from the collection using cross-entropy to rank the documents. The document model is represented by the Jelinek-Mercer smoothed probability of a word in a document and the probability of the word in a background document collection G.

The relation between the concepts has to be considered in the query model. As the document model is built based on a bag of words instead of multi-word terms as those linked to the concepts, so the terms have to be linked back to the individual words they are composed of. We propose the estimation of the $P(w_i|\mathcal{C})$ as a smoothed version of the concept model P_{CM} using the expansion P_R based on related concepts:

$$P(w_i|\mathcal{C}) = \lambda_r P_{CM}(w_i|\mathcal{C}) + (1-\lambda_r)P_R(w_i|\mathcal{C}) \quad (1)$$

3.3 Query Reformulation Approach

The relation $\mathcal{R}$ that defines the topic of the query is now part of the conceptual query. A linear combination of the related terms linked to the concepts and the relation terms is used:

$$P(w_i|\mathcal{C},\mathcal{R}) = \alpha P_{CM}(w_i|\mathcal{C}) + \beta P_R(w_i|\mathcal{C}) + \gamma P_{Rel}(w_i|\mathcal{R}) \quad (2)$$

$$\alpha + \beta + \gamma = 1 \quad (3)$$

$P_{Rel}(w_i|\mathcal{R})$ depends only on the terminology linked to this relation and the terminology linked to other relations in the ontology. In the case of a richer relation ontology the probability would as well consider the occurrence of the terms in the other relations.

The terms are collected from the tokenized documents in the training data set and cleaned from stop word terms. Information Gain (IG) measures the reduction in entropy and is chosen to rank the features. The entropy of a random variable X ($H(X)$) indicates the smallest number of bits needed on average to send a message from a stream of symbols drawn from X.

$$H(X) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (4)$$

Then, the information gain from the training examples T for the attribute a considers the entropy of the training examples and the conditional entropy of the attribute. In the estimation we will consider the different values of the class attribute to determine information gain of the attribute.

$$IG(T|a) = H(T) - H(T|a) \quad (5)$$

4. RESULTS

We have used the Lemur package to set up the experiments. A standard stop word list and the Krovetz stemmer are used. The parameters for the different models have been chosen empirically. The lambda values for the ontology query model (OQM) [8] have been set to 0.6. The randomization test for paired data is used to compare the methods († indicates $p < 0.01$).

The training set is based on the training queries that are used to retrieve the top-50 documents for each query. The positive documents are the documents relevant for the query while the non-relevant documents are considered as negatives. Since the number of negative documents overwhelms the number of positive ones, a random selection of negative documents is done to balance both classes. 5 times 2 fold cross validation [4] is used to sample the set of queries for each data set. The results are an average of the results obtained for each one of the partitions.

4.1 Datasets

We have used two datasets for our experiments. One considers the role of a gene in a disease and the second one the interaction of proteins. These two datasets are presented in turn.

4.1.1 PGN-disease data set

We have used the 2005 TREC Genomics collection because there is an interest on generic topics. This collection is made up of a subcollection of Medline, around 4M documents between years 1999 and 2004, and a collection of 50 queries. The queries in this benchmark are categorized into five groups defined by generic topic template (GTT) [1]. We have selected a set of queries which relate PGNs (proteins

[1] http://ir.ohsu.edu/genomics/2005protocol.html

and genes) to diseases. Queries are based on a topic template; i.e. the role of gene X in disease Y. From the TREC queries we have considered 20 queries related to the topic template.

4.1.2 PPI data set

Protein-protein interaction (PPI) databases rely on either experimental data or hand-curated analysis from the literature. Usually these systems rely on the retrieval of documents and on the extraction of relevant information in PPI systems (e.g. BioCreative II [2]). This task is relevant in the biomedical domain.

We would like to select relevant documents that indicate the relation between two proteins and are useful for curation. The DIP database [3] deals with protein-protein interaction on yeast and has pointers to Medline articles. In this database the yeast species has been more carefully curated than any other species.

In total 260 queries are prepared. The average number of relevant documents per query is two. The number of queries is larger than in our previous benchmark, so more significant results might be found. The relevance assessment is done based on documents collected for each one of the interacting proteins. The main difference is that the analysis is done based on the full text of the documents instead of abstracts. This means that some documents have been annotated with many interactions, which indicates that a high-throughput method has been used in the reported experiments, and it is unlikely that relevant information is contained in the abstract. These documents are discarded from the set of relevant documents. The document collection contains Medline citations till September 2004, about 15M Medline documents.

4.2 Identification of topical features results

4.2.1 PGN-disease dataset results

The configuration for the PGN-disease dataset is based on the results obtained from the relevance cleaning and refinement presented in [7] and [8]. From this set, the positive and negative documents are obtained and the dataset for the classification and feature selection processes is obtained. In Table 1 we can see the result obtained with the classifiers. The documents have been tokenized, stop words have been removed and no stemming is applied. The Naive Bayes classifier obtains the best result.

Algorithm	Precision	Recall	F-measure
J48	0.4429	0.4055	0.3992
NB	0.5624	**0.6457**	**0.5934**
SVM	**0.6235**	0.3452	0.4424
K-NN1	0.2231	0.0254	0.0454

Table 1: Document categorization results for PGN

In Table 2 we find the most relevant features in every fold. These features do not seem to be related in a general way to the relation denoted by the topic template. We have to remember that this dataset contains a small number of queries and the relevant documents vary according to the query.

[2]http://biocreative.sourceforge.net/bc2ws/index.html
[3]http://dip.doe-mbi.ucla.edu/

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
disease	gstm1	gstm1	apc	disease
mutation	polymorp...	genotype	polyposis	transforming
129	glutathione	null	adenomatous	mutations
onset	0	polymorp...	coli	major
mutations	genotype	0	gstm1	familial
alzheimer	null	cases	gene	alzheimer
lines	study	gluthatione	familial	genetic
familial	cases	genes	colon	onset
bard1	genotypes	controls	polymorp...	patients
allele	transferase	allele	0	linked

Table 2: Feature selection for PGN-disease (polymorp... stands for polymorphisms)

In Table 3 and Figure 1 we compare the baseline based on the categorization. As we can see, the best results are obtained with the baseline. This was expected since the classifiers presented above have a poor performance. This is because the training set does not allow finding a model that discards documents about the role of the PGN in the disease.

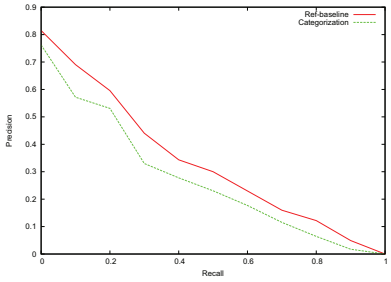

Figure 1: Precision-recall curve relation refinement for PGN-disease

TREC	Rel. Retr	MAP	R Prec
Baseline	747.2/1093.6	**0.3208**	**0.3608**
Categorizer	747.2/1093.6	0.2604	0.3033

Table 3: Refinement cleaning and categorization for PGN-disease

4.2.2 PPI data set results

Table 4 shows the results for document categorization. In contrast to the results for the PGN-disease data set we see that the classifiers have a better performance. Again, the documents are tokenized, stop words are removed and tokens are not stemmed. The SVM classifier obtains the best F-measure result in the cross-validation analysis.

In Table 6 we identify the tokens that are ranked by information gain in each fold. In contrast to the PGN-disease set, the tokens are more homogeneous in the different sets. From the list of terms, there are terms that clearly denote an interaction like *interaction, binding, complex and hybrid*, terms that are related to experiments done to verify the interaction between proteins. These terms have been found relevant in a similar study by Marcotte et al. [9] and Cohen et al. [3]. There are less obvious terms like *association* that have been found relevant in Rebholz et al. [10]. Almost all these features seem to be linked to the positive class, meaning as well that the features denoting other topics are more difficult to identify and the sub-topic analysis for this set may require more data than we have used.

Algorithm	Precision	Recall	F-measure
J48	0.7413	0.7828	0.7605
NB	0.6827	**0.9906**	0.8078
SVM	0.7953	0.8483	**0.8202**
K-NN1	**0.9611**	0.1558	0.2663

Table 4: Document categorization results for PPI

In Table 5 and Figure 2 we present the result comparing the baseline methods with the modified ontology query model. The baseline methods are the co-occurrences based on cleaning and refinement [7]. We have used the trained SVM model to perform the boosting of documents due to its performance in document categorization as found in Table 4.

As we can see in Table 5 and Figure 2, both approaches perform better than the baseline. Boosting based on the text categorizer provides a better performance. This means that there are specific arrangements that the model produced by the SVM captures better than the ontology query model.

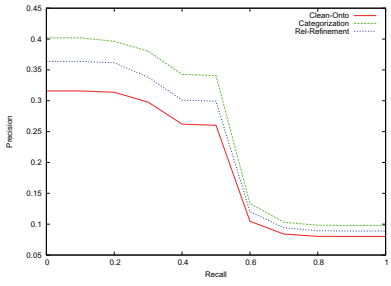

Figure 2: Precision-recall curve relation refinement for PPI

PPI	Rel. Retr	MAP	R Prec
Baseline	189.2/317.2	0.1873	0.1534
Categorizer	189.2/317.2	**0.2387**†	**0.1993**†
Refinement	**199.2/317.2**	0.2140 †	0.1926 †

Table 5: Baseline, categorizer and refinement for PPI

5. CONCLUSIONS

We have investigated into the topic template queries using two different techniques; a combination of ad-hoc information retrieval and text categorization and a query reformulation approach. We have seen that with enough training data we can effectively target relevant documents and identify terms that are typically used to identify the topic.

Furthermore, other topics could have been detected that are not related to the query and are being used to discard some of the non-relevant documents. This work requires the discovery of the topics existing in the document set [1]. In addition, proposals exist that integrate several topics based on language modeling [13].

6. REFERENCES

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
protein	interaction	interaction	protein	interaction
essential	protein	complex	interaction	essential
interaction	binding	protein	interactions	required
hybrid	hybrid	binding	binding	protein
proteins	proteins	hybrid	proteins	proteins
interacts	complex	interacts	association	complex
complex	vitro	required	vivo	binding
binding	vivo	vivo	domain	vivo
show	essential	association	required	hybrid
domain	required	proteins	complex	vitro

Table 6: Feature selection for PPI

[1] R. Berlanga-Llavori, H. Anaya-Sánchez, A. Pons-Porrata, and E. Jiménez-Ruiz. Conceptual subtopic identification in the medical domain. In H. Geffner, R. Prada, I. M. Alexandre, and N. David, editors, *IBERAMIA*, volume 5290 of *Lecture Notes in Computer Science*, pages 312–321. Springer, 2008.

[2] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *Text REtrieval Conference*, 1994.

[3] K. Cohen, M. Palmer, and L. Hunter. Nominalization and Alternations in Biomedical Language. *PLoS ONE*, 3(9), 2008.

[4] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[5] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. Witten, and L. Trigg. Weka-a machine learning workbench for data mining. *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314, 2005.

[6] M. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *SIGIR*, pages 76–84, 1996.

[7] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Ontology refinement for improved information retrieval. *Information Processing & Management: Special Issue on Semantic Annotations in Information Retrieval (submitted)*, 2009.

[8] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Terminological cleansing for improved information retrieval based on ontological terms. In *Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 6–14. ACM, 2009.

[9] E. Marcotte, I. Xenarios, and D. Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, 2001.

[10] D. Rebholz-Schuhmann, A. Jimeno-Yepes, M. Arregui, and H. Kirsch. Assessment of Modifying versus Non-modifying Protein Interactions. In *The Third International Symposium on Semantic Mining in Biomedicine*, 2008.

[11] P. Ruch, R. Baud, and A. Geissbühler. Learning-free text categorization. *9th Conference on Artificial Intelligence, in Medicine in Europe*, 2003.

[12] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[13] X. Wang and C. Zhai. A study of methods for negative relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 219–226. ACM New York, NY, USA, 2008.

Benchmark evaluation of context-aware Web search

Davide Menegon, Stefano Mizzaro, Elena Nazzi, Luca Vassena
Dept. of Mathematics and Computer Science — University of Udine
via delle Scienze, 206
Udine, Italy
mizzaro,menegon,nazzi,vassena@dimi.uniud.it

ABSTRACT

We discuss the issue of evaluating highly interactive and novel context-aware system with a methodology based on a TREC-like benchmark. Taking as a case study an application for Web content perusal by means of context-aware mobile devices, we present our approach to early stage evaluation, describing our aims and the techniques we apply. Besides presenting the results obtained, we discuss if our benchmark methodology can be an extensible and reliable tool for the evaluation of context-aware retrieval systems.

1. INTRODUCTION

Dynamism and evolving situations have become the central elements of the environment where Information Retrieval (IR) is asked to operate. The widespread diffusion of mobile devices and, with them, of real-world mobile users, have moved the static world of classical and Web IR towards an always changing context-based world. So, the notion of *context* (roughly described as the situation the user is in), and the information it conveys, are gaining increasing importance for the development of new IR systems. Context-awareness drives to a dynamic nature of the user needs, of the information available, and of the relevance of this information.

When combined with context-awareness, IR has been named Context-Aware Retrieval (CAR) [1]. Starting from considering only a low number of contextual features (location and time), current CAR systems entail such an amount of data that a new challenge for IR is how those data can enhance user satisfaction. How to evaluate the strategies and techniques that CAR systems use for this purpose is another challenge. Although CAR systems imply a high amount of interactivity with the user, and a user study seems the most sensible approach, our approach is to evaluate highly interactive and novel context-aware systems with a TREC-like benchmark based methodology. This paper presents the results obtained and discusses the methodology adopted.

We first briefly survey evaluation methodologies in IR and in CAR systems (Sect. 2), introducing our case study application. We then present our early evaluation approach (Sect. 3), describing aims, techniques, and results. Finally a discussion on our approach reliability and its usefulness is given in Sect. 4, while in Sect. 5 we draw some conclusions and present future work.

2. RELATED WORK

2.1 Context-aware retrieval and evaluation

With the spread of the concepts related to context-aware computing, IR has gained new and increasing importance. The newborn field of CAR, instead of concentrating only on topicality, encompasses contextual information into the retrieval process, aiming at discovering *"the query behind the context"*: to retrieve what the user needs, even if he/she did not issue any query [9].

In the development of a CAR system for mobile environments, evaluation plays an important role, as it allows to measure the effectiveness of the system and to better understand problems from both the system and the user interaction point of view. In [10, 3] the challenges for context-aware computing evaluation are illustrated: the need to identify meaningful metrics (each single application has peculiar aspects to be evaluated that differ from other types of applications); the difficulty to evaluate in small scale a system that is meant to be adopted by many users; the difficulties in explicit testing a novel system, meant to be integrated in everyday life and thus invisible (as it is aimed in the ubiquitous computing field) [13].

Depending on resources and aims, different evaluation metrics are chosen and different evaluation approaches are adopted (e.g., user studies and benchmark; lightweight and heavyweight [8]; etc.). As CAR applications are strictly related to users, the user-centered evaluation (live or in laboratory) seems the most natural one. In [12], for example, a framework for user evaluations of ubiquitous computing has been proposed. A detailed work about metrics for evaluating systems for information access has been done in [11]. More recently, in AmbieSense [7], a user centered, iterative, and progressive evaluation has been adopted combining IR evaluation methods with human-computer interaction development techniques.

Another largely adopted evaluation methodology is benchmark evaluation. Taking example from TREC (`trec.nist.gov/`) initiative for large scale evaluation of retrieval methodologies, benchmarks are defined by a collection of topics (expressions of user's information needs), a collection of resources to retrieve, and a set of relevance judgments about those resources for each of the topics. Within the CAR field, a benchmark evaluation has been used in our previous work [9]: we adopted a benchmark named MREC (MoBe Retrieval Evaluation Collection) to evaluate MoBe [2] a framework for CAR of applications. Since several alternative implementations for retrieval strategies were possible, it was important to compare their effectiveness. Besides, due to the still initial stage of development, it was not sensible to invest too many resources to get detailed results and even a rough and quick evaluation was useful.

Although the user evaluation approach is usually considered the

<contextDescriptor>
<title> Heathrow airport </title>
<description>
The user has just landed at London Heathrow international airport. He is looking at a flight timetable and at a timetable for connections to London. It is lunch time.
</description>
<narrative> . . . </narrative>
<relevance>
A Web page is relevant: it contains information about a flight, about the means of transport to reach town, about bars and fastfoods in the airport, or it allows to book a flight. A Web page that contains only one of these aspects is relevant; if it contains some links to relevant pages is partially relevant. If the judge is not able, for any reason, to judge the page, its value is "I don't know".
</relevance>
. . .
</contextDescriptor>

Figure 1: A (part of a) context descriptor.

most appropriate for CAR applications, it also has some, not negligible, drawbacks when compared with benchmarks. First, it is a high level evaluation, where the primary interest is to study how the system satisfies the user, and on how the user evaluates the system, rather than on how the system serves the information needs of its users. Benchmark evaluations, on the contrary, are system centered: they directly focus on the evaluation of implementation details, and they can evaluate different aspects of the retrieval process. Second, a mature prototype to test has to be implemented, including an effective user interface. This goes against the purposes of developers, as it forces significant implementation decisions uncovered by evaluation. Third, user evaluation is more complex to perform, requires more time than benchmarks, and it is more dependent on users' subjectivity. To summarize, both user evaluation and benchmark approaches have pro and contra, and it is not the case that one is better than the other.

2.2 The Context-Aware Browser

The Context Aware Browser (CAB) [4] is a general-purpose solution to Web content fruition by means of context-aware mobile devices. The main idea behind CAB is to empower a generic mobile device with a browser able to automatically and dynamically retrieve and load Web pages, services, and applications according to the current context. CAB extends MoBe: whereas in MoBe the resources being retrieved are Java Micro Edition applications, CAB works on Web contents, exploiting a Web browser, a daily tool used by most people.CAB acquires information related to the user and the surrounding environment by means of both sensors installed on the device and external servers. This information, combined with the user's personal history and the community behavior, is exploited to infer the user's current context, that is represented by a terms list. In the subsequent retrieval process, starting from this terms list, a query is automatically built and sent to an external search engine, in order to find the most suitable Web pages for the sensed context. In this paper we study different strategies to build in an automatic way the query sent to the external search engine.

3. EXPERIMENTAL EVALUATION

3.1 CREC: an incremental benchmark

The CAB application development stage is in its intermediary phase: a large part of its components are available, but the retrieval mechanism needs a more accurate implementation. An early evaluation of the strategies we would like to implement is needed, so, considering the successful approach of MREC and the useful insights obtained, we adopted again a TREC-like benchmark evaluation, named *CREC (CAB Retrieval Evaluation Collection)*. CREC focuses on the retrieval of Web pages and starts from the results obtained with MREC.

We were interested in the following questions: how to build in an automatic way the queries to be sent to the external engine? Which is the best strategy in terms of effectiveness? How does the retrieval effectiveness change on the basis of the increase/decrease of the number of terms in the query, or of different kinds of terms exploited in the query? How effective is an automatic query formulation when compared to a user manual search?

CREC is constituted, as usual, by three components: the statements of information needs, a collection of documents, and a set of relevance judgments. The statements of information needs are defined by context descriptors, which represent different examples of user's contexts in different domains (Fig. 1). CREC includes 10 context descriptors which differ for user activities, location, time, etc. The context descriptors have been designed in a similar way to the topics in TREC.

The documents collection consists of Web pages. The relevance judgments have been made by a unique judge using a four level relevance scale: relevant, partially relevant, not classified, or not relevant. The judgement operation is helped by the `<narrative>` and `<relevance>` fields in the context descriptor.

Due to the evolving behavior of the Web (pages are dynamically added, removed, or modified), we opted for a dynamic collection, that evolves during the tests. Moreover, and more importantly, if a new implementation of the CAB external search engine needs to be evaluated, CREC will not contain, in general, all the retrieved pages. Since this would make the evaluation not reliable, our approach is that the collection is not static: it will be extended by including the newly retrieved documents, and judging them.

We built two CREC versions so far. The first version has been constructed manually: for each context descriptor, a human operator created 5 queries to obtain the needed information in that context, as if requested to do it manually. We then collected the first retrieved documents, obtaining about 150 single documents per context. At this stage the collection was composed by 1229 pages (304 relevant, 211 partially relevant, 30 not classified and 684 not relevant). Starting from this first version of the collection, we adopted an "interactive search and judge" [5] approach to add more relevant documents. In particular we ran some queries, automatically built from context descriptors, and the first 10 retrieved Web pages, for each query, have been added to the collection and judged, obtaining its second version of 3634 pages (494 relevant, 596 partially relevant, 34 not classified and 2510 not relevant). The high number of relevant documents (contrary to a real settings where the number of relevant documents is much lower than the number of documents in the collection) is due to the fact that we searched for relevant documents, and it is not a problem since we assume that unjudged documents are not relevant. Moreover we are not interested in all the documents, but just in the first retrieved ones.

3.2 Strategies

We designed four automatic query construction strategies. These strategies do not exploit the whole context descriptor (as done by the human relevance judge), but work on a simple terms list, following the context representation exploited in the CAB. The terms list is automatically extracted from the `<description>` field by the usual stopword removal: for instance, the context descriptor in Fig. 1 is represented by the list "user just landed london heathrow

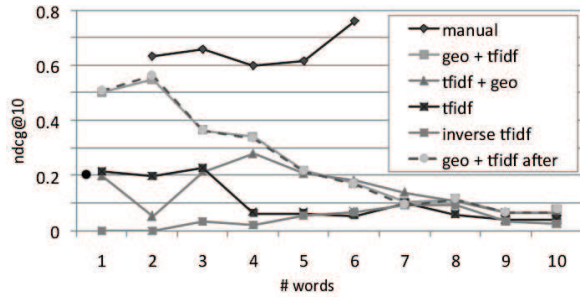

Figure 2: nDCG@10 for the five strategies on all contexts.

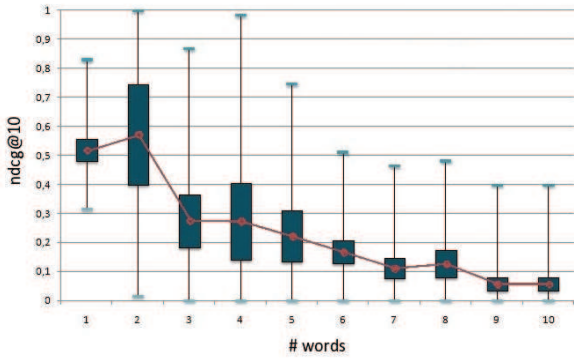

Figure 3: Detailed results for the *geo+tf.idf* strategy.

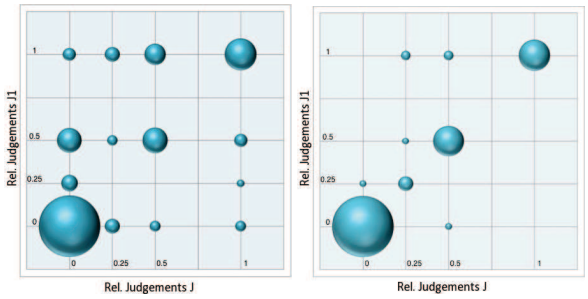

Figure 4: ADM values for the experiment involving J1.

international airport looking flight timetable timetable connections london lunch time". Also, strategies are based on two main indexes: *tf.idf* and *geoterms* (i.e., terms that refer to geographical information). We chose tf.idf as it is a classical, well known, and largely used metric, while we chose geoterms as location is the contextual dimension that probably is more informative of user's current context. These indexes, differently combined, are used to rank the term lists according to their importance (that will be different for the four strategies). For each strategy and context descriptor, 10 queries of different length (from 1 to 10 terms) are formulated, incrementally selecting the first 10 terms of the ranked lists. The manual approach, where a mobile user directly chooses terms and defines her query, is our upper reference strategy. In particular, for each context descriptor, 5 queries have been defined by a human operator. The strategies have been implemented in Java using Yahoo! as external search engine through the API provided (`http://developer.yahoo.com/search/web/`).

We measure strategies' effectiveness by means of a standard IR metric, nDCG@10, chosen as it emphasizes quality at the top of the ranked list. Moreover, nDCG@10 takes into consideration only the first 10 retrieved items, which is reasonable for CAB, since the user is unlikely to scroll long lists of retrieved items.

3.3 Results

Fig. 2 compares the four strategies, and the manual one, showing their effectiveness (nDCG@10, on the Y axis), averaged on all 10 contexts, for different query lengths (on the X axis). Apart from the manual one, the most effective strategy is the *geo+tf.idf*; Fig. 3 shows, beside its average, also min, max values and the variance. In this strategy, first all the geoterms are added to the query, then the other terms, ranked by decreasing tf.idf, follow. Further analysis of the data, not reported here, shows that the maximum performance is obtained when one tf.idf term is added after all the geoterms (each context contains 1, 2, or 3 geoterms), then nDCG@10 decreases. In general, the performance of long queries is very low.

All the proposed strategies have lower performances than the manual one (higher curve in Fig. 2), therefore there is still space for improvement. Moreover, the manual strategy has a different behavior: its performance tends to increase with query length. If performed by a human, a search session could improve with query refinements that add or change terms on the basis of the knowledge acquired by visualizing results. This is not performed by automatic strategies that construct the query incrementally. This is probably the reason why manual strategy has a performance improvement the more terms are used. Moreover, it is usually the case that CREC contexts are made up by more than one facet; this could be another reason for effectiveness degradation with query length.

4. RELIABILITY AND EXTENSIBILITY

We now turn to discussing the reliability of our benchmark, on the basis of three questions.

Does judge subjectivity invalidate the reliability of our benchmark? Relevance judgments have been made by a unique judge (henceforth denoted by J), using a four level scale. To verify that results would not change with a different judge, we performed an additional experiment, involving two more judges (J1 and J2), to measure inter-judge agreement. Given one context descriptor, each judge judged the first 30 pages retrieved by each strategy. The judgments by J and J1 were the same on 62% of the pages. The agreement between J and J2 is a slightly higher 69%. After a discussion between judges on the pages judged in a different way, the agreement grew to 90% and 95%, respectively.

In order to better understand the different judgments among judges, we also used the ADM measure [6]. The values in the relevance scale, relevant, partially relevant, not classified, and not relevant, correspond respectively to the values 1, 0.5, 0.25, 0. We obtained ADM values of 0.757 and 0.846 before the discussion, and 0.950 and 0.981 after the discussion. See Fig. 4 for a graphic representation. The X axis is J, the Y axis is J1; the figure on the left is before discussion, that on the right is after discussion. Therefore, subjectivity of the judge does not seem an issue for our benchmark, that seems reliable at least to a reasonable extent.

Can our incremental approach to collection construction cope with Web dynamics? The secondary judges J1 and J2 also manually formulated 3 different queries and judged the resulting pages. We then counted the pages retrieved by the secondary judges but not retrieved by the automatically constructed queries, and the relevant pages. J1 retrieved 63 new and never judged pages (21 relevant), while J2 retrieved 56 (34 relevant). This shows that new pages have appeared on the Web at a fast pace or, more probably, the initial collection was not complete.

However, the results shown in the previous section might not be influenced by these new pages. Indeed, measuring the effectiveness

of automatic strategies over the first 10 results, we notice that these new results usually are not in the first 10 positions of their ranked lists: automatic strategies effectiveness, measured with nDCG@10, is very similar, since the strategies retrieve documents already in the initial collection. The dashed line in Fig. 2 shows the effectiveness of geo+tf.idf strategy considering the new pages and judgments: it does not change significantly.

How much effort is needed to create and update the collection? As above said, we built two versions of the collection; to understand the effort necessary to build and extend it, we performed an additional experiment. We selected a context descriptor (not used in the benchmark reliability experiments) and two judges (J and a new J3) built a manual query and evaluated the retrieved pages (36 pages for person). While J3 had never done this experiment before, J is an expert of both the domain, as he was the designer of the contexts descriptors, and of the evaluation procedure, as he already judged pages and he is used to the 4 level relevance scale adopted.

The average time needed for the evaluation of a page is 48.9 seconds for J and 62.6s for J3, with a standard deviation of 14.3s for J and 32.3s for J3: the expert needs about 75% of the time needed by the not expert to perform the new judgments. Since the average time to evaluate a page is 55.7s and since there are 2405 different pages from the first and second version our collection, the time to increase the collection was about 37 hours.

5. CONCLUSIONS

In this paper we have presented our approach for the evaluation of a CAR application for the mobile environment. We evaluated different strategies for automatic query building based on users' current context. Despite user testing is the main evaluation technique adopted in this field, the early stage of development of our system and the need of measuring the effectiveness of different strategies guided us toward a TREC-like benchmark approach.

The CREC benchmark helped the development process giving good insights (e.g., in the best strategy, the best performance appears just adding a term after the "geoterms") and underlining weak points (e.g., adding more and more terms in the query does not necessarily increase performance). At the same time, at a general extent, this approach is useful as it can simplify also the user testing evaluation. For example, knowing which is the best strategy allows us to give to users just one prototype, instead of different prototypes, one for each strategy. Moreover, once the benchmark is configured, it can be reused to test new strategies or related features, in a semi-automatic way (new judgments are needed).

In this way, we have made our second step to refine a methodology whose aim is to become a general early stage evaluation technique for retrieval processes in context-aware systems. On the basis of our experience with MREC and CREC, we believe that early stage evaluations using a benchmark, followed by user studies, is an effective methodology to evaluate systems like CAB. The benchmark does not substitute the user testing evaluation. Rather, several early stage benchmark experiments could provide more solid basis for the subsequent user testing, that can thus be more focused.

Future work is needed on three aspects. Since with the manual strategy the effectiveness is higher, we will try to refine the geo+tf.idf strategy to make it more effective, e.g., by removing the constraint of incremental query construction (i.e., not requiring that a query having a length of i contains all the $i-1$ terms of the previous query). On a second and more general aspect, we will continue to work on the methodology of incremental benchmark construction, to better understand its reliability and usefulness. For example, although we did find that subjectivity and dynamics seem not to be an issue, we did not analyze what happens when two groups use two different versions of the evolving collection. Finally, it would be interesting to work towards a publicly available context test collection, containing a set of context descriptors (perhaps with a representation different from TREC topics, and maybe including a temporal longitudinal component as well, e.g., using Schank and Abelson's scripts), Web resources, and relevance judgments.

Acknowledgements

We are grateful to Nick Belkin, Jürgen Ziegler, and the referees for their useful remarks.

6. REFERENCES

[1] P. J. Brown and G. J. F. Jones. Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. *Personal and Ubiquitous Computing*, 5(4):253–263, 2001.

[2] A. Bulfoni, P. Coppola, V. Della Mea, L. Di Gaspero, D. Mischis, S. Mizzaro, I. Scagnetto, and L. Vassena. AI on the move: Exploiting AI techniques for context inference on mobile devices. In *Proc. of 5th Prestigious Applications of Intelligent Systems (PAIS 2008), colocated with ECAI08*, pages 668–672, 2008.

[3] M. Carter. Challenges for ubicomp evaluation. Computer Science Division, University of California, 2004.

[4] P. Coppola, V. Della Mea, L. Di Gaspero, D. Mischis, S. Mizzaro, I. Scagnetto, and L. Vassena. AI techniques in a context-aware ubiquitous environment. In A. Ella Hassanien, A. Abraham, and H. Hagras, editors, *Pervasive Computing: Innovations in Intelligent Multimedia and Applications*, Computer Communications and Networks. Springer, 2009.

[5] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 282–289, New York, NY, USA, 1998. ACM.

[6] V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):530–543, 2004.

[7] A. Göker and H. Myrhaug. Evaluation of a mobile information system in context. *Information Processing & Management*, 44(1):39–65, 2008.

[8] D. J. Harper and D. G. Hendry. Evaluation light. In *Proc. of the second MIRA workshop*, pages 53–56, 1997. Technical Report TR-1997-2, Department of Computing Science, University of Glasgow, Glasgow. `http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/`.

[9] S. Mizzaro, E. Nazzi, and L. Vassena. Retrieval of context-aware applications on mobile devices: how to evaluate? In *Proc. of Information Interaction in Context (IIiX '08)*, pages 65–71, 2008.

[10] J. Scholtz. Evaluation methods for ubiquitous computing. Ubicomp 2001 Workshop, September 2001.

[11] J. Scholtz. Metrics for evaluating human information interaction systems. *Interacting with Computers*, 18:507–527, 2006.

[12] J. Scholtz and S. Consolvo. Towards a discipline for evaluating ubiquitous computing applications. Technical Report IRS-TR-04-004, 2004.

[13] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–104, 1991.

Accessing Contextual Information for Interactive Novelty Detection

Wenyin Tang, Agus T. Kwee, Flora S. Tsai
School of Electrical & Electronic Engineering
Nanyang Technological University
Singapore
wenyintang@ntu.edu.sg, atkwee@ntu.edu.sg, fst1@columbia.edu

ABSTRACT

Novelty detection is a process of spotting novel yet relevant information for users. Contextual information from relevance feedback has been successfully utilized in relevant information retrieval systems. However, the usage of contextual information in novelty detection area has not been explored too much. One possible reason is that it is more difficult to describe the novel information by some certain key words, because any term becomes somewhat non-novel after its first occurrence. In this paper, we propose to use some high-level structures of words, such as number, time, location, person name, country name, etc, to accommodate the user's context in novelty detection systems. A GUI (graphical user interface) related to context acquiring is designed to facilitate an efficient user preference setting. Secondly, a rule-based novelty detection (rule-ND) algorithm is proposed to utilize the acquired contextual information. Moreover, a local evaluation method is proposed to evaluate the resulting system. The experiment show encouraging results of our method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information filtering and retrieval models

General Terms

Algorithms, experimentation

1. INTRODUCTION

Novelty detection (ND), as one step forward to the relevant sentence retrieval, is used to spot novel yet relevant sentences for users. Based on the appearance time order of the sentences, the information that has never appeared before is considered as "novel". The pioneering work of ND has been done for novel document retrieval by Zhang et al. [14]. Since it is more meaningful to spot novel sentences for the user instead of documents only, later studies paid more attentions on sentence-level novelty detection, such as those reported in TREC 2002-2004 Novelty Track [6, 9, 8] and those reported in other publications [1, 15]. Conventionally, the novelty of an incoming sentence is measured based on the occurrence of new words/terms in it.

Recent studies suggested several ways to improve the performance of novelty detection by integrating various natural language processing (NLP) techniques, such as part-of-speech (POS) tagging, named entity recognition (NER), WordNet, etc [13, 4, 2]. These NLP techniques facilitate the usage of additional contextual information, beyond the simple strategy using a-bag-of-words in novelty detection, but succeed of these methods still depends highly on how the system utilizes the additional information. One problem here is these methods utilized the contextual information implicitly, without responding to the user's preference. In an interactive novelty detection system, the user's preference about retrieved novel information is accessible and therefore, needs to be considered in the ND process.

This paper attempts to bridge a gap between the explicit user's preference and the advance of contextual information, by answering three questions: (i) what aspects of the user's context can be acquired and how to guide the user to interact with system efficiently? (ii) how could we utilize the acquired contextual information in novelty detection system? and (iii) how could we evaluate the effectiveness of the resulting system?

This paper is organized as follows. Section 2 analyzes the user context in ND system and the corresponding GUI design. Section 3 introduces a rule-based novelty detection (rule-ND) algorithm to utilize the acquired contextual information. The experimental study is presented in Section 4. Some concluding remarks are given in Section 5.

2. USER CONTEXT IN ND SYSTEM

In previous studies, contextual information accessed through relevance feedback has been successfully utilized for query expansion in relevant information retrieval [5, 3, 7]. In this case, the user may either provide the relevance feedback for documents/passages [5, 11, 12] where the useful terms are extracted, or select the key terms directly [10].

In novelty detection, however, accessing the context information through user feedback has not been explored too

much. One possible reason is that it is much more difficult to describe the novel information using key words. In relevant document/passage retrieval, if "football" is selected as a key word by the user, documents/passages with "football" are more likely to be retrieved by a search engine. But in novelty detection, the user cannot always use "football" as novel term because any term becomes somewhat non-novel after its first appearance.

Instead of terms feedback, which is impractical in the ND system, the high-level structures of terms such as number, time, location, person name, country name, etc are suggested in this paper. These high-level structures that show very natural aspects of information can be easily understood by the user and facilitate user context feedback in novelty detection. For example, readers of financial news may be more sensitive to different numbers. Given a pair of sentences s_1 and s_2 where s_1 is the nearest history sentence of s_2, s_2 contains novel information because they include different number information, i.e. US$50 and US$48.

Example 1:
s_1: Now the oil price is US$**50**/barrel.
s_2: Now the oil price is US$**48**/barrel.

However, s_2 may be predicted as non-novel by the conventional ND system because it is very similar to s_1, according to the similarity score calculated by cosine similarity [1].

Since a specific topic usually contains many aspects of information, the system allows the user to feedback their preferred aspects of novel information simultaneously, and hence to describe their contexts more completely. Accordingly, in GUI design, we use check-box with different aspects of information, as shown in Figure 1.

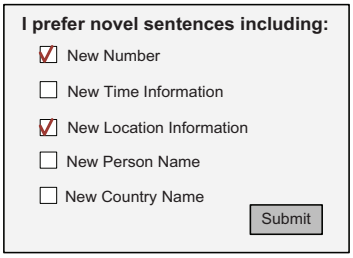

Figure 1: Snapshot of GUI for additional contextual information.

This snapshot of GUI is an initial design, where more optional aspects of novel information can be appended. Moreover, the user is able to specify his/her context for each topic separately via GUI.

3. INTERACTIVE NOVELTY DETECTION SYSTEM

3.1 Background of ND

In ND, the novelty of a sentence can be quantitatively measured and scored by a novelty metric. The most popular novelty metric, i.e. cosine similarity (see [1]), is adopted. This metric first calculates the similarities between the current sentence s_t and each of its history sentences s_i $(1 \leq i \leq t-1)$. Then, the novelty score is simply one minus the maximum of these cosine similarities, as shown in Eq. (1).

$$N_{cos}(s_t) = 1 - \max_{1 \leq i \leq t-1} cos(s_t, s_i) \quad (1)$$

$$cos(s_t, s_i) = \frac{\sum_{k=1}^{n} w_k(s_t) \cdot w_k(s_i)}{\|s_t\| \cdot \|s_i\|}$$

where $N_{cos}(s)$ denotes the cosine similarity score of the sentence s and $w_k(s)$ is the weight of k^{th} element in the sentence weighted vector s. The term weighting function used in our work is tf.isf (term frequency multiply inverse sentence frequency) as defined below.

$$w_k(s_i) = tf_{w_k,s_i} log\left(\frac{n+1}{sf_{w_k}+0.5}\right) \quad (2)$$

tf_{w_k,s_i} is the frequency of the word w_k in the sentence s_i; sf_{w_k} is the number of sentences, in which the word w_k appears in the collection; n is the number of sentences in the collection.

The final decision on whether a sentence is novel or not depends on whether the novelty score falls above or below a threshold. This document will be pushed into the history document list.

3.2 Rule-based ND

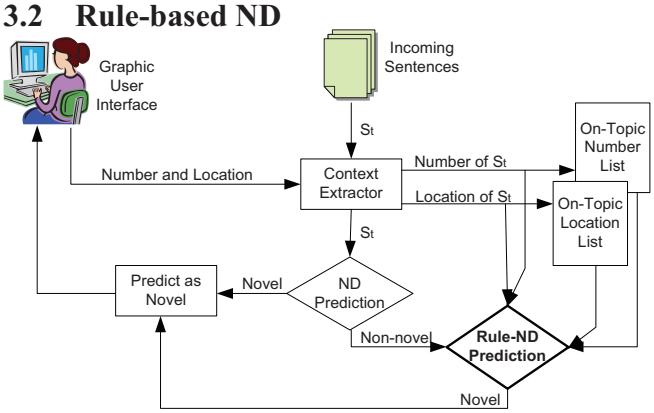

Figure 2: An interactive system with the user's preference of "number" and "location".

Figure 2 shows the framework of an interactive ND system with the user's preference of "number" and "location". This proposed system has a cascade framework, where the original ND algorithm classifies novel sentences at the first layer. The rule-ND algorithm retrieves the novel information missed by the original ND system at the second layer, based on a set of rules. In this system, any new incoming sentence s_t will be predicted in 4 steps, i.e.

step 1: Record the corresponding context information specified by the user, if any.

step 2: Run the original ND on s_t. Go to step 3 if the ND prediction is "non-novel", step 4 otherwise.

step 3: Run rule-ND on s_t. Go to step 4 if the rule-ND prediction is "novel".

step 4: Predict s_t as novel.

3.2.1 Number Information Extraction

To keep this paper short, we only illustrate the rule-ND involving different numbers. Unlike other named entities, numbers themselves are not that meaningful. Therefore, the first word after the number is also accessed by rule-ND using POS Tagging, while only those numbers in conjunction with noun (e.g. player, mile, etc) or cardinal number (e.g. billion, million, etc) are recorded as active terms. For example, "30 people", "42 millions" will be recorded and compared in rule-ND, instead of "30" and "42" only. Other named entities such as time, location, person and country show enough useful information and need not use some additional information.

3.2.2 One-To-One vs. One-To-All Comparison

After number extraction, rule-ND also compares the incoming sentence with its history sentences but predicts based on predefined rules. There are two strategies in sentence comparison, i.e. one-to-one and one-to-all comparison. The number-rule using one-to-one sentence comparison is defined as below.

`Number-rule-one2one: highly similar` s_1 `and` s_2 `have different NUMBER + same NOUN` $\Rightarrow s_2$ `is novel`

Using this rule, the sentence s_2 in Example 1 can be predicted as "novel". However, we found that the novel NUMBER+NOUN term will have already occurred in other on-topic history sentences, i.e. the context of s_2. Therefore, we accumulate all NUMBER+NOUN terms appeared in the history sentences, and propose an alternative rule, called number-rule-one2all, as below.

`Number-rule-one2all: satisfy number-rule-one2one AND` s_2 `has different NUMBER + NOUN term compared with on-topic number terms in history =>` s_2 `is novel`

The experimental result on TREC 2004 Novelty Track data shows that number-rule-one2all always outperforms number-rule-one2one in various situations. We also compare these two strategies in the original ND on the same data and get some interesting results: (i) The one-to-one sentence comparison largely outperformed one-to-all sentence comparison in the original ND. This indicates that one-to-one sentence comparison may be more effective for ND in regular situations, i.e. moderate similarities between sentences. (ii) One-to-all sentence comparison is more effective in rule-ND. This indicates that one-to-all sentence comparison may be more effective for detecting novel information in highly similar sentences. Obviously, one-to-all comparison strategy conducts a tighter criterion for novel sentence retrieval. This result shows, for highly similar sentences, the tighter criterion is necessary because only very significant novel terms will make a highly similar sentence novel.

3.2.3 How to Define Highly Similar Sentences

Another important issue is how to define the highly similar sentences. Rule-ND is attempted to retrieve the novel sentences that are missed by the original ND algorithm, as shown in Figure 2. According to the novelty threshold θ in the original ND, the simplest way to define the so-called "highly similar sentences" is let the high similarity threshold, ρ, equal to $1-\theta$, i.e. all the sentences predicted as non-novel sentence in the original ND are regarded as "highly similar sentences". We can also vary the high similarity threshold ρ from $1-\theta$ to 1, to get a stricter standard of high similarity and a smaller number of highly similar sentences.

4. EXPERIMENTAL STUDY

4.1 Data & Evaluation Issue

TREC 2004 Novelty Track data [8] are used as the experimental data here. This data are developed from AQUAINT collection. The NIST assessors created 50 topics for this data. For the evaluation of novelty detection systems, the novel sentences selected manually by the NIST assessor were considered as the truth data. In this experiment, we follow to use the evaluation measures proposed for TREC Novelty Track [8], i.e. set-based recall, precision and F score averaged across all 50 topics in this data. Let M be the number of matched sentences, i.e., the number of sentences selected by both the assessor and the system, A be the number of sentences selected by the assessor, and S be the number of sentences selected by the system. Then sentence set recall is M/A and precision is M/S. The F score is a function of precision (P) and recall (R), defined as $F = \frac{2\times P\times R}{P+R}$.

In fact, rule-ND (with number-rule) may only affect the predictions of sentences that involve numbers. In order to clearly observe the effectiveness of rule-ND, we use a local evaluation method, i.e. we still run the original ND and rule-ND on all sentences, but calculate the set-based recall, precision and F-score in the selective pool of sentences involving numbers. The local evaluation method can be generalized to other rules, by using different selective pools of sentences involving different types of named entities. The statistics of TREC 2004 Novelty Track data are summarized in Table 1.

Table 1: Statistics of TREC 2004 Novelty Track Data

	All sentences in TREC 2004 data	Sentences involving numbers
# Relevant	8343	2558
# Novel	3454	1081
N/R	41.4%	42.3%

4.2 Result

In this experiment, rule-ND (with one-to-all sentence comparison) algorithm is compared with the original ND algorithm. Since in rule-ND, one-to-all sentence comparison always outperforms one-to-one, we omit the later result.

In this experiment, we vary both novelty threshold θ in the original ND and the high similarity threshold ρ in rule-ND. The experimental result is shown in Table 2. From this table, We can observe that:

- Rule-ND performs better with a lower ρ. The best result of rule-ND for each run is obtained when ρ is set to 1-θ (shown in boldface in Table 2).
- Rule-ND improves recall 1.5% to 5% with little loss of precision (around -0.2%).

Table 2: Rule-ND vs. Original ND on TREC 2004 Novelty Track Data

	Algorithm	Parameter Settings	Precision	Recall	F-Score
Run1	Original ND	θ: 0.45	0.5132	0.9078	0.6343
	Rule-ND	θ: 0.45; ρ: **0.55**	**0.5116**	**0.9234**	**0.6370**
	(Number-rule-one2all)	θ: 0.45; ρ: 0.60	0.5112	0.9135	0.6346
	ρ=[0.55, 1]	θ: 0.45; ρ: 0.70	0.5114	0.9112	0.6340
		θ: 0.45; ρ: 0.80	0.5137	0.9100	0.6354
Run2	Original ND	θ: 0.55	0.5606	0.7772	0.6220
	Rule-ND	θ: 0.55; ρ: **0.45**	**0.5586**	**0.8063**	**0.6314**
	(Number-rule-one2all)	θ: 0.55; ρ: 0.55	0.5571	0.7928	0.6259
	ρ=[0.45, 1]	θ: 0.55; ρ: 0.60	0.5575	0.7829	0.6225
		θ: 0.55; ρ: 0.70	0.5580	0.7806	0.6218
		θ: 0.55; ρ: 0.80	0.5610	0.7793	0.6232
Run3	Original ND	θ: 0.65	0.6095	0.5212	0.5227
	Rule-ND	θ: 0.65; ρ: **0.35**	**0.5997**	**0.5706**	**0.5507**
	(Number-rule1-one2all)	θ: 0.65; ρ: 0.45	0.5978	0.5503	0.5398
	ρ=[0.35, 1]	θ: 0.65; ρ: 0.55	0.5961	0.5368	0.5321
		θ: 0.65; ρ: 0.60	0.5998	0.5269	0.5263
		θ: 0.65; ρ: 0.70	0.6009	0.5246	0.5249
		θ: 0.65; ρ: 0.80	0.6096	0.5234	0.5252

This experimental result of rule-ND is quite encouraging.

5. CONCLUSIONS & FUTURE WORK

This paper addressed the problem of accessing context information in novelty detection. Unlike relevant information retrieval, it is impractical to do the novelty feedback using key words, because any word becomes somewhat non-novel after its first appearance. Therefore, we proposed to use some high-level structures of words, such as number, time, location, person name, country name, etc, to accommodate the user's context in novelty detection systems. An interactive novelty detection system with a cascade structure was proposed. This system first performed original ND. Then, rule-ND algorithm was performed to retrieve novel information missed by the original ND based on the user's context. The experiment on TREC 2004 Novelty Track data showed that our method can improve recall 1.5% to 5% with little loss of precision (around -0.2%).

In our future work, we will further investigate the performance of the rule-ND system by using various individual or mixed types of named entities. The GUI for novelty feedback will be further improved by automatically annotating the named entities from a piece of text highlighted by the user, which may be more convenient and practical for the user.

6. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at sentence level. In *SIGIR 2003, Toronto, Canada*, pages 314–321. ACM, August 2003.

[2] X. Li and W. B. Croft. An information-pattern-based approach to novelty detection. *Information Processing and Management*, 44(3):1159–1188, May 2008.

[3] Y. Nemeth, B. Shapira, and M. Taeib-Maimon. Evaluation of the real and perceived value of automatic and interactive query expansion. In *ACM SIGIR 2004*, pages 526–527, 2004.

[4] K. W. Ng, F. S. Tsai, K. C. Goh, and L. Chen. Novelty detection for text documents using named entity recognition. In *Information, Communications and Signal Processing, 2007 6th International Conference on*, pages 1–5, December 2007.

[5] S. Robertson and I. Soboroff. The TREC 2002 Filtering Track report. In *The SMART retrieval system*, pages 313 – 323, 1971.

[6] S. Robertson and I. Soboroff. The TREC 2002 Filtering Track report. In *TREC 2002 - the 11th Text Retrieval Conference*, 2002.

[7] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *ACM SIGIR 2005*, pages 824–831, 2005.

[8] I. Soboroff. Overview of the TREC 2004 Novelty Track. In *TREC 2004 - the 13th Text Retrieval Conference*, 2004.

[9] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *TREC 2003 - the 12th Text Retrieval Conference*, 2003.

[10] B. Tan, A. velivelli, H. Fang, and C. X. Zhai. Term feedback for information retrieval. In *ACM SIGIR 2007, Amsterdam, The Netherlands*, pages 263 – 270, July 2007.

[11] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *ACM SIGIR 1996*, pages 4 – 11, 1996.

[12] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th international conference on information and knowledge management*, pages 403 – 410, 2001.

[13] H.-P. Zhang, J. Sun, B. Wang, and S. Bai. Computation on sentence semantic distance for novelty detection. *Journal of Computer Science and Technology*, 20(3):331–337, 2005.

[14] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *ACM SIGIR 2002, Tampere, Finland*, pages 81–88, 2002.

[15] L. Zhao, M. Zheng, and S. Ma. The nature of novelty detection. *Information Retrieval*, 9:527–541, 2006.

APMD-Workbench: A Benchmark for Query Personalization

Verónika Peralta
Laboratoire d'Informatique
Université de Tours
3, pl Jean Jaures
41000 Blois, France
+33-2-54552112
vperalta@univ-tours.fr

Dimitre Kostadinov
Alcatel-Lucent Bell Labs France
Route de Villejust
91620 Nozay, France
+33-1-30772116
Dimitre_Davidov.Kostadinov
@alcatel-lucent.com

Mokrane Bouzeghoub
Laboratoire PRiSM
Université de Versailles
45, av des Etats-unis
78000 Versailles, France
+33-1-39254057
mok@prism.uvsq.fr

ABSTRACT

Query personalization algorithms intend to deliver the most relevant data to each user according to their profiles. Validating efficiency and relevancy of such algorithms still remains a very difficult task as it requires a scalable dataset, a bunch of user profiles and queries and possibly user feedbacks. At the best of our knowledge, there is not a reference benchmark devoted to the validation of such algorithms. In most of the published papers, validation of personalized queries is done through ad hoc benchmaks whose features are not given and which are generally not provided to other authors to perform similar evaluations. In this paper, we present a benchmark for query personalization which aims to be a reference to validate query personalization systems. The benchmark is based on a large test bed derived from MovieLens and IMDb datasets, and provides, besides the data set itself, a large sample of queries and users as well as their corresponding good results.

1. INTRODUCTION

Query personalization is one of the main solutions to improve data relevance in information retrieval and database systems. Before being executed, user queries are reformulated on the basis of user profile preferences. This allows targeting user's domain of interest and thus delivering pertinent results and reducing result size.

In order to measure the relevance of results, we need to compare delivered results with those effectively preferred by users. In other words, we need a reference data set that contains several queries and the corresponding sets of query results that are relevant for each user. In this way, we can quantitatively evaluate the behaviour of a personalization algorithm using metrics such as precision, recall, result size, performance, etc.

In this paper we describe the construction of a benchmark for query personalization. There exist several benchmarks among which we can cite the TPC benchmarks for database server performances [9] or the TREC benchmarks for information retrieval systems [8]. However, as far as we know, there is no benchmark providing a validation framework to query personalization algorithms, at least in the database domain. A benchmark for query personalization should also manage different users and their preferences. Specifically, they should provide a large database, a set of users, a set of queries and the reference results associated to each user and query, i.e. they should provide collections of triplets {(query, user) → results}.

Obtaining such reference results is very costly because it involves asking users to explicitly evaluate query results. The TREC benchmark [8] was built in this manner and the task lasted several years. In addition, there was no notion of user profile in TREC, but users divided the task of judging if documents were pertinent or not in a global manner. For the proposed benchmark, instead of asking a set of users to manually classify query results according to their preferences and feelings, we reuse ratings already expressed by real users and published on the Web.

Our dataset is derived from two public databases, i.e. MovieLens [1] and IMDb [5]. Both databases deal with data about movies. The IMDb database contains rich information about films, actors, directors, the places where they are produced, their budgets, their categories and the average rank given by the users who had evaluated them. The MovieLens database contains very few information about films but provides a huge amount of evaluations given by users who have seen these films. The two databases are complementary as they almost target the same movies (actually the set of films referred in MovieLens is a subset of those referred in IMDb). The main advantage of using these databases is their large volume of data, which is freely available through Internet. In addition, movie data is very easy to understand, to use and to analyze.

The construction of the benchmark proceeds in three phases, as illustrated in Figure 1. The first phase consists in extracting, transforming and loading movie data from IMDb and MovieLens. The second phase consists in generating a set of queries and deriving the corresponding "good" results for each user. This can be done

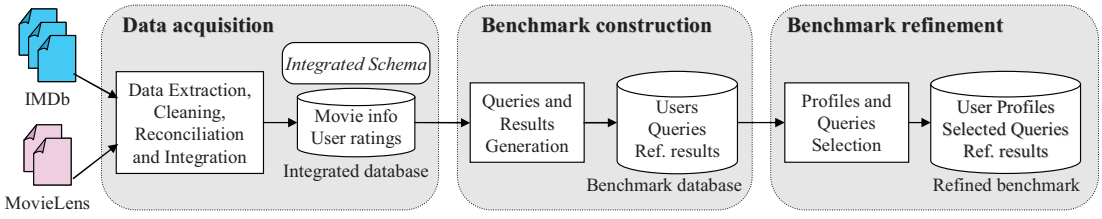

Figure 1 – Benchmark construction process

thanks to the content originated from MovieLens which provides a large set of user evaluations on different films they have seen. The benchmark database contains several thousand queries and users, and several millions reference results. The third phase consists in selecting the users and queries that are worth to be tested in a particular type of test. This selection is particularly necessary as the execution of the whole set of queries for the whole set of users necessitates tens of years although limiting the query evaluation process to a few seconds. Actually, the benchmark database serves as a basement to generate several specific benchmarks with different features, depending of the evaluation goal.

This paper describes the construction of the benchmark and its use for query reformulation. Section 2 describes the procedures for extracting and integrating IMDb and MovieLens data, for generating a set of queries, and for obtaining the reference results for each pair (query, user). Section 3 presents an example of use of the benchmark, illustrating the generation of user profiles and the comparison of results for a particular type of tests. Finally, Section 4 presents our conclusions.

2. BENCHMARK DESIGN

This section describes the challenges, the difficulties and the strategies used to define the benchmark.

2.1 Data Acquisition

The first phase consists in extracting, transforming and integrating IMDb and MovieLens data in order to build a relational database about movies, which includes movie descriptions and user ratings.

MovieLens data set consists of 3 text files, with tabular format, describing 1.000.209 anonymous ratings of 3.883 movies made by 6.040 MovieLens users. IMDb data set consists of 49 ad-hoc text files, called lists, which record different details about movies. At October 2006, list sizes varied from 25.000 to 5.000.000 tuples about more than 850.000 movies. We extracted data from 23 lists, representing the most relevant tabular features about movies.

For both data sets, data extraction consisted in several tasks including loading of text data into a relational database, transformation and normalization of data types, standardization of codes, filtering of inconsistent values and duplicate elimination. The matching of MovieLens and IMDb movie titles (which identify movies) was the most difficult task in the construction of the integrated database.

The integrated schema consists in 52 tables describing movies, companies and persons related to movies as well as the users that evaluated movies. It includes 1 table listing movies, 1 table containing user evaluations, 3 tables describing users, 20 tables describing movie features (e.g. genres), 23 tables relating movies to features and 4 auxiliary tables. We refer the interested user to [6] for further details on data acquisition.

2.2 Generation of Queries

In order to build the benchmark database we need to build a set of queries and the reference results for each pair (query, user). Instead of asking users to manually classify query results according to their relevance, we reuse movie ratings already given by MovieLens users. Specifically, each tuple of the *I_UserRatings* table of the integrated database corresponds to an evaluation of a movie, registered by a user, indicating a rating (in a 1-5 star scale).

As ratings qualify movies, we generate a set of queries returning movies, which are lately compared to the movies having a high user rating. Queries have a star-like form:

SELECT I_UserRatings.movieid
FROM I_UserRatings, <*additional tables*>
WHERE <*filtering conditions*> AND <*join conditions*>

Note that instead of selecting movies from the whole set of movies, we consider as space of solutions, the ones that the user has already evaluated (i.e. the I_UserRatings table). In this way, queries return only movies whose ratings are already known. We join the I_UserRatings table with some tables describing movie features and we add some filtering conditions on such features.

Filtering conditions are predicates of the form *feature operator value*, where *operator* $\in$ {=,≤,≥}, and *value* ranges in the domain of a movie *feature*. We randomly select a small number of predicates (from 1 to 5) for each query, which avoids generating monster queries that returns no data. However, the randomness of the selection allows obtaining result sets of different sizes, ranging from almost empty sets when queries have several restrictive predicates to almost all data when queries have few non-restrictive predicates. The additional tables are those referenced in the predicates and those necessary to join them to the *I_UserRatings* table. Details on query generation can be found in [7].

2.3 Computation of Reference Results

As the rating of each movie is known, we can easily build the set of reference results, i.e. those movies rated with 3, 4 or 5 stars.

Actually, we partition the I_UserRatings table into two subsets: (i) *training set*, available for the generation of user profiles, and (ii) *test set*, available for executing queries and measuring personalization results. This partitioning allows personalization algorithms to learn user preferences and derive user profiles without biasing the test results. To avoid side effects generated by the arbitrary choice of these two subsets, the process has been repeated for several subsets, randomly generated from the original dataset.

In order to compute partitions, five random attributes were added to the I_UserRatings table (namely, C1, C2, C3, C4 and C5), each one randomly filled with an integer between 0 and 9. Therefore, the test set is described by a condition on the values of one of the Ci attributes ($1 \le i \le 5$). We also parameterize the rating above which a film is considered to be good (from 3 to 5).

Consequently, in order to execute a query for a given user according to a partitioning and rating strategy, the query is restricted with three conditions on the I_UserRatings table: (i) a *TestSetCondition* of the form $Ci<N$, $1 \le i \le 5$, $0 \le N \le 9$, (ii) a *RatingCondition* of the form *rating* $\ge V$, $3 \le V \le 5$ and (iii) a UserCondition of the form *userid*$=U$, $1 \le U \le 6040$. As an example of restricted queries consider:

SELECT I_UserRatings.movieid
FROM I_UserRatings, I_MovieCountries
WHERE I_UserRatings.movieid =I_MovieCountries.movieid
AND I_MovieCountries.country = 'France'
AND *Ci*<*N* AND userid = *U* AND rating ≥ *V*;

2.4 Experiment and statistics

The benchmark was developed and stored in an Oracle 9i database. Preprocessing and parsing procedures where implemented in Java, loading was performed with the SQL-Loader utility and the remaining procedures where implemented in PL-SQL (stored procedures). In this section we describe some results and statistics obtained from the execution of those procedures, specifically, we describe parameter setting and we analyze the generated queries and their result sizes.

2.4.1 Setting parameters

As previously argued, we aim at generating different partitioning and rating strategies in order to obtain unbiased experimental results. We tested different parameters discarding those that produced too few tuples in the test set (e.g. when setting 'rating = 5').

We kept the 20 strategies shown in Table 1 (they are packed by 5, for $1 \le i \le 5$). These strategies combine two test set sizes, with approximately 50% (Ci≥5) and 70% (Ci≥3) of tuples respectively, and two rating conditions (*rating* ≥ *3* and *rating* ≥ *4*). Table 2 shows the average number of good ratings in the test set per user and type of strategy.

Table 1 – Combination of partition and rating strategies

Strategy id	Test set cond.	Rating cond.
1-5	Ci ≥ 5	Rating ≥ 3
6-10	Ci ≥ 5	Rating ≥ 4
11-15	Ci ≥ 3	Rating ≥ 3
16-20	Ci ≥ 3	Rating ≥ 4

Table 2 –Average number of good ratings in the training set per user and type of strategy

Test cond	All ratings	Rating≥3	Rating≥4
Ci ≥ 3	116	96	66
Ci ≥ 5	83	69	48

2.4.2 Obtained queries

Having set parameters, we proceeded to execute the query generation procedures. We obtained an initial set of 622.061 predicates for all queries, from which we randomly selected from 1 to 5 predicates per query. We generated 6040 queries and we added a query with no predicates. After generating queries, we executed them over several test sets (for all users) and we measured the size of the obtained results. Figure 2 illustrates result sizes for one particular strategy (with test size=70% and rating≥3). We took two measures: the average of user's result sizes, and the maximum result size (for the user obtaining the most of results). Note that most queries returns less than 20 tuples in average but they return more results for some users. Further details can be found in [7].

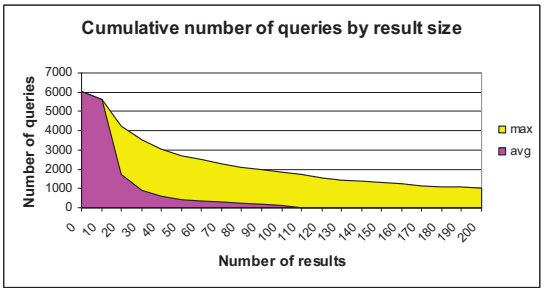

Figure 2 – Number of queries by result size

3. AN EXAMPLE OF BENCHMARK USE

The benchmark presented in the previous section can be used to evaluate various personalization approaches. This section shows its use for evaluating query reformulation approaches.

3.1 Query Reformulation Approaches

Query reformulation applies to the context of mediation systems. It consists in reinterpreting the user intention, expressed in his initial query, into a more complete query, considering at the same time the user profile and the descriptions of the data sources. Thus, query reformulation integrates two complementary tasks: (i) query enrichment which consists in integrating user profile content in the user query and (ii) query rewriting which transforms a query expressed on the virtual schema in expressions (rewritings) expressed on the data sources.

We evaluated three query reformulation approaches, proposed in [3]. Two of them are compositions of existing algorithms for query rewriting [2] and query enrichment [4]. They differ in the order of application of the rewriting and enrichment algorithms. The third approach is an interleaving of the two previous ones.

In order to evaluate and compare these query reformulation techniques, we adapted the benchmark presented in the previous section. Next sub-section presents the benchmark refinement process.

3.2 Benchmark Refinement

The query reformulation approaches require the presence of a distributed system and the availability of predicate-based user profiles. In order to fulfill these requirements, the benchmark have been refined by simulating a distributed system and by extracting profile predicates from the users' training sets.

To simulate a distributed environment, the integrated schema of the benchmark is taken as global schema. Then 52 views have been manually defined over this global schema in accordance with the following assumptions: (i) views should provide all data contained in the integrated database, assuring no information loss, (ii) views schemas should overlap, generating some redundancy in order to measure the capacity of an approach to select only relevant data sources, and (iii) some views definitions should contain selection predicates enabling to check if a reformulation approach selects data sources which are able to better satisfy the user preferences . Each view is considered as being a separate data source. User profiles are constructed as sets of predicates that state user preferences on movie features. Profile predicates have the form *feature=value*, where *value* ranges in the domain of the movie *feature*. For example, a certain user may prefer *movies spoken in French* or *action movies*; which is expressed by the predicates: *Language = French*, *Genre = Action*.

In order to extract a user profile from a set of user evaluations (those of the training set), we look for common features of the evaluated movies. For example, if most of the movies to which the user has assigned a great rating are filmed in France, we deduce that the user prefers *movies filmed in France*, and we propose the predicate *LocationCountry=France*.

We generated a large set of predicates where each predicate is associated with a *weight* representing the percentage of evaluated films that satisfy it. Weights allow conforming more or less restrictive user profiles by choosing the predicates with higher weights or accepting predicates with lower weights. Weighted predicates have the form *<table.attribute operator value (weight)>* where: *table* and *attribute* refer to an attribute of a table of the integrated schema (referencing a movie feature), *value* is an element of the attribute domain, *operator* ∈ {=,<,≤,>,≥} and *weight* represents the percentage of the evaluated films that satisfy the predicate. Examples of weighted predicates are "I_MovieLanguages.language = English (80)" and "I_Countries.continent = Europe (25)" which can be interpreted as *among the films the user has evaluated, 80% are spoken in English and 25% have been filmed in Europe*.

3.3 Query and Profile Selection

The execution of the whole set of available queries using the whole set of generated profiles necessitates tens of years although limiting the query reformulation process to a few seconds. Thus, the evaluation was performed on a subset of queries and profiles.

The main parameter which is taken into account for query selection is the response time of the query reformulation algorithms. Query rewriting is the most time-consuming phase of the reformulation process. According to [2], the response time of query rewriting algorithms depends on several parameters such as the number of virtual relations in the query, the schema type, the number of sources, the number of variables in the source schemas, etc. Most of these parameters do not vary in our benchmark; the only variable parameter is the number of virtual relations in the queries. We made some preliminary tests which shown that query rewriting takes about 1 minute for queries expressed on 9 virtual relations and more than 10 minutes for queries with 10 relations. As query reformulation is usually a real time process, we limited its response time to 1 minute. In addition, to enable the expansion of a query with additional virtual relations during query enrichment, we restricted to queries expressed on at most 5 virtual relations. Thus, a total of 13 queries was selected for the tests including the only available query expressed on 1 virtual relation and 3 queries for each other configuration (i.e. expressed on 2, 3, 4 and 5 virtual relations).
The selection of a subset of users (and their profiles) is guided by the following requirements: (i) a user test set should be large enough to get significant number of results when executing the queries on it, (ii) the predicates of a user profile should be expressed on different attributes (to simplify the query enrichment), (iii) profile predicates should have weights superior or equal to 80, and (iv) a user should have enough predicates to allow considering profiles with different cardinalities. To satisfy these requirements, we applied several filtering steps. First, users whose test set contain less than 100 movies have been pruned. Second, for each profile we pruned all predicates expressed on the same attribute but the one with the highest weight. Finally, only predicates which weights are superior or equal to 80 were selected. The filtering resulted in 747 profiles having from 2 to 10 predicates. For our tests we selected the 2 available profiles containing 10 predicates and 3 profiles containing 9 predicates. Each profile was then used to produce two new profiles by randomly selecting 3 and 6 of its predicates. Thus, the refined benchmark contains a total of 15 user profiles (3 profiles per user).

The benchmark allowed to compare the three query reformulation approaches and to highlight the contexts where each one performs better. Evaluations show that introducing personalization into query reformulation improves the precision of the obtained results but increases response time and can lead to the loss of relevant results. A more complete description of the tests and the obtained results can be found in [3].

4. CONCLUSIONS

This paper proposes a benchmark for query personalization based on movies rated by real users. The benchmark database includes a large set of users, queries and reference results.

The benchmark can be refined to support the evaluation of various personalization approaches. We showed an example of refinement that we used for comparing three query reformulation approaches. In this example we generated simple user profiles. The benchmark can also be used for testing and comparing profile generation algorithms, both for individual users and communities. Our hope is that this platform serves as a reference test in the database community in order to federate the sparse evaluations around a common data set as done for example in the TREC protocol.

A detailed description of the APMD-Workbench can be found at: http://apmd.prism.uvsq.fr/SubProject4/TestPlatform/ ([10]).

5. REFERENCES

[1] GroupLens Research: "movielens: helping you to find the right movies". Web site, ULR: http://movielens.umn.edu, last accessed on July 9th, 2007.

[2] Halevy, A., Pottinger, R.: "MiniCon: A scalable algorithm for answering queries using views", Very Large Data Bases Journal, Vol. 10, p. 182-198, 2001.

[3] Kostadinov, D.: "Data Personalization: an Approach for Profile Management and Query Reformulation", PhD thesis, University of Versailles, France, 2007.

[4] Koutrika, G., Ioannidis, Y. E.,: "Personalization of Queries in Database Systems", In Proc. of the 20th Int. Conference on Data Engineering, Boston, USA, p. 597-608, 2004.

[5] Internet Movie Database, Inc.: "The Internet Movie Database", Web site, URL: http://www.imdb.com/, last accessed on July 9th, 2007.

[6] Peralta, V.: "Extraction and Integration of MovieLens and IMDb Data". Technical Report, Laboratoire PRiSM, Université de Versailles, France, July 2007.

[7] Peralta, V.: "Generation of a Reference Data Set for Query Personalization". Technical Report, Laboratoire PRiSM, Université de Versailles, France, October 2007.

[8] Text REtrival Conference (TREC). URL: http://trec.nist.gov/, last accessed on September 2007.

[9] Transaction Processing Performance Council. URL: http://www.tpc.org/, last accessed on September 2007.

[10] URL of the APMD-Workbench: A Benchmark for Query Personalization Systems: http://apmd.prism.uvsq.fr/SubProject4/TestPlatform/.