A CONTEXT-BASED MODEL FOR WEB QUERY REFORMULATION

Ounas Asfari, Bich-Liên Doan, Yolaine Bourda
Department of Computer Science, SUPELEC, 3 rue Joliot-Curie, Gif-sur-yvette, France

Jean-Paul Sansonnet
LIMSI-CNRS, University of Paris 11, Orsay, France

Keywords: Query Reformulation, Context, Task modeling, Information Retrieval, Personalization.

Abstract: Access to relevant information adapted to the needs and the context of the user is a real challenge in the Web Search, owing to the increase of heterogeneous resources on the web. In most of cases, user queries are shortened and ambiguous, thus we need to handle implicit needs or intentions that are behind these queries. For improving user query processing, we present a context-based method for query expansion that automatically generates context-related terms. Here, we consider the user context as the current state of the task that the user is undertaking when the information retrieval process takes place, thus State Reformulated Queries (SRQ) are generated according to the user task state and the ontological user profile to provide personalized results in a particular context.

1 INTRODUCTION

In the present, it has become common to seek daily information on the web, including such tasks as using information retrieval system for shopping, travel booking, academic research, and so on. Understanding the user task is critical to improve the processing of user needs. The increase of mobile devices (such as PDA, cellular phone, laptop…) including diverse platforms, various work environments, have created new considerations and stakes to be satisfied. Thus, it is now expected to use the mobile device anywhere to seek information needed to perform a current task, but the problem is that the classic information retrieval systems provide same results for different needs, contexts, intentions and personalities, so too many irrelevant results are provided, it is often difficult to distinguish context-relevant information from secondary information or even noise. Thus the results provided for mobile users to perform tasks must be related to the context.

The user context can be assimilated to all factors that can describe his intentions and perceptions of his surroundings (Mylonas et al., 2008). These factors may cover various aspects: environment (light, services, people…), spatial-temporal (location, time, direction…), personal (physiological, mental, professional …), social (friends, colleagues…), task (goals, information task), technical etc. Many studies (Mylonas et al., 2008), (Sieg et al., 2007) try to take into account the user context but the problems to be addressed here include how to represent the context, how to determine it at runtime, and how to use it to influence the activation of user preferences. It is very difficult to take into consideration all the contextual factors in one information retrieval system, so the researchers often define the context as certain factors (location for example). In this paper our definition of the context is that the context describes the user current task, its changes over time and its states, i.e. we take into account the task which the user is undertaking when the information retrieval process occurs.

Recent studies have tried to dynamically enhance the user query with the user's preferences by creating a user profile for providing personalized results (Micarelli et al., 2007). However, a user profile may not be sufficient for a variety of queries of the user. One disadvantage of automatic personalization techniques is that they are generally applied out of context. Thus, not all of the user interests are relevant all of the time, usually only a

Asfari O., Doan B., Bourda Y. and Sansonnet J..
A CONTEXT-BASED MODEL FOR WEB QUERY REFORMULATION.
DOI: 10.5220/0003100802880293
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 288-293
ISBN: 978-989-8425-28-7

subset is active for a given situation, and the rest cannot be considered as relevant preferences.

User query is an element that specifies an information need, but queries, especially short one as mobile user queries, do not provide a complete specification of the information need. Many relevant terms can be absent from queries and terms included may be ambiguous, thus queries must be processed intelligently to address more of the user's intended requirements. Typical solution includes expanding query representation that refers to methods of query reformulation, i.e., any kind of transformation applied to a query to facilitate a more effective retrieval. This paper presents a method to reformulate user queries depending on the user profile, containing his interests, together with the user context which is considered as the actual state of the user current task in order to provide personalized results in context. We combine linguistic knowledge about query using WordNet and semantic knowledge using ODP ontology (Open directory Project www.dmoz.org) and knowledge about user (user profile and user task context) into a single framework in order to provide the most appropriate answer for a user's information needs in the search time and task state.

The rest of the paper is organized as follows: Section 2 shows the related work; Section 3 presents our models to reformulate user's queries; section 4 presents the system architecture; Section 5 shows the experimental study; Finally, Section 6 gives the conclusion.

2 RELATED WORK

Many studies have been employed to expand the user query in information retrieval, as far as we know these studies don't depend on the user task, in this paper we depend on a task model for expansion the user query, thus in section 2.1, we describe related work where the query expansion had been investigated. In section 2.2, we review studies where task model had been used.

2.1 Query Expansion

Query expansion is the process of augmenting the user's query with additional terms in order to improve results by including terms that would lead to retrieving more relevant documents. Many works have been done for providing personalized results by query reformulation. Approaches based on the user profile for query enrichment have been proposed, this process consists in integrating elements of the user profile into the user's query (Koutrika et al., 2004). The limitation of these approaches is that they do not take into consideration the user context to activate elements from the user profile.

Studies on query reformulation by relevance feedback are proposed, the aim is to use the initial query in order to begin the search and then use information about whether or not the initial results are relevant to perform a new query (Lv and Zhai, 2009). Because relevance feedback requires the user to select which documents are relevant, it is quite common to use negative feedback. Furthermore the techniques of disambiguation aim to identify precisely the meaning referred by the terms of the query and focus on the documents containing the words quoted in the context defined by the corresponding meaning. But this disambiguation may cause the query to move in a direction away from the user's intention and augment the query with terms related to the wrong interpretation.

Many approaches, like (Bhogal et al., 2007), try to reformulate the web queries based on semantic knowledge by using ontology in order to extract the semantic domain of a word and add the related terms to the initial query, but sometimes these terms are related to the query only under a particular context. Others use sense information (WordNet) to expand the query (Navigli and Velardi, 2003). In this paper, we propose a hybrid query expansion method that automatically generates query expansion terms from the user profile and the user task. In our approach we exploit both a semantic knowledge (ODP Ontology) and a linguistic knowledge (WordNet) to learn the user's task, and we exploit UML states diagram to represent the user current task.

2.2 Task Model

One aspect of characterizing user's contexts is to consider the tasks which have led them to engage in information retrieval behavior. Users use documents to understand a task and solve a specific problem. Thus, when a user begins a task, he searches the information that will help solve the problem at hand. Various researchers have demonstrated that the desired search results differ according to types of tasks. According to (Terai et al., 2008) two types of tasks: Informational task which involves the intent to acquire some information assumed to be present on one or more web pages; transactional task which is based on the intent to perform some web-mediated activity. The approach (Freund et al., 2005) proves that the nature of the task has an impact on decisions of relevance and usefulness. In the approach

(Luxenburger et al., 2008) a language model of a user task is defined as a weighted mixture of task components: queries, result sets, click stream documents, and browsed documents. Most existing systems do not integrate user needs with the characteristics of the relevant task state as the execution of the task progresses.

3 MODELS AND ALGORITHMS

Our aim is to provide context-based personalized results by improving the user web-queries intelligently. We consider the user current task as a contextual factor. Here we will describe our models for detecting the user current task, constructing an ontological user profile and generating the reformulated queries.

3.1 General Language Model

We construct here a new general language model for query expansion including the contextual factors and user profile in order to estimates the parameters in the model that is relevant to information retrieval systems. In the language modeling framework, a typical score function is defined in KL-divergence as follows (Bouchard and Nie, 2006):

$$Score(Q,D) = \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) \propto -KL(\theta_Q \| \theta_D) \quad (1)$$

Where: θ_D is a language model created for a document D, θ_Q a language model for the query Q, generally estimated by relative frequency of keywords in the query, and V the vocabulary.
$P(t|\theta_D)$: The probability of term t in the document model, $P(t|\theta_Q)$: The probability of term t in the query model.

The basic retrieval operation is still limited to keyword matching, according to a few words in the query. To improve retrieval effectiveness, it is important to create a more complete query model that represents better the information need. In particular, all the related and presumed words should be included in the query model. In these cases, we construct the initial query model containing only the original terms, and a new model SRQ (state reformulated queries) containing the added terms. We generalize this approach and integrate more models for the query. Let us use θ_Q^0 to denote the original query model, θ_Q^T for the task model, θ_Q^S for the contextual state model, and θ_Q^U for a user profile model. θ_Q^0 can be created by MLE (Maximum Likelihood Estimation). Given these models, we create the following final query model by interpolation:

$$P(t \mid \theta_Q) = \sum_{i \in X} \alpha_i P(t \mid \theta_Q^i) \quad (2)$$

Where: X= {0, T, S, U} is the set of all component models and α_i (with $\sum_{i \in X} \alpha_i = 1$) are their mixture weights, Thus the (1) becomes:

$$Score(Q,D) = \sum_{t \in V} \sum_{i \in X} \alpha_i P(t \mid \theta_Q^i) \log P(t \mid \theta_D) = \sum_{i \in X} \alpha_i Score_i(Q,D) \quad (3)$$

Where: $Score_i(Q,D) = \sum_{t \in V} \alpha_i P(t \mid \theta_Q^i) \log P(t \mid \theta_D)$ (4)
is the score according to each component model.

3.2 Constructing Task Model

The task model is used to detect and describe the task performed by the user, when he submits his query to the information retrieval system. In this paper we depend on study questionnaires (W. White and Kelly, 2006) which were used to elicit tasks that were expected to be of interest to subjects during the study. A generic classification was devised for all tasks identified by all subjects, producing the following nine task groupings:

Academic Research; News and Weather; Shopping and Selling; Hobbies and Personal Interests; Jobs/Career/Funding; Entertainment; Personal Communication; Teaching; Travel.

We generate a UML states diagram for each task in order to detect the changes in the task-needs over time and for describing all the sequences of the performed task. This generated diagram contains the task states and at least one attribute for each one. Accordingly, an index is built for: the terms of the tasks, the terms of its states including the state attributes, and the related task concepts from ODP. Thus this index consists of r terms. We will use this index when using the term vector model.

The user task can be identified automatically by taking advantages of existing linguistic resources (WordNet) and semantic resources (ODP Ontology) for assigning a task to user query. For that, we apply the following algorithm:

Let *q* a query submitted by a specific user at the current task denoted A_*. This query is composed of n terms; it can be represented as a term vector:

$$\vec{q} = \langle t_1, t_2, \ldots, t_n \rangle$$

For this query $\vec{q}$ a current task A_* is built by a single term vector:

$$\vec{A_*} = \langle a_{s1}, a_{s2}, \ldots, a_{si} \rangle$$

Where: a_{S1}, a_{S2}, …a_{Si} the terms that represent the state attributes of the task states s_1, s_2, …s_i for the

current task A_*. For example, if the actual state is "Find a Restaurant", then the state attribute will be "Restaurant" and a value from the user profile (such as vegetarian) will be assigned to this state attribute in order to personalize the query.

The initial query q is parsed using WordNet in order to identify the synonymous terms:

$$\vec{q}_w = \langle t_{w1}, t_{w2}, \ldots, t_{wn} \rangle$$

The query $\vec{q}_w$ is queried against the ODP ontology in order to extract a set of concepts ($c_1, c_2 \ldots, c_m$ with $m \geq n$) that reflect the semantic knowledge of the user query. These concepts of the user query and its sub-concepts are represented as a single term vector:

$$\vec{C}_q = \langle c_1, c_2, \ldots, c_m \rangle$$

Then the concepts are compared with the previous nine tasks, to do this, we compute the similarity weight between $\vec{C}_q$ and the proposed nine tasks, depending on the task index which is previously explained:

$$SW(A_1) = \text{Cos}(\vec{C}_q, \vec{A}_1)$$

$$SW(A_2) = \text{Cos}(\vec{C}_q, \vec{A}_2)$$

$$SW(A_9) = \text{Cos}(\vec{C}_q, \vec{A}_9)$$

Finally, the task A_* corresponding with the maximum similarity weight ($Max\ (SW\ (A_*))$) is automatically selected as the current task. Figure 1 illustrates the various vectors.

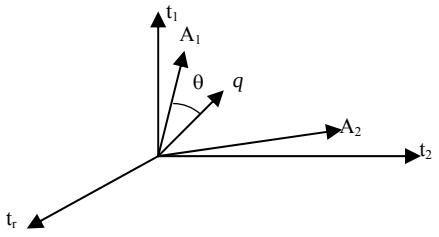

Figure 1: Representation of the tasks and the query as term vectors.

Where: query terms: $t_1, t_2, \ldots, t_n$. Terms of task index: $t_1, t_2, \ldots, t_r$. Terms of task state attributes: $a_{S1}, a_{S2}, \ldots, a_{Si}$. Each term's weight is computed using tf * idf weighting scheme.

3.3 Contextual State Model

The contextual state model is responsible for determining and analyzing the actual state of the current task. We suppose that the different states of the current task are modeled using an UML state diagram. There is at least one relevant attribute a_{si} for each detected state S_i. Because mobile device moves with the user, it is possible to take into account the actual task state in which the user is in when submitting certain queries to the information retrieval system IRS. Such contextual information may come automatically from various sources such as the user's schedule, sensors, entities that interact with the user; it may also be created by the user.

According to our assumption, we have defined 9 UML state diagrams for the main pre-defined tasks. After the user's query is submitted to our platform, the related task is assigned automatically to the user query and a set of SRQ (State Reformulated Queries) related to each state is presented to the user. The user is then asked to choose the appropriate SRQ according to his state. Finally, the contextual model will follow the UML state diagram to present the next SRQ.

3.4 Ontological user Profile Model

Ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. Concepts (or classes or categories) appear as nodes in the ontology graph. A user profile is a collection of personal data associated to a specific user. *The Ontological user profile* is constructed by the representation of the user profile as a graph of related concepts of the ODP ontology, inferred using an index of user documents. Here, a dynamic ontological user profile is considered as semi-structured data in the form of attribute-value pairs where each pair represents a profile's property. The properties are grouped in categories or concepts using ODP Ontology. In the proposed ontological user profile the annotating of each concept is done by giving value for each attribute in the ontology concept based on an accumulated similarity with the index of user documents, a user profile is created consisting of all concepts with non null value. Using ontology as the basis of the profile allows the initial user behavior to be matched with existing concepts in the domain ontology and relationship between these concepts. When the ontological user profile is created, its query-related concepts must be activated. This is done by mapping the query context $C_q = \langle c_1, c_2, \ldots, c_m \rangle$ on this ontological user profile (note that, the query context is calculated during the construction of the task model). This allows activating for each query context concept its semantically related concepts from the ontological user profile, following our contextual approach depending on the relevant propagation (Asfari, 2008). Hence, the relevant user profile attributes that are determined by the previous activated concepts are found. This attributes with its values are used to reformulate the user query.

3.5 SRQ Model (State Reformulated Queries)

Query expansion is the process of adding relevant terms to the original query (Asfari et al., 2009). However, in a more general sense, it also refers to methods of query reformulation, Thus we look for a relevant terms to use it in query expansion, that means we look for terms that are related to the query, the user, and the task state in the same time and don't contain unrelated terms. The initial user query is reformulated depending on these relevant terms in order to produce SRQ (State Reformulated Query) to improve the retrieval performance. The two aspects for producing SRQ are: query expansion and query refinement.

Query Expansion: the initial query is expanded with tow type of generated terms:

- The terms that represent the state attributes, from UML state diagram, for the current task A_* (denoted a_{S1}, a_{S2}, …,a_{Si}) One state attribute for each task state.
- The query-relevant attributes from the ontological user profile with its values. ($<a_{u1}$, $va_{u1}>$, $< a_{u2}$, v $a_{u2}>$, …,$<a_{uj}$, v $a_{u3}>$)

Query Refinement: Sometimes irrelevant attributes may be present in the selected user profile concepts. In order to keep only the relevant user profile attributes for the current task state S_i, we compare between these generated attributes and the current state attributes, next we exclude from the generated user profile attributes these non similar with the state attributes. We must also exclude the duplicated terms if they exist in the resulting SRQ. Finally SRQ is built according to the syntax required by the used search engine in order to submit the SRQ and to provide back results to the user.

Let q= $\{t_1, t_2..., t_n\}$ initial query which is related to task at hand. The state reformulated query in the task state S_i and for a specific user profile P_j is: $S_iRQ<Q,P_j,S_i>$, The relevant results D_i in the states S_i are produced by applying $S_iRQ<Q,P_j,S_i>$ on an information retrieval system. We expect that the results D_i in the task state S_i are more relevant than the normal results produced by using the initial query q in S_i.

4 SYSTEM ARCHITECTURE

Figure 2 illustrates the system architecture. It combines the several models described in the previous section: the task model, the contextual state model, the ontological user profile model and the SRQ model.

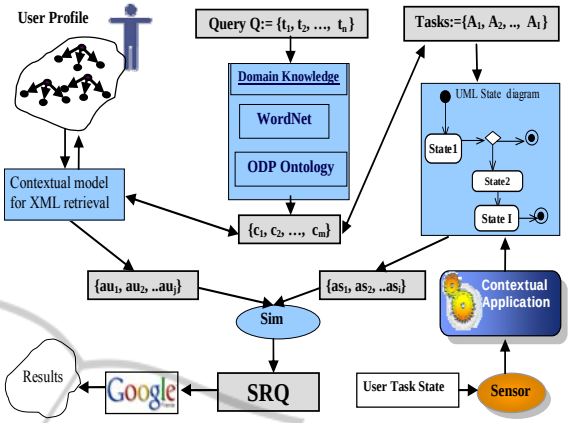

Figure 2: System architecture.

5 EXPERIMENTAL STUDY

Here we first suppose that the queries we are considering are related to current task at hand and secondly, the tasks are modeled by UML state diagrams. Our system works depending on the following practical consequent steps:

When the user submits his query in our platform, the task model will assign a task for this query as the first step. Next, the UML state diagram for this task is retrieved. The system then uses the attributes associated with each state (in UML) and the user profile attributes for producing the relevant terms, next the irrelevant terms are excluded, finally, the reformulated query denoted SRQ is submitted to Google to retrieve the relevant results. For example, let *q*: "*Trip to Paris*", the task model assigns the task "*Travel*" to this query, and then the contextual state model allows the proposition of several task states that are represented in UML state diagram as shown in the figure 3. Next the system can present the following SRQ:

- S_1RQ :{Paris + Airline OR Book ticket OR Inexpensive}
- S_2RQ :{Paris + hotel+2 star OR single}
- S_3RQ :{Paris+ Monuments OR Weather OR plan OR Metro}
- S_4RQ :{Paris+ restaurant OR Italian Cuisine OR Vegetarian Food}
- S_5RQ:{Paris + Photos}
- S_6RQ: {Paris+ News + Weather}

The evaluation of such systems is complicated due to the dynamic aspect of the system environment. So, we performed two manual evaluations, one to evaluate the detected task and

another to evaluate the SRQ (State Reformulated Queries):

We asked 10 different users to submit 3 queries (for doing different tasks), the system then detects the task for each query. Next the users are asked if the tasks were similar to their tasks or not. We then got nearly 21 out of 30 positive responses (70%). To evaluate the SRQ queries we asked the 10 users to submit different queries and we applied each one to the Google search engine at the different states of the task which was proposed by our task model. We reformulated these queries by adding the relevant terms and then we reapplied them at the states using the same search engine. We compared the first 20 retrieval results produced in the two cases (by queries q and queries SRQ).

Results: we calculated the average number of relevant pages by queries *q* and SRQ on the first 20 results (P@20). We noticed that the precision of the relevant results using the initial query *q* is 0.17 and 0.59, respectively, by using SRQ queries which were reformulated depending on the current task state and user profile.

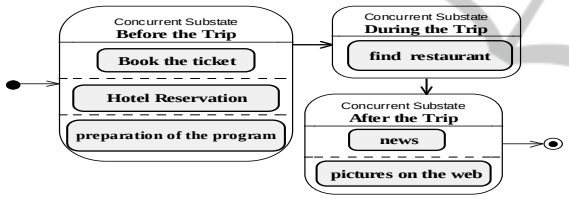

Figure 3: Shows an example of a "travel task" that is modeled by UML state diagram.

6 CONCLUSIONS

In this paper, we have proposed a hybrid method to reformulate user queries depending on a dynamic ontological user profile and user context for producing State Reformulated Queries (SRQ). The user context is considered as the actual state of the task that he is undertaking when the information retrieval process is performed. We have constructed a general architecture that combines several models for query expansion: the task model, the contextual model, the user profile retrieval model and SRQ model. We exploit both a semantic knowledge (ODP Ontology) and a linguistic knowledge (WordNet) to learn user's task, and we exploit a UML states diagram for this task to learn user current state. We have also constructed a new general language model for query expansion including the contextual factors and user profile. We have illustrated on an experimental study that the results obtained by SRQ queries are more relevant than those obtained with the initial user queries in the same task state. As a future work, we plan to evaluate this method by creating a test collection.

REFERENCES

Asfari, O., Doan, B. L., Bourda, Y. Sansonnet, J. P.: Personalized access to information by query reformulation based on the state of the current task and user profile, In: *The Third International Conference on Advances in Semantic Processing*, Malta, 2009.

Asfari, O.: Modèle de recherche contextuelle orientée contenu pour un corpus de documents XML. In: *CORIA 2008*: 377-384, France.

Bouchard, H., Nie, J. Y.: Modèles de langue appliqués à la recherche d'information contextuelle, In: *Conf. en Recherche d'Information et Applications (CORIA)*, Lyon, 2006.

Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion, *Information Processing and Management. In: an International Journal*, v.43 n.4, p.866-886, July, 2007.

Freund, L., E. G. Toms, and Charles, L. A. Clarke: Modeling task-genre relationships for IR in the workplace, *In: SIGIR 2005, at Salvador*, Brazil.

Koutrika, G., Ioannidis, Y. E., Personalization of Queries in Database Systems, In: Proceedings of 20th International Conference on Data Engineering, 2004.

Lv, Y., Zhai C.: *Adaptive Relevance Feedback in Information Retrieval, IN: CIKM, 2009*, Hong Kong.

Luxenburger, J., Elbassuon, S., Weikum, G.: Task-aware search personalization, In: Proceedings of the 31st annual international *ACM SIGIR, Singapore, 2008*.

Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S.: Personalized Search on the World Wide Web. P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), *In: The Adaptive Web*, LNCS 4321, Berlin, 2007.

Mylonas, Ph., Vallet, D., Castells, P., Fernandez, M., and Avrithis, Y., Personalized information retrieval based on context and ontological knowledge. In: Knowledge Engineering Review, *Cambridge University, 2008*.

Navigli, R., Velardi, P.: An Analysis of Ontology-based Query Expansion Strategies, In: Workshop on Adaptive Text Extraction and Mining at the *14th Conference on Machine Learning*, Croatia 2003.

Sieg, A., Mobasher, B., and Burke, R.: Representing context in web search with ontological user profiles. In: Proceedings of the *6th International Conference on Modeling and Using Context, Denmark*, August 2007.

Terai, H., Saito, H., Takaku, M., Egusa, Y., Miwa, M., and Kando, N.: Differences between informational and transactional tasks in information seeking on the web. In: *Proceedings of the Second Symposium IIiX*, 2008.

W. White, R., Kelly, D.: A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance, In: *CIKM'06, 2006*, USA.