Inferring the user interests using the search history

Lynda Tamine, Mohand Boughanem, Nesrine Zemirli
IRIT-SIG
118 Route de Narbonne
31062 Toulouse CEDEX 06 France
tamine@irit.fr, bougha@irit.fr, nzemirli@irit.fr

Abstract

Personalization involves the process of gathering user-specific information during interaction with the user, which is then used to deliver appropriate results to the user's needs. This paper presents a statistical method that learns the user interests by collecting evidence from his search history. The method focuses on the use of both user relevance point of view on familiar words in order to infer and express his interests and the use of a correlation metric measure in order to update them.

1 Introduction

It is widely assumed nowadays that because of the explosive growth of web documents, keyword based search technologies are not effective, in the sense that they are not able to deliver appropriate results in response to specific user's information needs. The major reason is that they don't take into account the user profile in the retrieval process [Numberg, 2003; Budzik and Hammond, 1985].

Although, relevance feedback techniques [Rocchio, 1971] improve the retrieval accuracy by considering the user's preferences, they are not effective in real world applications [Kelly and Teevan, 2003]. [Budzik and Hammond, 1985] In order to tackle this problem, contextual information retrieval emerged recently as an active area. It explores the impact of context, viewed as a set of social, cultural and task features, on human information behaviour. Our interest in context is namely in defining user's profiles in order to constraint the semantic space of information determining the topical relevance. In this sense, several approaches explored techniques for building user's profile using implicit feedback [Pazzani and Billsus, 1997; Mc Gowan, 2003; Lieberman, 1995; Pretshner and Gauch, 1999; Liu and Yu, 2004; Budzik and Hammond, 1985]. Most of them model the user long-term interests as retrieval contexts represented by word vectors [Pazzani and Billsus, 1997; Lieberman, 1995; Budzik and Hammond, 1985],class of concepts [Mc Gowan, 2003] or a hierarchy of concepts [Liu and Yu, 2004; Pretshner and Gauch, 1999].

This paper presents a new technique for building and learning the user interests accross past search sessions. Comparatived to previous work, our approach has the following new features:

- related and unrelated user's interests are dynamically inferred from the search history using a statistical rank-order correlation operator,
- rather than using a basic Tf-Idf word weighting scheme in the user profile representation, we propose a new measure to estimate the relevance of the words according to the user interests.

In section 2, we present the strategy of collecting and modeling the user's search history . In section 3, we explain how they are used to learn the user's interests.

2 Building the user profile using search history

In our point of view, a user profile expresses the user long-term interests. It contains two related components: an aggregative representation of the user search history and a library of user contexts reflecting his interests when seeking information. More precisely, our approach uses the evidence collected across successive search sessions in order to track potential changes in the user's interests. At time s, the user profile is represented as $U = (H^s, I^s)$ where H^s and I^s represent respectively the search history and a set of interests of the user U at time s. Our method runs in two main steps. The first one consists of representing the user search history by collecting information from user feedback at each retrieval session, and then gathering this information in order to infer the user contexts expressed using a set of weighted dominant keywords. The second step consists of learning the user interests by using the contexts discovered during the previous step. The learning algorithm is based on a correlation measure used to estimate the level of changes in the user interests structure during a period of time.

2.1 Representation of the user search history

Let q^s be the query submitted by a specific user U at the retrieval session performed at time s. We assume that a document retrieved by the search engine with respect to q^s is relevant if it generates some observable user behaviours (reading during a reasonable duration, saving, printing etc) or it is explicitly judged as relevant by the user. Let D^s be the related set of assumed relevant documents during session S^s, $R_u^s = \cup_{s_0..s} D^s$ represents the potential space search of the user across the past search sessions. We use matrices to represent both user search session and search history. The construction of this matrix, described below, is based on the user's search record and some features inferred on the user relevancy point of view. The user search session is represented by a Document-Term matrix S^s $D^s * T^s$ where T^s is the set of terms indexing D^s (T^s is a part of all the representative terms of the previous relevant documents, denoted $T(R_u^s)$. Each row in the matrix S^s represents a document $d \in D^s$, each column represents a term $t \in T^s$. In order to improve the accuracy of document-term representation, we aim at introducing in the weighting scheme a

factor that reflects the user's interest for specific terms. For this purpose, we use term dependencies as association rules checked among T^s [Lin *et al.*, 1998] in order to compute the user term relevance value of term t in document d at time s denoted $RTV^s(t,d)$:

$$RTV^s(t,d) = \frac{w_{td}}{dl} * \sum_{t' \neq t, t' \in D^s} cooc(t,t') \quad (1)$$

w_{td} is the common Tf-Idf weight of the term t in the document d, dl is the length of the document d, $cooc(t,t')$ is the confidence value of the rule $(t \rightarrow t')$, $ccoc(t,t') = \frac{n_{tt'}}{n_t}$, $n_{tt'}$ is the number of documents among D^s containing t and t', n_t is the number of documents among D^s containing t. $S^s(d,t)$ is then determined as:

$$S^s = (RTV^s)^t \quad (2)$$

The user search history is a $R_u^s * T(R_u^s)$ matrix, denoted H^s, build dynamically by reporting document information from the matrix S^s and using an aggregative operator combining for each term its basic term weight and relevance term value computed across the past search sessions as described above. More precisely, the matrix H^s is built as follows:

$$H^0(d,t) = S^0(d,t)$$

$$H^{s+1}(d,t) = \begin{cases} \alpha * w_{t,d} + \beta * S^{s+1}(d,t)\ if \\ t_j \notin T(R_u^{(s-1)}) \\ \alpha * H^s(d,t) + \beta * S^{s+1}(d,t) if \\ t_j \in T(R_u^{(s-1)}) and\ d \in R_u^{(s-1)} \\ H^s(d,t)\ otherwise \end{cases} \quad (3)$$

$(\alpha + \beta = 1), s \succ s_0$

2.2 Learning the user's interests

The goal of this step is to extract the user contexts from his search history in order to learn his long-term interests. For this purpose, we propose a statistical method that constructs and updates a set of user's interests I^s. This method induces at each learning period, a set of beliefs on the user contexts represented each one as a set of weighted key words. At learning time s, an ordered vector denoted c^s reflecting a query context, is built using the formula:

$$c^s(t) = \sum_{d \in R_u^s} H^s(d,t) \quad (4)$$

$c^s(t)$ is normalised as follows: $c^s(t) = \frac{c^s(t)}{\sum_{t \in T^s} c^s(t)}$. In order to track the changes in the user's interests, we compare the current context cc^s and the previous one pc^s using Kendall rank-order correlation operator $\circ$:

$$\Delta I = (cc^s \circ pc^s) = \sum_{t \in T(R_u^s)} (cc^s(t) - pc^s(t)) \quad (5)$$

The coefficient value ΔI is in the range [-1 1], where a value closer to -1 means the query contexts are not similar and a value closer to 1 means that the query contexts are very related each other. Based on this coefficient value, we apply the following strategy in order to learn the user's interests and so update the set of user interests I^s :

1. $\Delta I > \sigma$ (σ represents a threshold correlation value). No potential changes in the query contexts, no information available to update I^s;
2. $\Delta I < \sigma$. There is a change in the query contexts. In this case we gauge the level of change: is this reflects a refinement of a prior detected user interest or the occurrence of a novel one? In order to answer this question we do as follows:
   - select $c^* = argmax_{c \in I^s}(c \circ cc^s)$,
   - if $cc^s \circ c^* > \sigma$ then refine the user interest c^*, update the matrix H^s by dropping the rows representing the least recently documents updated, update consequently R_u^s,
   - if $cc^s \circ c^* < \sigma$ then add the new tracked interest in the library I^s, try to learn a period of time c^*: set $H^{s+1} = S^s$, $s_0 = s$

3 Conclusion and future work

In this paper, we described a new approach for user profiling using statistical methods to gather the search history and track changes in user's interests. Unlike most previous related work, we focus on the updating of the search history representation using user relevance point of view on familiar words, in order to build and learn different user's interests. The design of an experimental evaluation of our approach requires a large scale of quantitative data on user search sessions and accurate contexts provided by the related queries during a reasonable period of testing a particular search engine. We currently develop an evaluation methodology which includes the construction of such collections test and the definition of accurate performance measures.

4 Ackowledgments

This research was partially supported by the French Ministry of Research and New Technolologies under the ACI program devoted to Data Masses (ACI-MD), project MD-33.

References

[Budzik and Hammond, 1985] J. Budzik, K.J Hammond. Users interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th international conference on intelligent user interfaces*, pp 44-51, 2000.

[Mc Gowan, 2003] J.P Mc Gowan. A multiple model approach to personalised information access. Master Thesis in computer science, Faculty of science, University College Dublin, February 2003.

[Kelly and Teevan, 2003] D. Kelly, J. Teevan. Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, 2003

[Lieberman, 1995] H. Lieberman. Letizia, *"An agent that assists web browsing"*. In *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI'95), pp 924:929, Montreal, August 1995

[Lin *et al.*, 1998] S.H. Lin, C.S. Shih, M.C. Chen, J. Ho, M. Ko and Y. M. Huang. Extracting classification knowledge of Internet documents with mining term-associations: A semantic approach. In *the 21th International SIGIR Conference on Research end Development in Information Retrieval*, 1998

[Liu and Yu, 2004] F. Liu, C. Yu. Personalized Web search for improving retrieval effectiveness. IEEE Transactions on knowledge and data engineering, 16(1), pp 28-40, 2004

[Numberg, 2003] G. Nunberg As Google goes, so goes the nation, New York times, May 2003

[Pazzani and Billsus, 1997] M. Pazzani, D. Billsus. Learning and revising user profiles : The identification of interesting Web sites, Machine learning, Vol 27, pp 313-331, 1997

[Pretshner and Gauch, 1999] A. Pretshner, S. Gauch. Ontology based personalised search, In *Proceedings of the 8th IEEE International Conference, Tools with Artificial Intelligence* (ICTAI), pp 391-198, 1999

[Rocchio, 1971] J. Rocchio. Relevance feedback in information retrieval, In G. Salton editor, The SMART retrieval system - experiments in automated document processing. Prentice-Hall, Englewood Cliffs, NJ, 1971