Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 385 – 388 Sept. 2006 ISSN 1003—7985

Ontology-based framework for personalized recommendation in digital libraries

Yan Duanwu Cen Yonghua Zhang Wei Mao Ping

(Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: To promote information service ability of digital libraries, a browsing and searching personalized recommendation framework based on the use of ontology is described, where the advantages of ontology are exploited in different parts of the retrieval cycle including query-based relevance measures, semantic user preference representation and automatic update, and personalized result ranking. Both the usage and information resources can be exploited to extract useful knowledge from the way users interact with a digital library. Through combination and mapping between the extracted knowledge and domain ontology, semantic content retrieval between queries and documents can be utilized. Furthermore, ontology-based conceptual vector of user preference can be applied in personalized recommendation feedback.

Key words: digital library; personalized recommendation; ontology; content retrieval; user preference

One of the most important resources for supporting users in a distance e-learning environment is the possibility of accessing a digital library, which allows the users to collect and organize the necessary information for achieving their particular goals. In general, people have two ways to find the information they are looking for: searching and browsing. Searching engines index millions of documents and allow users to enter keywords to retrieve documents that contain these keywords. Browsing is usually done by clicking through a hierarchy of subjects until the area of interest has been reached. Searching and browsing algorithms are essentially the same for all users.

Indeed, in terms of searching, about one half of all retrieved documents have been reported to be irrelevant[1]. The main problem is that there is too much information available, and that keywords are not always an appropriate means of locating the information in which a user is interested[2]. An effective personalization system can decide autonomously whether or not a user is interested in a specific webpage or document and, in the negative case, prevent it from being displayed.

Ontologies achieve a reduction of ambiguity, and bring powerful inferencing schemes for reasoning and querying[3]. The use of ontologies for describing the possible scenarios of use in a digital library enables the possibility of predicting user requirements in advance and to offer personalized services ahead of expressed need.

1 Ontology-Based Personalized Recommendation Framework

Two elements determine the functionalities of the desired personalization system: first, the user's profile, including navigational history and user preferences; secondly, the information collected from the navigational behavior of the digital library users. The user profile should include all the information relevant to users: personal information, which is publicly made available by each user in order to facilitate the discovery of similar interests, and navigational history and behavior records, which will be used together with the personal information to build the set of recommendations. This information should help the user to improve his or her searching, by obtaining additional information when searching or browsing.

Generally speaking, information retrieval deals with modeling information needs, content semantics, and the relationships between them[4]. This involves modeling and capturing such user interests, and relating them to content semantics in order to predict the relevance of content objects, considering not only a specific user's request but the needs of all the users. When it comes to the representation of semantics (to describe content, user interests, or user requests), ontologies provide a highly expressive ground for describing units of meaning and a rich variety of interrelations among them. Ontologies achieve a reduction of ambiguity, and make available powerful inferencing schemes for rea-

Received 2006-04-20.

Foundation item: The Young Teachers Scientific Research Foundation (YTSRF) of Nanjing University of Science and Technology in the Year of 2005—2006.

Biography: Yan Duanwu (1976—), male, doctor, lecturer, yanwu_nju@163.com.

soning and querying. Not surprisingly, there has been a growing body of literature in the last few years that studies the use of ontologies to improve the effectiveness of information retrieval[5-7] and personalized search[8].

In this paper, we present a comprehensive personalized retrieval framework where the advantages of ontologies are exploited in different parts of the retrieval cycle: query-based relevance measures, semantic user preference representation, automatic preference update, and personalized result ranking. The framework is set up in such a way that the models benefit from each other and from the common, ontology-based grounding. In particular, the formal semantics are exploited to improve the reliability of personalization.

2 Ontology-Based Content Retrieval

Our ontology-based framework assumes the availability of a corpus D of text or multimedia documents, annotated by domain concepts (instances or classes) from an ontology-based knowledge base (KB) O. The KB is implemented using any ontology representation language for which appropriate processing tools (query and inference engines, programming APIs) are available. In our semantic search model, D rather than O is the final search space.

There are two phases in our retrieval model (see Fig. 1). In the first one, a formal ontology-based query is issued by some form of query interface (e. g. natural language processing-based) which formalizes a user information need. The query is processed against the KB using any desired inferencing or query execution tools, outputting a set of ontology concepts that satisfies the query. From this point, the second retrieval phase is based on an adaptation of the classic vector space information retrieval model, where the axes of the vector space are the concepts of O, instead of text keywords. As in the classic model, the query and each document are represented by vectors $\boldsymbol{q}$ and $\boldsymbol{d}$, so that the degree of satisfaction of a query by a document can be computed by the cosine measure:

$$\mathrm{sim}(d, q) = \frac{\boldsymbol{d} \cdot \boldsymbol{q}}{|\boldsymbol{d}| \cdot |\boldsymbol{q}|} \tag{1}$$

The problem remains to build $\boldsymbol{d}$ and $\boldsymbol{q}$ vectors, which is summarized next.

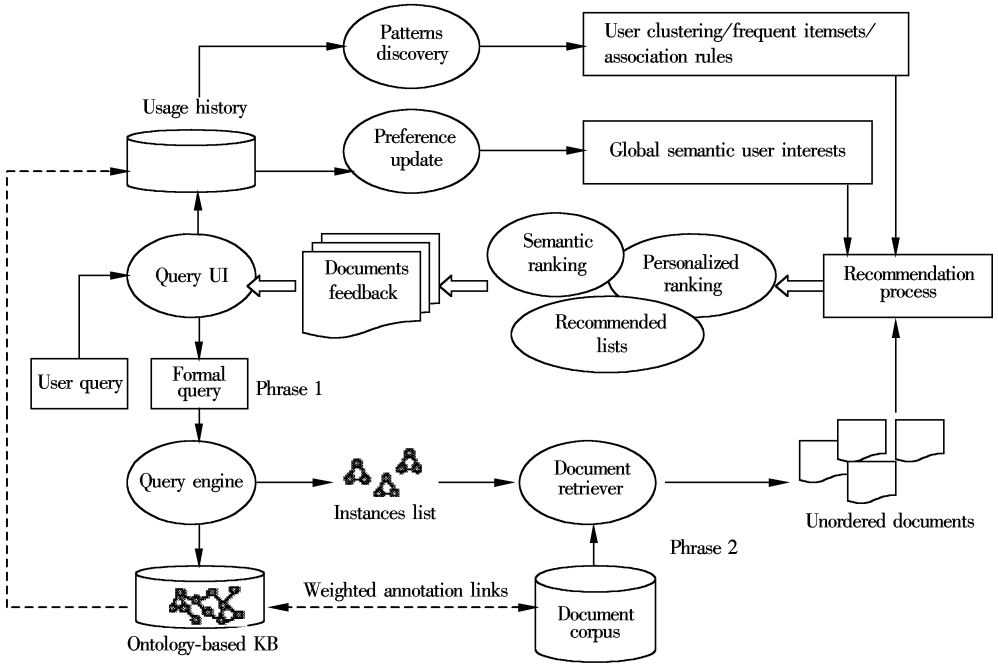

Fig. 1 Ontology-based framework for personalized recommendation

Document vectors Each content item in the search space D is represented by a vector $\boldsymbol{d}$ of concept weights, where for each domain concept $x \in O$ annotating d, d_x represents the importance of the concept x in the document (if x does not annotate d, then $d_x = 0$). The weight of annotations can be assigned by hand or automatically. If the document contains text, d_x can be computed automatically by a TF-IDF algorithm[9] as

$$d_x = \frac{f_{x,d}}{m_y f_{y,d}} \log \frac{|D|}{n_x} \tag{2}$$

where $f_{x,d}$ is the number of occurrences of x in d, m_y $f_{y,d}$ is the frequency of the most repeated instance in d, and n_x is the number of documents annotated by x. This requires that an appropriate mapping of concepts to text keywords be available, whereby the number of occurrences of a concept in a document can be defined as

the count of concept keywords in the text. What an appropriate mapping is in this context, and how it can be automated are subjects of active research.

Query vector The proposed construction of the query vector defines $q_x = 1$ if x appears in some tuple of the query result set, and 0 otherwise. The weights q_x can be further refined with a TF-IDF scheme, as suggested by[9]

$$q_x = 0.5 + 0.5 \frac{f_{x,q}}{m_y f_{y,q}} \log \frac{|D|}{n_x} \tag{3}$$

where we define $f_{x,q}$ as the number of tuples of the result set where x occurs.

3 Ontology-Based Personalized Recommendation

Personalization is a means to improve the performance retrieval (e. g., measured in terms of precision and relevance) as subjectively perceived by users. The key aspects involved include the representation of user interests (beyond a specific one-shot query), the dynamic acquisition of such interests by the system, and the exploitation of user preferences.

In our personalization framework, the semantic preferences of a user are represented as a vector $\boldsymbol{u} \in [0, 1]^{|O|}$ of concept weights, where for each domain concept $x \in O$, $u_x \in [0, 1]$ represents the intensity of the user interest for x. With respect to other approaches, where user interests are described in terms of preferred documents, words, or categories, here an explicit conceptual representation brings all the advantages of ontology-based semantics, such as reduction of ambiguity, formal relations and class hierarchies.

3.1 Automatic preference update

The extraction of preferences for semantic concepts is achieved by applying clustering algorithms on usage information data. The considered usage data consists of documents selected by the user for viewing them, or explicitly marked as relevant in relevance feedback sessions. Our approach for extracting preferences from the history of user interaction consists of the clustering of documents based on the semantic annotation that matches concepts to documents, by which common topics implicit in clusters of concepts are detected.

The concept-vector representation of documents described in section 2 can be reformulated to an equivalent interpretation of a document d as a normal fuzzy set on the set of concepts. Based on this set, and the knowledge contained in the form of available relations between the concepts, we aim to detect the degree to which a given document d is indeed related to a topic t. We will refer to this degree as $R(d, t)$. In other words, we attempt to calculate a relation $R: D \times T \rightarrow [0, 1]$, where D is the set of available documents and T is the set of topics. In designing an algorithm that can calculate this relation in a meaningful manner, it is necessary for the algorithm to be able to determine which of the topics are indeed related to a given document.

The topics that interest users, and should be classified as positive interests are the ones that characterize the detected clusters. Degrees of preference can be determined based on the cardinality of the clusters, i. e., clusters of low cardinality should be ignored as misleading and the high weights of topics in the context of the clusters indicate intense interest. The notion of high cardinality is modeled with the use of a large fuzzy number $L(\cdot)$, where $L(t)$ is the truth-value of the proposition "the cardinality of cluster t is high". Therefore, each of the detected clusters t is mapped to positive interests by

$$u_x = \sum_{t \in T} \mu(x, t) L(t) K(t) \tag{4}$$

for each $x \in O$, where $\mu(x, t)$ denotes the degree of membership of the concept x to the cluster t, and

$$K(t) = \bigcap_{d \in t} R(d, t) \tag{5}$$

3.2 Personalization effect promotion

Once a semantic profile of user preferences is obtained, either automatically as described in the previous section, and/or refined manually, our notion of preference-based content retrieval is based on the definition of a matching algorithm that provides a personal relevance measure $\mathrm{prm}(d, u)$ of a document d for a user u. This measure is set according to the semantic preferences of the user, and the semantic annotations of the document, weighted as explained in section 2. The procedure for matching d and u is based on a cosine function for vector similarity computation:

$$\mathrm{prm}(d, u) = \frac{\boldsymbol{d} \cdot \boldsymbol{u}}{|\boldsymbol{d}| \cdot |\boldsymbol{u}|} \tag{6}$$

In order to bias the result of a search (the ranking) to the preferences of each user, the measure above has to be combined with the query-based score without personalization $\mathrm{sim}(d, q)$ defined in section 2, to produce a combined ranking. The combination of several sources of ranking has been the object of active research in the field of IR. We have adopted the method of such combination, by which the two rankings are merged by a linear combination of the relevance scores:

$$\mathrm{score}(d, q, u) = \lambda \mathrm{prm}(d, u) + (1 - \lambda) \mathrm{sim}(d, q) \tag{7}$$

where $\lambda \in [0, 1]$. The choice of the λ coefficient in the

linear combination above is critical and provides a way to gauge the degree of personalization, from $\lambda = 0$ producing no personalization at all, to $\lambda = 1$, where the query (local user interests) is ignored and results are ranked only on the basis of global user interests.

Given the inherent ambiguity of user actions upon which user preferences are automatically inferred, the automatic preference extraction techniques have an unavoidable risk of guessing wrong preferences, the negative effects of which increase with λ. Even when the extraction is the most successful, there is considerable risk of contradicting explicit user requests if λ is too high, and λ should, therefore, be set with great care. It is commonly agreed that the user should have the means to turn personalization off ($\lambda = 0$), or even tune λ as a free parameter (see Ref. [10]). Other than this, a fixed moderate value for λ can be typically set by experimental tuning and can be automatically self-adjusted in the context of a search.

4 Conclusion

In this paper we have described a personalized recommendation framework, which uses all information relevant to the process of searching and browsing a digital library to build a complete navigational profile for each user. All these profiles are then combined with the help of an ontology that establishes the possible relationships between the elements presented in a typical scenario of use in a digital library integrated in an e-learning environment.

Ontology is a powerful tool for describing complex scenarios of use such as a digital library, where several concepts and relationships between these concepts can be identified and formally represented. The use of ontologies promotes the integration of new services into existing ones, and provides mechanism of user preference-oriented semantic recommendation feedback.

References

[1] Casasola E. ProFusion PersonalAssistant: an agent for personalized information filtering on the WWW [D]. Lawrence, KS: The University of Kansas, 1998.

[2] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems [J]. *ACM Transactions on Information Systems*, 2004, **22**(1): 5 – 53.

[3] Sadeh T, Walker J. Library portals: toward the semantic web [J]. *New Library World*, 2003, **104**(1184, 1185): 11 – 19.

[4] Micarelli A, Sciarrone F. Anatomy and empirical evaluation of an adaptive web-based information filtering system [J]. *User Modelling and User-Adapted Interaction*, 2004, **14**(2, 3): 159 – 200.

[5] Guha R V, McCool R, Miller E. Semantic search [A]. In: *Proc of the* 12*th Intl World Wide Web Conference* [C]. Budapest, Hungary, 2003. 700 – 709.

[6] Kiryakov A, Popov B, Terziev I, et al. Semantic annotation, indexing, and retrieval [J]. *Journal of Web Semantics*, 2005, **2**(1): 47 – 49.

[7] Vallet D, Fernández M, Castells P. An ontology-based information retrieval model [A]. In: *The Second European Semantic Web Conference* [C]. New York: Springer, 2005, **3532**: 455 – 470.

[8] Gauch S, Chaffee J, Pretschner A. Ontology-based personalized search and browsing [J]. *Web Intelligence and Agent Systems*, 2003, **1**(3, 4): 219 – 234.

[9] Baeza-Yates R, Ribeiro-Neto B. *Modern information retrieval* [M]. Translated by Wang Zhijin. Beijing: China Machine Press, 2005. (in Chinese)

[10] Google personalization[EB/OL]. (2004-11-05) [2006-02-10]. http://labs.google.com/personalized.

数字图书馆中基于本体的个性化推荐框架

颜端武　岑咏华　张　炜　毛　平

(南京理工大学信息管理系,南京 210094)

摘要:为了提高数字图书馆信息服务的能力,描述了一个基于本体的用户浏览和搜索个性化推荐系统框架.该框架将本体的优点应用于检索周期中,包括提问相关测度、语义化的用户兴趣表达和自动更新、以及个性化的检索结果排序等.在用户访问数字图书馆的交互过程中,可通过本体来构造用户提问和文档内容的匹配机制以实现语义化的内容检索,并可进一步使用本体来构造用户兴趣偏好的概念向量以实现面向用户的个性化推荐反馈.

关键词:数字图书馆;个性化推荐;本体;内容检索;用户偏好

中图分类号:TP391.3