An Unified User Dynamic Interest Description Model

Cheng Zeng, Hui Wang, Haiping Cao, Ying Zhang
State Key Lab of Software Engineering, Wuhan University, Wuhan, China, 430072
zengc@whu.edu.cn

ABSTRACT

The amount of information available world-wide and its network-based linkage will continue its rapid growth in the foreseeable future. It is urgent that we need a new retrieval mode to span the gap between information explosion and user requirement. This paper reports an user interest mining method based on cross-media and a unified model which describes user's dynamic interest changing through time to overcome the contradiction between short-time interest and long-time interest description in traditional methods. We also present a developing personalized retrieval system FizZ which mines user's interest by installing IE plug-in.

Keywords

UDM, MIS, Cross-media, interest community

1. INTRODUCTION

Current web search engines are very successful and are increasingly big business. They sharply shorten the time that people obtain information and bring lots of enterprises huge profits. But we have to recognize that these existing technologies could not satisfy users' particular information requirement. As evidenced in a recent study, "Contrary to many enterprises' expectations, search technology hasn't settled into a general satisfying phase. The biggest increases in the relevance of web search results to users will come neither by fine-tuning existing search algorithms nor from defining completely new algorithms based on web content and structure alone, the biggest increase in user' satisfaction will come when the results returned are tailored to the individual as well as to the question asked". This information deluge calls for new semantic extraction and the most innovative retrieval and exploration techniques.

The first problem of improving user satisfaction for retrieval results is to mine and understand user interest, and describe it with a reasonable mode. This paper presents a new method which mines user's interest and preference by analyzing various types of media documents that user browsed based on cross-media technology. Besides, we add time axis into user interest model to distinguish short-time interest from long-time interest and discovery the periodic interest change.

2. RELATIVE RESEARCHES

The preparation for constructing user interest description model(UM) is to collect user interest which includes explicit and implicit collecting. The former has been broad applied to lots of retrieval systems. The latter denotes that the system mines user implicit interest by analyzing Web log[Jou0a], such as LOGSOM[Jou0b], and tracking user's manipulation. Shahabi[Jou0c] sends agent to collect user interest in remote computers. It overcomes some problems such as error translating of Web cache and IP address etc., but it is possible to be limited if the client has security mechanisms. In addition, lots of systems [Jou0d] combine explicit and implicit methods in practice.

Current UM mainly includes vector space model, classes or layers model, catalogue structure model and so on. In [Con0a], Rachid adds user profile and weighted index item expression based on vector space model to realize personalized retrieval. Lin[Jou0e] presents a new mechanism describing and updating user interest model to realize personalized information service.

In addition, many researchers describe user interest with a tree structure based on the semantic relationship among interest lemmas. In [Con0b], ontology is utilized to improve the efficiency of personalized information retrieval and a framework is put forward which includes relationship measuring, user interest expressing and automatically updating. Yang[Jou0f] provides a complex preference operation for personalized retrieval and recommending in digital library.

3. DYNAMIC INTEREST DESCRIPTION

These methods introduced above emphasize particularly on text content mining or user action analyzing, but most of information in our daily life exist in multimedia form and the content that user is interested in begins to extend to more wonderful media types such as video, image, audio, etc.. The traditional interest mining method based on text will be not able to exactly and adequately collect user preference. Most of UM are inclined to describe long-time

This work is supported by the Doctor Subject Fund of Education Ministry No.20070486064 and the Hubei Province Natural Science Foundation of China No.2007ABA038 and National Basic Research Program of China(973 Program) No.2007CB310800 and the High School 111 Project of China No. B07037

interest, but ignores the instance that user interest fluctuates in short time, e.g. a user is used to browse investigative Web sites at work, but watch online movie during leisure time. The traditional UM is unfit for describing this kind of dynamic user interest. So this paper put forwards the user dynamic interest description model(UDM) based on multi-granularity interest space to resolve the problem.

Multi-granularity Interest Space

The multi-granularity interest space(MIS) is based on an interest concept hiberarchy(ICH) which is similar to ontology tree and is the tree catalog of all user interest points. Fig.1 shows the ICH and a 5+1 dimensions MIS based on a subset of ICH in terms of some application requirement, where d_i, t respectively represents the codes of interest concept and time dimension.

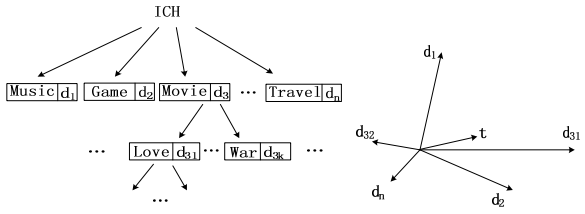


Figure 1. ICH and MIS

The goal of constructing MIS is to compute the similarity between different users' interest or between user interest and retrieval result. In actual application, we could casually choose concept nodes in appropriate ICH layer to construct a certain MIS. So the similarity computation not only considers the time consuming, but also is more desirable in real user interest.

The time dimension is added into MIS so that each user interest is represented by a dynamic twisted n- dimensions cube which corresponds to a interest sub-space. We look the cube as the time-space representation of UDM.

Mining User Interest Based on Cross-media

This paper mines user interest with the strategy combining explicit and implicit methods. The difference with traditional method is that cross-media technology is applied in user interest mining which discoveries cross-media relationship in different aspects. So different types of media with similar semantic could express each other.

For explicit user interest mining, we use the manner that user appoints local folders as analyzed objects other than inputting keywords in the course of registration. Users usually arrange the stored manner and types of local files according with their interest so that we could mine them in terms of some certain rules, e.g. the deeper path the file is stored, the lower interest user will show.

Then each document will be analyzed to mine the latent user interest. For visual media, semantic templates [Con0c] are utilized to extract semantic concepts. So text analysis could be used to cluster semantic concepts and discovery interest points. Audio is analyzed directly based on header information. By analyzing Web pages, online movie that user has browsed, user interest could be implicitly obtained. The text circling round them, or existing hyperlink will be utilized to indirectly mine the semantic information in video or image. It is similar that analyzing relative annotation in Web pages or header information in local cache will gain the semantic content of online music.

UDM

3.3.1 User background information

User background information contains some user private information that is possibly obtained such as sex, IP, blood type, age and so on. The information couldn't be directly utilized to describe user interest, but they have certain regular relationship each other. For example, woman is always interested in cosmetic and those whose ages are between 25 and 35 seem more interested in it.

3.3.2 Interest sub-space(ISS)

ISS is the kernel of UDM which is the projection of user interest in multi-granularities interest space. In traditional UM based on vector space, user interest is represented as multi-dimensions vector. It is convenient to compute user interest similarity with vectors angle. But it ignores user interest changing in short time and will spend much time if interest dimension is very high.

By interest community discovering, we could construct the most suitable multi-granularities interest sub-space for each user. An ICH sub-tree can be chosen to construct collective interest community sub-space. For each user interest point, we record the curve function constructed by the interest rank changing through time. By union process, we could calculate the fluctuating curve function $f_{d_i}^{k}(T, D_i)$ of the interest point for the user k in general one day. So user's ISS could be represented by a function set:

$$ISS(k) = \{f_{d_1}^{k}(T, D_1), f_{d_2}^{k}(T, D_2), \dots, f_{d_i}^{k}(T, D_i), \dots\} \quad (1)$$

Where there are n user interest points, namely the number of leaf nodes in ICH sub-tree. ISS corresponds to an irregular cube (UIC) in n+1 dimensions interest space. All cubes and their relationship will construct a huge user interest network which provides some latent knowledge.

3.3.3 Interest community information share

There is inevitable interest relationship among different users which is inclined to develop in the community manner. Those UIC corresponding to different users who have similar interest generally gather together. By clustering in different MIS, it is able to discovery different degree of interest communities.

Interest point set of each user's corresponds to an ICH

sub-tree. A unified deputy sub-tree is constructed for these users by similarity computation of tree structure, which is the base of the public interest space. Each user interest point corresponds to a function $f_{d_i}^{k}(T, D_i)$. The curve function of their interest father point based on ICH could be calculated by average curve energy function:

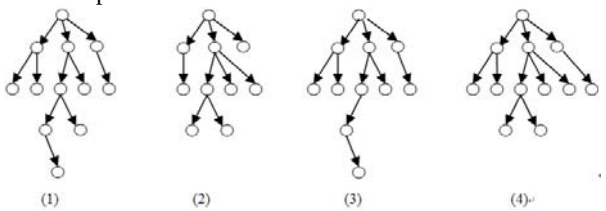
$$f_{d_f}^{k}(T, D_f) = \sqrt{\sum_{u=i}^{j} E(f_{d_u}^{k})/n^2} \quad (i < j) \qquad (2)$$

where $E(f_{d_u}^{k}) = \int f_{d_u}^{k} dt$

d_f denotes the father point of set $d_i,\ldots,d_j$. The public interest space ensures that all UIC could be clustered in the same measure system to realize interest community discovery.

The clustering algorithm chosen for discovering community is DBSCAN where the representation of UIC plays an important role in the determination of different clusters. Usually, centroid or other key point is used to replace UIC. It is possible that the high layers of different users' ICH sub-tree have enormous difference, although they have very similar interest and the space dimension corresponding to sub-tree is high. Using UIC centroid directly is not efficient in this status. A set of core points P_{cor} is used as the representatives. Let $CorC_i^x \subseteq C_i^x$ (cluster C_i^x in a set of cluster C^x at interest subspace x) be a set of the core points belonging to this cluster. The definition of P_{cor} is as follows: $P_{cor}C_i^x \subseteq CorC_i^x$ is a set of core object iff ($\forall p_k, p_l \in P_{corC_i^x}: p_k \neq p_l \Longrightarrow p_k \notin N_{\epsilon x}(p_l)$) and ($\forall c \in CorC_i^x, \exists p \in P_{corC_i^x}: c \in N_{\epsilon x}(p)$ with $N_{\epsilon x}(d_i)$ of a point d_i is defined as $\forall d \in N_{\epsilon x}(d_i): \|di - d\| \leq \epsilon_x$.

For the incipient clustering result, we mine the public portion of interest sub-tree corresponding to all UIC in class A, which will be used to construct the interest space for next time clustering, shown in Fig.2. The integrating course accords with the rule that the nodes in the same layer use union operation but those in different layers use intersection operation. The rule covers as more user interest as possible, but ignores each user's personalized interest which will be advised in the course of constructing interest space of sub-class after class A is clustered.



(1)-(3) 3 different ICH sub-tree;(4) The common interest tree

Figure 2. The conformity of interest sub-trees in class A

The clustering in the interest spaces of different granularities realizes different levels of descriptions for the same user. So the scale of interest clustering could be adjusted by selecting the size of similar interest community in terms of requirement. If the granularity is coarse, lots of users would have similar interest. On the contrary, if the granularity is fine, there would produce interest divaricating among those users so that we will obtain more interest communities. The shared interest in a community is represented by a set of information pair:

$$CIS(k) = \{(\text{Subtree-ID, C-centroid})\} \qquad (3)$$

Where Subtree-ID denotes the ID number of ICH sub-tree corresponding to the granularities space and C-centroid denotes the interest community centroid.

Computing Interest Relationship

Interest relationship includes that among different users' interest and that between user interest and retrieved results. The former is calculated by distance between different ISS centroids. However, the complexity for calculating the centroid of multi- dimensions geometric object does not satisfy for the demand of dynamic relationship among large numbers of online users. This paper uses the method of regular exampling that transforms ISS to a set having some exampled points, where each point denotes the interest status for a user at a moment. So the points set corresponding to a moment could describe user interest in short time. The sparseness or denseness of exampled points reflects the extent that user has the tendency for such interest region in long time. The user interest description is integrated into a unified mode. The centroid of exampled points set will substitute for the user interest, where the centroid coordinate is the mathematical expectation of n exampled points in every dimension x_i:

$$\overline{x_i} = \sum_{j=1}^{n} x_j / n \qquad (4)$$

The Euclidean distance between centroids is calculated to estimate the difference between users.

The retrieved results, such as Web page, image and so on, exist as objective entities which generally contain fixed information. So they are also able to be represented as points in interest space and the relationship between user interest and retrieved result is calculated by space distance.

4. EXPERIMENTS AND ANALYSIS

We construct 148 UDM by analyzing browsed Web pages and submitted interested folder of each user's based on cross-media technology. We track users' online action for two days. Because the first aspect needs a rather big data size to validate the algorithm efficiency, the data of 148 users' are distinctly insufficient. So we simulate the data of 3000 users' based on existing data and the sampling frequency is every 2 minutes. At last there are about 600 interest describing points for each user. Each point is represented by a tuple {ID, Username, Time, Space, Cor}

where Space and Cor respectively denote current interest space and point coordinate. The algorithm complexity is n*log(n).

We random select 8 users of different communities in low granularity interest space and track their clustering results in high granularity interest space which is repeated several times. The result proves the upward restriction of user interest relationship in different granularity spaces. By counting and analyzing the clustering results, we obtain some funny relationship information, shown as follows:

Specialty 1	*Sum1*	*Specialty 2*	*Sum2*	*COM(%)*
Medicine	19	Biology	15	75.6
Computer	25	Commerce	21	69.2
Law	18	History	14	42.9
Art	8	Computer	25	33.9
Biology	15	Law	18	16.7

Table 1. Comparability between different specialties

Age 1	*Sum1*	*Age 2*	*Sum2*	*COM(%)*
30-35	24	40-45	16	57.3
25-30	19	30-35	24	39.6
25-30	19	20-25	26	35.3
Under 10	2	Above 60	3	28.7
15-20	6	40-45	16	13.9

Table 2. Comparability between different ages

Type 1	*Sum1*	*Type 2*	*Sum2*	*COM(%)*
O	28	B	43	41.2
AB	29	A	48	32.1
B	43	A	48	15.3
O	28	AB	29	10.2

Table 3. Comparability between different blood types

The result shows that specialty background has more effect for forming an interest community, but blood type plays a lower effort. In table 1, users of medicine specialty have 75.6% interest similarity with those of biology specialty. In table 2, users whose ages are 30~35 have more interest similarity with those between 40 and 45. Besides, users under 10 have not lower interest gap with those above 60 than our imagination. In table 3, users of blood type O have the most interest comparability with those of blood type B, and the most interest gap with those of blood type AB. Although the size of experiment data is not big, the analyzing results approximate the factual circs to a certain extent.

The recommended information based on interest communities is also displayed in result page of FizZ(shown in Fig.3). After users install the FizZ IE plug-in and startup it, the Web pages they browsed will be tracked and analyzed to mine user interest. Besides, user could appoint local folder with different media documents to analyze user interests. If user agrees to share his own interest, these private information will be able be utilized to discovery interest communities. By user feedback, we could evaluate whether retrieval results accord with the user interest.

Fig. 3 Retrieval result page of FizZ

5. CONCLUSIONS

The key problems of realizing personalized service are how to obtain and describe user interest. In this paper we present an user interest mining method based on cross-media. Besides, we put forward a unified UDM which is suitable to describe user's short-time and long-time interest, and discovery share information in interest communities. The experiment evaluates the feasibility and efficiency of our method. Especially, the developing prototype system FizZ achieves a fairly high user satisfaction.

6. REFERENCES

[Jou0a] Guo Yan, et al. Analyzing Scale of Web Logs and Mining Users' Interests**,** Journal of Computers. Vol.28, No.9, 2005

[Jou0b] Smith.KA, Ng.A. Web Page Clustering Using A Self-organizing Map of User Navigation Patterns. Decision Support Systems, 2003 ,35 :245-256

[Jou0c] Shahabi C, et al. An Adaptive Recommendation System without Explicit Acquisition of User Relevance Feedback. Distributed and Parallel Database, 2003, 14 :173-192

[Jou0d] Smyth.B, Bradley.K, Online Recruitment Services. 2002, 45(5) :39-40

[Con0a] Rachid Arezki, et al. Information Retrieval Model Based on User Profile. AIMSA 2004, p.490-499

[Jou0e] LIN Hong Fei, YANG Yuan Sheng. The Representation and Update Mechanism For User Profile. Computer Research and Development, Vol.139, No.17, p.843-847, 2002

[Con0b]Pablo Castells1, et al. Self-tuning Personalized Information Retrieval in an Ontology-Based Framework. OTM Workshops 2005, p. 977-986, 2005.

[Jou0f] YANG Yan, LI Jian-Zhong, GAO Hong. Ontology-Based Preference Model in Digital Library**.** Journal of Software, Vol.16, No.12, pp. 2080-2088, 2002

[Con0c] Cheng ZENG, et al. Cross-media Database Retrieval System Based on TOTEM. The 7th Web Information Systems International Conference. 2006, p182-193