International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 1, March 2005 pp. 88-93

User modeling based on fuzzy category and interest for web usage mining

Si-Hun Lee, Jee-Hyong Lee

School of ICE, Dept of ECE Sungkyunkwan University

Abstract

Web usage mining is a research field for searching potentially useful and valuable information from web log file. Web log file is a simple list of pages that users refer. Therefore, it is not easy to analyze user's current interest field from web log file. This paper presents web usage mining method for finding users' current interest based on fuzzy categories. We consider not only how many times a user visits pages but also when he visits. We describe a user's current interest with a fuzzy interest degree to categories. Based on fuzzy categories and fuzzy interest degrees, we also propose. a method to cluster users according to their interests for user modeling. For user clustering, we define a category vector space. Experiments show that our method properly reflects the time factor of users' web visiting as well as the users' visit number.

Key words : Web usage mining, fuzzy category, user modeling, fuzzy interest

1. Introduction

Data mining can be defined as searching high-capacity database for useful but unknown information that cannot be drawn by simple queries[1][6][8][9]. Web mining is a searching for useful patterns in data stored in web site or web-site usage data. Usually, web mining includes web structure mining, web contents mining and web usage mining[5][7].

Web usage mining is a research field for searching potentially useful and valuable information from web log file or web usage data. One of the most interesting data to find out through web usage mining is web users' interest fields and the models of users who have similar interest fields. Web log file is usually used for web usage mining. Web log file is a simple list of pages that users refer. Since web pages usually include several topics, it is not easy to find user's interest fields by only the list of the web pages user visited.

We must also look at the content of web pages to grasp user's interest fields. Also, we must consider one of a time factors because user's interests may change as time goes on. The previously proposed mining methods did not reflect those properly[3][4].

We introduce fuzzy category to describe the contents of web pages since a web page may contain several topics and even if a page includes a topic, it may not belong to a single category. In addition, we also use a factor of time when a user visited web pages for mining users' interest fields. That is, out method mines web user interest fields using fuzzy category of web contents and a time factor of web page visits. We also propose a method for clustering users who have similar interest field and creating a model for each user group.

2. User's fuzzy interest

In this section, we propose a method to grasp user's interest fields from user's web log using fuzzy category and one of visiting time factors.

2.1 Fuzzy Category

In order to find user's interest fields from the web pages user visited, we first have to know what contents the pages contain. For describing the contents of pages, we introduce fuzzy categories. A topic in a page may belong to a single category or several categories. For example, sports shoes may belong to sport category as well as shoes category, thus so is the contents of a web page for sports shoes. We denote the degree to which a page belongs to a category. For example, a page P1 for sports shoes may belongs to sports category with a degree of 0.3 and to shoes with 0.7. We will represent those as $\mu_{\text{sports}}(P) = 0.3$ and $\mu_{\text{shoes}}(P) = 0.7$ and as a vector form (0.3, 0.7) for (sports, shoes).

Table 1. Membership degrees to categories

	C_1	C_2	C_3	C_4	C_5
P_1	0.1	0	0	0.3	0.6
P_2	0.4	0	0	0.1	0.5
P_3	0.1	0	0.3	0.1	0.5
P_4	0	0	0.4	0.6	0
P_5	0	0	1	0	0
P_6	0.4	0.6	0	0	0
P_7	0	0	0	0.2	0.8
P_8	0	0	0	0.9	0.1
P_9	0.2	0.8	0	0	0
P_{10}	0.8	0.2	0	0	0
P_{11}	0	0	0	0.1	0.9

Manuscript received Aug. 23, 2004; revised Jan. 27, 2005.
This work was supported by Sungkyunkwan university.

Since the contents of a page are fuzzily categorized, we call it fuzzy category. Before mining users' interest fields in a web site, we should have the possibility (or membership degree) of each web page to each category. A web administrator may choose categories of interest and assign membership degrees to each page according to its contents. Table 1 is an example of the fuzzy category of a web-site with 11 pages. The administrator chooses five categories from C_1 to C_5. In this web site, page P_1 contains topics which belong to C_1 with a degree of 0.1, C_4 with 0.3 and C_5 with 0.6.

2.2 Fuzzy Interest

For mining user's interest fields, we have to look at the pages he visited. If a user has an interest in a certain field, he may frequently visit the page containing it. We have to investigate what contents the pages, a user often visits, include from the investigation, we have to infer user's interest field. Instead of choosing a field as user's interest, we evaluate user's interest degrees to the fields. It a user visit web pages containing topics of a certain field, his interest degree of the field will be high, vice versa.

One of important factors in grasping user's interest field is a time factor. User's interest degree changes as time goes on. However, most existing web mining methods do not consider factors of visit time. We reflect one of time factors representing when a user visits pages for finding user's interest fields.

First, we define the category counter as follows :

$$\mathrm{Count}(C) = \sum_{t=1}^{T} \mu_{T_t}(C) \tag{1}$$

It represents how many the pages a user visits include contents of category C. T is number of transactions a user have made. $\mu_{T_t}(C)$ is defined as follows, which represents how many pages included in the t-th transaction belongs to category C.

$$\mu_{T_t}(C) = \frac{\sum_{p \in T_t} \mu_C(p)}{\text{number of pages included in } T_t} \tag{2}$$

Interest(C) is a user's degree of interest in category C. It is defined as follows :

$$\mathrm{Interest}(C) = \frac{\mathrm{Count}(C)}{\sum_{t=1}^{T} t - \sum_{t=1}^{T} \{(t * \mu_{T_t}(C)) + T\}} \tag{3}$$

It assigns higher degree to the categories included in recently visited pages.

For example, a user visited a web site of Table 1 twice. He made two transactions $T_1=\{P_1, P_7, P_{11}\}$ and $T_2=\{P_2, P_3, P_7\}$. That is, the user visited pages P_1, P_7 and P_{11} at the first visit, and P_2, P_3 and P_7 at the second. Tables 2 and 3 show the fuzzy category of the pages in each transaction and the degrees that the transactions include each category, i.e. $\mu_{C_i}(T_j)$ for i=1,2,...,5 and j=1,2.

Table 2. Transaction 1

	C_1	C_2	C_3	C_4	C_5
P_1	0.1	0.0	0.0	0.3	0.6
P_7	0.0	0.0	0.0	0.2	0.8
P_{11}	0.0	0.0	0.0	0.1	0.9
$\mu_{C_i}(T_1)$	0.03	0.0	0.0	0.20	0.77

Table 3. Transaction 2

	C_1	C_2	C_3	C_4	C_5
P_2	0.4	0.0	0.0	0.1	0.5
P_3	0.1	0.0	0.3	0.1	0.5
P_7	0.0	0.0	0.0	0.2	0.8
$\mu_{C_i}(T_2)$	0.17	0.0	0.10	0.13	0.60

Then, user's degree of interest in C_5 can be evaluated as follows :

$$\mathrm{Count}(C_5) = \sum_{t=1}^{2} \mu_{T_t}(C_5) = 0.77 + 0.60 = 1.37$$

$$\mathrm{Interest}(C_5) = \frac{\mathrm{Count}(C_5)}{\sum_{t=1}^{T} t - \sum_{t=1}^{T} \{t * \mu_{T_t}(C_5)\} + T} = \frac{1.37}{3 - \{1 * 0.77 + 2 * 0.60\} + 2} = 0.452$$

We may say that the user has interest in C_5 with a degree of 0.452.

2.3 Attribute of fuzzy interest

User's interest degree may change as time goes on, but it may have a tendency in changes. We have investigated it and identify four basic attributes of changes through time.

1. If a user does not refer pages including a field, he does not interest in that field.
2. If a user refer only pages including only a field, he has the most interest in that field.
3. The more a user visits a page, the more interest he has in the field of the page.
4. Even if a user equally visits two pages which include one topic each, he has much interest in the topic of the pages recently visited than the other.

Attribute 1 and 2 are the boundary condition, attribute 3 is the monotonicity and attribute 4 is the recentness. Our definition of the interest degree also satisfies the above attributes.

Following are the proof :

Attribute 1.

$\forall T_{n,}\ \mu_{T_n}(C)=0 \rightarrow \text{Interest}(C) = 0$

Proof 1.

It is clear by definition.

Attribute 2.

$\forall T_{n,}\ \mu_{T_n}(C)=1 \rightarrow \text{Interest}(C) = 1$

Proof 2.

It is clear by definition.

Attribute 3.

If a user makes only a transaction T including category $C_1, C_2, \ldots, C_c$, Interest(C_i) increases as the number of transactions increases for i = 1, 2, ... , c .

Proof 3.

Since a user always makes the same transaction, Count(C_c) is independent of the number of transactions.

Thus For m<n,

$$(\sum_{t=1}^{n} t - \sum_{t=1}^{n} t * \mu_{T_t}(C_i) + n) - (\sum_{t=1}^{m} t - \sum_{t=1}^{m} t * \mu_{T_t}(C_i) + m) > 0$$

Therefore, Interest(C_i) for m transactions is smaller than Interest(C_i) for n.

Attribute 4.

Let us suppose that a user makes two transactions T_1 including only category C_1 and T_2 including only C_2, and $\mu_{T_1}(C_1)=\mu_{T_2}(C_2)$

If a user makes n transaction T_1 first and n transaction T_2 next, Interest(C_1)<Interest(C_2)

Proof 4.

Since $\mu_{T_1}(C_1)=\mu_{T_2}(C_2)$, Count($C_1$)=Count($C_2$).

Thus $\sum_{t=1}^{T} t * \mu_{T_2}(C_2) > \sum_{t=1}^{T} t * \mu_{T_1}(C_1)$

Therefore, Interest(C_1)=Interest(C_2).

Through those analysis, we can know that the interest degree also satisfies the basic attributes which user's interest may have.

3. User Modeling

We have described a method to find out one user's interest degree. To provide better services to users in a web site it needs to understand user groups with the similar interest as well as a single user. This section proposes a method to create user model based on user's fuzzy interest degree.

3.1 Category Vector Space

This section proposes a method that creates user model for a users who have similar interest fields. We first need to group users having similar interest degree.

In order to do grouping users who have similar interest field we supposes that each category conceptually is independent from each other. That is, the interest degree of a category cannot be inferred from other categories. For example, if web pages are fuzzily categories into 3 categories : C_1, C_2 and C_3. we cannot infer interest degree of C_3 from interest degree of C_1 or C_2. This is analogue to a vector space whose axes C_1, C_2 and C_3 vertically meets. We call this space category vector space[10]. All users' fuzzy interest degree can be mapped on a point in category space. If two users' interests are similar, those will be located near from each other in category vector space. Thus, distance between two point can represent the similarity of two users' interests.

For example, there are category C_1 and C_2, and User 1's fuzzy interest degree is (0.5, 0.7) and User 2's fuzzy interest degree is (0.7, 0.4). Then, we make a space where C_1 and C_2 are axes. The interest degrees of each user can mapped on to a point each in the space as shown in Fig .1.

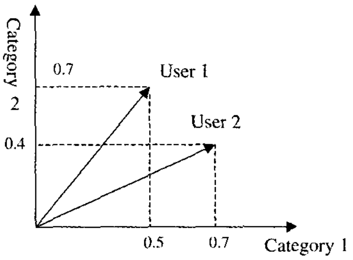


Fig. 1 Two categories space

3.2 Clustering and user model

All users' interests can be mapped into category vector space. After we map each user's fuzzy interest degree to category space, we need to group the point which locates near from each other, which means grouping users with similar interest. In order to group the point, we use clustering algorithm.

Method that we use is k-means algorithm. Algorithm k-means works as follows[2].

```
Step 0: Select k objects as initial centroids.
Step 1 (Assignment):
  - For each object compute distances to k centroids.
  - Assign each object to the cluster to which it is the
    closest.
Step 2 (New Centroids): Compute a new centroid for each
    cluster.
Step 3 (Convergence):
  - Stop if the change in the centroids is less than the
    selected criterion.
  - Otherwise, repeat Step 1.
```

By applying the algorithm, we can get several user groups. Next, we creat user models of the user groups as follows :

$$\text{model}(G) = \frac{1}{|G|}(\sum_{u \in G} \text{Interest}_u(C_1), \ldots, \sum_{u \in G} \text{Interest}_u(C_n))$$

G : User group

|G| : The number of users in G

$Interest_u(C_i)$: User u's interest degree to category C_i

That is, the model of group G is defined by the average of all users' interest in the group. Such user models can be used for estimating a new user's preference or selecting new service for a new user.

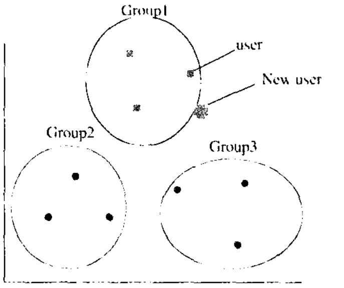


Fig. 2 Example of user clustering

For example, let us suppose that three user groups and a new user's interest is mapped onto the big point as shown Fig. 2. The first step is to identify which user group has similar interest to the new user. We evaluate the distances between the new user's point and the points of user groups in category vector space. Since the new user is near to Group 1, it is expected that he may see the interest subjects or pages that users in Group 1 mainly visit. Thus, we can recommend those to the new user.

4. Experiment

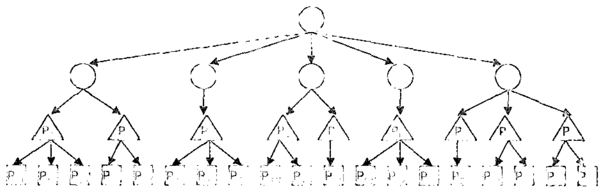


Fig. 3 Sample web pages

Fig. 3 shows the structure of the sample web pages to use in experiments. There is one main page which has five links to the second level pages. Each page in the second level has 1 to 3 links to the third level pages(P_1,P_2, ... , P_9). The third level pages also have links to the fourth level pages. Since we have interest in user's visiting of the third and fourth level pages, we will focus on those pages. We define fuzzy category membership degree of each web page as given in Table 4.

We perform two experiments to confirm that fuzzy interest degree properly reflects the time factor of visits.

Let us suppose that there are three users who made 9 transactions each as shown in Table 5. User 1 is a model of a user whose interest changes. He mainly visited P_3, P_{31}, P_{32} and P_{33} in category C_3 at first, but later P_6, P_{61}, P_{62} and P_{63} in category C_6. His interest has changed but the number of visits to pages in category C_3 is still more than that of pages C_5. If we consider only the visit count, his interest is mined as Fig. 4. Fig. 4 shows the degree of interest in each category. The topic which has visited much, C_3, will be considered as user's main interest. However our method says that user's main interest is C_6, since our method consider the time factor of visits as well as the visit count, category C_6, which has the visit counter slightly smaller than C_3 but one visited more recently than C_3, is mined as the main interest.

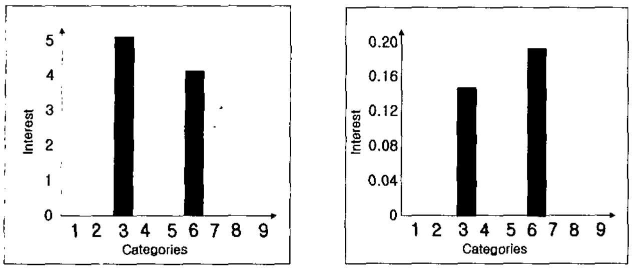
User 2 and 3 have visited the same pages through 9 transactions, but the visit order is reverse. If we consider only how many times they visit a page, their interest fields are the same. However, since User 2 recently visits P_1, P_{11}, P_{12} and P_{13} and User 3 visits P_9, P_{91} and P_{92}, we may say that User 2 has much interest in category C_1 than other categories and User 3 in C_4. Fig. 6 and Fig. 7 show User 2 and 3's degrees of interest in the categories each. Since the visit order is reverse though the visit counts are the same, so are the interest degree.

Table 4. Fuzzy categories of the sample pages

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
P_1	0.7	0.3	0	0	0	0	0	0	0
P_2	0.3	0.7	0	0	0	0	0	0	0
P_3	0	0	1	0	0	0	0	0	0
P_4	0	0	0	0.8	0.2	0	0	0	0
P_5	0	0	0	0.2	0.8	0	0	0	0
P_6	0	0	0	0	0	1	0	0	0
P_7	0	0	0	0	0	0	0	0	0
P_8	0	0	0	0	0	0	0	0	0
P_9	0	0	0	0	0	0	0.7	0.2	0.1
P_{10}	0	0	0	0	0	0	0.1	0.8	0.1
P_{11}	0	0	0	0	0	0	0.1	0.1	0.8

Table 5. Three users' transactions

	User 1	User 2	User 3
T_1	$\{P_3,P_{31},C_{32}\}$	$\{P_9,P_{91},P_{92}\}$	$\{P_1,P_{11},P_{12},P_{13}\}$
T_2	$\{P_3,P_{32},P_{31}\}$	$\{P_8,P_{81},P_{82}\}$	$\{P_2,P_{21},P_{22}\}$
T_3	$\{P_3,P_{31},P_{33}\}$	$\{P_7,P_{71}\}$	$\{P_3,P_{31},P_{32},P_{33}\}$
T_4	$\{P_3,P_{32},P_{33}\}$	$\{P_6,P_{61},P_{62},P_{63}\}$	$\{P_4,P_{41},P_{42}\}$
T_5	$\{P_3,P_{33},P_{32}\}$	$\{P_5,P_{51}\}$	$\{P_5,P_{51}\}$
T_6	$\{P_6,P_{61},P_{62}\}$	$\{P_4,P_{41},P_{42}\}$	$\{P_6,P_{61},P_{62},P_{63}\}$
T_7	$\{P_6,P_{61},P_{63}\}$	$\{P_3,P_{31},P_{32},P_{33}\}$	$\{P_7,P_{71}\}$
T_8	$\{P_6,P_{62},P_{63}\}$	$\{P_2,P_{21},P_{22}\}$	$\{P_8,P_{81},P_{82}\}$
T_9	$\{P_6,P_{63},P_{62}\}$	$\{P_1,P_{11},P_{12},P_{13}\}$	$\{P_9,P_{91},P_{92}\}$



(a) Fig. 4 (b) Fig. 5

Fig. 4 : Existing method of User 1's degree of interest.
Fig. 5 : Proposed method

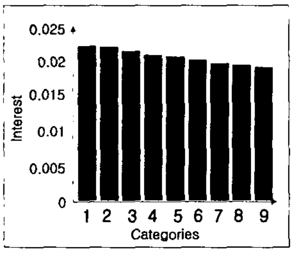


Fig.6 User 2

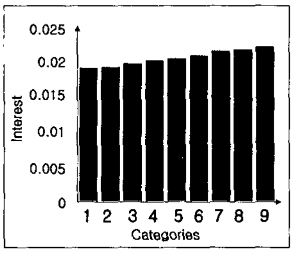


Fig.7 User 3

Also, we perform on experiment to verify whether our user modeling also properly reflects the time factor of page visits. Table 6 shows 10 users' interest categories.

Table 6. Ten users' interest categories

User	Interest Category
1	Society
2	Society, Culture
3	Society, Performance
4	Society, Sports
5	Politics → Society
6	Politics
7	Politics, Economy
8	Politics, International
9	Politics, Digital
10	Society → Politics

We consider the case where some users has fixed interest field but we have changed their interest. User 2 is interested in society and culture but User 5 was interested in politics but is in society present, which is represented as politics → society in Table 6. User 10's interest is also change from society to politics. For the experiment, we virtually generate transactions for each user so that their interest are mined as Table 6. If we divide the users into two groups by our method, we have two groups shown as Table 7.

Table 7. User clustering results

Group	User
1	1, 2, 3, 4, 5
2	6, 7, 8, 9, 10

Group 1 is a group of users who are interested in society, and group 2 is for users interested in politics. Therefore, although User 5 was interested in politics, User 5 must belong to group 1 because user 5 is currently interested in society. Also, it is reasonable that User 10 belongs to group 2, because he was interest in society, but presently in politics. Therefore, we may say that proposed user modeling method we proposed properly reflects user's interest changes through time.

5. Conclusion

This research is on web user modeling using use fuzzy category and fuzzy interest. We defined fuzzy category, and presented a method to find out fuzzy interest degrees reflecting the time factor of transactions. We also identified the attributes of interest changes through time, and proved that our fuzzy interest satisfied those. Also we investigated how fuzzy interest reflected the time factor through the experiment and compared with the existing method. We also examined whether our modeling method properly reflected interest change.

The experiments show that our method properly mines user interest changing through time, and thus generates a reasonable user models. Now, we continue our research by using real web usage data to find out valuable information from a real web site.

References

[1] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. of VLDB Conference*, pp.487-499, 1994.

[2] C.C. Aggarwal, S.C. Gates, P.S. Yu, "On the merites of building categorization systems by supervised clustering," *In Proc. of the ACM SIGKDD conference*, pp.352-356, 1999.

[3] H. Yi, Y.C. Chen, L.P. Chen, "Enabling Personalized Recommendation on the Web Based on User Interests and Behaviors," *Proc. of 11th International Workshop*, IEEE, pp.1066-1077, 2001.

[4] A. Gyenesei, "A Fuzzy Approach for Mining Quantitative Association Rules," *TUCS Technical Reports*, no. 336, 2000.

[5] J.S. Jang, S.H. Jun, K.W. Oh, "Fuzzy Web Usage Mining for User Modeling," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 2, no. 3, pp.204-209, 2002.

[6] R. Cooley, B. Mobasher, J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns,“ *Journal of Knowledge and Information System*, vol. 1, no. 1, pp.8-19, 1999.

[7] R. Cooley, B. Mobasher, J. Srivastava, “Web mining : Information and Pattern Discovery on the World Wide Web,“ *Proc. of the 9th IEEE International Conf. on*

Tools with Artificial Intelligence, pp.61-62, 1997.

[8] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation," *Communications of the ACM*, 43, pp.127-134, 2000.

[9] B. Mohasher, R. Cooley, J. Srivastava, "Automatic personalization based on Web usage mining.," *Communications of the ACM*, vol. 43, pp.142-152, 2000.

[10] O. Zamir O. Etzioni, "Web document clustering : A feasibility demonstration," *Proc. of the ACM SIGIR Conference*, pp.46-53, 1998.

Si-Hun Lee
received the B.S degree in Department of Electronic and Communication Engineering from Korea Maritime University, Busan, Korea, in 2003, and M.S degree in Department of Computer Engineering from Sungkyunkwan University, Suwon, Korea, in 2005. His research interests include Web-mining, User modeling and fuzzy.

Phone : +82-31-290-7987
Fax : +82-31-290-7211
E-mail : c1soju@skku.edu

Jee-Hyong Lee
received the B.S, M.S and Ph.D in Department of Computer Science from Korea Advanced Institute of Science and Technology(KAIST) in 1993, 1995 and 1999, respectively. He joined Sungkyunkwan University, Korea in 2002 and is currently an associate professor at Computer Engineering. His research interests include data mining, fuzzy theory and application, neural network, and ubiquitous computing.

Phone : +82-31-290-7154
Fax : +82-31-290-7211
E-mail : jhlee@ece.skku.ac.kr