A novel user-interest model based on mixed measure

Junkai Yi¹, Yacong Zhang², Mingyong Yin³,* and Xianghui Zhao⁴

¹ College of Information Science and Technology, Beijing University of Chemical Technology, No. 15, Beisanhuan East Road, Beijing, China.
Email: yijk@mail.buct.edu.cn
² College of Information Science and Technology, Beijing University of Chemical Technology, No. 15, Beisanhuan East Road, Beijing, China.
Email: zhangyacong84@163.com
³ Institute of Computer Application, China Academy of Engineering Physics, No. 64, Mianshan Road, Mianyang, China. Email: yinmingyong@126.com
⁴ China Information Technology Security Evaluation Center, No. 8, Shangdi West Road, Beijing, China. Email: zxhitsec@sina.com

Abstract. In this paper, we present a user-interest model based on mixed measure for finding users’ interests, and in this model we can find the users’ interests in webpages to improve the accuracy of users’ interests and provide better personalized service for users. Information gain and mutual information are applied in this model for dimensional reduction, then find users’ interests from the content of webpages by classifying-clustering. The results of experiments show that the model we propose has better performance, and the accuracy of the model is satisfactory.

1. Introduction

With the popularity of network, the Internet has become an indispensable information source for human. Meanwhile, our need to the information also increases gradually, but it is difficult to find the information we need exactly and quickly. Personalized service can meet users’ demand by providing different services to the different users [1]. We focus on problems such as feature selection and user-interest exploration, and propose a user-interest model based on mixed measure in this paper. We divide the content of webpage into several words in the model, and apply information gain and mutual information for dimensional reduction. Then we use classification method for coarse grained partitioning to obtain several large groups, and apply clustering algorithm to each group for users’ interest. The user-interest matrix is denoted by IR (interesting) based on calculating the degree of interest by the analysis of users’ behaviours.

At present, many researchers study the user-interest modeling and achieve abundant research results. He L et al. proposed LDA-based theme mining model, and this model divided users’ interest into initial interest and forwarded interest [2]. Jayarathna S et al. used four different user-interest theme model to generate fine-grained user-interest model, and made the degree of attention of all passages and fragments for visualization [3], which was in accordance with users’ interests. Liu Z, Chen X et al. focused on the noisy of message and the complexity of words in users’ microblogs, and proposed a keyword-mining approach to dig users’ interests based on the combination of conversion and frequency [4]. Kumar S et al. extracted the users’ interest from the history of webpages, recognized the importance of users’ interest through computing the number of keyword transmission, and ranked the webpage according to the importance of interests [5]. Liu D et al. proposed an approach to analysis users’ interests using fingerprint method. The fingerprint was binary strings obtained from the high-dimension feature vector.
based on Simhash algorithm, and used fingerprint sequence and distance between fingerprints to detect user-interest features [6]. Giri R et al. used an unsupervised theme model, deduced users’ interests according to the files users browsed. [7]. LI C et al. proposed an election approach to extract high-dimensional keywords, and used the high-dimensional keywords to reflect users’ interest. [8]. Li S et al. proposed a hierarchical clustering of Chinese words based on graph. The author created users’ interest into a structure of hierarchical trees, and used maximum-matching mapping method to match the users’ interests in the algorithm [9]. Ma Y et al. proposed a user-interest modeling approach of personal information integration and semantic reasoning based on multiple sources [10].

2. The User-interest Exploration
The main process of creating a user-interest model is to analyze the webpage content users browsed. We obtain the users’ interests and compute the degree of users’ interest. Moreover, we combine the users’ interests and the degree of users’ interest to denote users’ interest. The procedure is shown as follows.

![Figure 1. The procedure of user-interest model.](image)

2.1. The Preprocessing of Webpage Texts
A better way of describing the webpage is vector space model (VSM). VSM uses feature words \((t_1, t_2, ..., t_n)\) to denote text \(d\), and each feature words \(t_i\) has its weight \(w_i\) according to the importance in the text. So all texts could be denoted by the feature vector \((w_1, w_2, ..., w_n)\). Before describing the webpage texts, we need to partition the words. We remove all tags, which are unrelated to the texts. We also extract webpage title and content as the object of partitioning words in this model.

To avoid the problem of curse of dimensionality, we combine information gain and mutual information to dimensionality reduction for the initial feature word sets. Information gain is used to describe the influence of feature to the whole classifying system. The formula is shown as follows.

\[
IG(t) = \sum_{j=1}^{k} P(C_j) \log \frac{P(C_j)}{P(t)} + P(t) \sum_{j=1}^{k} P(C_j|t) \log \frac{P(C_j|t)}{P(C_j)} + P(t) \sum_{j=1}^{k} P(C_j) \log \frac{P(C_j)}{P(t)}
\]

Mutual information is used to evaluate the influence of feature to specific category. Mutual information is shown as follows.

\[
MI(t,c_i) = \log \frac{p(t|c_i)}{p(t)} \times \frac{p(c_i)}{p(t|c_i)} = \log \frac{p(t|c_i)}{p(t)}
\]

We use the weighted sum for the information gain with the mutual information, and sort the weighted result. The weighted sum need to save \(k\) features to obtain the final feature word sets. We assume that \(k\) groups are \(C(j=1, ..., k)\), and \(t\) is the feature words for one group. \(P(C_j|t)\) denotes the probability of feature words \(t\) belonging to the group \(C_j\) in training set. Let \(p(t,c_i)\) denote the probability of feature words \(t\) belonging to group \(c_i\) in the training set. \(p(t)\) denotes the feature \(t\) probability of text in training set. \(p(c_i)\) is the probability of text belonging to group \(c_i\). If feature \(t\) in group \(c_i\) has higher probability, feature \(t\) has lower probability in other groups. It means that feature \(t\) and group \(c_i\) has large correlation and obtain higher value of mutual information.

In the mixed measure approach of this paper, we use maximal value of mutual information and the value of information gain to the weighted sum for one feature words \(t\). The formula is shown as follows, and the formula is the degree of feature importance to whole data set or one group.

\[
\text{Degree}(t) = \alpha IG(t) + \beta MI(t)
\]
where \( \alpha \) and \( \beta \) is 0.5. The words will be saved if the degree of value is 80% above in all training sets. We reduce dimension of words largely in this approach, and increase the efficiency of clustering. After obtaining feature word set, we use formula \( tf-idf \) to calculate the weight of feature words in this approach. In addition, the text can be described by vector space model with feature words and weight. Formula \( tf-idf \) is shown in formula (4).

\[
\text{w}(t,d) = \frac{(1+\log \text{tf}(t,d)) \times \log (N/n_t)}{\sqrt{\sum_k [(1+\log \text{tf}(t,d)) \times \log (N/n_t)]^2}} \tag{4}
\]

\( \text{tf}(t,d) \) is the number of feature \( t \) in text \( d \); \( N \) is the number of texts, which includes feature \( t \). We use vector space model \( d_i = (w_{i1}, w_{i2}, ..., w_{in}) \) to denote each text \( d_i \) from all text set \( D = \{d_1, d_2, ..., d_m\} \).

### 2.2. The Exploration of User-Interest Cluster

We obtain preliminary user-interest theme groups through sampling observing the history of webpages in this paper. User-interest theme groups mainly includes seven theme classes such as sport, shopping, video, game, military, technology, and education. We grab these theme webpages from some famous websites, and label these classes. Moreover, we divide the theme webpages into training set and testing set, and the ratio of two sets is 7:3. We use these webpages to train the classifying model through decision tree C4.5 algorithm. After classifying the webpages, we can obtain the user-interest distribution roughly.

We cluster each class next in this model. In the process of clustering, the common similarity function is Euclidean distance, Dice coefficient, Jaccard coefficient, cosine distance and so on. We use cosine distance to calculate the similarity of samples in this paper. The formula is shown as follows.

\[
\text{Sim}(D,Q) = \frac{D \cdot Q}{||D|| \times ||Q||} = \frac{\sum_i (d_i \times q_i)}{\sum_i d_i^2 \times \sum_i q_i^2} \tag{5}
\]

\( D \) and \( Q \) is the feature vector of different texts in the same theme, \( D = (d_1, d_2, ..., d_m) \), \( Q = (q_1, q_2, ..., q_m) \), and \( \text{Sim}(D,Q) \) means text \( D \) and \( Q \) has high similarity if the value of \( \text{Sim}(D,Q) \) is large. We apply K-means algorithm to process fine-grained clustering in this paper. K-means clustering algorithm has high efficiency, and deals with big text sets effectively. We determine \( K \) through the multiple experiments in this paper. In addition, we can obtain the user-interest clusters after finishing the clustering, and next step is to express users’ interests directly.

### 3. The User-Interest Expression

The most direct way to describe the extent of users’ interests is the time of users’ staying and the content in the webpages. Generally, small amount of content and long-time of browsing has large value of interests. So we calculate the degree of users’ degree through analyzing staying time and the content amount in this paper. Furthermore, we obtain the time of users staying in the webpages and the size of webpages from browsing information in this model, and calculate the degree of user’ interests for each page.

\( T \) is the time of users staying in the webpages. Length is the size of webpages. The formula of calculating users’ interests to the webpages is shown as follows.

\[
\text{P}_d = \frac{T}{\text{Length}} \tag{6}
\]

If one user has \( n \) webpages in interest class \( I_i \), and the degree of users’ interests of each webpage is \( p_{ij} \). The formula of the degree of interests in this class is shown as follows.

\[
\text{IR}_{ij} = \frac{\sum_{j=1}^{n} p_{ij}}{n} \tag{7}
\]

We need to normalize the result of formula (7), and obtain the degree of users’ interests of one user-interest class. In addition, we construct matrix \( \text{IR} \) to describe the users’ interests and the degree of users’ interests in this model. \( \text{IR} \) is shown as follows.
We assume that we can obtain $k$ user-interest webpage clusters after classifying and clustering in this model. $IR_i$ is the degree of users’ interests in each interest class. Moreover, we choose the feature words of $m$ interest classes to denote the keywords of this class in this model. The interest matrix is denoted by $IR$.

Let $t_{ij}$ denote $j$-th keywords of $i$-th interest subclass. The Figure 2 shows users’ interests to the degree of interest in hierarchical structure according to the theme division of users’ interests.

\[
IR = \begin{bmatrix}
IR_1 & t_{1,1} & \cdots & t_{1,m} \\
IR_2 & t_{2,1} & \cdots & t_{2,m} \\
\vdots & \vdots & \ddots & \vdots \\
IR_k & t_{k,1} & \cdots & t_{k,m}
\end{bmatrix}
\]

(8)

Figure 2. User-interest structure.

4. Experiment Result and Analysis

4.1. Data Set and Evaluation Standard

The experiment includes 13491 webpages. For evaluating the effectiveness of dimensionality reduction, this experiment divides the webpages into training set and test set in the ratio of 7 to 3. In addition, we use these webpages to train classifying model, and classify the webpages. After classifying the data, we finish the process of the user-interest exploration and the user-interest expression.

In the process of clustering, we apply $F$-measure to evaluate the result of clustering in this paper. $F$-measure is defined by accuracy rate and recall rate. For a given class, $TP$ denotes the number of samples, which are classified correctly, $FP$ denotes the number of samples, which are classified incorrectly, and $FN$ denotes the number of the samples, which belong to this class but are classified to the other class. The accuracy rate $P$ and recall rate $R$ are defined as follows.

\[
\text{Accuracy rate: } P = \frac{TP}{TP+FN} 
\]

(9)

\[
\text{Recall rate: } R = \frac{TP}{TP+FP} 
\]

(10)

\[
\text{F-measure: } F = \frac{2PR}{P+R} 
\]

(11)
4.2. Experiment Result

We preprocess the user browsing webpages in this paper, and obtain text feature. Furthermore, we use C4.5 algorithm to classify the texts according to the theme. The accuracy rate and recall rate of classifying is 82% and 85% respectively, and we use the model, which has been trained by theme class webpages, to classify the user browsing webpages. We need cluster the webpages in each class, and we obtain the user-interest cluster in this theme. Table 1 is the result of clustering, which MAC address is 00-22-15-7F-4A-D0, clustering in game theme.

| name         | title                      | class     | keywords                           |
|--------------|----------------------------|-----------|------------------------------------|
| 10700.htm    | Secretly makeup            | cluster0  | Dress up, Makeup,                  |
| 10953.htm    | Bobbi doll's student outfit| cluster0  | Fashion, Operation, management     |
| 11078.htm    | Cosmetics shop             | cluster0  |                                    |
| 10736_3.htm  | Ultimate Parkour           | cluster1  | Racing, driving, sports, off-road, riders |
| 11675.htm    | Altman motor racing        | cluster1  |                                    |
| 10896.htm    | security police            | cluster2  | Adventure, challenge, fight, gun battle, shooting |
| 10904.htm    | Siege cannon               | cluster2  |                                    |
| 10519.htm    | Escaping Chemistry Laboratory | cluster3 | Leisure, puzzle, picture matching, the chamber of secrets, maze |
| 10541.htm    | Pet picture matching       | cluster3  |                                    |
| 30110.htm    | Mermaid fault              | cluster3  |                                    |

Table 1 shows the name, title, cluster number of webpages users browsing. We select five feature words as the keywords of each class in this paper. User game theme class has four interest clusters in Table 1, and each interest class has several keywords. For evaluating the result, we calculate F-measure of clustering and classifying before clustering. The result is shown in Table 2.

| Clustering | Classifying-clustering |
|------------|------------------------|
| class      | N | TP | FP | FN | P  | R   | F   |
| 1          | 331| 201| 142| 130| 0.61| 0.59| 0.60|
| 2          | 452| 324| 134| 128| 0.72| 0.71| 0.71|
| 3          | 401| 296| 98 | 105| 0.74| 0.75| 0.74|
| 4          | 437| 301| 143| 136| 0.69| 0.68| 0.68|
| 5          | 302| 198| 137| 104| 0.66| 0.59| 0.62|
| 6          | 420| 294| 104| 126| 0.70| 0.74| 0.72|
| 7          | 436| 325| 132| 111| 0.75| 0.71| 0.73|
| 8          | 320| 213| 122| 107| 0.67| 0.64| 0.65|
| 9          | 468| 365| 104| 103| 0.78| 0.78| 0.78|
| 10         | 231| 182| 51 | 49 | 0.79| 0.78| 0.78|

The approach of classifying before clustering has higher accuracy rate and recall rate than clustering in Table 2. F-measure in classifying-clustering is almost 80% above. So we choose classifying-clustering to increase the accuracy rate. Figure 3 shows the result of classifying-clustering.
Figure 3. The comparison of $F$-measure.

Figure 4. The change of users’ interest.

Table 5 shows the interest distribution of the user $a$ and $b$.

| user  | Interest class                                         |
|-------|--------------------------------------------------------|
| User $a$ | Shopping (makeup, facial mask, eye cream, 0.37), (floral skirt, shirt, 0.25), (computer, cell phone, charger 0.2), {video (romance, love, 0.4), (science fiction, exploration, space, 0.25)} |
| User $b$ | {sports (rockets, basketball, home, 0.45), (swimming, diving, world championships 0.3)}, {video (detective, suspense, darkroom, 0.32), (terror, horror, crime, 0.28)} |

From Table 3, we can know that user $a$ like shopping and video. In these two interests, user $a$ likes makeup best. In one interest video, the type of user $a$ and $b$ has obvious differences. User $a$ likes romance, and user $b$ likes suspense videos. In addition, the experiment studies the change of users’ interests over time. Figure 4 shows the change of user $a$ and user $b$ over time. We can know that interest 1 has little changes, but the change of interest 2 is large in a certain period. So we can recognize interest 1 as stable interest and long-term interest, and interest 2 is the short-term interest. The division of long-term interest and short-term interest can offer better personalized service to the users.
5. Conclusion
The accuracy of user-interest model decides the quality of personalized service directly. We propose an information gain and mixed measure of mutual information approach for dimensionality reduction of webpages in this paper. In the process of user-interest exploration, we need classify the webpages into several classes. This approach narrows the range of clustering, and increases the accuracy of users’ interests. Moreover, we calculate the degree of interest using user-staying time and the webpage content, and describe users’ interests with the content of users’ interests in this paper.

6. Acknowledgments
This work is supported by China Academy of Engineering Physics Project 2015A0403002, National Industrial Technology Development Program JCKY2016212C005, and National Natural Science Foundation of China Project U1636208 and U1536116.

7. References
[1] Organero M M, Kloos C D, Merino P M. Personalized service-oriented e-learning environments [J]. Internet Computing. IEEE, 2010, 14(2): 62-67.
[2] He L, Jia Y, Han W, et al. mining user interest in microblogs with a user-topic model [J]. Communications, China, 2014, 11(8): 131-144.
[3] Jayarathna S, Patra A, Shipman F. Mining user interest from search tasks and annotations[C]//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013: 1849-1852.
[4] Liu Z, Chen X, Sun M. mining the interests of Chinese microbloggers via keyword extraction [J]. Frontiers of Computer Science, 2012, 6(1): 76-87.
[5] Kumar S. Mining User Interests from Web History[C]//Web Intelligence/IAT Workshops. 2013: 282-283.
[6] Liu D, Wu Q, Han W. Mining micro-blogging users’ interest features via fingerprint generation[C]//Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Atlantis Press, 2013.
[7] Giri R, Choi H, Hoo K S, et al. User Behavior Modeling in a Cellular Network Using Latent Dirichlet Allocation[M]//Intelligent Data Engineering and Automated Learning–IDEAL 2014. Springer International Publishing, 2014: 36-44.
[8] LI C, WANG J, GUO Z, et al. User Interest Modeling Based on Terminal User Behavior Analysis [J]. Journal of Computational Information Systems, 2015, 11(1): 349-356.
[9] Li S, Wu G, Hu X. Hierarchical user interest modeling for Chinese web pages[C]//Proceedings of the Third International Conference on Internet Multimedia Computing and Service. ACM, 2011: 164-169.
[10] Ma Y, Zeng Y, Ren X, et al. User interests modeling based on multi-source personal information fusion and semantic reasoning [M]//Active Media Technology. Springer Berlin Heidelberg, 2011: 195-205.