Discovery of a User Interests on the Internet

†Fang Li, Yihong Li, Yanchen Wu, Kai Zhou, Feng Li, Xinguang Wang
Dept. of Computer Science & Engineering
Shanghai Jiao Tong University,
No.800 Dong Chuan Rd. Shanghai 200240, P.R. China
†fli@sjtu.edu.cn

Abstract

This paper proposes a system for finding a user's interests on the Internet. It is based on his browsing behaviors and the contents of his visited pages. The system has two features. One is building user's browsing interests implicitly, multiple keyword vectors, one per interest. The other is that it can generate interests by selecting different time periods. Dynamical generation can adapt to the change of user interests. Experiments show that most of generated interests are matched to user's real interests. The system finds their interests automatically and dynamically.

1 Introduction

User interests (or user profiles) can be collected by two ways: explicit and implicit collection [5]. Explicit collection is predefined or feedback by user's ratings through an interface. The users tell the system what their interests are and what they think about the information that they have received. Many users are not willing to tell the system what their true intentions are, they do not want to spend time on filling forms or rating items. A less intrusive method (Implicit collection) is to use an automatic way to find the interests of a user, instead of obtaining it directly from the user. There are roughly two kinds of automatic way to capture a user's interest implicitly: behavior-based and history-based. The behavior-based research [1] proves that the time spent on a page, the amount of scrolling on a page and the combination of them has a strong positive relationship with user interests. Browsing histories capture the relationship between user's interests and his click history in which sufficient contextual information is already hidden in the web log.

We proposed a method to find a user's interests with the combination of browsing behaviors and contents. User interests can be automatically generated by applying clustering methods on visited web pages, while the degree of his interest can be analyzed based on his browsing behaviors.

The rest of this paper is organized as follows. Section 2 describes our methods. Section 3 gives a running example. Section 4 presents the experiments and analysis.

2 Finding User Interests

Page contents are important for finding user interests. Given a set of visited pages, clustering algorithm is applied to divide the pages into several clusters. Based on the clustering results, some keywords are extracted to represent user interests in each cluster.

Given the corresponding feature vectors $X = \{x_1, ..., x_n\}$ of n visited pages $P = \{P_1, P_2, ..., P_n\}$, where $x_i = (x_{i1}, ..., x_{id})^T \in R^d$ is the feature vector of the i^{th} page, x_{ij} is the value of the j^{th} feature in the i^{th} page. The features can be words or phrases after the POS tag[1] is applied on the contents. Our clustering algorithm (shown in Algorithm 1) first selects seeds based on Kaufman approach (step 1 to 10) [4], then it uses the selected m seeds as the initial centroids and use the Spherical K-Means algorithm (SK-means) [2] to cluster pages into m clusters (step 11 to 20). The Spherical K-Means has the main advantage of requiring a linear number of comparisons while still guaranteeing good quality cluster.

Based on the clustering results, a user's interest is represented as a set of keywords which are the top 3 features of the centroids vector of each cluster. The degree of a user's (I_G) interest is defined as the sum of the corresponding interested degrees of pages in a cluster.

$$\mathrm{I}_{G_i} = \sum_{P_j \in G_i} \mathrm{I}_{P_j} \tag{1}$$

where G_i is a cluster, which represents a user's interest, P_j is a web page belonging to the cluster. I_{Pj} is the

[1] From http://www.hyland.com

interested degree of the page calculated based on the user-interest model we proposed in [3]. We use Gaussian Process Regression model to capture the relationship between user interests and browsing behaviors.

3 A Running Example

We have realized the system based on the proposed methods. The system is called "family safe"[2] because it can help parents to find the browsing interests of their child automatically and implicitly. By choosing different time periods, the interests of the period can be generated dynamically. A user was asked to surf the web by using the IE 7.0 that had embedded with our plug-in. We obtained his 2-month browsing log including browsing behaviors and contents. Then we used the system to find his interests automatically. Some of the results are given in the following:

- Figure 1 shows the result of page clustering. Three keywords of each cluster represent each interest on the left. All generated keywords with translations are compared with his real interests (Table 1). The detailed information about the first interest is shown on the right side, which consists of the keywords, the number of the related pages, its summary and the time spent on this interest. The largest interest is shopping. The number of pages in the cluster is 139 pages. The system also extracts some sentences from the viewed pages to provide an overview of the interest. These sentences are shown on the right side of the window.

Figure 1. User Interests Generated

- The user was asked to list his interests during the period. Table 1 shows the comparison between the user-predefined and the system-found interests. All the keywords can be matched to his interests.

- Figure 2 illustrates the distribution of the user's interests. Each of the interests is represented by 3 keywords, the percentage of the interest, the degree of the interests and the view time. For example, the great interest is shopping guidance. Three keywords are: shopping guidance, discount and Baishen (name of the

[2]The project was funded by the Intel China Lt.Co. and the UDS-SJTU joint research lab for language technologies.

Algorithm 1 Page Clustering Algorithm based on KA Initialization and Spherical K-Means

Input: 1. Feature vectors $X = \{x_1, ..., x_n\}$ of n pages visited.
2. The predefined number of page clusters m.

Output: The set of page clusters.

$Seeds \leftarrow \{x_{center}\}$ /* x_{center} is the most centrically located page instance in X, regarded as the seeds initially. */
repeat
  for each $page x_i \notin Seeds$ **do**
    for each $page x_j \notin Seeds$ **do**
      $C_{ji} = \max(D_j - d_{ji}, 0)$, in which $d_{ji} = \|x_i - x_j\|$ and $D_j = \min_{s \in Seeds} d_{sj}$.
    end for
    Calculate the gain of selecting x_i as a seed by $\sum_j C_{ji}$.
  end for
  $Seeds \leftarrow Seeds \cup \{x_{seed}\}$ where $seed = \arg\max_i \sum_j C_{ji}$.
until $|Seeds| = m$
$Centroids^{(t)} \leftarrow Seeds$ /* $Centroids = \{c_1^{(t)}, ..., c_m^{(t)}\}$, $c_j^{(t)}$ denotes the centroid vector of the j^{th} page cluster, t is the iterative times and $t = 0$ initially. */
repeat
  for $j \leftarrow 1$ *to* K **do**
    $C_j^{(t+1)} \leftarrow \emptyset$ /* $C_j^{(t+1)}$ denotes a cluster with the centroid $c_j^{(t)}$. */
  end for
  for each $page \in \mathbf{X}$ **do**
    $C_j^{(t+1)} \leftarrow C_j^{(t+1)} \cup \{x_i\}$, where $j = \arg\max_l x_i^T \cdot c_j^{(t)}$
  end for
  Recalculate $c_j^{(t+1)} = \frac{u}{\|u\|}$, where $u = \sum_{x_i \in G_j^{(t+1)}} x_i$
until $\left\| c_j^{(t+1)} - c_j^{(t)} \right\| < \varepsilon$
Output the set of page clusters $\mathbf{C} = \{C_1, ..., C_m\}$
return $\mathbf{C}$;

shopping mall). The interested degree is 79.41, the percentage is 23.57%. The time spend on viewing pages of shopping is 4744 seconds.

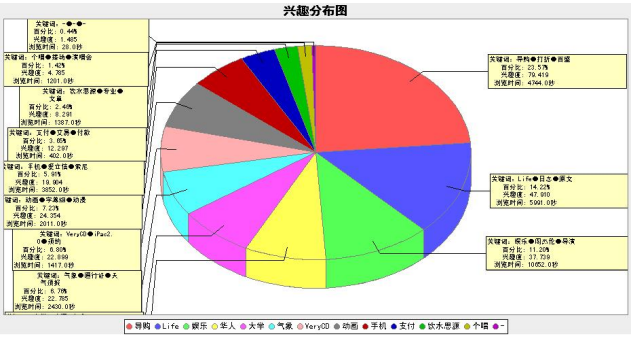

Figure 2. User interests Distribution

- During the period, the change of interests of per day, per week, or per month can be analyzed using a time series chart. Figure 3 shows the evolution of the user's interests per day from Nov.11, 2007 to Dec.15, 2007. It is easy to observe that his shopping interests of 15, Nov. is the greatest, and then he gradually lost interest in shopping. Some interests such as education (university, Jiaotong, course), Blog (life, blog, original) are constantly keeping.

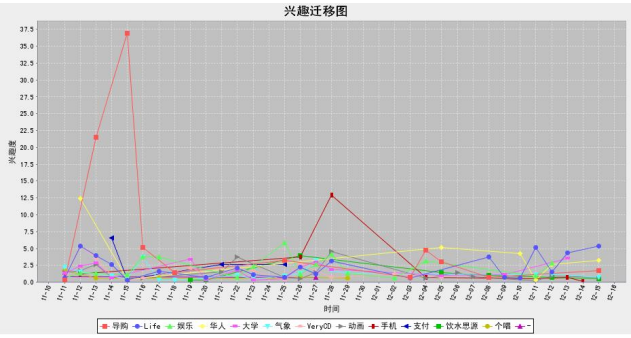

Figure 3. User interests evolution from Nov.11 to Dec,15. 2007

4 Experiments

4.1 Clustering Evaluation

We choose two test sets: sohu (news.sohu.com) and sina (sina.com.cn) as references (the ground truth). Each set has

Table 1. Comparison of interests found

User predefined interests	System-found interests
shopping	导购(Shopping guidance), 打折(discount), 百盛(BaiSheng: the name of a shopping mall)
blog	Life, 日志(Blog), 原文(original)
Entertainment	娱乐(Entertainment), 周杰伦(Jielun Zhou: a famous popular singer),导演(director)
TV movies	华人(Overseas Chinese), 电视剧(TV movies),综艺(all kinds of art)
Education	大学(University), 交通(Jiao Tong), 课程(Course)
weather	气象(Meteorologic), 通行证(pass), 天气预报(Weather forecast)
book	VeryCD, iPac2.0, 预约(Reservation)
Cartoons	动画(motive), 字幕组(caption), 动漫(cartoons)
Mobile phone	手机(Mobile phone), 爱立信(Ericsson), 索尼(Sony)
shopping	支付(pay), 交易(transaction), 付款(pay money)
BBS	饮水思源(BBS of SJTU), 专业(major), 文章(article)
Entertainment	个唱(Singer), 捧场(Welcome for singer), 演唱会(Music show)

10 categories and 100 pages of each category. We use K-means from Weka and SK-means as baseline. K is set to 10, it is a reasonable assumption based on our experiments. Precision and recall are widely used in information retrieval. We use cluster precision and recall to evaluate the correctness of clustering results. They are defined in the following:

$$precision = \frac{1}{|P|}\sum_{p_i \in P}\frac{|\ C(result, p_i) \cap C(reference, p_i)\ |}{|\ C(result, p_i)\ |}$$

$$recall = \frac{1}{|P|}\sum_{p_i \in P}\frac{|\ C(result, p_i) \cap C(reference, p_i)\ |}{|\ C(reference, p_i)\ |}$$

Where C represents "$Cluster$", P is the set of pages, $|P|$ is the number of pages. $Cluster(result, p_i)$ is the system generated cluster where page p belongs; $cluster(reference, p_i)$ is the cluster where p_i is according to the reference result. We also use Entropy and Purity to evaluate the clustering results. The entropy is a more comprehensive measure than purity. It considers the entire distribution. Both purity and entropy are biased to fa-

vor large number of clusters. The results of clustering web pages from Sohu and Sina are shown in the Table 2.

Table 2. Result of Clustering

		Entropy	Purity	Precision	Recall
Sohu Net	K-means	2.8337	0.2460	0.1675	0.2802
	SK-means	0.8587	0.7698	0.6686	0.6988
	Our method	0.7464	0.8121	0.7109	0.7342
Sina Net	K-means	2.9084	0.2410	0.1674	0.2364
	SK-means	0.9583	0.7260	0.6320	0.6796
	Our method	0.6902	0.8105	0.7460	0.7965

Both results show that our method outperforms the baselines. The algorithm has a high purity of 81%, with the average precision of 72.5%, recall of 76%. Results show that there are lots of different noises on the web pages, such as advertisements. Some advertisements are regarded as the contents of the pages, which has reduced the precision.

4.2 Human reviews

There are 10 voluntary students jointed in our experiments. Each participant was installed our system for half a year. We use their two months data as the test set. There are 8621 pages which cover different topics including politics, culture and so on. We ask each student to rate the interests generated with the rating scale of 1 to 5. Figure 4 is the average score of keywords rated by each user.

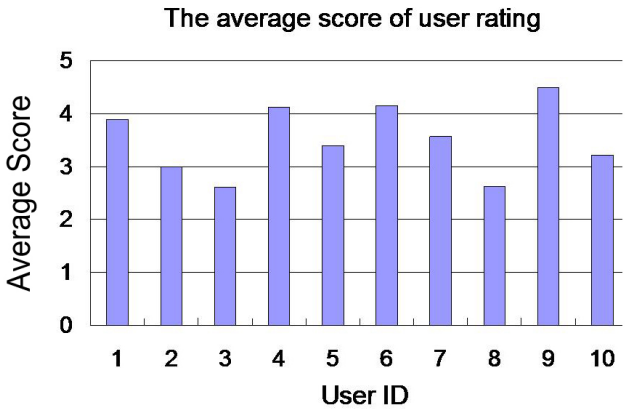

Figure 4. The average score of user rating

The percentage of different rating scale for keywords is show in Table 3. Rating 5 means three generated keywords are correct user interests, rating 4 means two of three keywords are correct. Based on the results, about 59.14% interests (keywords vectors) generated are rated as 5 and 4. Only 13.98% interests generated by the system are proved irrelevant with user's interests.

Table 3. Human Evaluation score percentage

Score	5	4	3	2	1
Keyword vectors	25.81%	33.33%	19.35%	7.53%	13.98%

5 Conclusion

In this paper, we propose a system to investigate the problem of finding user interests. Our system utilizes the implemented plug-in to collect the data of the pages visited by a user and track his browsing behaviors. The system combines the page content and browsing behavior analysis to find and generate the user's interests automatically. By selecting different time periods, user interests can be generated dynamically. The change of interests can be analyzed. One of the applications of our system is to be installed in a home PC. Parents can know their child's browsing interests implicitly and relieve their worries for unhealthy information on the Internet.

References

[1] M. Claypool, M. Claypool, D. Brown, D. Brown, P. Le, P. Le, M. Waseda, and M. Waseda. Inferring user interest. *IEEE Internet Computing*, 5:32–39, 2001.

[2] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001.

[3] Y. W. K. Z. F. L. X. W. Fang Li, Yihong Li. Discovery of a user interests on the internet. *Autonomous Systems – Self-Organisation, Management, and Control*, 2008.

[4] n. J. M. Pe J. A. Lozano, and n. P. Larra˙ An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recogn. Lett.*, 20(10):1027–1040, 1999.

[5] H. R. Kim and P. K. Chan. Learning implicit user interest hierarchy for context in personalization. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, pages 101–108, New York, NY, USA, 2003. ACM.