http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (4): 770-776, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.770.776


# Personalized Video Search Based on User Log Mining

Ming Jiang, PeiSi Cen, JingFan Tang and Xing Qi Wang
Institute of Software and Intelligent Technology, Hangzhou Dianzi University,
Hangzhou, 310018, China

**Abstract:** Since user log is an important carrier for recording the user's behavior of Web search engine, by mining and analyzing user log the user's laws and interest could be obtained effectively, which could be used to improve the accuracy of searching results when they use search engine. This study aimed to present an effective method for personalized video search based on user log mining. It first analyzed the related video log in user log which was free provided by Sogou company, classified them based on the contents, judged user's intention by using sentence similarity matching algorithm and returned the videos which were closest to user's search intention. The proposed method was verified through the experiment finally in this study.

**Key words:** User log, data mining, search engine, video search, similarity match

## INTRODUCTION

With the rapid development of the Internet, the network has become an important way for people to get information. However, the amount of information on the Internet also is on an exponential increase with the development of the Internet. Getting the information which is needed quickly and accurately from huge amounts of data becomes more and more difficult. In order to find the content wanted quickly, search engine becomes an indispensable tool (Spink *et al*., 1999). Users only need to submit query phrase to search engine, the search engine will return users related search results within a very short time. But the performance traditional search engine shows is always unsatisfactory because in many times when users enter a query word, search engine returns a thousands of search results. However, few users are willing to skim the results more than three pages. Users often fail to get the result they want. The reasons causing the result are as follows: first, the query phrase users input is short. Existing research shows that the input query phrase is always one or two words, so it is difficult to express the users' real query intention. Second, the phenomenon of ambiguity in language is unavoidable, and the technology of search engine is still stay at the stage of keywords matching now. For example, if the users input query word apple. The search engine is unable to judge the users' intention is whether apple company or fruit apple. If the users want to search an apple company but the search engine returns a lot of results related to fruit apple, which is apparently unsatisfied.

Since, traditional search engine has so many problems, it is a trend to develop personalized search service. Mining user log is an important method to discover users' interest and provide personalized search. Search engine user log is the record of the interaction information between users and search engine. These log files contain a large number of users' accessing information (e.g., IP address, visited URL, access date and time and access path). By mining search engine user log, it can find out the common discipline (accustomed search words) of users' search activity.

This study focused on video search based on user log mining. Through mining video sites visited in the user log of Sogou search engine, it could get the users' common laws and interested areas when they visited video sites, and return the best satisfied results to users when they visited next time.

## RELATED DATA

Data mining receives great attention in recent years, domestic and foreign scholars have done a lot of research and experiment on it. The authors put forward an algorithm (Mecca *et al*., 2007) for clustering search results by analyzing and mining user's search behavior, and return search results according to their interest characteristics clustered. As a result, it reduces a large number of unrelated matching and improves the speed of retrieval. The method (Ruthven, 2003) was used for evaluating search results value. It first provided discriminating and scoring the correlation of the search results by users and then the system generated a new

**Corresponding Author:** Ming Jiang, Institute of Software and Intelligent Technology, Hangzhou Dianzi University, Hangzhou, 310018, China

query based on this user returned information to improve the accuracy of the final result (Speretta and Gauch, 2005), it built a new user interest model by analyzing and handling users' browsing behavior and server log data, calculating the relationship between the model and the resulting document when searching and returning the search result based on the sort of the correlation. The authors provided a WS-LDA (Weakly Supervised Latent Dirichlet allocation) model (Guo *et al*., 2009), got the probability mode of a query word which consisted of some entity and divided it into some class by training, and judged the query with entity type by using training mode at the query side. Huang *et al*. (2003) built a correlation matrix between query words by analyzing the query words in user log, and measured the similarity between the query words by using three calculation similarity methods. Finally, it intercepted some query words whose similarity were bigger and took them as semantically related words. Huang's experimental results showed that this method based on query were words better than the method based on clicking the document.

Now user log mining (Huidan and Zeping, 2009) was used in the following areas: first, it could be used to count log data simply mainly from a view of statistics with the goal of analyzing the capability of Web site. The returned results were the pages which user frequently visited, the number of visiting times in one unit time and the figure of the access data following time. Second, with the aim of improving Web site design, it could be used to mine users' frequent access paths and user clustering, reconstruct the connection between the pages of the sites, better adapt to the user's access habit and provide users with personalized information services. Third, with the goal of understanding user's intent, it could get access habit and frequently visited domain by analyzing user log, return user's interested content and avoid returning redundant information.

There were few examples on applying it to video search right now, user would still get a mounts of videos which did not fit their intention when they used search engine to search video. This study would apply user log mining to the video search. Search engine judged user's intention by user log, then returned videos related user's intention rather than returned videos which were only query word matched and improving its properties.

## THEORY AND ALGORITHM

User log mining was generally divided into three phases: data preprocessing, pattern discovery and pattern analysis.

User log file stored the original records of the user's access site information, so mining these data directly was difficult. Before using algorithms or tools for analyzing it, pretreatment was not only conductive to the subsequent mining algorithms but also important to ultimate accurate and reliable user browsing mode.

As the core of Web log mining, the aim of pattern discovery phase was to use a variety of data mining techniques for discovering rules and pattern hidden behind data. Pattern discovery mainly included methods such as generating sequential patterns, association rule, clustering and classification.

The analysis of pattern was the final step in web log mining. Its purpose was to filter the useless or not interested parts in pattern generating process by the rules. And then change useful patterns and rules into knowledge and apply it to detail events in reality. The process was shown in Fig. 1.

**Data preprocessing:** In this study, the data for experiment was provided free of charge from Sogou company's user log. The format was as follows: access time \user ID\query word\the rank of this URL in the returned results\the sequence number user clicked\the URL user clicked, where user Id was assigned automatically according to the Cookie information when user uses browser to access search engine, it was that when using the same browser to input different query correspond the same user ID. Some of the data was shown as Fig. 2.

This study involved video search, so the records which were unrelated video site must be removed firstly.

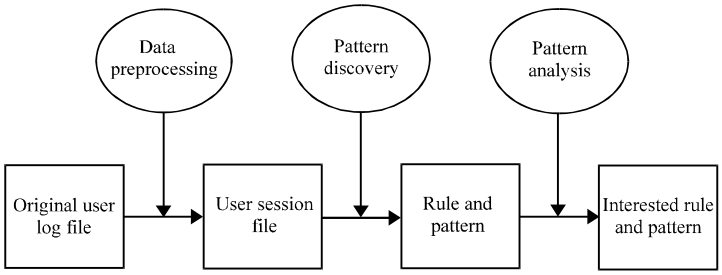


Fig. 1: The process of user log mining

00:00:00 00717725924582846 [闪字吧] 1 2 ww.shanzi ba.com/
00:00:00 41416219018952116 [霍震霆与朱玲玲照片] 2 6 bbs.gouzai.cn/thread-698736.html
00:00:00 9975666857142764 [电脑创业] 2 2 ks.cn.yahoo.com/question/1307120203719.html
00:00:00 21603374619077448 [111aa图片] 1 6 www.fotolog.com cn/tags/aalll
00:00:00 7423866288265172 [豆腐的制成] 3 13 ks.cn.yahoo.com/question.1406051201894.html
00:00:00 06168777764073358 [t udou. com+禁播电影] 29 topic.bindou.com/1487/
00:00:00 3933365481995287 [最佳受孕时间] 6 3 ks.cn.yahoo. com/question/1407051001276.html

Fig. 2: Example of original user log

少年英雄方世玉 www.youku.com/playlist_show/id_937166.html 电视剧频道 少年英雄方世玉
警察学校全集 www.youku.com/playlist_show/id_1049937.html 电影频道 警察学校(全集)共7部
唐山黑恶势力 v.youku.com/v_show/id_cb00XMzczNzc2OA==.html 原创频道 看看唐山黑恶势力多么残忍1
影视歌曲 www.youku.com/playlist_show/id_886465.html 音乐频道【音乐】影视歌曲mv收录
死神 mv www.youku.com/playlist_show/id_202885.html 动漫频道 死神
倩女幽魂 www.youku.com/playlist_show/id_551240.html 电视剧频道 倩女幽魂 (全集)
新结婚时代 www.youku.com/playlist_show/id_193709.html 电视剧频道 电视剧 新结婚时代

Fig. 3: Example of latest user log

The process of identifying user's session could be reduced if the whole file was considered as one record. When analyzing the user log, the user ID could be removed and the access time, the rank of the URL in the results and the sequence number user clicked which were not required could also be removed. Finally, there were only the information of query words and their related URL video site in the record of user log. Because of insufficient information provided by the current user log for the requirements of following analysis, it was needed to dig out more information based on URL. It could get the title and type of the video by its URL because one URL belonged to one video address and one video included the information such as title, play time, type, number of hits, comments. Here only the record from YouKu site would be extracted and a program would be wrote to realize the above work. Some example of the data was shown as Fig. 3.

**Classification of the user log:** After preprocessing from the previous section, the data would only included recodes related YouKu site. YouKu site divided the video into categories such as movies, TV shows, music, animation and original, etc. Previously, the video's type which corresponded to the category was extracted. It could divide the data into about twenty types.

Now the user query words in each type would be classified: access circulating the user query words in this type, if the query word existed in the previous record, this query word's information would be merged with the former query word's information. Here only one query word would be kept because they were the same. But the videos' titles must be all kept no matter whether they were equally or not because the more same query words with same title means the more this video was accessed. If this query word did not exist in the previous record, this query word was classified as a new type. Synonyms were treated as two words because video search was different from the general text search. After the previous steps, the user log now was divided into twenty types and each type was divided into several small types with query word as symbol. The structure was shown in Fig. 4.

**Chinese word segmentation algorithm:** The length of user's query word was not fixed. If the query word was only a single word, it would be used to match the results directly, but in the case where query word was long and includes many words, the effect was poor and fewer results would be returned when using the query word to match especially for the Chinese words. So it was really required to do the word segmentation for the Chinese query word.

There were many concrete algorithms for Chinese word segmentation (Huang and Zhao, 2007) currently. In this paper, it adapted forward maximum matching method algorithm (Ruilei *et al*., 2011). The concrete steps were as follows: it assumed that the length of the longest entry in the automatically segmentation dictionary was m, then it took the previous m characters of the current String in the text as a matching field, found the automatic word segmentation dictionary. If there was a word whose length is m in automatically segmentation dictionary, the match succeeded, and it would cut the match field as a word. Otherwise, there was no word whose length was m in the automatically segmentation dictionary, it would delete the last character of the match field and remain the m-1 characters as a new match field for matching until a word was cut from the String. A word would be cut after it has been matched. Then it did these steps until all words were cut and then save the results.

**Judge user's intention:** When users input the query word for searching video, the query word could be obtained for the video site search engine to search the videos related with the query word. There were different types of videos at present such as film type, TV show type or original type.

The user's intention could be judged according to the user log. If user input a new query word, which did not exist in current user log, it would return user with the video which was got from the previous video site. If user's query word existed in user log and the query word was not a single word, it would first perform Chinese word segmentation for the word. Then it judged video type according to sentence similarity matching algorithm (Chuan-Peng and Zhi-Gang, 2012) on the relevance between the query word and video type. Last, it would return the video with the biggest relevance to the user. For sentences of A and B, there were definitions as follows:

**Definition 1:** length (A) was used to represent the number of words in sentence A after doing Chinese word segmentation and length (B) was used to denote the number of words in sentence B after doing Chinese word segmentation. Same word (A, B) represented the number of same words in sentences A and B. Then the word form similarity word same (A, B) using the following formula:

$$\text{Word same}(A,B)=2\times\frac{\text{Same word}(A,B)}{\text{Length}(A)+\text{length}(b)}$$

**Definition 2:** Once word (A, B )was used to represent the collection of words which only once together existed in sentences A and B. A order (A, B) represented the vector consisted of the location number which was the word in Once word (A, B) appeared in sentence A. Rev num (A, B) represented the vector consisted of the components which came from A order (A, B) and the components was sorted by the sequence which was the word corresponding to the component appeared in sentence *B*. Rev num (A, b) represented the reverse number of every adjacent component of Border (A, b). Then the word order similarity Ard same (A, B) of sentences A and B was defined by the following formula:

$$\text{OrdSame}(A,B)=\begin{cases}1-\dfrac{\text{RevNum}(A,B)}{\text{Onceword}(A,B)-1}, & \text{当 OnceWord}(A,B)>1,\\ 1, & \text{当 Once Word}(A,B)=1,\\ 0, & \text{当 Once Word}(A,B)=0\end{cases} \quad (1)$$

Then the similarity of sentence A and B which was defined as Same (A, B) using the following formula:

$$\text{Same}(A,B)=\lambda_1\times\text{Word same}(A,B)+\lambda_2\times\text{Ord same}(A,B) \quad (2)$$

where, $\lambda_1+\lambda_2=1$.

Here, was a simple example to explain the process mentioned above.

The different between Chinese sentence and English sentence was Chinese sentence need doing Chinese word segmentation before calculating the similarity while English sentence needn't. For simplicity, English sentence was used to explain the process here.

"Children like hearing stories before sleep" was treated as sentence A and "before sleep children like reading stories" was treated as sentence B.

The sentence A had 6 words and sentence B also had 6 words. Children, like, stories, before, sleep were the same words between them. So:

$$\text{Word same}(A,B)=2\times\frac{5}{6+6}=\frac{6}{6}$$

Once word (A, B)was a collection shown as follows: ("Children", "like", "stories", "before", "sleep").

A order (A, B) could be got from correspondence:

| Children | Like | Hearing | Stories | Before | Sleep |
|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 |

The word corresponded the component appeared in sentence B was shown as follows:

| Before | Sleep | Children | Like | Reading | Stories |
|---|---|---|---|---|---|
| 4 | 5 | 0 | 1 | - | 3 |

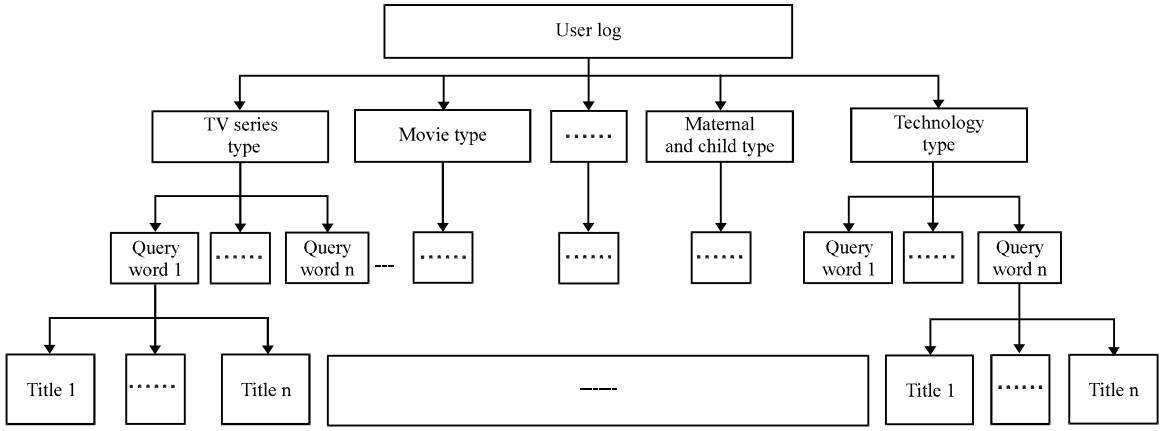


Fig. 4: The types in user log after classifying

Border (A, B) = (4, 5, 0, 1, 3) Here, was only one reverse ((5, 0)) in Border (A, B). So:

$$Ordersame(A,B)=1-\frac{2}{5-1}=\frac{3}{4}$$

$$Same(A,B)\frac{9}{10}\times\frac{5}{6}+\frac{1}{10}\times\frac{3}{4}=\frac{33}{40},\ \lambda1=\frac{9}{10},\ \lambda2=\frac{1}{10}$$

With the sentence similarity matching algorithm mentioned before. The similarity between query word and all kinds of type could be calculated. As follows, key represented query word and class $type_i$ represented video type, where i = 0....., 20, there were n titles in class $type_i$: $title_1$, $title_2$....., $title_n$, the similarity between query word key and video type class $type_i$ using the following formula:

$$Same(key,classtype_i)=\frac{\sum_1^n same(key,title_n)}{n} \quad (3)$$

Now considering calculating one type video's similarity was smaller than other video types, but the number of this type was more than the others. The attention of this type video was also large. If considering the effect of the video number. The formula becomes as follows:

$$Same(key,classtype_i)=\alpha_1\times\frac{\sum_1^n same(key,title_n)}{n}+\alpha_2\times n \quad (4)$$

where, $\alpha_1+\alpha_2 = 1$.

If user's query word was only a single word, the similarity between the query word and various types of video could not be judged according to the algorithm mentioned above. The similarity of them were judged by using the hit number of query video in every type .The entire process mentioned above was expressed in Fig. 5.

## EXPERIMENT AND RESULTS ANALYSIS

The user log for experiment was provided by Sogou company (http://www.sogou.com/labs/dl/q.html), and the sentence of "我与苹果的故事" was as query word. There were 18 title records in user log distributed in type animation class, education class, original class, maternal and child type. The calculation results of these titles and query word using sentence similarity matching algorithm were shown in Table 1.

The two similarities between the all kinds of types and query word were calculated and the results were shown in Table 2.

It was judged that the maternal and child type' video had the biggest similarity with the query word. So the videos related with maternal and child type would be returned to users, rather than all videos were returned to users.

From the analysis of the experimental results ,the phenomenon that the more same word after doing Chinese word segmentation of the query word and videos' title have, and the closer order they were in their sentences, the higher similarity they have was found. In addition, considering video number or not was also related to the judgment of similarity, extreme case maybe appear if this element was not considered. For example, a video type had only one video and its title equaled to the query word. At the same time, its access record also existed in the user log. So the similarity between it and query word

Table 1: The similarities between query word and video titles

| Video type | Video title | Similarity |
|---|---|---|
| Education type | 儿童故事 小苹果医生 | 0.36 |
| | 尚文华讲述苹果的故事 | 0.64 |
| | 会爆炸的苹果 益智故事 | 0.59 |
| | 会爆炸的苹果 寓言故事 | 0.59 |
| Animation type | 听啥嘀啥讲故事02苹果熟了 | 0.33 |
| | 秋头甲鸟故事比西瓜还大的苹果 | 0.41 |
| | 苹果大师的故事 | 0.55 |
| Original type | 苹果和梨子的爱情故事 | 0.72 |
| | 俩苹果，圣诞节的小故事 | 0.50 |
| | 苹果故事 朋友 | 0.72 |
| | 舞台剧 一个苹果的故事 | 0.64 |
| Matemal and child type | 红苹果的故事 | 0.70 |
| | 幼儿故事表演《红红的苹果》 | 0.36 |
| | 成成和苹果的故事 | 0.70 |
| | 给苹果里的猫咪讲故事 | 0.46 |
| | 苹果想要听故事 | 0.50 |
| | 苹果与方向盘的故事 | 0.65 |
| Technology | [苹果核肖发_Siri给你讲故事 | 0.42 |

Here it took $\lambda_1 = 0.9$, $\lambda_2 = 0.1$

Table 2: The similarity without considering video number and the similarity considering video number

| Video type | Similarity without considering video number | Similarity considering video number |
|---|---|---|
| Education type | 0.545 | 0.891 |
| Animation type | 0.429 | 0.686 |
| Original type | 0.645 | 0.981 |
| Matemal and child type | 0.561 | 1.105 |
| Technology | 0.420 | 0.478 |

Here it took $\alpha_1 = 0.9$, $\alpha_2 = 0.1$

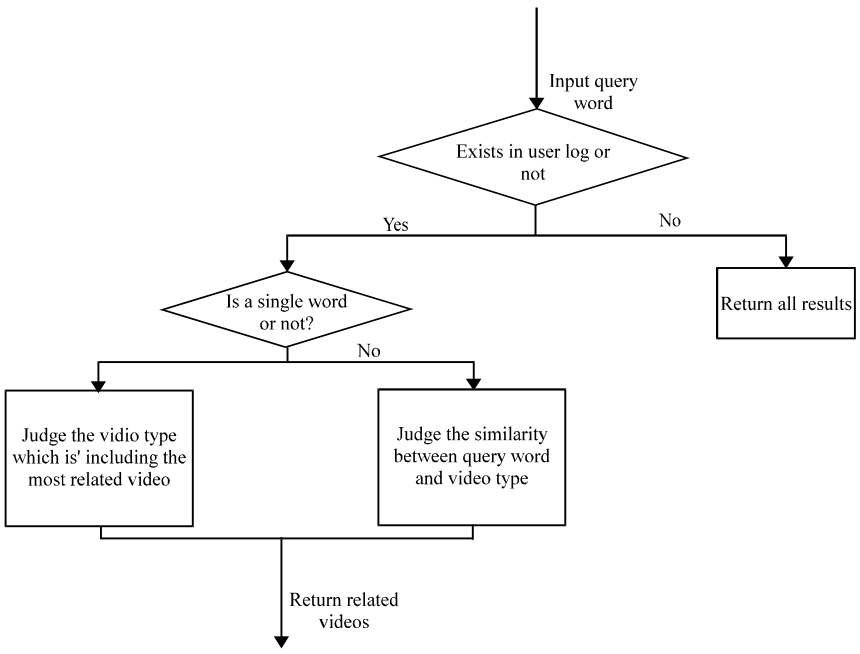


Fig. 5: The whole process flow

was 1 which was higher than other similarities. But the only one video could not be returned to user because it did not meet the actual situation. If considering the factor of the numbers of video, it could effectively avoid this situation from happening.

## CONCLUSION

This study firstly introduced the research of web log mining in recent years, and got the laws when user visited the video site by analyzing the user log provided by

Sogou company. Then it put forward a method which judged user frequently access domain by using the similarity of user's query word and the video title. Last, the paper studied the possibility of this method by taking YOUKU site as a carrier.

The problem of judgment the similarity of query word which was only single word would be considered as our key work in future. Integrating other video site to expand the usage scope of this method was also regard as key work.

## ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China (No. 61070213), the Zhejiang Provincial Natural Science Foundation of China (No. Y1111192, LY12A01019) and the Zhejiang Provincial Technical Plan Project (No. 2011C13008).

## REFERENCES

Chuan-Peng, C. and W. Zhi-Gang, 2012. A method of sentence similarity computing based on hownet. Comput. Eng. Sci., 34: 172-175.

Guo, J., G. Xu, X. Cheng and H. Li, 2009. Named entity recognition in query. Proceedings of the 32nd Acm SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2009, Boston, Massachusetts, USA., pp: 267-274.

Huang, C.K., L.F. Chien and Y.J. Oyang, 2003. Relevant term suggestion in interactive Web search based on contextual information in query session logs. J. Am. Soc. Inform. Sci. Technol., 54: 638-649.

Huang, C.N. and H. Zhao, 2007. Chinese word segmentation: A decade review. J. Chin. Inform. Process., 21: 8-19.

Huidan, X. and L. Zeping, 2009. Research on web log mining. Comput. Digital Eng., 37: 17-19.

Mecca, G., S. Raunich and A. Pappalardo, 2007. A new algorithm for clustering search results. Data Knowl. Eng., 62: 504-522.

Ruilei, W., L. Jing, P. Xiaohua and L. Xiupei, 2011. An improved forward maximum matching algorithm for Chinese word segmentation. Comput. Applic. Software, 28: 195-197.

Ruthven, I., 2003. Re-examining the potential effectiveness of interactive query expansion. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-August 1, 2003, Toronto, ON, Canada, pp: 213-220.

Speretta, M. and S. Gauch, 2005. Personallized search based on user search histories. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, September 19-22, 2005 France, pp: 622-628.

Spink, A., J. Bateman and J.B. Jansen, 1999. Searching the Web: A survey of EXCITE users. Internet Res., 9: 117-128.