Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(12):851-858

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Blog posts recommendation based on PLSA and Naive Bayesian classification algorithm

Lin Cui, Caiyin Wang and Xiaoyin Wu

Intelligent Information Processing Lab, Suzhou University, Suzhou, Anhui, China

---

ABSTRACT

As one of the important applications of Web2.0 technology, blog attracts more and more users. Writing and browsing blog has become a popular hotspot of network culture, which promotes the development of blog search service. But, the current blog search engines are mostly only based on matching query keywords; lack the ability of automatically extracting users' interests and recommendation. Really Simple Syndication (RSS) is a format of describing website and keeping synchronization with website content. Using RSS to aggregate blog posts has the advantage of letting users get the latest update of blog posts. However, the posts collected by RSS don't always attract users; users still need to browse every subscription post to find the interesting posts. To address this problem, the time spent by users on reading blog posts is viewed as a key factor to measure the users' interests. In this paper, we firstly used probabilistic latent semantic analysis (PLSA) to discovery the topics of blog posts, then adopted Naive Bayesian algorithm to classify the blog posts which was primarily connected with the users' reading time, and lastly ranked and recommended the unread interesting posts to users. Experiments showed that our proposed method could recommend the favorite blog posts to users according to the users' browsing interests.

Key words: Blog Posts Recommendation, Probabilistic Latent Semantic Analysis, Naive Bayesian Classification Algorithm, Really Simple Syndication, Reading Time

---

INTRODUCTION

With the rapid development of Web 2.0 technology, blog has become one of the important information sources of the Internet. Users can share their information and write personal views via blog which breaks the single mode that users can only passively receive Internet information. Users are very eager to find and explore their interesting blog, so as to promote the development of blog search service. At present, there have been many blog search engines, such as Technorati, Day Pop and so on. Google and Baidu also developed their blog search engines. Through using these blog search engines, we find that their working principles are mostly based on the input keywords matching and their search methods are only the extension of traditional search engines. Moreover, precision ratio and recall ratio of most blog search engines are not satisfied to users.

RSS is an abbreviation of Really Simple Syndication which is also named as polymerization RSS and is a simple way of online content sharing. RSS subscription can quickly obtain the most comprehensive, up-to-date information. More and more people of all ages enjoy RSS. Now, most blog websites provide convenient RSS service.

In this paper, a blog posts recommendation method based on PLSA and Naive Bayesian classification algorithm was introduced. When users read blog posts collected by RSS, by recording the users' reading time, we could estimate whether users were interested in blog posts and recommended the interesting posts to users. The model obtained the topic of blog posts by analyzing blog posts collected by RSS Reader with PLSA and users needn't sum up the key words about interests. Then using Naive Bayesian classification algorithm, our proposed method realized automatically mining users' interests related to reading time, and recommended the popular blogs that were

consistent in the interesting topics to users. Our ultimate design goal is extracting blog posts corresponding to the topics of users' interests from RSS Reader for users.

RELATED WORK

Personalized recommendation service is the developmental trend of information service. Through the study of different users' interests, personalized recommendation service is provided for different users. The model of users' interests is the foundation and core of the personalized recommendation service, which has two main ways that are explicit users' interests modeling and implicit users' interests modeling [1]. Now, most of the existing personalized recommendation systems use implicit methods, which are obtained by capturing the users' behavior of browsing webpage contents.

In 1994, Morita and Shinoda proposed that reading time could reflect the users' interests [2]. In 1997, Nichols introduced the implicit feedback user behavior type [3]. In 1998, Oard and Kim investigated the implicit feedback technology in the information filtering system, on the basis of feedback behavior given by Nichols, they summarized three kinds of implicit feedback behavior which were review, retention and reference, in which the review behaviors included choice, reading time, editing, repetitive operation and the purchase power [4]. In 2003, Kelly and Teevan summarized and analyzed the existed important papers about implicit feedback, they discovered that implicit feedback had a very good influence on the user modeling and reading time was one of the important factors [5].

In fact, we find that the correlation degree between the length of the article and reading time is not high, because users may not read the full text of articles, particularly for not interesting articles. After users only read the first few lines, they would make decision whether or not to continue. So, according to users' reading behavior, this paper makes the following assumptions:

(1) Compared with the articles which users are not interested in, users will spend more time on the articles which users are interested in.
(2) For not interesting article, no matter how long the length is, reading time is always less. For the articles that users are interested in, users spend more reading time on long articles than on short articles. But users also tend to read the full text of short articles, not full text of long articles. As the article's length increases, the reading time will increase, but don't have a fixed growth rate.
(3) Because users tend to finish reading a very short article, it is very difficult to estimate whether users are interested in this article through reading time.
(4) Generally speaking, the blog website which is more consistent with the users' interests is more likely to output high quality articles that users are interested in.

INTRODUCTION OF BLOG AND RSS

WHAT IS BLOG?

Blog, namely "Web log", is a website in which users can write their posts without any limits of time and is the fourth network exchange way after E-mail, BBS and ICQ [6]. Blog posts are managed mainly by time sequence. As a new social media and a personalized information release platform, Blog attracts an increasing number of Internet users. Users can publish their posts easily and timely, and also conveniently online exchange with others. Additionally, blog can make readers leave comments in an interactive format. Blog is an effective combination of private and public nature. Compared with BBS, blog reveals even more personalized and more targeted support.

INTRODUCTION OF RSS

In 1997, Netscape developed RSS; the concept of "push" technology was born [7]. Later, RSS technology is used to create summary for blog website content which promotes the development of blog. RSS is an information aggregation technology and blog posts can be sent to the users' desktop according to the requirements of users. The core idea of blog is to realize knowledge sharing, to exert the function of blog, it is inseparable from RSS. Blog which supports RSS can produce XML language codes in the background, the codes are known as RSS Feed [8]. RSS Feed is marked with different labels, the user needn't enter the name of blog website, and the address of a blog website's RSS Feed is added to RSS Reader. When subscribing to a blog website, the blog posts are automatically updated to the user's computer with the help of RSS Feed, and make users read posts very easily. A full RSS Feed codes is as follows:

```
<?xml version="1.0" encoding="utf-8"?>
<rss version=version number>
<channel>
<title><!--webpage name-- ></title>
<description><!--brief description-- ></description>
```

<link><!--page's url-- ></link>
<item>
<title><!--post's title -- ></title>
<link><!--post's URL-- ></link>
<description><!-- post's content-- ></description>
<author><!-- author-- ></ author>
<pubDate><!-- pubdate -- ></pubDate>
</item>
<item>……</item>
</channel>
</rss>

The posts collected by RSS don't always attract users; users still need to browse syndication to find the interesting posts in RSS Reader. If RSS posts were ranked by users' interests, users would be liberated from the heavy subscription reading. The user's preferences can be obtained from the feedback analysis, but explicit feedback, such as score, label and so on, can bring additional burden to users; implicit user feedback can achieve the users' preferences by monitoring the users' natural behavior. This paper analyzed the implicit user feedback about reading time and models personal hobbies. The time spent on reading RSS subscription by users was used to be a key factor to measure the interests of users.

INTRODUCTION OF PROBABILISTIC LATENT SEMANTIC ANALYSIS AND NAVIE BAYESIAN CLASSIFICATION ALGORITHM

PROBABILISTIC LATENT SEMANTIC ANALYSIS

In 1999, Thomas Hofmann proposed Probabilistic Latent Semantic Analysis (PLSA) which effectively overcame the defects of LSA. The method uses probability model to simulate potential semantic space, computing the relationship between documents and latent semantic and the relationship between latent semantic and words. Documents and words are mapped to the same semantic space, which can identify synonyms and polysemy [9]. In topic modeling, it is generally believed that a document is a combination of the words belonging to many topics. Given topic Z, we assume that the word w is conditionally independence of the document d. Compared with LSA, PLSA introduces a latent class. From the view of probability model, the model of PLSA is shown below:

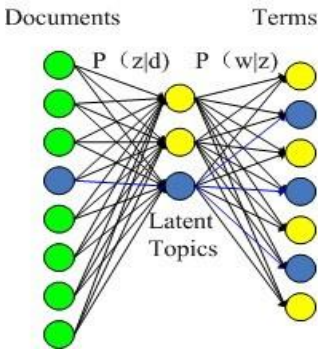

Fig. 1. Representation of PLSA model

In this probabilistic model, latent variable $z_k \in \{z_1, z_2, \cdots, z_k\}$ corresponds to latent topics. $P(d_i)$ represents the probability of the document d_i discovering in the document set. $P(w_j \mid z_k)$ represents the number of the probability of occurrence of the relevant term. When determining the semantic z_k, $P(z_k \mid d_i)$ represents the semantic distribution of a document.

According to $P(d_i)$, randomly sampling select a document d_i; after selecting the document d_i, according to $P(z_k \mid d_i)$, we sampling select the semantics z_k expressed by documents; after selecting the semantic z_k, according to $P(w_j \mid z_k)$, choose the words of the document; such, we get an observation (d_i, w_j). Repeating this process many times, we obtain a similar N co-occurrence matrix. The following two formulas are used to describe the joint distribution (d_i, w_j).

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i) \quad (1)$$

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \quad (2)$$

In the formula (2), $P(w_j \mid z_k)$ is a distribution probability of potential concept in words, through sorting $P(w_j \mid z_k)$, we can get a visual representation of words. $P(z_k \mid d_i)$ represents latent semantic distribution probability in the document d_i.

PLSA adopts expectation maximization algorithm (EM) to fit latent semantic model. After initializing random number, E step and M step are alternately implemented to perform iterative computation. On the E step, calculates the prior probability of (d_i, w_j) to potential concept z_k:

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)} \quad (3)$$

On the M step, using the following two formulas to re-estimate the model:

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{n} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{n} n(d_i, w_j)P(z_k \mid d_i, w_j)} \quad (4)$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{m} n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(d_i)} \quad (5)$$

When expected increase in volume is less than a threshold, iteration is stopped. At this time, an optimal solution is obtained:

$$E(L) = \sum_{i=1}^{n}\sum_{j=1}^{m} n(d_i, w_j)\sum_{k=1}^{k} P(z_k \mid d_i, w_j)\log[P(w_j \mid z_k)P(z_k \mid d_i)] \quad (6)$$

Because the training process of PLSA is the process of parameter estimation in fact, EM algorithm is the body of PLSA. EM algorithm needs multiple iterations, each iteration respectively includes that E step calculates posterior probabilities of implicit variables and M step solves the parameters of maximization of the targeted function, these two steps are the amount of calculation for PLSA. Because EM algorithm needs huge computation, in order to reduce the complexity of the algorithm, in this paper, we adopted the calculation mode of iterating one thousand loop times to get an approximate value.

NAVIE BAYESIAN CLASSIFICATION ALGORITHM

Naive Bayesian method is a classification algorithm based on Bayesian theorem and characteristic conditions independent assumptions. When using Naive Bayesian classifier for classification, we need training firstly, then estimate the priori probability of the category and the posterior probability of features, and lastly achieve the classification [10].

Define the category of document $d = \{w_1, w_2, \cdots w_j\}$ belongs to $C = \{C_{\text{int}eresting}, C_{not-\text{int}eresting}\}$, in the case of mutually independent features, considering the weight values of feature words, the classification method is illustrated using the formula (7) as follows.

$$c_{NB} = \arg\max_{c_j \in C}\left\{P(c_j)\prod_{i=1} P(w_i, c_j)wt(w_i)\right\} \quad (7)$$

Where, $P(c_j)$ is the priori probability of class c_j, $P(w_i \mid c_j)$ is the posterior probability of feature word w_i in

category c_j, $wt(w_i)$ is the weight value of word w_i in test set, when adopting Boolean type, weight value $wt(w_i)=1$.

The priori probability $P(c_j)$ is a probability of occurrence of the pre-given class. If $P(c_j)$ is not estimated, the probability of occurrence of each class can be considered to be equal. Priori probability $P(c_j)$ can be pre-estimated through manually directly computing or through probability calculation based on training set. Based on training set, the estimation method of $P(c_j)$ is shown in the formula (8).

$$P(c_j)=\frac{Doc(c_j)}{\sum_{c_j\in C} Doc(c_j)} \tag{8}$$

Where, $Doc(c_j)$ is the number of documents belonging to category c_j.

The posteriori probability $P(w_i \mid c_j)$ refers to the probability of the feature word w_i in category c_j, which can be estimated through calculation on training set. The estimation method is shown in the formula (9):

$$P(w_i,c_j)=\frac{Weight(w_i,c_j)}{\sum_{i=1}^{n} Weight(w_i,c_j)} \tag{9}$$

Where $Weight(w_i \mid c_j)$ is the sum of weights that the word w_i belongs to the category c_j. If the word w_i in training set doesn't exist in all the classes, then $P(w_i,c_j)=0$. However, if this happens, the result of posterior probability is 0, and then another class occupies a dominant position. If the number of word w_i in all categories is 0, there is no way to re-classify. To avoid $P(w_i,c_j)=0$, we adopts Laplace transform, thus the improved formulas of posterior probability calculation are shown in the formula (10), (11) and (12).

$$P(w_i,c_j)=\frac{Weight(w_i,c_j)+\delta}{\sum_{i=1}^{n} Weight(w_i,c_j)+\delta|V|} \tag{10}$$

$$V=\sum_{c_j\in C}\sum_{i=1}^{n} Weight(w_i,c_j) \tag{11}$$

$$\delta=1/|V| \tag{12}$$

When adopting Laplace transform, δ is generally taken 1, constant *V* takes the sum of weights of all the words. However, when $\delta=1$, there exist some problems that the existence probability of feature words which doesn't appeared in training set is increased, and the probability of the existed words is reduced. In order to solve this problem, we take $\delta=1/|V|$, which is equivalent that the posterior probability is a minimal existence probability, when the feature words doesn't exist. The existence of feature words has also not much influence on original probability.

BLOG POSTS RECOMMENDATION BASED ON PLSA AND NAVIE BAYESIAN CLASSIFIER

In this paper, mining users' interests is based on the following two assumptions: the first one is that users' interests is stable and blog contents are associated with users' interests; the second one is that if a user browses a blog post longer, we can speculate that the topic of this blog relates to this user's interests. Based on the above assumptions, we firstly extracted the posts from RSS Reader, pre-treated blog posts, and then used PLSA to identify the topics of posts and adopted Naive Bayesian classification algorithm to dig out the users' interests, lastly ranked blog posts in accordance with the number of occurrences of each interests. The number of occurrences of interests reflected the extent of users interesting to blog posts in RSS Reader; different interests should be calculated separately. To obtain a better personalized service, users were required to actively participate in evaluating posts. The detailed steps based on PLSA and Naive Bayesian classification algorithm are described as follows:

FEATURE SELECTION

When preprocessing the posts of RSS Reader, the first step is word segmentation. In this paper, Chinese word processing part adopted the word segmentation software Chinese Lexical Analysis System (ICTCLAS) developed by Institute of Computing Technology in Chinese Academy of Sciences, ICTCLAS has good segmentation effect relatively. The blog posts from RSS Reader are regarded as a series of words. By eliminating the stopping words and calculating TF-IDF weights and so on, the posts are further preprocessed. Select the top number of words of blog posts as the key words, and then use them for PLSA model training, we can get $P(z|d)$ matrix, which is a expression that document is in the topic space.

PARAMETER LEARNING

Scheduling problem can be seen as a classification problem, the blog posts are divided into two classes: interesting class and not-interesting class. The probability of the posts which are divided into interesting class ranked higher should be in the front position, C is used to express interesting class, $P(C|d)$ is what we get. For the same topic of many posts, the posts' weights are incorporated learned weights when learning each post. The global topic weight is represented with $P(c|z_k)$. Let z_k represent the before sum of all $P(z_k|d)$, then the formulas of updating $P(c|z_k)$ are as follows[11]:

$$P(C|z_k)_{new} = \frac{P(C|z_k)_{old} \cdot z(k)_{old} + P(z_k|d) \cdot P(C|d,t)}{z(k)_{old} + P(z_k|d)} \tag{13}$$

$$z(k)_{new} = z(k)_{old} + P(z_k|d) \tag{14}$$

Where, $P(C|z_k)_{old}$ represents weights before combination, $P(C|z_k)_{new}$ stands for key words weights after combination. For every new post, repeat this process. $z(k)$ is also such so.

RANKING POSTS

If the posts are divided into interesting class and not-interesting class, then the posts which have a greater probability of being classified into the interesting class should be discharged to the front of the list. Because $P(\text{int}eresting|d) + P(not-\text{int}eresting|d) = 1$, on the basis of Naive Bayesian classification algorithm, there are:

$$P(d|C) = \prod_{k=1}^{K} P(z_k|C)^{P(z_k|d)} \tag{15}$$

$$P(z_i|C) = \frac{P(C|z_i) \cdot P(z_i)}{\sum_{k=1}^{K} P(C|z_k) \cdot P(z_k)} \tag{16}$$

Let C be an interesting class and $\neg C$ be a not- interesting class. The likelihood ratio of $P(C|d)/P(\neg C|d)$ is:

$$\frac{P(C|d)}{P(\neg C|d)} = \ln\frac{P(C)}{P(\neg C)} + \sum_{k=1}^{K} P(z_k|d) \ln\frac{P(z_k|C)}{P(z_k|\neg C)} \tag{17}$$

EXPERIMENTAL PRINCIPLE ANALYSIS

Data are derived from RSS subscription blog posts of Sina blog (http://blog.sina.com.cn/). Using a simple RSS Reader to record users' reading time with seconds, we assume that reading time is a key factor to reflect the users' interests and the post's length is not always associated with reading time. Therefore, reading speed can not be used to measure the users' interests. The revised reading time can be used to measure the attitude of users to some posts. Firstly, we selected 410 blog posts, after users finished reading, we required users to rank the posts explicitly. The score was divided into three levels: not-interesting, general, interesting. For 410 blog posts, 190 blog posts were ranked not-interesting, 95 blog posts were ranked general, and the rest 125 blog posts were ranked interesting. Through careful analysis, we can conclude that for the posts which users are not interested in, reading time distribution is more concentrated; but for the posts which users are interested in, reading time distribution is more dispersed.

EXPERIMENTAL DATA AND RESULTS ANALYSIS

In view of pre-reading RSS posts, we provided users, items and scores. In this paper, the selected data set contained 100 users and 857 RSS items about blog posts. Each user scored at least 15 blog posts. During the evaluation process, the data set of RSS items was divided into training set and test set according to 4:1.

Experiment is a classification task test actually. To obtain more objective evaluation, this experiment adopted 5-fold cross validation strategy. Recommendation quality metrics used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [12], their computation formulas are as follows:

$$MAE = \frac{\sum_{u \in T} | R_{u,j} - R'_{u,j} |}{| T |} \quad (18)$$

$$RMSE = \frac{\sqrt{\sum_{u \in T} | R_{u,j} - R'_{u,j} |^2}}{| T |} \quad (19)$$

Where $R_{u,j}$ represents the evaluation score of user u to resource j, $R'_{u,j}$ represents the prediction score of the user u to resource j. T is the current test set, $|T|$ represents the number of test samples. For MAE and RMSE, it should be noted that the smaller MAE and RMSE are, the higher recommendation quality is.

As a classification task test, the experiment verified the effectiveness of the proposed blog posts recommendation based on PLSA and Naive Bayesian algorithm (Topic + NB) compared with the other three methods which are the random method (Random), the term matching method only based on key words (Term) and the potential topic method only based on PLSA (Topic). Experimental results are illustrated in the figure 2, in which, for MAE index, compared with the baseline of random methods, term model obtains 28.9% (0.872-0.583=0.289) performance improvement; topic model decreases by 43.3%(0.872-0.439=0.433); more evidently, MAE index in topic + NB model decreases by 59.4%(0.872-0.278=0.594) and further arrives at 27.8%, which shows the performance of topic + NB model is the most optimal in MAE index. About RMS index, test shows the similar change trends as the figure 2 shows.

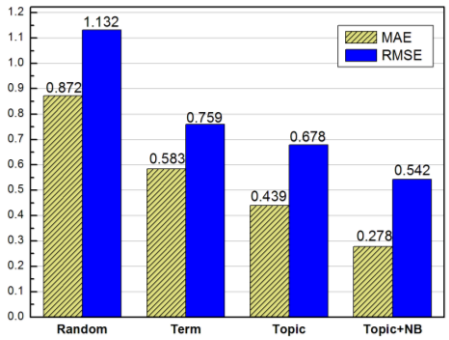

Fig.2: Contrast of MAE evaluation indicators and RMSE evaluation indicators

CONCLUSION

In this paper, we propose the blog recommendation method based on PLSA and Naive Bayesian algorithm which is a blog recommendation way for users' interests. Studies have found that the crucial factor of this method is to obtain the time of users' reading blog posts. The interests mining problem was converted into the discovery of blog topics and blog posts classification problem. PLSA was used for the discovery of blog topics and Naive Bayesian algorithm was applied to classification of RSS subscription posts and ranked them. Experiments showed that through learning a certain amount of posts and users' feedback, this model can provide reasonable prediction. We conclude that, in the same experiment condition, the blog recommendation method based on PLSA and Naive Bayesian algorithm outperforms the other three methods greatly. In summary, our proposed method is an effective and consistent method for blog posts recommendation.

Acknowledgements
This work was supported by Ordinary Project of Anhui Province Colleges and Universities Natural Science Foundation of China (No.KJ2013B283, No.KJ2012Z401) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2013YKF14) and Project of Anhui Province Higher Education Revitalization Plan of China (No.2013zytz074).

REFERENCES

[1] L. Talia, S. Michal, O. Ilit, I. Ohad, M. Joachim. *International Journal of Human-Computer Studies*. **2010**, 68(8): 483-495.
[2] M. Masahiro, S. Yoichi. Information filtering based on user behavior analysis and best match text retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, **1994**:272-281.
[3] Xin Fu. *Journal of the American Society for Information Science and Technology*. **2010**, 61(1):30-49.
[4] R. W. White, J. M. Jose, I. Ruthven. *Information Processing and Management: an International Journal*. **2006**,42(1):166-190.
[5] D. Kelly, etc. *SIGIR Forum*. **2003**, 37(2): 18-28.
[6] Jie Jiang, Yule Deng, Kate He, etc. A Blog Personality Recommender System Based on Cloud Computing Infrastructure. Proceedings of the **2012** International Joint Conference on Service Sciences. **2012**:1-5.
[7] G.T. Fekade, T. Joe, C. Richard, etc. Semantic-based Merging of RSS Items. World Wide Web. **2010**,13(12):169-207.
[8] G. H. Young, H. L. Sang, H. K. Jae, etc. A new aggregation policy for RSS services. Proceedings of the 2008 international workshop on context enabled source and service selection, integration and adaptation. **2008**:1-7.
[9] Lingfeng Niu, Yong Shi. **2010** *IEEE International Conference on Data Mining Workshops*. **2010**:1196-1203.
[10] Harry Zhang, Jiang Su. *Journal of Experimental & Theoretical Artificial Intelligence*. **2008**,20(2):79-93.
[11] Sen Liu. Analysis and application of probability latent semantic [D]. Hangzhou:Zhejiang University, **2011**.
[12] A. Erman, F. Faramarz. A belief propagation based recommender system for online services. Proceedings of the fourth ACM conference on Recommender systems. **2010**:217-220.