Chapter 80
Application of Text Data Mining to Education in Long-Distance

Jianjie Song and Hean Liu

Abstract Based on Web technology and Text Data Mining, focusing on Personalized service of Long-distance Education, this chapter aims to apply Content Mining Algorithm into Personalized recommendation of learning content, use the simulation results to verify the effectiveness of the fusion algorithm, and apply the tested algorithm to the construction of higher vocational teaching website.

Keywords Text data mining · Long-distance education · Content mining algorithm

80.1 Introduction

If a physical link is recommended from two aspects: novelty and information. Being away from the website current user is visiting should be a priority target. The physical link path length is determined by the topology of a directed graph. Each node of directed graph represents a site in the corresponding page URL [1]. If there exists a physical link from page X to page Y, there is a directed edge from corresponding node X to node Y there exists. The path distance of two web sites (i.e., u 1 and u 2) with physical link is defined as: a directed graph on the site, from u 1 to u 2 the minimum access path length.

J. Song
Hunan vocational college of science and Technology, Changsha 410004, Hunan, China

H. Liu (✉)
Hunan City University, Yiyang 413000, Hunan, China
e-mail: liuheanlaoshi@sina.com

X. He et al. (eds.), *Computer, Informatics, Cybernetics and Applications*, Lecture Notes in Electrical Engineering 107, DOI: 10.1007/978-94-007-1839-5_80,

Assuming sliding window size W is 3 [2], the operating sequence of current sliding window visitors is W = <A,B,C> , according to W and | W |. When we visit Pathset sequence database, firstly we focus on the search four the top three as A, B, C sequence, and put the last figure of sequence meeting the requirements into the recommended set. If the element in recommended set is greater than 1, such as the recommended including the following elements of {D, E, F, M}, it is advisible to choose the longest distance as the recommended webpage while concentration of greater than 1 if the recommended concentration of elements, then the study page C links to recommended concentration of the physical path distance, choose from the largest of the recommended page. Assuming M meets the requirements page, M is recommended as the next visiting page. When a user visited the M, the user access operations sequence in new sliding window changes into a <B,C,M>, then completing a recommended operation. If the four items do not meet the requirements of the sequence of focus, then search the three items begin with B, C sequence, and the last figure satisfying the requirement is added to recommended set. The other operation is the same the same as above. If the concentration does not meet the sequences requirements, then search the two items set begin with C until you find series satisfying the requirements, and other operations is the same as above procedure.

80.2 Vector Space Model

The basic concept of vector space model is as follows [3]:

(1) Documentation: refers to an article or a part of the text or fragment.
(2) Feature items: the contents of any document or fragment to be simple as it contains the basic morpheme units (characters, words, phrases, or phrases, etc.) posed by the collection, these basic characteristics of language units are collectively referred to as, namely the set of documents with term list can be expressed as D (t_1, t_2, …, t_i, …, t_n), where it is the first i-feature items, $1 \leq i \leq n$.
(3) Characteristics of the weights of items: one item for the D-document containing the n (t_1, t_2, …, t_i, …, t_n), ti feature items are often given a certain weight w_i, and they were importance in the document, namely: D = (w_1, w_2, w_i,…, w_n). Similarly, the user information requirement can also be expressed with the vector form.
(4) The vector space model: given a document D = (<t_1, w_1>, <t_2, w_2>,…, <t_i, w_i>, <t_n, w_n>), because t_i is repeated in the document and must also have priorities and the relationship have some difficulty. In order to simplify the analysis, it may temporarily not be considered in order of t_i in the document and requested the t_i can not be repeated. Then the t_1, t_2,…, t_n can be regarded as an n-dimensional coordinate system, w_1, w_2, w_i,…, w_n as the corresponding coordinate value, the D (w_1, w_2, w_i,…, w_n) can be seen as n-dimensional space

(feature items document space, i.e. TD space) in a vector, we call (w_1, w_2, w_i,…, w_n), namely the document vector D.

(5) Similarity: it is used to measure the related degree between documents or between the user's information needs (content). This method use the similarity information retrieval or information filtering, you must first be able to document the characteristics of individual items weighted and operation of collection, and then calculated the transmission document vector space with the information needs of users of the similarity between vectors, and finally provides users with a set of documents in descending order by similarity list.

80.3 The Recommendation of the Corresponding Web Page Hyperlink Method

To achieve the navigation, you must understand the interest of objects so as to targeted. First of all, it should be recommended hyperlink consistent with user interests; Second, it must a hyperlink on a Web page and user interest in the match. We use the vector space model to achieve this match. Application in the text content analysis, we give the representation of the characteristics of web pages, the page p can be expressed as a k-dimensional vector, which feature items that the page p in fi weight. Because each corresponds to a Web page hyperlink, the hyperlink to each corresponding to a k-dimensional vector p. Suppose that a user browsed the web in the current m-hyperlinks, we interested users to calculate the k-vector and the m-dimensional vector of similarity, that is, trials, set a threshold α, according to the set threshold Value [4], we have the first three with the largest similarity to users interested in hyperlink recommendation to the user, completing the recommended action.

80.4 The Recommended Method Based on Cooperation

Based on cooperation, the recommended approach is also known as collaborative filtering, a person's interest is not isolated, it is a relative concern in the interest of a group. Under normal circumstances the information received by people around the crowds is a particular result. Based on the above factors, we can group similar information to evaluate through their recommendations to other groups of users. Under normal circumstances, the use of groups of people can be divided into two types of active and passive, active people can make full use of the initiative to provide feedback, which feedback will be applied to filter non-active population, its drawback is that information resources must be considered characteristics and can not find information of interest to the user, when the system uses the early, less education information resources, the use of proximity between objects is not easy to be investigated through the evaluation.

Information customization module, information needs analysis module, similar to the matching module together form the system based on collaborative filtering. Custom modules and information needs of information analysis module of these two methods and we said, before, like content filtering technology service system, so without in-depth analysis here, mainly for the third similarity matching module to expand the analysis. This module is the first to use clustering methods, by contrast it uses a user profile object clustering, you must first use the clustering method used. Clustering can then be mapped to the user profile concept hierarchy in the multidimensional space vector form several separate feature vector, and then calculate the distance method of vector space or vector space model approach to calculate the similarity between user profiles degree, you can arrive at a similar target group. Similar to the target group can also be mentioned in the article, we apply to user groups and the clustering algorithm to obtain. According to different rating and scoring documents, this target group users were the results of the situation to get a list of recommendations that can then be recommended list and information resources. In the past, analyzing the matching calculation, the user will receive the recommended information. Recommended based on content and cooperation the two methods, we can put him in combination. First, content-based methods have the user interest model, this model shows that the contents of each user for the level of interest, similar to the interested users of its feature vectors, so the user based on feature vectors are classified by content, known content class. However, to be able to recommend to the user information of interest, we must first consider the unity of the user evaluation will be divided into two category, known as co-class. Purpose of doing hope to use the evaluation to use objects and not within the given recommendation. Two effects must be considered in an integrated similarity to the user information corresponding to recommend. According to the evaluation of the user, dynamically adjust the user types and the adaptation of various parameters in order to improve the recommendation accuracy.

Denote by μ the overall average rating [5, 6]. A baseline estimate for an unknown rating r_{ui} is denoted by b_{ui} and accounts for the user and item effects:

$$b_{ui} = \mu + b_u + b_i \tag{80.1}$$

The parameters b_u and b_i indicate the observed deviations of user u and item i, respectively, from the average.

In order to estimate b_u and b_i one can solve the least squares problem:

$$\min_{b*} \sum_{(u,i)\in\kappa} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1 \left(\sum_u b_u^2 + \sum_i b_i^2 \right) \tag{80.2}$$

Here, the first term $\sum_{(u,i)\in\kappa} (r_{ui} - \mu - b_u - b_i)^2$ strives to find b_us and b_is that fit the given ratings. The regularizing term—$\lambda_1 \left(\sum_u b_u^2 + \sum_i b_i^2 \right)$—avoids overfitting by penalizing the magnitudes of the parameters.

An easier, yet somewhat less accurate way to estimate the parameters is by decoupling the calculation of the $b_i's$ from the calculation of the $b_u's$. First, for each item i we set:

$$b_i = \frac{\sum u : (u,i) \in \kappa (r_{ui} - \mu)}{\lambda_2 + |u|(u,i) \in \kappa|} \tag{80.3}$$

Then, for each user u we set:

$$b_u = \frac{\sum i : (u,i) \in \kappa (r_{ui} - \mu - b_i)}{\lambda_3 + |i|(u,i) \in \kappa|} \tag{80.4}$$

Averages are shrunk towards zero by using the regularization parameters, $\lambda_2 \cdot \lambda_3$, which are determined by cross validation. Typical values on the Netflix dataset are: $\lambda_2 = 25,\ \lambda_3 = 10$.

Central to most item–item approaches is a similarity measure between items. Frequently, it is based on the Pearson correlation coefficient ρ_{ij}, which measures the tendency of users to rate items i and j similarly. Since many ratings are unknown, it is expected that some items share only a handful of common raters. Computation of the correlation coefficient is based only on the common user support. Accordingly, similarities based on a greater user support are more reliable. An appropriate similarity measure, denoted by s_{ij}, would be a shrunk correlation coefficient:

$$s_{ij} \stackrel{def}{=} \frac{n_{ij}}{n_{ij} + \lambda_4} \rho_{ij} \tag{80.5}$$

The variable n_{ij} denotes the number of users that rated both i and j. A typical value for λ_4 is 100. Notice that the literature suggests additional alternatives for a similarity measure.

This set of k neighbors is denoted by $S^k(i;u)$. The predicted value of r_{ui} is taken as a weighted average of the ratings of neighboring items, while adjusting for user and item effects through the baseline estimates:

$$\begin{aligned}
\hat{r}_{ui} &= b_{ui} + \frac{\sum j \in S^k(i;u) s_{ij}(r_{uj} - b_{uj})}{\sum j \in S^k(i;u) s_{ij}} \\
&= b_{ui} + \sum_{j \in S^k(i;u)} \theta_{ij}^u (r_{uj} - b_{uj}) \\
&= \mu + b_u + b_i + |R(u)|^{-1/2} \sum_{j \in R(u)} (r_{uj} - b_{uj}) q_i^T x_i + |N(u)|^{-1/2} \sum_{j \in N(u)} q_i^T y_i \\
&= \mu + b_u + b_i + q_i^T \left(|R(u)|^{-1/2} \sum_{j \in R(u)} (r_{uj} - b_{uj}) x_i + |N(u)|^{-1/2} \sum_{j \in N(u)} y_i \right)
\end{aligned} \tag{80.6}$$

Model parameters are learnt by gradient descent optimization of the associated squared error function. Our experiments with the Netflix data show that prediction accuracy is indeed better than that of each individual model. For example, with 100 factors the obtained RMSE is 0.8966, while with 200 factors the obtained RMSE is 0.8953.

80.5 Conclusion

According to tests, we may draw the conclusion that we can make improvements in web site design by this algorithm. The main measures are as followings:

We can take the optimization of the WEB site linkage structure into account from two respects. First, WEB log files can be realized for the users to become more convenient to use the resources of the website, and to strengthen the close link between pages by adapting the relevance of them. Second, if the location that a user links actually is lower than expected location, we can apply through deeper web log files to find the application and we can optimize web site pages through the establishment navigation between practical and expected users.

We can improve the site by modifying some property of pages. These methods may include the following three aspects. First of all, the probability of any hypertext links being selected in a page, depend on the number of hypertext links there contain in a page. And if a web page the page contains a lot of hypertext links, the relative probability of the link being selected will be reduced. Secondly, compared with those hypertext links behind, those before them will be easily selected. Therefore, the position is proved to be very important. Besides, under equal conditions, the regional size is also another important factor of being selected. Lastly, another factor of being selected is related to the clearness of contents and the availability of recognition of the words in a hypertext. If they convey a clear and clean meaning between the words in a hypertext and the link page, then the probability of being selected will be larger.

References

1. Huang XJ, Croft WB (2009) A unified relevance model for opinion retrieval[C]. In: Proceeding of the 18th ACM conference on information and knowledge management, HongKong, ACM, pp 947–956
2. Kim SM, Hovy E (2010) Determining the sentiment of opinions[C]. In: Proceedings COLING-04, Geneva, Association for Computational Linguistics, 1, pp 267–1376
3. Bo P, Lillian L, Shivakumar V (2009) Thumbs upon sentiment classification using machine learning techniques, presented at the 2002 conference on empirical methods in natural language processing (EMNLP'2009), pp 79–86
4. Gelan Y, Xue X, Gang Y, Jianming Z (2010) Semi-supervised classification by local coordination lecture notes in computer science, vol 6444, Neural information processing. Models and applications, pp 517–524

5. Gelan Y, Xue X, Gang Y, Jianming Z (2010) Research of local approximation in semi-supervised manifold learning. J Inf Comput Sci 7(13):2681–2688
6. Borges J, Levene M (1999) Data mining of user navigation patterns. In: Proceedings of the workshop on web usage analysis and user profiling (WEBKDD'99)