Concept-Based Pages Recommendation by Using Cluster Algorithm

Chen-Chung Chi, Chin-Hwa Kuo, Ming-Yuan Lu, Nai-Lung Tsao
Can Lab., Dept of CSIE Tamkang University, Taiwan, R. O. C.
ryanjih@mail.tku.edu.tw, chkuo@mail.tku.edu.tw,694190637@s94.tku.edu.tw, beaktsao@mail2000.com.tw

Abstract

In this research, we used a proxy server to search for information related to the user's browsed webpages. From the records of the proxy server we constructed a profile of the user's browsing habits. At the end of the user's search subsystem, we will use content based concept to extract keywords to obtain the article's characteristics' description. From the recommendation system, the webpages will be classified using the hierarchical grouping method, and through collaborative filtering, the recommendation webpages will be chosen to provide further readings for students language learning.

Keywords: content filtering, collaborative filtering, recommendation system

1. Introduction

The most popular search engine, Google [1], uses a link and authority mechanism, to calculate the user's searched webpages with PageRank [2]. The information content with the highest relevance is recommended to the user.

The search engine's recommended webpages compares the user's search terms with the keywords in the article with the most content and highest number of links. This paper intends to exploit the idea of sharing to design a method different from common recommendation system; we use the concept of user-to-user recommendation system. Using a grouping method, the user can receive groups of high interest and other user's related browsing groups.

2. Related work

The expansion of a recommendation system can be consolidated into an independent type of research as there are many theories and research one in this topic. Since the item evaluation system is the main backbone of the system, in normal theories the calculation is done on webpages that have not been viewed by the user. The idea is to decrease the problems encountered when these evaluation scores are calculated.

2.1 Content-Based Filtering

Content-based filtering [3, 4] is to the function value of the user's past browsed webpages and score function, unviewed webpages then can be recommended to the user. the data is separated into Content(d) and ContentBasedProfile(c). From item j's content d_j we can use TFIDF to extract keywords such that:

For the user c, a vector can be defined for different keywords:

$$Content(dj) = < W_{1j}, W_{2j}, ..., W_{kj} >$$

From the two vectors above, we can obtain a function score such that:

$$ContentBasedProfile(c) = < W_{1j}, W_{2j}, ..., W_{kj} >$$

Then we can use TFIDF to get the respective vectors, and then in normal similarity value calculation, we can determine the recommendation list. For instance, we use cosine similarity measure:

$$u(c,s) = \cos ine(\overrightarrow{Wc}, \overrightarrow{Ws}) \bullet \frac{\overrightarrow{Wc} \bullet \overrightarrow{Ws}}{\left\|\overrightarrow{Wc}\right\|^2 \bullet \left\|\overrightarrow{Ws}\right\|^2} = \frac{\sum_{i=1}^{k} W_{i,c} W_{i,s}}{\sqrt{\sum_{i=1}^{k} W_{i,c}^2} \sqrt{\sum_{i=1}^{k} W_{i,s}^2}}$$

However, this application proves to be quite problematic. The computer is unable to process textual meaning automatically, which results in a decrease in the quality of recommendations. [5] points out that a user's browsed content is the main

978-0-7695-3167-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICALT.2008.214

basis for recommendations, but it is very difficult for the user to find the underlying meaning. The system is unable to determine the quality, style, and perspective of each item.

2.2 Collaborative Filtering

The main concept behind collaborative filtering [6, 7] is social filtering, which uses the item content as its basis. It decides the recommendation through the evaluation of the item by members of the group, then estimates the user c's evaluation score r(c,s) for the item s to make a decision.

$$r_{c,s} = \underset{c' \in \hat{C}}{aggr}\, r_{c's}$$

And $r_{c,s}$ can be determined as follows:

$$r_{c,s} = \frac{1}{N} \sum_{c' \in \hat{C}} r_{c',s}$$

$$r_{c,s} = k \sum_{c' \in \hat{C}} sim(c, c') \times r_{c',s}$$

$$r_{c,s} = \overline{r_c} + k \sum_{c' \in \hat{C}} sim(c, c') \times (r_{c',s} - r_{c'})$$

In the first equation, the item N represents the value of similarity between user c and user c'. Equation 2 calculates the similarity value and adds the weight calculation. The purpose of equation 3 is to solve the different evaluation criteria of each user, especially apparent when evaluating scope becomes greater. Therefore, some rules must be set for the evaluation criteria, to adjust the importance so that it fits the actual situation. In particular,

$$\overline{r_c} = \frac{1}{|S_c|} \sum_{S \in Sc} r_{c,s}$$

$S_c = \{ S \in S \mid r_{c,s} \neq 0 \}$

S_c defines the item number after evaluation

Some further approachs suggest an item-based collaborative filtering method and calculate the similarity between items to make recommendation. User-based collaborative filtering does not have the problems of being unable to differentiate between items and item type restrictions that content-based filtering encounters, yet it does have a problem with scalability. Although scholars have suggested that collaborative filtering method have solved this problem, but in general, collaborative-filtering method has a few more problems: *new user problem, new item problem, and sparsity*.

3. Concept-Based Webpage Recommendation System

This proposal uses a clustering algorithm to improve the user-based recommendation of the collaborative filtering. When it is about to recommend webpages that the user has not viewed yet, it first groups the webpages that the user has already viewed, then it focuses on the unviewed webpages to proceed with the recommendation score calculation.

3.1 User search subsystem

To accurately search through the user's browsing history, the system uses a proxy server, simultaneously record the user's actions. Also, to protect the user's privacy, the system will recapture the webpage's content.

There are different ways of grouping webpages on the Internet, depending on what perspective you are looking at. From a second language learner's point of view, using this recommendation system to search for a webpage, hopefully the webpage will have more content for reading and studying.

3.2 Article content preprocessing system

There are three processes in the preprocess subsystem:

(1) Elimination of punctuation marks and numbers.

(2) Target webpage content extraction.

(3) Keyword generation:

- *Eliminate stop word*
- *Combine synonyms or similar words*
- *Convert all text to lower case letters*
- *extract keywords by TF-IDF method.*

3.3 Article Evaluation Recommendation Subsystem

In this subsystem we use a clustering algorithm to determine the level of relativity of the recommended groups from the webpages, and then recommend these webpages to some users with the same interest for reading. We use the algorithm and user's interest level to produce a score. If the score is greater enough, then the webpage is recommended.

(1) Hierarchical K-means: It sets the number of webpages divided by 1000 as the k value.

(2) Recommendation mechanism: If an unviewed webpage is to be recommended to the user, it then calculates the number of times the webpage has been viewed in the cluster and the ratio of the webpages viewed by the user to the entire cluster.

$$R_{u,i} = \frac{\sum_{n \in Gj} \left| i \in P_{nj} \right|}{\sum_{n \in Gj} \left| P_{nj} \right|} * \frac{P_{uj}}{P_u}$$

4. System Evaluation

This research made use of the viewing history of the 30 days in ten computer labs, and after eliminating unreliable webpages, we get experiment shown in table 1:

Table 1. Cluster eliminates result

	Numbers of Webpages	Size of Clusters	Allocation Speed(minutes)
10 days	12355	12	2.12
20 days	23849	15	4.20
30 days	42633	19	7.60

The original target of size of cluster was number of webpages divided by 1000, but the empirical data shows that since many webpages were too similar, so the size of the cluster was far lower than the original expections. Also, once a large number of webpages undergoes clustering calculation, since there are more clustering group centers that need to be calculated, so the speed of allocation decrease greatly.

Comparing this system's recommendation speed and item-to-item we get result shown in table 2:

Table 2. Recommendation speed

	Concept-based time (sec.)	Time for item-to-item (sec.)
10 days	12	2.12
20 days	15	4.20
30 days	19	7.60

In table 2, while the increase of days, number of webpages, and the total recommendation time for concept-based webpage recommendation system increases as well, but less than the increase for a traditional item-to-item recommendation system. However, A concept-based webpage recommendation system requires more time than an item-to-item recommend- dation system.

5. Conclusion and Future work

This approach investigates the results [8], makes use of the advantages of content-based recommendation. uses the collaborative filtering method and hierarchical k-means clustering algorithm to produce the user-to-user recommendations, then recommend new information with improving the collaborative-filtering. In terms of recommendation speed, it improves the performance and obtain recommended webpage information immediately.

In the future, we hope this approach can be approved to increase its reliability in these ways:

(1) Automatic image annotation system
(2) Strengthen the search mechanism
(3) Language learning

Reference

[1] Google, http://www.google.com

[2] L. Pages, S. Brin, R. Motwani, & T. Winograd, "The pagerank citation ranking:Bringing order to the web". *Technical report*, Stanford University, CA, 1998.

[3] B. Chen, P. C. Tai, R. Harrison & Y. Pan, "Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis", *In Proceedings of Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts,* IEEE, 2005, pp. 105-108.

[4] M. Balabanovic and Y. Shoham, "Fab: Content-Based Collaborative Recommendation", *Communications of the ACM,* Vol. 40, No. 3, 1997, pp. 66-72.

[5] U. Shardanand, and P. Maes, "Social information filtering: Algorithms for automating 'word of mouth'", *In Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems,* pp. 210-217, 1995.

[6] L. Niu, X. W. Yan, C. Q. Zhang, & S. C. Zhang, "Product Hierarchy-Based Customer Profiles for Electronic Commerce Recommendations", *Proceedings of the First International Conference on Machine Learning and Cybernetics*, Beijing, China, 2002, pp. 1075-1080.

[7] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom. and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews, ", *Proceedings of the Computer Supported Collaborative Work Conference*, Chapel Hill, 1994, pp. 175-186.

[8] C. C. Chi, C. H. Kuo, and C. C. Peng, "The Designing of a Web Page Recommendation System for ESL", *In Proceedings of Advanced Learning Technologies,* 2007, pp. 730-734.