Design and Implementation of Chinese Information Filtering System

Xiumei Zhang
School of Software,
University of Science and Technology Liaoning,
Anshan, 114051, China
E-mail:aszxm2002@yahoo.com.cn

***Abstract*—With the development of the network information, information processing is vital in all aspects. Information filtering is more important research aspect. Especially Chinese information filtering is urgent affairs. According to research of the domestic and abroad, in the article vector space method and hyB+ tree index method is combined to filter text. Experimental results show that, this method is feasible.**

Keywords- information filtering;vector space method; hyB+ tree index

I. INTRODUCTION

Since ninety times, Internet expanding dramatically, it contains various types of original information, including text, voice information, image information and so on. How to find and obtain the information from the vast amounts of information quickly that becomes the primary issue. Information filtering is to meet this requirement. Information filtering is the basis of the information needs of users, searched user interested information in the dynamic flow of information, shielded useless information. A good information filtering system must meet the following three conditions: (1) Useful information is provided to the user effectively. (2) Useful information is provided to the user quickly enough. (3) The throughput of information is dealt with effectively. Here we are dealing with information of text form.

II. TEXT REPRESENTATION

At present, in the information processing, text representation mainly uses the vector space model (VSM)[1]. The main idea of Vector space model is that space document is viewed as a group of orthogonal entries, generating vector space. Each document D is normalized vector V(d)=($t_1, w_1(d); t_2, w_2(d);$ …;$t_n, w_n(d))$.In which, t_i is term, $w_i(d)$ is the weight of t_i in the D. To say, t_i is all terms or all phrases that is appeared in D. That can improve the accuracy of the content representation. $w_i(d)$ is the frequency function of t_i appearing in D. In the VSM, TF-IDF is a common term weighting method. The following formula is used as calculating the term weight.

$$W_{ik} = \frac{tf_{ik}\ \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \times \log^2(\frac{N}{n_k} + 0.01)}}$$

The N is the total number of documents for all the samples, n_k is the document frequency of term t_k, tf_{ik} is the occurrence frequency of t term t_k in the document.

III. FEATURE EXTRACTION

The space is a fixed delimiter of English sentence, while the Chinese has not, that brings great obstacles to Chinese information processing. For example, the computer was unable to distinguish the "bat bought "between “the bat is bought "and" ball, is auction ". Therefore, the document feature representation is necessary preprocessed. Text is dealt with here. 1.Find out stop words of the document reference to controlled thesaurus, that is some function words and prepositions. 2.Implement the document segmentation through word segmentation procedures. 3.The Feature term is extracted from the document.

The document feature representation is composed of two parts: one is feature vocabulary of a document is chosen, two is the weight of document feature in the document is calculated. By Zipf [2]law, in a document, any word frequency is multiplied by its weight serial number that is approximately equal to the constant, namely Frequency*Rank ≈ Constant. It concludes that the best performance of document feature is that the frequency of occurrence of moderate vocabulary.

IV. DESIGN FRAMEWORK OF INFORMATION FILTERING

Due to the added, deleted or modified at any time, namely collection literature is dynamic. In view of this phenomenon, we put forward a kind of new method, in order to more effectively solve the above problems. The method can use the space effective, find the information associated with user interest quickly.

The data structure is the most important for a good system. So to determine its data structure, the structure of this model is similar to the inverted index structure, but it has a pointer to the vector space file. For the concentration of

Published by Atlantis Press, Paris, France.

any literature, vector space is the vector for the preservation of a fixed number of entries, namely statistical information is stored.

A. Hyb+tree structure

The index is similar to the B+ tree structure, but that is not completely identical to the B+ tree, here it is called the hyB+[3] tree. N is its depth, its left subtree has the following properties: (1) In addition to the root node and the leaf nodes, $\lceil n/2 \rceil$ nodes at least, there are n children at most. (2) At least 2 root. (3) The path of any one node to the is the same except the leaf node. They store up n-1 term,T1, T2, ... , Tn-1 and N pointer P1, P2, ... , Pn (if i<j, then Ti<Tj) .

Leaf nodes are stored in a term file. The file is divided into the same space. Each leaf node L preserves a pointer and at most n-1 item information, pointer P points to the next file block. The information is composed by Four-dimensional array form (T,DF,F, L), T is a string type term, DF is document frequency that contains the T document number, F is a pointer to the first containing T documents, L is a pointer to the current containing T literature. The concrete structure is as follows:

P1	T1	P2	T2	…	Tn-1	Pn

T1	DF1	F1	L1	…	Tn-1	DFn-1	Fn-1	Ln-1	P

Figure 1. hyB+ tree structure

B. Vector space structure

It consists of the term vector of each document. These vectors are stored in a vector file, M terms are saved in each vector, M is a fixed parameter that is predefined. These terms are identified the literature and differenced from other literature, so there should be no high-frequency and low-frequency items appeared. First, a stoplist is used, and a parameter M is determined. Let us assume that the vector and the K1 correspondence. Its structure can use three-dimensional array representation (TFi, Bi, Ni). TFi is appearing times of items in the literature, Bi is a pointer, pointing at three-dimensional array of the K2 prior to insertion of the set of K1 (containing the same item Ti); Ni is a pointer, pointing at three-dimensional array of the K3 after insertion a set of K1 (containing the same item Ti). If K1 is the first inserted into the literature, Bi points to the hyB+ tree that has the same feature terms.

ID	TF1	N1	B1	TF2	N2	B2	…	TFm	Nm	BM

Figure 2. Vector space structure

C. User Template

Inverted index structure is applied to text filtering user template by David Goldberg [5] in text retrieval, namely that is the establishment of the index structure of the user template ,for a word x that forms an inverted list of all templates containing x , which is the mapping of the inversion table from a word to its corresponding memory location by hash table, called directory .General Setting in the inverted list has the two domains, the number of templates and the word weight of the template. In addition, including two arrays, one is used to represent a threshold value for each template, known as the threshold value table, another is used to represent the results of the current text and each template , which is a temporary work unit, called the score list.

The user template modification used Rocchio model iteration [6]. Rocchio model is as follows:

$$P_{k+1} = P_k + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} + \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2}$$

Which P_{k+1} is a new template, P_k is the old template, R_k vector representation of the underlying text, S_k is text vector of unrelated texts, n_1 is the number of related texts, n_2 is the number of unrelated texts, β and γ is the weighted of positive and negative feedback contribution rate.

D. The match of User Template and query

The user template for feature extraction based on keyword is $U = (u_1, u_2, ..., u_m)$, u_i is the weight of the user input query items t1,t2,…,tm.

The feature vector of the text T is $V = (v_1, v_2, ..., v_m)$, v_i the keyword weights.

The similarity is computed for User template and the text filtered vector ,calculated as inner product operation :

$$Sim(U,V) = \frac{\sum_{i=1}^{m} u_i v_i}{\sqrt{\sum_{i=1}^{m} u_i^2 \sum_{i=1}^{m} v_i^2}}$$

And similarity threshold θ is selected,if $Sim(U,V) > \theta$,it is considered the matching of text V and template U,finally the text is displayed to the user.

E. Key Algorithm

Algorithm of the problem is given:the insert and delete of hyb+tree and vector file.

Insert:

1 A space is assigned to the vector file.

2 ID of the current literature pluses 1.

3 The first position value of the document vector is saved.

4 For i=1 to M do

The item of Article I is inserted into the hyB+ tree.

If the item is found, goto the step 5; otherwise, goto the step 8;

5 if DF=0, goto step 6; otherwise, goto the step 7;

6 Let the pointer F[i] point to the I vector, let B[i] point to the location of the leaf node of the hyB+ tree.

7.B[i]=L, let L point to the I vector, DF+1, let pointer N[i] point to the location of the leaf node of the hyB+ tree.

8 The I term is inserted into the hyB+ tree, DF=1, and let B[i] and N[i] point to leaf nodes of the location of the hyB+ tree.

Delete:

1Integer variables F=0, the array P[M]of the M elements is saved

2 For g=1 to M do

Let P[g]=N[g], if P[g] points to a file, DF=DF-1, at the same time, F=F+1;

If P[g] points to a vector of three-dimensional array that is not in the cache, then the vector is assigned to buffer[g].

If the pointer doesn’t point to the hyB+ tree, let N[i] point to a three-dimensional array as B[j]=B[i]; otherwise, goto the step 3.

3 let L=B[i];

4 if the B[i] does not point to a hyB+ tree, let be B[i] points to a three-dimensional array of N[k]=N[i]; otherwise, the step 5;

5 Let L=N[i];

6 While (F<M), if P[g] points to a file, DF-1 and F+1; otherwise, goto the step 7.

7 If the P[g] points to a vector of three-dimensional array that is in the cache, let P[g]=N; otherwise, the vector is inserted into cache buffer[g];

Finally the vector file block is released .

F. *Example of Application*

Information retrieval commonly metrics is applied in Information filtering evaluation ,in such as precision, recall rate.The accuracy is P= $\frac{Nf}{N}$ (Nf is the number of documents filtered, N is the whole text document number).

The part of the experimental results is given about 500 texts from Chinese Internet network information data .

TABLE I. PART EXPERIMENTAL RESULTS

	1	2	3	4	5	Average accuracy
Accurate rate of new method of	0.4539	0.4460	0.4250	0.3970	0.4610	0.4366
Accuracy rate of vector space model	0.4112	0.3680	0.4100	0.3629	0.4215	0.3947

V. CONCLUDING REMARKS

At present,the common use of the vector space model, this paper uses vector space and hyB + tree to deal with the text filter. The purpose is to have certain improvements rather than take advantage of the vector space model, the test data shows that has3% -5% improvement, and the filtering speed has also been improved to some extent, to achieve the initial purpose. In the subsequent test its performance will be studied further.

REFERENCES

[1] G.Salton et al., A Vector Space Model for Automatic Indexing, Communications of the ACM ,Vol.18,No.1,1995

[2] Zipf,H..P.Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, Massachusetts, 1949.

[3] Ioannis Caragiannis, Alexandros Katsoulis. Paul Spirakis. Bsil Tampakas. Indexing and Retrieval in Large Dynamic Document Collections.1995.

[4] Douglas W.Oard, User Modeling for Information Filtering. http://www.ee. umd. edu/medlab/filter/papers/umir.html

[5] David Goldberg, et al., Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM ,Vol.35,No.12,61-70,1992

[6] J.J.Rocchio.Relevance feedback in information retrieval. In The SMART Retrieval System—Experiments in Automatic Document Processing, 313—323, Englewood Cliffs, NJ,1971.

[7] Steve Gant, A Sample Information Filter for the Web, http://ils.unc.edu /gants/report.html

[8] A. Moffat, J. Zobel, and S.T, Klein. Improved inverted file processing for large text databases. In R.Sacks-Davis and J. Zobel, editors, Proc.6th Australasion Database Conference, pages 162-171, Adelaide, January 1995.

[9] Tak.W.Yan and Hector Garcia-Molina, Index Structure for Information Filtering Under the Vector Space Model, In Proceeding of the Third International Conference on Parallel and Distributed Information system,1994,89-98.