Profile Based Information Retrieval

Athar Shaikh, Pravin Bhjantri, Shankar Pendse,V.K.Parvati

Department of Information Science and Engineering,
S.D.M.College of Engineering & Technology, Dharwad

Abstract-This paper present Profile Based information retrieval system(PBIR). This system provide the user to register with it and based on the users registered areas of interest the system searches the related and efficient information form the world wide web using the technique of web text mining and arranges the unstructured data into structured format and present it to the user. This system also stores the previously searched data and based on users areas of interest and rating awarded to the interest by the user his profile will be updated at particular scheduled time.

I. INTRODUCTION

In recent years, as a consequence of the emergence of the internet search meta-engines, the information acquisition process has provided several kinds of tools to collect information from the internet. However, the quality of the collected information is still a problem. In the internet, for instance, millions of users accesses search engine every day and they have different profile and accesses different types of URLs. Based on this, it has emerged the idea of developing webbased systems that are not only responsible for searching information, but also finding good quality information to the users. In this sense, a user would not spend hours accessing unimportant URLs and could extract URLs which are interesting for its preferences. In the literature, several web-based search information systems have been proposed, using some type of filter techniques. However, to improve the quality of these systems, it is expected that the required information has not been only retrieved, but also is has also to fulfill the personal needs of a user, through a user profile. Some systems involving user profile to collect information from the internet can be found in [3],[4]. Most of the existing systems use a search in all possible pages in order to detect important pages for the users. However, even the use of filter techniques may become impractical when accessing and checking a large amount of pages. In the proposed system, based on the user areas of interest related and efficient information is retrieved from the web server and is provided to user [1].

II. PROPOSED SYSTEM

Profile Based Information Retrieval (PBIR) is an application program which enables user to edit his profile and specify his area of interest and award rating to it. Our application provides the login support and lets him to edit his profile in which he will be specifying his area of interest. Our application scans the user profile and extracts the area of interest as the key word (this part of our application is referred to as Term Extractor). The key word is taken as input and the information related to that is searched first in the repository. If the information is found in the repository it will be organized in a proper format and dumped back to the profile. The next time user logs in to his account, he will find the information related to his area of interest which he has specified in his profile. If the information is not found in the repository then search will be carried out in web and the information is gathered and redirected to the user in proper format. The searched information is also stored in our repository so that any second user who enters the same area of interest can be served by searching in the repository itself rather than searching in the web again and again if. If the rating awarded by the other user is less than that of the actual quantity of data that is stored in the repository, it is processed and required quantity of data is fetched and is redirected back to the user in proper format. Other way round if the rating awarded by the user is higher than that of the actual quantity of data that is stored in the repository, the extra needed information is searched in the web and is merged with the existing information in the repository, processed and redirected to the user in the proper format. Our application also lets user for any instantaneous search by providing the search tab in the profile page.

III. FUNCTIONALITY

The major functions that the software performs are:

A. Keyword extracting

Keyword extractor, extracts the user areas of interest as the keywords and provide those keywords as input to the system search engine.

B. Searching for the information in the system data base and the internet

Using the keywords extracted by the Extractor Search Engine searches the related information in the system data base and if the search is successful than the user profile is provided with the retrieved information and if the search is not successful than the search engine performs the search over web and retrieves the information from the web and update the system data base and also the user profile.

C. Scheduling the update checker

Update checker is scheduled by the user itself according to his requirements. Update checker checks for the latest updates in the web server at the scheduled time and retrieves the updates if any and helps the system in keeping user profile updated with latest information and also system data base

D. Storing of the previously searched information

This function provide the previously searched information in the user profile and helps the user to get quick overview of his previously searched information.

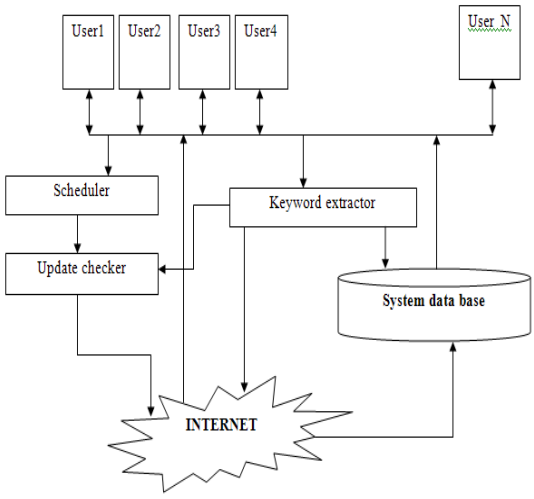

Fig 1 Block diagram for PBIR

IV. WEB TEXT MINING PROCESS

Compared with the traditional data and the data warehouse, the information on Web is semi structured and/or unstructured, dynamic state data[5]. So it is very difficult to directly carry on data mining on the Web page. The data on Web have to be through necessary data processing. The processing process of the typical Web mining includes four step showed as following figure 2:

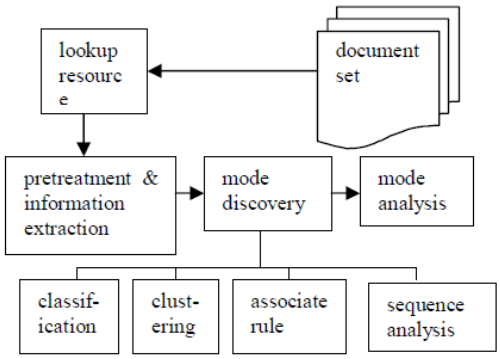

Fig 2 Block diagram for web text mining

1. To lookup resources: its task gets data from the target Web document. It is remarkable that information resources sometimes not only is limited by on-line Web document, but also include E-mail, electron document, newsgroup, perhaps the log data of website, even is the information in the bargain database which passes Web formation [2].

2. Pretreatment and information extraction: its task is to get rid of useless information and carry on information necessary sorting from the acquired Web resources. For example automatically clean advertisement conjunction, clean surplus format marking, automatically identify paragraph or field, get characteristic item, and form to data rules, some logic form or even relation table [2].

3. Mode discovery: it is automatic and can be carried on in the same one website or among several websites. It is the non-ordinary process in which we discover valid, novel, latent, available and comprehending knowledge including forms of concept, mode, rule, regulation, restriction and visualizing etc.. Its adoptive technique includes classification, clustering, associate rule and sequence analysis etc [2]..

4. Mode analysis: It verifies and explains the mode produced in top one step. It can be automatically completed by machine or be mutually completed among analytic personnel and machine [2].

V. WEB INFORMATION EXTRACTION

Information extraction is the task of finding specific pieces of information from unstructured or semi-structured document. Most of the web pages in Internet are HTML document or XML document. The document pretreatment initially need throw away irrelevant marking with text mining to the contents of web page by using web page information extraction module. Then web page information extraction module carries on quantization characteristic item that is extracted metadata. It describes document information in structure format. It converts unify a format of TXT text and save in a folder for latter processing. Text characteristic is divided into the description characteristic(text name, date, size, type...etc.) and the semantic characteristic(text author, title, organization, contents...etc.).

The model used by text characteristic has Boolean Logic Model, Vector Space Mode1 (VSM), Latent Semantic Indexing (LSI) and Probability Mode1 etc. Now we discuss Vector Space Model method which use is more and its effect is better in text mining system.

Vector Space Model was put forward in 60's by Salton. It is the earliest and also the most famous mathematics model in information extraction[6]. The pretreatment & information extraction mode discovery mode analysis classification clustering associate rule Figure 2. Web text mining common process lookup resource document set sequence analysis basic thought of Vector Space Model is to use Bag.of- Word to express document. There is a key hypothesis of this representation: the lemma appearing early or late sequence is unimportant in the article. Each characteristic item corresponds with a dimension of the characteristic space. Then a document is expressed as a vector, namely a point of characteristic space. Such as a document di is showed as following :

$$V(di)=(t1, wi1; \ldots; tk, wik; \ldots; tn, win) \quad (1)$$

Among them, tk is characteristic item or lemma, wik is weight value of tk in di. The weight value is usually a appearance frequency function of the characteristic item in the document. The weight value function is showed as:

$$Wik= tfk\ (di\ ).lg(N/Nk+0.5) \quad (2)$$

The tfk(di) denotes the appearance frequency of the characteristic tk in the document di. N is total number of the training document set. Nk is document number of appearance lemma tk in training document set. After document is disparted lemma by its program, lack contributive lemma to classification is taken out by using halt-use-lemma-list. It can also adopt strategy of characteristic lemma relativity analysis, clustering, thesaurus or approximate word merging etc.. It is expressed as text vector as formula (1) in the end. While using vector space method to express document, the dimension of text characteristic vector usually attains to count 100,000. Even through deleting halt-use-lemma by halt-use-lemma-list and deleting low frequency lemma applied ZIP rule, there are still tens of thousands dimension characteristics to be left. Finally, it general choice certain amount of the best characteristic to carry on text mining. So further carrying on characteristic to reduce seem to be exceptional importance. Usually, the choice of characteristic subset is to construct a characteristic valuation function, to evaluate each characteristic in characteristic set, to acquire a valuation score for each characteristic,

to carry on compositor all characteristics by valuation score, to choice the best characteristic of scheduled number as the characteristic subset. The valuation function of the text characteristic choice extends from the information theory. It is used for getting valuation score to each characteristic lemma. The valuation score need nicely reflect related degree between lemma and of every sort. There are common valuation function: information gain, expected cross entropy, mutual information, the weight of evidence for text, word frequency etc..

For example, a word frequency matrix which expresses word frequency of a document is shown as following table 1. Among them, row is corresponding with characteristic item t, column is corresponding with document d, the vector value reflects the related degree between characteristic item t and document d, so each document is regarded as pace vector V. Table 1. a word frequency matrix

TABLE 1
A WORD FREQUENCY MATRIX

	d1	d2	d3	d4	d5	d6
t1	305	80	40	75	18	310
t2	30	145	75	202	17	325
t3	26	35	165	50	220	360
t4	381	90	75	58	14	25
t5	322	85	35	69	15	315

VI. WEB TEXT CLASSIFICATION ALGORITHM

As for the vector space model (VSM) is adopted in the algorithm. The similitude degree method of literature search technique is adopted in the system to classification mining namely carry on characteristic vector match. Suppose that the sample information is U, needed to be classified information is V, cosine of vector angle can be used to measure both of the similitude degree, it is shown as formula (3).

Sim(V,U) = cos (V ,U)=

$$\sum_{k=1}^{n}(W_{vk} * W_{uk}) \Big/ \sqrt{\sum_{k=1}^{n} W_{vk}^{2} \sum_{k=1}^{n} W_{uk}^{2}} \qquad (3)$$

Text classification is a kind of typical model directive machine learing problem. It is generally divided into training and categorizing two stages. Its training process has already come to decide the classification ability that the system have and this classification ability is fixedly constant in the classification process in future. The great majorities of current text classification system don't have ability of continuous study . Owing to the problem above existed, this paper puts forward a new algorithm that can carry on a feedback processing to classification result. The new algorithm joins the process of feedback on The traditional foundation frame "training →Categorizing " algorithm. It expands the algorithm process as "Training → Categorizing → feedback judgment → feedback". This kind of method is more close the real meaning machine learning. It makes the algorithm has certain degree cognition self- determination. Its concrete algorithm is described as follows:

Training stage:

(1) C={c 1, c 2, ……, cn} // Define the category set

(2) S ={s 1, s 2, ……, sm} // Give training text set

For i =1 to m

Training text si is marked as the sign cj that is belonged to category V(si) characteristic vector of si

Endfor

(3) For j =1 to n

Cj[wj1,wj2,…,wjk]centroidal characteristic vector is representative of each category Cj by characteristic vector of all training text belonged to category cj

Endfor

Categorizing stage:

(4)Threshold[1…n] threshold of information similitude degree for each category D[w1,w2,…,wk] characteristic vector of new text D to wait for classification

(5) For j=1 to n Do

Sim information similitude degree between D[w1,w2,…,wk] and Cj[wj1,wj2,…,wjk] If sim Threshold[j] Then // Categorizing and feedback judgment Add D, D[w1,w2,…,wk] and sim to classification form of corresponding category cj // Categorize
(6) Query about characteristic vector Cj [wj1, wj2,…, wjk] of category Cj and its characteristic item number K
For i=1 to k Do

$$wji' = \frac{n * wji + wi}{n + 1}$$

Endfor
Cj[wj1,wj2,…,wjk] Cj[wj1',wj2',…,wjk'] // feedback
Endif
Endfor[2]

VII. CONCLUSION

In recent years, as a consequence of the emergence of the internet search meta-engines, the information acquisition process has provided several kinds of tools to collect information from the internet. However, the quality of the collected information is still a problem. But through this paper we provide a efficient way of retrieving the related and efficient information from web using the web mining techniques. This system also helps its user to get updated regularly and stay tuned with efficient knowledge.

REFERENCES

[1] Carcara: A Multi-agent System for Web Mining using Adjustable User Profile and Dynamic grouping Manuel F Gomes Junior and Anne Magaly Canuto
[2] Research and Realization of Text Mining Algorithm on Web1 Shiqun Yin Yuhui Qiu Jike Ge
[3] F Menczer. Complementing search engines with online web mining agents. Decision Suport Systems, 35(2):195-212, 2003.
[4] H. Chen, A.L Houston, R. R Sewell and B.R Schatz. Internet browsing and searching: User evaluations of category map and concept space techniques. Jounal of the American Society for Information Science, 49(7):582-603, 1998.
[5] Soumen Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002
[6] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. "Fast discovery of association rules." Advance in knowledge discovery and data mining, AAAI Press/The MIT Press, 2004, pp. 307-328