On Multi Class Vector Space Model-based Information Retrieval

Gokul L. Patil
M tech Scholar, Central Indian Institute of Technology, Indore.

Arif Khan
Assistant Professor, Central Indian Institute of Technology, Indore.

Deepak Khulare
Assistant Professor and Head, Central Indian Institute of Technology, Indore.

ABSTRACT

A new model of information retrieval algorithms, multi class vector space model, is proposed in this paper based on traditional vector space model. Web document has semi structured characteristic. The keyword or terms that are used for indexing purpose in any location, so content of this location represent important information in the web documents. Vector space model ignores the importance of these terms with respect to their position while calculating the weight of the indexing terms. The experimental result shows that this method can further improve the performance of vector space model, save storage space and speed up the retrieval speed with high precision and recall rate.

General Terms

Web information retrieval,stemming,Extraction.

Keywords

Web text mining; Text Classification; Characteristic Vector; Similitude Degree, Vector Space Model

1. INTRODUCTION

The web has become the largest available repository of data. Therefore it becomes more and more difficult for people to acquire needed data accuracy and quickly nowadays. People usually use the search engine yahoo, Google etc. To browse the information mainly but search engine involves so wide range whose intelligence level is low. Search engine focus on recall ratio instead of precision ratio so it is very difficult to mine data further. There is lot of similarity between search engine and web based information gathering. Search engine mainly focus on three parts web scraper, index database & search service but web information extraction mostly aim at a certain and concrete profession

Many different retrieval models have been proposed, studied and empirically validated. Web information retrieval models are mainly two categorized in two type, Traditional or classic information model and modern web information retrieval model.

In classic information retrieval model, the documents are typically transformed into a suitable representation to make the retrieval efficient. Classic information model aim to rank the documents based on the content of the collection.[7]

Modern web information retrieval exploits the link structure of the Web and log information of web. The links provide a positive critical assessment of a Web page's content which originates from outside of the control of the web page's author. The hyper link structure is exploited by two of the most frequently used Web information retrieval methods HITS (Hypertext Induced Topic Search) and Google Page Rank algorithm [7].

In this paper, Information retrieval method is put forward based on the vector space model [9]. This model is referred as multi class vector space model. The multi class vector space model applied for information retrieval which is better suited for dynamic environment. The theoretical analysis and experimental result shows that suggested method improves the performance.

2. LITERATURE SURVEY

Though Web information extraction and classification is face to particular domain, but its principle and processes are so resemblant. Therefore, in this section we will design a system model of Web information extraction and classification. The system can be divided into four major modules according to the dissimilarity of functions: data pretreatment module, information extraction module, data classification module and information output module.

Data pretreatment module

Data pretreatment plays major role in the information extraction process. With the data pretreatment done better, data quality is higher, the process of information extraction will be more and more valid and applicable, and educing result can also be more successful. As types of data source are so many, the characteristic attribute of various data can not always satisfy the demand of topic, so main function of the data pretreatment is to define data source, to format data source and filter initially data source on Web. The module needs that the data and its type of structure, half-structure and non-structure in the Web page will be reflected to object database

Information extraction module

Information extraction module is responsible for further collecting, processing, analyzing, comparing, sampling, storing whole in warehouse to making raise of native data. It can consider extraction information by setting up a Web page model, or using methods such as probability statistics, database The Web page mainly constitutes with two fractions of Tag label, links and display content. By creating Web page model, information extraction module resolve Tag label, construct the label tree of the Web page. Then it can analyze the structure of display content. After getting the Web page structure, it carries reserving or deleting of the data by taking content blocks as a unit. Getting data finally need to carry

eliminating repeat work before putting in database and creating index.

Data classification module

Data classification module adopts classification technique with content of web pages to classify and organize organically the information by category, and then information can be used hereafter. It can consider classifying data by using methods such as decision-making tree, KNN algorithm (the K-Nearest Neighbor) namely K the closest neighbors algorithm, SVM algorithm (Support Vector Machine), VSM algorithm (Vector Space Model), Bayes algorithm, NN (Nerve Network) algorithm etc.

1) **Decision-making tree***:*

Decision-making tree's induce is a classical classification algorithm. It adopts the way from top to bottom and the method called defeat in detail to construct the decision-making tree. It uses the information gain measurement to select test attribute on each node of the tree. Then it can pick up rule from the born decision-making tree.

2) **VSM algorithm***:*

VSM algorithm, what is also named Vector Space Model method, was put forward in 60's by Salton. As the earliest and also the most famous mathematics model in information search, its basic thought mean the text to adding authority characteristic vector:

D=D(T ,W ;T ,W ; ;T ,W)

Then it confirms categories of samples that need to be divided by method of computing text similitude degree. When the text is denoted as a model of space vector, the similitude degree of text can be denoted recurring to inner accumulate among the characteristic vector. In practical application, VSM algorithm commonly establishes category vector space according to the training sample in the database and the sort system in advance. As carrying classification to a sample waiting to sort, it only need to compute the similitude degree what is namely inner accumulate to the sample waiting to sort with each category vector. Then we can select similitude degree the biggest category act as the corresponding category of the sample waiting to sort. As computing space vector of the category in advance in the VSM method is needed, the establishment of the space vector very great degree depends on the characteristic items which are included in the category vector. According to the research discover, there are more excessive non zero characteristic items which are included in the category, the expression capability to the category in inclusive each characteristic item is more weak. Therefore compared with other classification method, VSM algorithm is fitter for the classification of the professional literature

Information output module

The data output module presents data to consumers after processing data in the object database. The module belongs to follow-up work in process of grab at data. The module's responsible for degree can depend on the need of consumers. The basic function is turning data with structure mode, and then presenting to the consumer. Otherwise, it can also increase statistics function such as report forms, diagram mark etc. When amount of data attains certain degree, it can set up data model, carry time sequence analysis and relativity analysis, discover mode or relation between each the concept and the rule. Then it makes the biggest efficiency use of data.

2.1 Extraction Process

Text extraction process is most decisive to the Web text extraction. The customer is required by the system to fill in information parameters, which describe the whole process of text extraction. The text extraction process flow table is showed as table 1.

Step	Title	Illuminate
1st	Basic set-up	To put the appointed website.
2nd	List setup	To classify list and set up pagination
3rd	Link setup	To peel off each link
4th	Text setup	To take out text's title, text, author, time etc.
5th	test	To extract one information
6th	Attribute setup	To setup attribute

TABLE I *Extraction Process*

3. MULTI CLASS VECTOR SPACE MODEL

In vector space model, use tf-idf weight as a statistical measure to evaluate how important a term is to a document in a collection. The importance increases proportionally to the number of times a term appears in the document but is offset by the frequency of the word in the collection. The baseline method for computing the weight of a term in a document is to count the number of times the term occurs in the document. This is referred as term frequency (tf). Term frequency does not exploit structural information present in the web page. For exploiting web page structure,terms frequency as well as position of term is also considered. This is referred as feature frequency. The term, that appearing in the special locations represents more important information in the web document. Algorithm for classification (classifier algorithm) is as follows. Algorithm can divided in two stages. First stage is training stage and second stage is categorizing stage.

Let D = { D1, D2, D3 , ... Dn } be the Document set

Training stage:

(1) Define the category set

C= {C1, C2,….CN}.4 category set are defined.

Eg. C= (Business, National, Education, Sports)

(2) Information classification tree is created. Classification tree is implemented using parent-child.

(3) Give training text set Pr-processed data is given as input.

Eg. S= (Java, Computer, Program, Language, sinu, doctor, patient, treatm)

For i =1 to m

Training text si is marked as 0 or 1

End for

m stands for total no of words in dataset

Categorizing stage:

(4)For each term in training text set weight is calculated.

For i=1 to n Do

Deg: Similarity values i.e. cosine which means similitude degree between Info and each category is calculated. Similitude degree between category and info decides the category of news. Threshold value is decided as 0.08. Condition for categorizing the news depending on threshold and coding is as follows

```
double thresholdValue=0.080000; // set the threshold value
    if ((doc1Degree > doc2Degree) && (doc1Degree > doc3Degree) && (doc1Degree > doc4Degree))
    {
    simDegreeDiff=doc1Degree-thresholdValue;
    System.out.println("Threshold Value is =>"+simDegreeDiff);
            if(simDegreeDiff<=doc2Degree)
            {
            News belongs to 2nd and 1st document
             }
    else if(simDegreeDiff<=doc3Degree)
            {
            News belongs to 3rd and 1st document
            }
    else if(simDegreeDiff<=doc4Degree)
            {
            News belongs to 4th and 1st document
            }
    else
            {
             News belongs to only 1st category.
            // set news number and document number to it for further use
            }
}
```

(5) If degree is matching with more than one category i.e. particular news is categorized in two categories.

(6) *After each class have two sub categories. For getting subcategory again the procedure applied for main classification is used i.e. by using similitude degree it is classified.*

4. IMPLEMENTATION

While implementing multi class vector space model,stemming is used. A stemming is a process of linguistic normalization, in which the variant forms of a word are reduced to a root form Apply Porter's Stemming algorithm[11] to remove morphological variants of all terms

4.1 Portor Algorithm

The algorithm works in following steps[5].

Step 1

• Remove "es" from words that end in "sses" or "ies"– passes --> pass, cries --> cri.

• Remove "s" from words whose next to last letter is not an "s"– runs --> run, fuss --> fuss.

• If word has a vowel and ends with "eed" remove the "ed"– agreed -->agre, freed -->freed .

• Remove "ed" and "ing" from words that have no other vowel– dreaded --> dread, red --> red, bothering --> bother, bring --> bring.

• Add "e" is word has a vowel and ends with "ated" or "bled"– enabled --> enable, generated --> generate.

• Replace trailing "y" with an "I" if word has a vowel–ex. satisfy -->satisfi, fly --> fly.

Step 2

With what is left, replace any suffix on the left with suffixon the right ex.-tionaltion conditional --> condition.

Step 3

With what is left, replace any suffix on the left with suffix on the right ex.-icateic fabricate --> fabric.

Step 4

Remove remaining standard suffixes al, ance, ence, er, ic,able, ible, ant, ement, ment, ent, sion, tion, ou, ism, ate, iti, ous, ive,ize, ise.

Step 5

Remove trailing "e" if word does not end in a vowel ex-hinge-->hing

4.2 Example

The vector space model procedure can be divided in three stages.

- The first stage is the document indexing where content bearing terms are extracted from the document text.
- The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user.
- The last stage ranks the document with respect to the query according to a similarity measure.

Salton's classic weighting is given by the following

$$W_i = tf_i * \log(D/df_i)$$

tfi = term frequency (term counts) or number of times a term i occurs in a document.

dfi = document frequency or number of documents containing term i .

D = number of documents in the database.

For getting clear idea of similitude degree one example is explained

Suppose we query an IR system for the query "gold silver truck". The database collection consists of three documents (D = 3) with the following content.

- D1:"Shipment of gold damaged in a fire"
- D2:"Delivery of silver arrived in a silver truck"
- D3: "Shipment of gold arrived in a truck"
- Q, D1, D2 and D3 are query, document1, document2 & document 3 respectively.
- In the respective columns it specifies number of occurrences of particular word.
- dfi specifies total number of occurrences of particular words in either D1 ,D2 & D3.
- D/ dfi is total number of documents divided by dfi .IDfi is the inverse log of D/dfi.
- Weights of each Q, D1, D2& D3 is calculated by the product of counts and IDfi.

As $|D_i| = \sqrt{\sum W^2_{ij}}$

$|D1| = \sqrt{0.4771^2 + 0.4771^2 + 0.1761^2} = \sqrt{0.5173}$ =0.7192

$|D2| = \sqrt{0.4771^2 + 0.9542^2 + 0.1761^2 + 0.1761^2} = \sqrt{1.2001}$ =1.0955

$|D3| = \sqrt{0.1761^2 + 0.1761^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.1240}$ =0.3522

And $|Q| = \sqrt{\sum W^2_{Qj}}$

$|Q| = \sqrt{0.1761^2 + 0.1761^2 + 0.4771^2} = \sqrt{0.2896}$ =0.5382

Next we compute all dot products (zero terms are ignored)

$Q.D_i = \sum(W_{Qj} * W_{ij})$

$Q.D_1 = 0.1761*0.1761 = 0.0310$

$Q.D_2 = 0.1761*0.1761 + 0.4771*0.9542 = 0.4862$

$Q.D_3 = 0.1761*0.1761 + 0.1761*0.1761 = 0.0620$

Now we will calculate the similarity values.

$\text{Cosine}\theta_{D1} = \frac{Q.D1}{[|Q|*|D1|]}$ = =0.0801

$\text{Cosine}\theta_{D2} = \frac{Q.D2}{[|Q|*|D2|]} = \frac{0.4862}{[0.5382*1.0955]}$

=0.8246

$\text{Cosine}\theta_{D3} = \frac{Q.D3}{[|Q|*|D3|]} = \frac{0.0620}{[0.5382*0.3522]}$

=0.3271

Since $\text{Cosine}\theta_{Di} = \text{sim}(Q, D_i)$

And $\text{sim}(Q, D_i) = \frac{\sum_i W_{Q,j} W_{i,j}}{\sqrt{\sum_j W^2_{Q,j}} \sqrt{\sum_i W^2_{i,j}}}$

Finally sorting and ranking of the documents is done in descending order according to the similarity values.

Rank 1: Doc 2 = 0.8246

Rank 2: Doc 3 = 0.3271

Rank 3: Doc 1 = 0.0801

Thus given query is closer to Doc 2 as a similarity value of Doc 2 is higher [8].

5. CONCLUSION

There is a large quantity of various information on the

Internet. However, it is quite difficult for us to search and use the information because of two reasons. One is due to the large quantity, variety and disorder of the information; another is the different structure of information source. It has mainly discussed some key algorithms about text classification. The whole process is based on parameters and configurations of management control, to provide graphics interface edit or the guide operation. Due to its excellent result of the collection, it can be used to create knowledge database, and it can also be treated as a small scaled vertical search system. The system will provide the exact information to the final consumer. It turns extensive information from non-preface to a preface in the process of classification during the process so that the system is very adaptable and convenient to make the information more generally and easily used by the consumers, and it can provide different personalized services according to different consumers' requirements. It is different from search engine and intelligence technique. In addition, there are some specific requirements for customers who should be professional and familiar with the HTML language to operating and maintaining the system.

6. REFERENCES

[1] Shiqun Yin Gang Wang Yuhui Qiu Weiqun Zhang. " Research and Implement of Classification Algorithm on Web Text Mining". IEEE.(2007)446-449 .

[2] M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone "A Web Text Mining Flexible Architecture". World Academy of Science, Engineering and Technology 32 2007 .

[3] Catarina Silva, Bernardete Ribeiro "Margin-based Active Learning and Background Knowledge in Text Mining".Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)IEEE.

[4]. Weiguo Fan1, Linda Wallace, Stephanie Rich, Zhongju Zhang "Tapping into the Power of Text Mining".

[7] Yin Yuhui Qiu Jike Ge, Xiaohong Lan."Research and Realization of Extraction Algorithm on Web Text Mining". (2007)278-281. Workshop on Intelligent Information Tech nology Application

[9] Micah J. Crowsey, Amanda R. Ramstad, David H. Gutierrez, Gregory W. Paladino, and K. P. White, consultancy. "An evaluation of unstructured Text Mining software" IEEE

[10] Shiquin Yin Yuhui Qiu ,Chengwen Zhong Jifu Zhou. "Study of Web Information extraction and Classification.Method".IEEETransaction(2007)5548-5552 .