A Refined TF-IDF Algorithm Based on Channel Distribution Information for Web News Feature Extraction

Mingmin Xu
Computer Science and Technology Department
East China Normal University
Shanghai, China
mmxu@ica.stc.sh.cn

Liang He*
Computer Science and Technology Department
East China Normal University
Shanghai, China
lhe@cs.ecnu.edu.cn

Xin Lin
Computer Science and Technology Department
East China Normal University
Shanghai, China
xlin@ cs.ecnu.edu.cn

Abstract—**TF-IDF algorithm is widely used in text feature extraction, in which IDF value demonstrates the importance of a term. While applying to the procession of web news, the traditional IDF doesn't work well, especially in a collection divided according to channels. In order to solve this problem, a refined IDF schema is proposed, named Channel Distribution Information (CDI) IDF, which is based on the information among the IDF values of each channel collections. According to the statistical features, the Top terms and the meaningless terms could be identified. Experiments on a manual labeled test set indicated that, related to the traditional TF-IDF, the CDI TF-IDF increases the Recall, Precise and F0.5 measure by 2.71%, 3.07% and 3.00%.**

Keywords- feature extraction; TF-IDF; channel distribution information

I. INTRODUCTION

There are many news domains, and online versions of newspaper offices and TV stations appearing on the Internet. The ability to access, organize and think critically about the information is becoming both more important and more difficult. Often a search engine is used, both the customized and the universal ones. Besides that, many methods have been applied to analyze the mass news, including news classification, topic detection and special event generation. Consequently, the way to represent a news page with proper features is the prerequisite.

In a common information processing system, two factors are used to weight a term [1]: 1) TF, the frequency of the term in the text segment, and 2) IDF, which is used to indicate the distinction of the term. This results in larger weights for terms that appear more frequency, and larger weights for the unusual terms. Several terms with the largest weight are returned as the features.

Inversed Document Frequency (IDF) [2] is a measurement of the general importance of a term. It's based on the statistical conclusion that, the importance is offset by the frequency of the term in the collection, and it's an index reduction. Suppose we have a text set with one page contains the phrase "the brown cow". It appears in the IDF that, the term "the" is not good feature because it's too commonly used, and useless to indicate the page.

Since the IDF measure is based on a very large collection in every aspect, it's laborious to generate and maintain. Often, it's get from a much smaller set of one or several aspects. Taking the news processing for instance, even though we can collect enough materials, some problems will accrue.

In a news page collection, the frequency doesn't reflect the importance very well. The top word, as a special character, makes it unreliable to treat IDF value as the measurement of importance. Often, the top word is a phrase of famous people or an important event, which receives extensive concern. Nearly there are many top words in each aspect. For example, in the sports news, terms related to leagues, teams and stars will frequently appear, so are terms about countries, organizations and politicians in current-affair news. They may have a high proportion, while it doesn't mean those terms are meaningless. But it's not easy to draw the boundary between the top words and the meaningless ones. If the traditional TF-IDF algorithm is applied, most likely, these top words will be lost.

The pages of a news domain are usually well cataloged. Generally, most domains share a similar structure, we call it channel. The channel IDF (which is an IDF value calculated in the collection from one particular channel) is taking into account in [3]. Unfortunately, things will become even worse. Simply, when the range is reduced, the frequency of top words increases. Those top words will have much lower IDF values, and they're more probable to be dropped.

We exploit the term's distribution between different news channels, and then, propose an algorithm to refine the IDF value. The text collection is got from Sogou.COM, and it's divided into 14 channels according to the catalog of SOHU.COM. At first, the IDF value of terms is calculated in every channel-limited collection. Then, the distribution parameters of a term in different channels are computed. Finally, the original value is refined to get a more reasonable measurement. A series of experiments has been done to

This work is supported by National Natural Science Foundation of China (60903169), National Science and Technology Ministry of China (2007BAH09B04), Significant Scientific and Technological Project of Shanghai Science and Technology Commission (08DZ15001(10)), and Key Scientific and Technological Project of Shanghai Science and Technology Commission(08511500303)

978-0-7695-3987-4/10 $26.00 © 2010 IEEE
DOI 10.1109/ETCS.2010.130

compare the efficiency. Experimental results indicate that, the proposed approach could recognize the top words and the meaningless ones, and achieve a higher accuracy of feature extraction than the traditional TF-IDF algorithm.

The rest of this paper is organized as follows: In Section 2 some related works are surveyed. In Section 3 detailed discussion of the approach is given. In Section 4 the efficiency evaluation is shown. Finally, we conclude the paper with a short summary and directions for future work in Section 5.

II. RELATED WORKS

Applied to the web, many specific features of web pages are taken into account to improve the result performance, such as, hyper-link structure, HTML labels, DOM Structure, user search log and some other features. A web page is a semi-constructed text document. It can be converted into a DOM tree which reflects the nodes hierarchical instruction. And a web page can be divided into two parts, the head and the body. In the head part, there is some Meta information, which is usually un-visible, while there are many anchors jump to a related page. Often, we emphasize some important words and phrases using special HTML labels like <b>, <i>, <h1>, or highlighting it. In the following part, some related works about feature extraction for web pages are described.

In [4], Kazunari et al. proposed a refined TF-IDF scheme which exploits the hyper-links of neighboring pages. In that paper, three models were introduced to refine the terms, with contents from both backward direction links and forward direction links, up to several levels. In [5], Zhang et al. presented a primary feature model for term weighting. According to their work, all features in a page space are divided into two parts. One is named primary feature space containing phrases with a specific HTML label. The others are classified into the sub feature space. The two parts vary in the weighting coefficient. While, instead of the traditional IDF factor, document frequency (DF) is taken into account. In [6] Wen-tau Yih et al. described a system that learned how to extract keywords from web pages for advertisement targeting. In the same system, some new features of web pages is introduced, such as the meta section features in the header of an HTML document and user search log, The user search logs show the overall behaviors of users. Henzinger et al. [7] discussed finding news articles on the web which are relevant to news currently being broadcast topics. The tf-IDF2 algorithm is used to weight a term. The motivation to use IDF2 is that rare words, like named entities are particularly important for issuing focused queries. Thus, the IDF component is more important than DF. Also, they tried stemming to aggregate the weight of terms that describe the same entity.

III. PROPOSED METHOD

In this paper, we propose a new algorithm CDI IDF (Channel Distribution Information IDF schema) to refine the IDF measurement for a term based on the characters of channel-IDF values. The news is divided by its catalog structure, and the focus points of different channels are not the same. Then it's easier to distinguish the tops words from the meaningless ones with the help of statistical method. After that, an operator is added to the traditional IDF value, to represent the importance of a term better.

A. IDF values in multi-channels

First of all, a large news page collection divided by channels is needed. A directive crawler is helpful to collect news pages from the Internet. A directive crawler is similar to the common ones, but it only fetches pages from one or several limited domains. Usually, it's better to construct the collection from multi-domains in order to get enough materials. And multi-resources will reduce the ill-effect of the expression habit. Generally, news published at a domain has its own format details which make the words frequency unreliable. When a page is fetched, there are too many noises in the original file. Some pre-processing works should be applied, such as title extraction and content extraction. After that, it is partitioned into different channels. Generally, each news domain has its catalog structures, and even more, these catalogs are very similar. With the help of manual interaction, a public catalog structure could be generated, and a mapping relationship is build up with the news domain.

For a term T, the IDF value in multi-channels is a set of all the channel-IDF values. The definition of each channel-IDF value is very similar to the traditional IDF. Suppose there is a collection with N channels, the IDF values of term T could be represented as follows:

$$IDF(T) = \{IDF_{C_i}(T) | C_i \in C\} = \left\{ \log_2 \frac{|D_i|}{|1 + \{d_i : T \in d_i\}|} \right\} \quad (1)$$

Where C is the whole collection, and C_i is the collection of channel i, D_i is the number of pages of C_i, $\{d_i : T \in d_i\}$ is the number of pages where the term T appears. If the term is not in the collection, this will lead to a division-by-zero. So we use $1 + \{d_i : T \in d_i\}$ here.

B. Channel Distribution Information (CDI)

After all the terms' IDF values in multi-channels have been calculated, we will have a detail analysis about their distribution among channels. We comprehensively considered every factor of web news. It's assumed that:

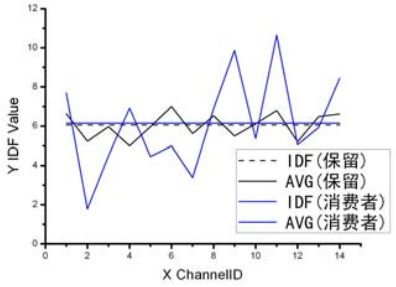
- A top word is channel-related. To be specific, for any term T, there may be one or several channels, among which, T is frequently mentioned. But it's nearly impossible for a term to be mentioned frequently in every channel. Otherwise, the term T is more likely to be a noise word;
- The frequencies of terms in news collections don't coincide well with the trend of IDF. Regularly, there are many meaningless terms that may have a high channel-IDF value because of the editors' expression custom. They may be used to expressing with other phrases, or they may alternative use any one from a synset as the art of rhetoric.

Based on these findings, the standard deviation is used to describe the amount of information for a term. Giving a set of channel-IDF values of term T, the standard deviation is expressed as

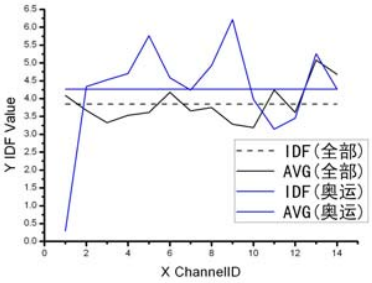
$$\sigma(T) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(IDF_{C_i}(T) - \mu)^2}, \quad \mu = \frac{1}{n}\sum_{i=1}^{n} IDF_{C_i}(T), C_i \in C \tag{2}$$

Where μ is the average IDF value of the set, not the IDF value of term T in the whole collection.

As shown in Figure One (a) and Figure One (b), term “全部”(total) and “奥运”(Olympic) both have lower channel-IDF values, and term “保留”(reserve) and “消费者”(customer) have higher channel-IDF values. The term “全部” and “保留” are meaningless terms. Abnormally, the channel-IDF values, the average and the global one of term “保留” are not low. On the other hand, the term “奥运” is a non-noisy term; it has a low IDF value especially in some channels.



(a)



(b)

Fig. 1 sample terms of top words and meaningless words. *These data is from Experiments in Section 4

In order to describe the importance of a term, the standard deviation is normalized, divided by its average channel-IDF value. The ratio is proposed as follows:

$$SD\_Rate(T) = \sigma(T) \Big/ AVG(T) \tag{3}$$

For the terms presented in Figure One, the rates are in Table One. It can be seen from that, the importance increases along with its ratio.

Table 1. the rate of standard deviation and average IDF value for terms in Figure One

Term	SD_Rate
保留(reserve)	0.274468467
全部(total)	0.360548305
奥运(Olympic)	0.838805933
消费者(customer)	1.019760522

C. *Refined IDF*

Giving a term T, the importance cannot specify the difference between channel-IDF values in different channels. Seeing the term “奥运” in Figure One, it's not a good way to refine all the channel-IDF values in the same way. Intuitively, in the channels that are much related to the top word, the channel-IDF values should be increased more than the others. And for those meaningless terms, it's better to decrease the channel-IDF values, despite of the original ones.

In order to locate the differences of IDF values between the particular channel and the other ones; we made a comparison between the channel-IDF value sets, while one includes the particular channel and the other doesn't. Without the particular channel, the normalization rate is expressed as follow:

$$SD\_Rate_i(T) = \sigma_i(T) \Big/ AVG_i(T) \tag{4}$$

Where $\sigma_i(T)$ is the standard deviation, and $AVG_i(T)$ is the average IDF value without IDF value in channel i.

And the ratio, dividing SD_Rate by SD_Rate_i, is used to identify the correction for each channel i. For the set of channel-IDF values, the adjustment method is expressed as follow:

$$IDF_i'(T) = IDF_i(T) \cdot \left[\frac{SD\_Rate(T)}{SD\_Rate_i(T)}\right]^2 \tag{5}$$

The adjusted channel-IDF value could reflect the importance of the term in an individual channel collection. That is, for a top word with a high standard deviation, all the channel-IDF values will be increased; and among the multi-channels IDF values, the one far away from the mean will be assigned with a much higher value. Conversely, the meaningless terms which are centralized distributed will be depressed.

D. *The Algorithm of CDI TF-IDF*

The procedure of CDI-TF-IDF is presented as follows:

Input : Collection C, Page P, rate

Output : Features

1 FOR $C_i \leftarrow C$

2 $N(C_i) \leftarrow number(C_i)$

//get the number of pages in channel i

3 FOR term $\leftarrow$ Dict

4 $IDF_i(T) \leftarrow \log 2\left(N(C_i) \Big/ number(term, C_i)\right)$

//calculate channel IDF

5 FOR term $\leftarrow$ Dict

6 FOR $C_i \leftarrow C$

7 $ID_{Fi}'(T) \leftarrow IDF_i(T) \cdot \left(SD\_Rate(T) / SD\_Rate_i(T) \right)^2$

//adjusted channel IDF

8 $Terms_{title} \leftarrow split(P_{title}), Terms_{content} \leftarrow split(P_{content})$

9 $i \leftarrow channel(P)$

//text split

10 $Terms \leftarrow merge(Terms_{title}, Terms_{content}, rate)$

//merge the same terms

11 FOR $term \leftarrow Terms$

12 $W_{term} \leftarrow W_{term} \cdot getIDF(i, term)$

// multiply the TF with the IDF value

13 sort(Terms) // sort terms according to weight

14 $Features \leftarrow getFirst(Terms, num)$

// return the num – biggest terms as the features

Similarly to the traditional TF-IDF algorithm, the channel-IDF values could be calculated in advance. In the stage, which constructs the channel-IDF values, the additional time requested is O(T*N) where T is the number of terms and N is the number of channels in the collections. The time requested to get the precise and recall rate is the same as the traditional algorithm. For the space complexity, both the new CDI-TF-IDF and the channel-TF-IDF algorithms are N times larger than the traditional one.

IV. EXPERIMENTS

In this section, an experiment is executed to evaluate the efficiency of the proposed method. The news page collection is taken from Sogou.COM, a pure text collection which contains about 2 millions pages distributed in 14 channels in the year of 2008. Each news page consist two parts, the title and the content. And then we choose 178 news pages randomly as the test set from the whole collections. Each page is labeled with some features manually. The number of features is unfixed, from 4 to 16.

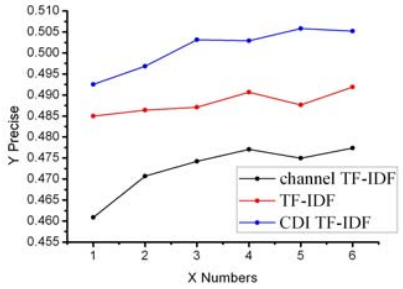
The two criteria, the Precise-Recall and the F-measure, are used to evaluate the efficiency in the experiments. The Precise can be seen as a measurement of exactness or fidelity, whereas the Recall indicates the completeness [8]. For a test page, the Precise is defined as the number of features of auto-generated results included in the manually labeled set divided by the total number of auto-generated features. The Recall is defined as the number of feature of auto-generated results included in the manually labeled set divided by the total number of manually labeled features. The F-measure was derived by Van Rijsbergen [9] so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". The general F_β measure (for non-negative real values of β) is expressed as follow:

$$F_\beta = (1+\beta^2) \cdot \frac{precise \cdot recall}{\beta^2 precise + recall} \quad (6)$$

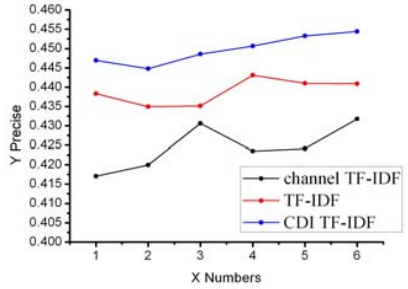
Two commonly used F-measures are the F2 measure, which weights recall twice as much as precision, and the F0.5 measure, which weights precision twice as much as recall. Here, based on the consideration that the loss of un-correct features is larger than the omission of correct features, the F0.5 measure is used.

The average rate of all the pages in the test is used as the final evaluation standards.

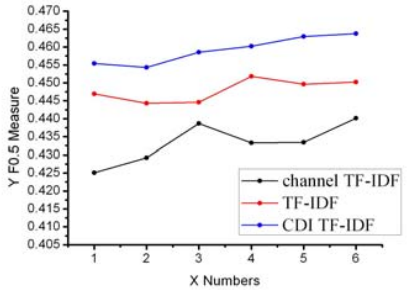
In this experiment, the traditional TF-IDF algorithm is used as the baseline. We evaluate the three algorithms, the traditional TF-IDF algorithm, the channel TF-IDF algorithm and the CDI TF-IDF algorithm with respect to the Recall, Precise and F-measure. The results are presented in Figure Two.



(a)



(b)



(c)

Fig. 2 the Recall in (a), Precise in (b) and F-measure in (c) for the three algorithms, channel TF-IDF, general TF-IDF and CDI TF-IDF

It can be seen from the two graphs that, 1) when the traditional TF-IDF algorithm is applied simply to a channel-divided collection, both the precise rate and the recall rate are decreased; 2) the CDI TF-IDF algorithm eliminates the negative influence of channel-division, and gets a better result. It can be seen that, the precise rate and recall rates have

increase by 2.48% and 2.66%, respectively. To sum up, the CDI TF-IDF algorithm is more efficient for the feature extraction in web news.

V. Conclusions and Future Work

In this paper, a new algorithm is proposed to refine the IDF scheme for feature extraction in web news. The channel IDF value for a term is adjusted based on the distribution characteristics among all the channels. The refined channel IDF value could represent the importance of a term better by recognize the Top terms, the meaningless terms and the related channels for each one. The experimental results show that the algorithm is effective to improve the accuracy of feature extraction in news web pages. In order to get a better performance, 1) a more effective model will be used to recognize the Top Word terms and the meaningless terms, 2) the way to maintain the channel IDF values dynamically will be another important matter in the future.

References

[1] Ricardo BY, Berthier RN. Modern Information Retrieval, New York: Addison-Wesley, ACM Press, 1999. Page 19-34

[2] Kenneth W. Church, William A. Gale Inverse Document Frequency (IDF): A Measure of Deviations from Poisson, Natural language processing using very large corpora, Kluwer Academic Press, Boston (1999), pages 283–295

[3] Meng Qiu, Liang He et al. A New Keyword Extraction Algorithm CDLC for Chinese News Web Pages

[4] Kazunari Sugiyama et al. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages, 14th ACM Conference on Hypertext and Hypermedia 2003, Pages 198-207

[5] Zhang Min et al. DF or IDF? On the Use of Primary Feature Model for Web Information Retrieval, Journal of Software 2005, Pages 1012-1020 (in Chinese)

[6] Wen-tau Yih et al. Finding Advertising Keywords on Web Pages, Proceedings of the 15th international conference on World Wide Web 2006, Pages 213-222

[7] Monika Henzinger et al. Query-Free News Search, Proceedings of the 12th World Wide Web Conference 2005, pages 1–10

[8] Ralph Grishman, Beth Sundheim, Message Understanding Conference-6: a brief history, Proceedings of the 16th conference on Computational linguistics, August 05-09, 1996, page 446-471

[9] Van Rijsbergen, C.V., Information Retrieval. London;Boston. Butternworth, 2nd Edition 1979. ISBN 0-408-70929-4