DOMAIN KEYWORD EXTRACTION TECHNIQUE: A NEW WEIGHTING METHOD BASED ON FREQUENCY ANALYSIS

Rakhi Chakraborty

Department of Computer Science & Engineering, Global Institute Of Management and Technology, Nadia, India
rakhi.chakraborty84@yahoo.in

ABSTRACT

On-line text documents rapidly increase in size with the growth of World Wide Web. To manage such a huge amount of texts,several text miningapplications came into existence. Those applications such as search engine, text categorization, summarization, and topic detection are based on feature extraction.It is extremely time consuming and difficult task to extract keyword or feature manually.So an automated process that extracts keywords or features needs to be established.This paper proposes a new domain keyword extraction technique that includes a new weighting method on the base of the conventional TF•IDF. Term frequency-Inverse document frequency is widely used to express the documentsfeature weight, which can't reflect the division of terms in the document, and then can't reflect the significance degree and the difference between categories. This paper proposes a new weighting method to which a new weight is added to express the differences between domains on the base of original TF•IDF.The extracted feature can represent the content of the text better and has a better distinguished ability.

.KEYWORDS

Text mining,Feature extraction,weighting method, Term Frequency Inverse Document Frequency (TF•IDF), Domain keyword extraction.

1. INTRODUCTION

For rapidly development of computer network technology various aspects of electronic documents also grows rapidly and most of the enterprise information saved as a text form. Hence text mining has become an increasingly popular and also important field in the research of data mining. Text mining is different from the traditional data mining has been pointed out by Han. J, and Kamer. M[1]. The conventional data mining defines the relationship, things and structured data as the research target. While the text mining defines the text database as the research targetwhich consists of a large number of documents from the various data sources, including research papers, news articles, books, journals, reports, patent specifications, conference paper, e-mail messages, web pages and so on. Text mining is aninfantile interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics, Fig.1 shows that how the text mining is interconnected with other. Text mining would like to solve problems such as the uncertainty and ambiguity in the text information.Text mining or knowledge discovery from text (KDT) mentioned for the first time by Feldman et al. [2] which is deals with the machine supported analysis of text. Text mining uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics.

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 109–118, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3211

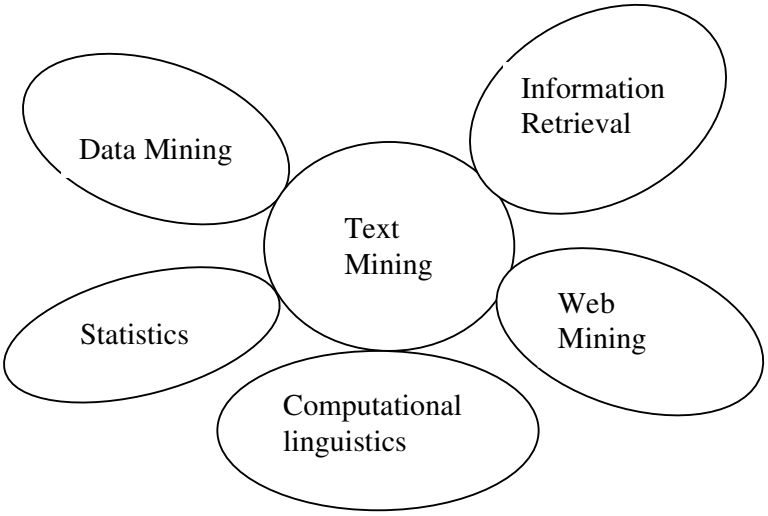


Figure1. Interdisciplinary domain

2. AUTOMATED DOMAIN KEYWORD EXTRACTION TECHNIQUE

Keyword extraction is an extremely time consuming and difficult task, when it is done manually. Due to large volume of published news articles, it is almost impossible to extract keywords manually.Forestablish an automated process that extract keywords from news articles,anunsupervised keyword extraction technique namely Automated Domain keyword Extraction Technique is introduced in that research paper. Here a new weighting method is also introduced on the base of the conventional TF•IDF [3] [4].

2.1. Keywords

Keywords are a set of significant words in an article that gives high-level narrative of its contents to readers. To produce a short summary of news articles identifying the keyword from a large number of online news data is very useful. Keyword extraction technique is used to extract main features in studies such as text classification, text categorization, information retrieval,topic detection and document summarization.

2.2. Different Methods forextracting keyword

The main methods of the keyword extraction are TF*IDF (Term Frequency Inverse Document Frequency), mutual information, information gain, NGL coefficient, chi-square and so on. TFIDF and mutual information are the conventional methods. It's a old weighting method but making it popular by recent algorithm Salton, G. & Buckley, C. [5]

The significant effect has been proved in the practical applications using TFIDF formula to acquire single text keyword. Mutual information is commonly used in statistical language models to evaluate the correlated degree between strings. Bigger mutual information between strings indicates the stronger correlation in the viewpoint of statistics. But the small mutual information does not always means that there is weaker correlation between strings and in computing the string requires minimum number.

2.3. Single Text Keyword Vs Domain keywords

TF-IDF (Term Frequency-Inverse Document Frequency) weightingmodel [6]is a statistical model that evaluates the degree of importance of a word in a single document, but it is not suitable for extracting domain keywords.The keywords of single text are difficult to accurately reflect text domain knowledge and user interests, which will cause the Web more difficult to provide high-quality personalized services for readers.

Reader want extract multi-texts keywords to reflect the domain knowledge of texts, in order to provide automatically text clustering, classification and summarization and topic detection.

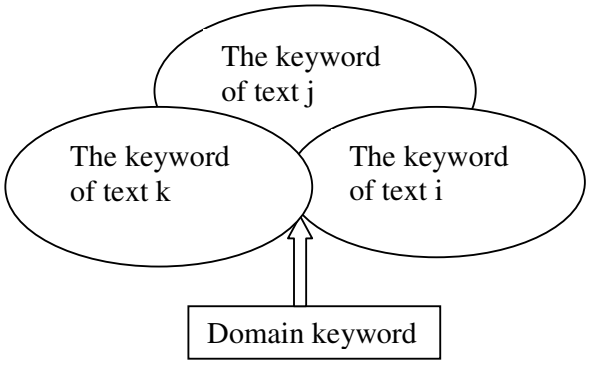


Figure2.Relation between single text keyword and domain keyword

2.4 Conventional TF-IDF Formula

The conventional, TF-IDF weight scheme invented byBerger, A et al [7] which is as follows:

TF: $$TF(t, d_i) = \frac{n_{t,i}}{\sum_{k=1}^{|T|} n_{k,i}}$$

Where, $TF\ (t, d_i)$ =term frequency of word t in document d_i
$n_{t,i}$ = number of accurances of term t in d_i
$n_{k,i}$ = number of accurances of all terms in d_i

IDF: $$IDF_t = log \frac{M}{m_t + 0.01}$$

Where, M = total number of documents in the corpus
m_t = total number of documents in the corpus where term t appears.

TF-IDF: $$w(t, d_i) = TF(t, d_i) \times IDF_t$$

Where, $w\ (t, d_i)$ = weight of term t in document d_i.

TF-IDF value is composed of two components: TF and IDF values. The rationale of TF value is that more frequent words in a document are more important than less frequent words. TF value of a particular word in a given document is the number of occurrences in the document. This count is usually normalized to prevent a bias towards longer documents to give a measure of the importance of the termt_i within the particular documentd_j, like a TF equation given in above.The second component of TF-IDF value, IDF, represents rarity across the whole collection. This value is obtained by dividing the number of all documents by the number of documents containing the

term, and then taking the logarithm of that quotient, like an IDF equation given above. For example, a word, 'today' appears in many documents and this word is weighted as low IDF value. Thus it is regarded as a meaningless word.

2.5 Problem of Conventional TF-IDF Weighting

TF-IDF is generally accepted as an effective way of feature extraction for a single document.

TF•IDF is based on the assumption that

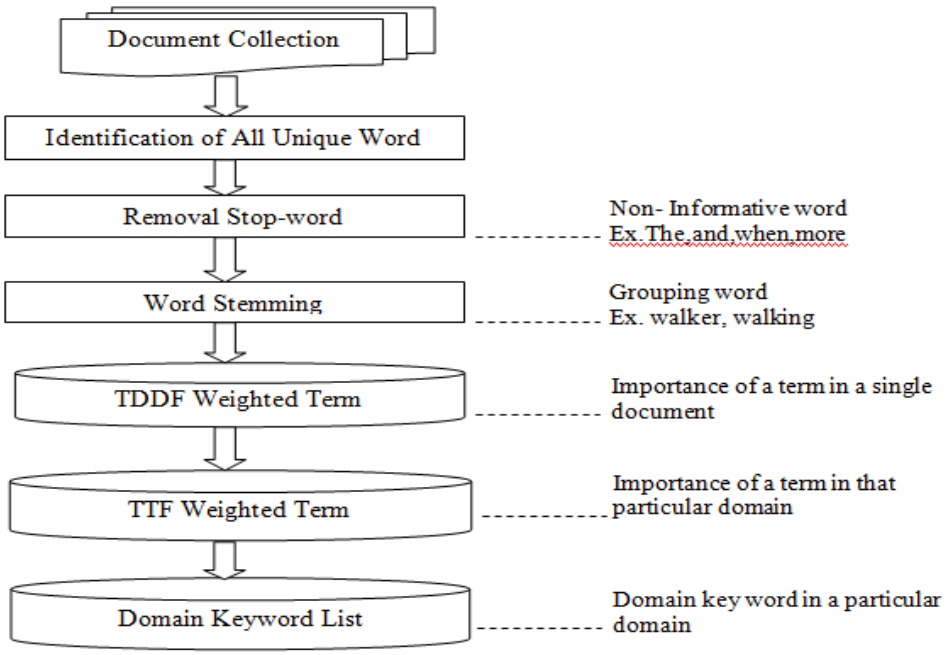
- ✓ The term has a high frequency in the document, i.e., a large TF value;
- ✓ The term has a low document frequency in the whole document collection, i.e., a small DF value.

It is embodied in the following two aspects:

- × If the TF of a term in a document is low but high in certain category (not all the document), the term can also represent the feature of the text very well.
- × According to TF-IDF, the terms which have low document frequency in the whole document collection can represent the feature of the text. But for a highly important term of a certain category should have high document frequency in that category.

The two aspects neglect the frequency in the terms of a certain category. Because of the problems we have to add a weight to the original TF•IDF. The added weight considers the frequency of the term, which is in a particular category in the whole text collection, rather than simply consider the frequency of the term which is in the other documents of the whole text collection.

2.6. Block Diagram of Domain Keyword Extraction Technique



2.7. Process of Domain Keyword Extraction Technique

2.7.1. Problem Definition

In that research work, features are also referred as keywords, a set of significant words or terms in a text which gives high-level description of its contents. The problem to be solved in thispaper is to extract significant features for each news domain.

Let, $D = \{d_1, d_2, .. , d_M\}$ be a set of news documents that belong to various news domain. $T_j = \{t_{j1}, t_{j2} , \ldots, t_{jn}\}$ be a set of n terms extracted from a single document d_j. Then T, a set of terms extracted from a document set that belong to a particular domain is a union of $T_1, T_2, \ldots , T_C$. M = Total number of documents in the whole corpus. C = Total number of documents belong to a particular domain or category. m_t = The number of documents that term t occurrences in the documents set D called document frequency. c_t = The number of documents that term t occurrences in the same domain documents set is called document domain frequency.

To extract significant keywords from T, weighting the term of each document in a domain is done. Then collect n% terms of each document. After that count term frequency of each term from term collection and sort the terms by this weight and extract top-N terms with high weight. Then a word list namely 'keyword list 'is formed with these terms for that domain.

2.7.2 Document Preprocessing

- **Remove Stop words**

The irrelevantset of words is even appearing frequently. For example; a, the, of, with, for etc… (I.e. articles, preposition, pronouns)

- **Stemming**

Identify groups of words where the words in a group are small syntactic variants of one another and collect only the common word stem per group. For example, "development", "developed" and "developing" will be all treated as "develop".
Stop-list, stemming reduces the amount of dimensions and enhances the relevancy between word and document or categories. So, we remove stop words and word stemming.

2.7.3 TDDF Formula

The propose TDDF (TF-IDF based Document domain frequency) formula for the domain keyword extraction as follows:

$$W(t, D, C) = \left(W_{d_i}^{t}(t, d_i) + 0.01\right) \times \frac{c_t}{C}$$

Where, $W_{d_i}^{t}(t, d_i) = tf(t, d_{i)} \times \log\frac{M}{m_t}$, is the weight of word t in document d_i

$tf(t, d_{i)}$, is the normalized term frequency of word t in document d_i

$\log\frac{M}{m_t}$, is the inverse document frequency of t

$\frac{c_t}{C}$, is word common possession rate

$W(t, D, C)$, weight of term *t* in document d_i .with respect to the domain in which document d_i belongs

C = total number of documents in a particular domain in which d_i belongs

c_t = total number of documents containing term *t* in a particular domain in which $d_i$belongs

The original TF-IDFis TF multiplies by IDF, where TF and IDF are short for term frequency and inverse document frequency respectively. The addition of a weight to the original TF-IDF which is done to overcome the problem of TF•IDF, it is known as common word possession rate. This common word possession rate considers the frequency of the term, which is in a particular category in the whole text collection, rather than simply consider the frequency of the term which is in the other documents of the whole text collection.

Common word possession ratereflecting the possibility of word *t* becomes the domain keyword. Higher *c* means that the word *t* becomes keyword with greater possibility, vice versa.

The 0.01 in $\left(W_{d_i}^{t}(t, d_i) + 0.01\right)$ is to prevent the word *t* frequency from becoming zero in d_i which causing the denominator is also zero.

2.7.4 Table Term Frequency

In this step, we calculate a threshold value. Then we collect those terms, whose weight is above the threshold value.

Here, the threshold is the most important n% terms from each document according to TDDF value calculated in the previous step.

Then we count term frequencies from the term collected. We name this term frequency as “Table Term Frequency” because the terms collected are stored in a temporary table.

2.7.5 Generate Keyword List

From the table where we stored terms with these TTF, collect top n% terms and ranking them according to the highest table term frequency.

Those are the keyword list for that domain.

2.8. Experiments and Result

- **Step 1**

The first step to extract keywords is downloading news documents from Internet portal site. Internet portal site provides news pages by domain, such as flash, politics, business, society, life/culture, world, IT/science, entertainment, column, English, magazine and special. Here sports news is choosing for experiment. About 22 news documents for sports are taken. The news pages of HTML is written in a fixed structure, Using this structure of HTML page and extract pure news article and stored into text files.

- **Step 2**

After preprocessing each document,i.e. stop-word removing and stemming,then store the each term and its occurrences against its document id into relational database.

Document_id	Word	Occurance	TF
sp1	asian	7	0.049
sp1	athletes	1	0.007
sp1	believed	1	0.007
sp1	billion	1	0.007
sp1	bodybuilding	1	0.007
sp1	ceremony	2	0.014
sp1	check	1	0.007
sp1	china	3	0.021
sp1	cities	1	0.007
sp1	city	2	0.014
sp1	clifford	1	0.007
sp1	close	1	0.007
sp1	cohosting	1	0.007
sp1	compared	1	0.007
sp1	competition	1	0.007
sp1	continent	1	0.007
sp1	controversies	1	0.007
sp1	cost	1	0.007
sp1	countries	2	0.014

- **Step 3**

Then take a “dictionary” table in relational database to store all the terms exist in the document corpus and its document domain frequency.

Word	Doc_domain_fr...	Word_common...
account	1	0.333
advani	1	0.333
ahead	1	0.333
ahmedabad	1	0.333
air	1	0.333
alok	1	0.333
annu	1	0.333
aoti	1	0.333
asian	2	0.667
athletes	1	0.333
bagging	1	0.333
bat	1	0.333
begin	1	0.333
believed	1	0.333
belligerence	1	0.333
billiards	1	0.333
billion	1	0.333
bodybuilding	1	0.333
bowlers	1	0.333
bronze	1	0.333
captain	1	0.333
century	1	0.333

- **Step 4**

Then calculate the weight by using TDDF formula.

Document_id	Word	Occurance	TF	Doc_domain_fr...	Word_common...	Weight
sp1	asiad	2	0.014	1	0.333	0.004662
sp1	asian	7	0.049	2	0.667	0.032683
sp1	athletes	1	0.007	1	0.333	0.002331
sp1	believed	1	0.007	1	0.333	0.002331
sp1	billion	1	0.007	1	0.333	0.002331
sp1	bodybuilding	1	0.007	1	0.333	0.002331
sp1	ceremony	2	0.014	1	0.333	0.004662
sp1	check	1	0.007	1	0.333	0.002331
sp1	china	3	0.021	2	0.667	0.014007
sp1	cities	1	0.007	1	0.333	0.002331
sp1	city	2	0.014	1	0.333	0.004662
sp1	clifford	1	0.007	1	0.333	0.002331
sp1	close	1	0.007	2	0.667	0.004669
sp1	cohosting	1	0.007	1	0.333	0.002331
sp1	compared	1	0.007	1	0.333	0.002331
sp1	competing	1	0.007	1	0.333	0.002331
sp1	competition	1	0.007	1	0.333	0.002331
sp1	competitive	1	0.007	1	0.333	0.002331
sp1	continent	1	0.007	1	0.333	0.002331
sp1	controversies	1	0.007	1	0.333	0.002331
sp1	cost	1	0.007	2	0.667	0.004669
sp1	countries	2	0.014	1	0.333	0.004662

- **Step 5**

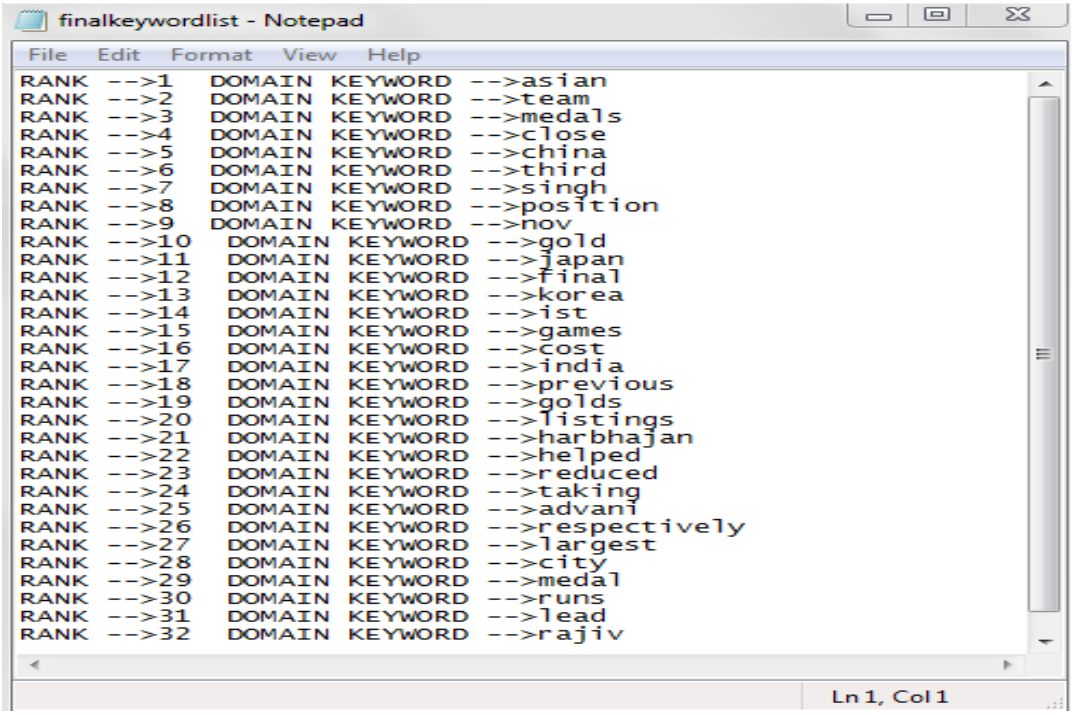
Take top high-weighted term which are the above threshold value and store them in a table called “ntopterms” table.

Document_id	Word
sp1	games
sp1	asian
sp1	medals
sp1	china
sp1	guangzhou
sp1	gold
sp1	set
sp1	events
sp1	cost
sp1	team
sp1	close
sp1	third
sp1	position
sp1	japan
sp1	korea
sp1	ceremony
sp1	local
sp1	previous
sp1	largest
sp1	countries

- **Step 6**

Then count the term frequency from the collected term and rank them according to the high term frequency whose value are above the threshold value.

Those are the "KEYWORD LIST" for sports domain.



3. CONCLUSIONS

In that research workkeyword extracting technique that can extract domain keywords is proposed.These keyword extracting techniques can be used to extracting main features from specified document set and applied to document classification. The domain keyword of news document is one of the basic elements of text classification, clustering, summarization; topic detection etc.The experiment shows that the proposed TDDF formula can extract "multi-text" domain keyword more effectively. The quality and quantity of domain keyword can be flexibly controlled by "Word common possession rate".

Further experimental work is needed to test the generality of this result, although news articles can be considered as a representative of various types of documents.

ACKNOWLEDGEMENTS

I would like to express my warmest gratitude for the inspiration, encouragement and assistance that I received from my esteemed guidesMr. Apurba Paul, throughout the research work. It is because of his continuous guidance encouragement and valuable advices at every aspect and strata of the problem from the embryonic to the developmental stage, that research has seen the light of this day.

REFERENCES

[1] Han.J, Kamer.M. .Data Mining Concepts and Techniques.BeiJing:Higher education press, 2001. 285-295.
[2] R. Feldman and I. Dagan.Kdt - knowledge discovery in texts. In Proc.of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117,1995.
[3] Robertson, S. E., "Term specificity [letter to the editor]", Journal of Documentation, Vol. 28, 1972, pp. 164-165.
[4] Stephen Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60, No.5, 2004, pp 503-520
[5] Salton, G. & Buckley, C. (1988). Term-weighingapproaches in automatic text retrieval.In Information Processing & Management, 24(5): 513-523.
[6] Thorsten, J., 1996. Probabilistic Analysis of the Rocchio Algorithmwith TFIDF for Text Categorization.Proceedingsof 14th International Conference on Machine Learning.McLean. Virginia, USA, p.78-85.
[7] Berger, A et al (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In Proc. Int. Conf. Research and Development in Information Retrieval, 192-199.

AUTHOR

RakhiChakrabortyis Currently working as an Assistant Professor in Global Institute of Management & Technology,Krishnanagar, Kolkata, India. Prior toshe obtained her Bachelor degreeand M.tech degree in Computer Science and Engineering in the West Bengal university of Technology.Her areas of interest is in Data Mining and computer architecture