International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.1, No.7, November 2012 DOI:10.15693/ijaist/2012.v1i7.18-21

An Empirical Study on Term Weights for Text Categorization

N.Naveenkmar[1], K.Batri[2]
[1]*Assistant Professor, SBM College of Engineering & Technology, Dindigul, India.*
[2]*Professor, PSNA College of Engineering & Technology, Dindigul, India.*

***Abstract* - In this paper, various term weighting methods for text categorization has been discussed. The terms represent the words, queries, phrases and indexing units and identify the texts. The supervised and unsupervised weighting methods to represent the prior information (supervised) or not (unsupervised) in the membership of training documents of categories were discussed. The review of various term weights approach under the text-based information processing presented will provide the necessary information for the researchers. This research is to provide a useful approach on the relationship among various term weight methods as well as to exploit the research domain.**

Keywords-Term weight, Text categorization, Information retrieval, Discriminative terms.

I. INTRODUCTION

A text categorization is the task to classify the group of documents into a set of predefined categories from the natural language documents [1, 3]. In vector of term weight methods assign suitable weights to improve the performance of text categorization. The terms are representing the vector. It is used to perform entire collection of documents. Hence, data modeling puts term weights in a historical point of view rooted in statistics, numerical, and term analysis usually. The term vector is appeared in the content of document as $D = \{t_1, t_2, t_3 \ldots t_k\}$, the number of terms as n in the document D where $t_1, t_2, t_3 \ldots t_k$ are individual terms. Vector space model of documents are represented by vector of words. The set of words or bag-of-words represents the documents of words and is converted into vector approach. And each word is considered a unique feature [2, 10]. A term weighting focuses on quantify the importance of relative in different terms, and takes them as subsets. Then in feature selection, discriminative terms are assigned larger weights [7].

Deopole and Sebastiani discussed the construction of three phases to represent the text classifier. If consists of internal representation of document based on indexing term selection, term weight and term classifier. First, in the phase of term selection, related terms are used for dimensionality reduction that consists of term selection on identified text classifier. Second, the term weights represent the document weight for the selected terms that are computed. These two phases are internal representation of document indexing. Finally, the term classifiers are the creation of learning from generated weights of classifier induction [1]. In general, the single term weights used appropriately for content representation and become compatible of words that extracted from the text of documents by information retrieval field. This term weighting is used the components in combination of term frequency, collection frequency, normalization length of components applied in various text fields [7, 3]. The preprocesses are the stemming, stop word removal, different term weighting scheme, feature selection and text classifier based on distinct benchmark data collections in various parameters and in spite of evaluation methods in micro-averaged and macro-averaged precision, recall, break-even point depends on different researchers used to text data processed. Micro- average gives equal weights with more positives on every document. Macro-average gives equal weights to every category in the face of frequency.

The semantic approaches of categories are explored the words appearing in the category labels in exploiting the WordNet. And this category proposed that the novel semantic term weighting is better than TF-IDF scheme to improve the performance of indexing terms. Semantic information of document classifies the terms into two classes based on positive and negative scores in the documents to verify the comment of sentiments identified [8, 10]. A Clustering based feature weighting approach for text classification proposes that each class in the training collection is known as cluster, and uses to enhance the text classification. In follows TF-IDF feature weighting is not an account in weight of features, and is not related to the document one another, but when it based on classes of relevant to the documents [2].

The remainder of the paper is organized as follows: Section 2 discusses the popular term weights on heuristic and statistical based on supervised term weighting methods. In Section 3, reviews an unsupervised term weighting method. Section 4, enlightens of some applications, finally, we conclude the paper in Section 5.

II. TERM WEIGHTING METHODS

In this study, major part of term weights falls under the two categories according to supervised term weighting and unsupervised term weighting [3]. The supervised term weights acts on supervised learning and it uses category label of training documents then contributes text categorization. An unsupervised term weighting methods borrows from information retrieval of the traditional term weights (binary, tf, tf.idf) based on calculated and does not use the information category of training documents. These cases used to represent the content of document $D= \{d_1, d_2, d_3 \ldots d_n\}$ known as set of documents and the term vector is appeared in the form $D =$

$\{t_1, t_2, t_3 \ldots t_k\}$, as known as a set of distinct terms. The parameters *n* and *k* are the total number of documents and terms, respectively. The vector space model of documents is represented by vector of words. In every function f (t_k, c_i), t_k, the order of term, set of categories (c_1, c_2, $c_3 \ldots c_i$) are distributed. Let c_i is specific category in the collection of documents.

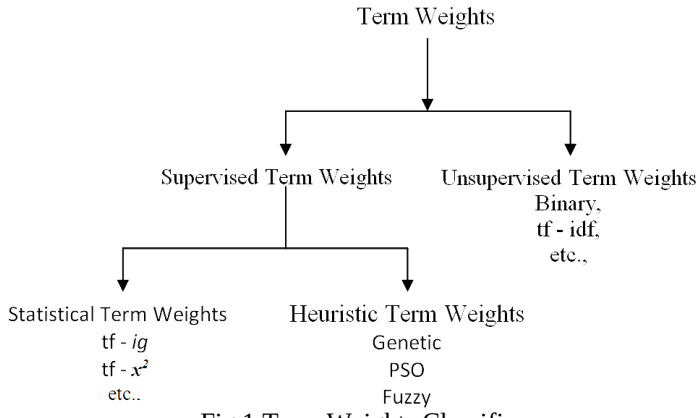

Fig.1 Term Weights Classifier

III. SUPERVISED TERM WEIGHTING METHODS

The supervised learning process performs the text classification, in this task distinct form of category labels are pre-defined based on new documents recommended by a training set of labeled documents. And this weight of a given term is computed an occurrence of probability terms in the training data of positive and negative classes [4]. And a new term weighting method, on statistical confidence of ConfWeight, with respect to each category c_i assigned the labels. The assigned labels are with prior knowledge of labeled training data set [14]. In general, the each term performs uses traditional information theory and statistical functions of feature metrics namely information gain, chi-square, odd ratio and gain ratio [1]. The feature selection is mainly concerned about the problem of reducing high dimensionality of the feature space in text categorization. In this task, term weights of two categories statistical-based and heuristic-based.

A. Statistical-based Term Weights

The weight terms are related to feature selection metrics applying by replacing *idf* factors with the term frequency weight of $tf.x^2$. It is more effective than tf.idf [3]. Based on this concept, several methods replaced on information gain (tf -*ig*), chi-square $(tf - x^2)$, relevance frequency (tf-*rf*), etc.

1) Information Gain (*ig*)

Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in a document. The information gain of word lying in the group of retrieved documents into sequence of organizing the documents. In relevant structure of documents into the weight of each word (term), we adopt the information gain ratio (IGR) of the probabilistic distribution of each word as a term weight [11]. The information gain and gain ratio is computed for term t and category c_i

$$\mathrm{ig}(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t,c) \log_2 \frac{p(t,c)}{p(t)p(c)} \quad (1)$$

$$\mathrm{gr}(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_i, \bar{t}_k\}} p(t,c) \log_2 \frac{p(t,c)}{p(t)p(c)}}{- \sum_{c \in \{c_i, \bar{c}_i\}} p(c) \log_2 p(c)} \quad (2)$$

2. Chi- Square (x^2)

The x^2 is the common statistical test, and it measures the difference in distribution that expects a feature occurrence in the class value. The x^2 – test measures the lack of independence between term t_k and category c_i. It is given by:

$$x^2(t_k, c_i) = \frac{\left[P(t_k, c_i)p(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)p(\bar{t}_k, c_i)\right]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \quad (3)$$

3. Term frequency-relevance frequency (tf - rf)

The relevance frequency performs the asymmetric term weighting and does not take into account the total number of documents in different classes. The supervised term weightings are concentrated in the positive category than in the negative category. Here it gives more contributions and selects the positive samples from the negative samples. It assigns the negative class relatively a lower weight [4, 14]. The documents set of relevance judgments for each request on the relative distribution of terms [6]. The discriminating power is involved the terms *a*, *b*, *c*, *d* and a imposing by a number of relevant documents that contain only terms *a* and *b*. The term frequency with relevance frequency to combined with the relevance frequency of term weight t_i is defined as,

$$\mathrm{tf.rf} = tf * \log\left(2 + \frac{a}{c}\right) \quad (4)$$

B. Heuristic-based term weights

The term weights have a good threshold, it is difficult to determine and no theoretical association exists between the threshold and the expected categorization performance. Therefore in this case heuristic-based adjustment and enhancement to the feature selection and term-weighting functions are required to improve the performance.

1) Genetic Algorithm

Genetic algorithms are heuristic optimization methods, to represent terms in document that can be recognized by a part of speech technique and concept and information extraction used. The documents are discrimination of the distribution of two concepts in the two documents that diminish the similarity between two documents and term weight. In the term weight calculated by Weighted Topic Standard Deviation (WTSD). So, it performs the concentrate of document on its identified topic [16].

2) Fuzzy Logic

Fuzzy logic based term weighting scheme, using TF-IDF method for determining the weight of a term. Automatic term weighting basis of described parameters in a set of rules depends on input to a fuzzy logic engine which considers the

relative weight of each one and returns a value which corresponds to proposed term weight. This method is to improve the automation and to conclude the suitable terms in information retrieval [15].

3) Particle Swarm Optimization (PSO)

This technique proposes the term extraction of relevant terms is contained in to the input documents automatically. In a number of particles that constitute a swarm moving around in the search space finds the best solution. The term extraction is to generate list of terms that are relevant to the domain of input domain. In this PSO technique improves the accuracy of term extraction results [17].

IV. UNSUPERVISED TERM WEIGHTING METHODS

The unsupervised learning for text categorization is usually borrowed from information retrieval on this traditional term weighting method, such as learning models without training data. In this measure of statistical based term weight represents an avoid information-theoretic interpretation of traditional term weight (*tf-idf)* as the amount of information [13]. In a text categorization, an imbalanced data is processed by new term weighting scheme of term frequency with term feature (*tf. term feature*) compute well [12]. In the TF-ICF scheme generates document representations independently without the knowledge of streaming documents in linear time [5]. As stated under, term weighting techniques are binary, term frequency (*tf*), inverse document frequency (*idf*), term frequency – inverse document frequency (*tf- idf*) and related variants depend on categories of information. Also, the traditional weighting scheme related theoretical view is given below.

A. Binary weighing

This simplest approach is to let the weight term to represent the two values (0, 1) that denote the weights equal to 1 if the term occurs in the document (present) or weight equal to 0 if the term does not occur in the document (absent). An ignoring the occurrence of term values learning algorithms (SVM, KNN, and Boosting) can be used especially in Naïve Bayes, Decision tree where the real number format of term weight cannot be used [14].

B. *tf* - weighting

Another simplest approach of term frequency *tf(t, d*) , the number of times the term *t* occur in the document *d* adopted the raw frequency of term *t* in the document by *f(t, d)* also has defined as *tf(t, d) = f(t, d).*The term frequency *(tf)* allows the factors, such as *log (1+tf), itf* and different variants. The normalization factor is common to all the terms in the document and without reflection for discriminative document.

C. tf × idf weighting

The tf-idf weighting is the most important method in information retrieval. Koren Sparck Jones had detailed on terms are frequent matching a substitute, and certainly retrieve non-relevant rather more than relevant documents [9]. In the weighting scheme of form *tf-idf*, two statistics measures multiply the term frequency and inverses document frequency. Here *n* is representing the total number of documents contained

$$\text{tf-idf}\ (t_k, d_j) = \text{tf}\ (t_k, d_j) \times \log\ (n/\ n\ (t_k)) \qquad (5)$$

V. APPLICATIONS

A. Text categorization

The term weights need to handle and classify the documents in which prominent of textual data is the dominant component. The scheme of categorical difference weights supports the sentiment of text based information to categorize the positive or negative classes based on sentiment classification and performed an integrated support vector machine (SVM) [8]. The text categorization tasks of machine learning algorithms, such as kNN, decision tree, Naïve Bayes, Neural Network, Linear Regression and Boosting etc. [3, 4, 10, 14]. The term weight scheme integrates a text classification algorithms used to evaluate the performance measure.

B. Information Retrieval

In common support of application an IR is used in weighting scheme for term frequency – inverse document frequency (tf–idf), the different variants of them applied nowadays. Documents are generally identifying the set of terms or keywords that are collectively used to represent their contents used information retrieval. The information derived from the set of retrieved documents in order to summarize the documents. A query based similarity information among the original documents retrieved [11].

VI. CONCLUSIONS

In this study, the differences in term weighting methods of supervised and unsupervised methods are presented. The functional form major term weighting methods designed with text categorization and text filtering will be of move focus for many applications. It is concluded that the probabilistic distribution for each word using a term weight will enhance the categorization.

REFERENCES

[1] Franca Debole and Fabrizio Sebastiani, Supervised Term Weighting for Automated Text Categorization. In Proceedings of SAC-03, 18th ACM Symposium on Applied Computing, Melbourne, US, pp. 784–788, 2003.

[2] Lin Zhu, Jihong Guan and Shuigeng Zhou, CWC: A Clustering-Based Feature Weighting Approach for Text Classification. In Springer-Verlag Berlin Heidelberg, pp 204-215, 2007.

[3] Man LAN, Sam-Yuan SUNG, Hwee-Boon LOW, Chew-Lim TAN, A Comparative Study on Term Weighting Schemes for Text Categorization, IEEE International Joint Conference on Neural Networks, pp. 546-551, 2005.

[4] Zafer Erenel, Hakan Altinçay and Ekrem Varoğlu, Explicit Use of Term Occurrence Probabilities for Term Weighting in Text Categorization. Journal of Information Science and Engineering, Volume 27, pp. 819-834, 2011.

[5] Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump1, Mark T. Elmore, and Ali R. Hurson, TF-ICF: A New

Term Weighting Scheme for Clustering Dynamic Data Streams, 5th International Conference on Machine Learning and Applications, pp. 258 - 263 2006.
[6] S.E.Robertson, K.Spark Jones, Relevance Weighting of Search Terms, Journal of American Society of Information Science, Volume 27, No 3, pp. 143-160, 1976.
[7]Gerard Salton, Christopher Buckley, Term-Weighting Approaches In Automatic Text Retrieval, Information Processing & Management, Volume 24, Issue 5, pp513–523, 1988.
[8]Long-Sheng Chen and Chia-Wei Chang, A New Term Weighting Method by Introducing Class Information for Sentiment Classification of Textual Data. In proceedings of IMECS 2011, Hong Kong, March 16-18, pp.394-397, 2011
[9]Karen Sparck Jones, IDF term weighting and IR research lessons, Journal of Documentation, Volume 60, 2004.pp. 521-523
[10] Qiming Luo, Enhong Chen, Hui Xiong, A Semantic Term Weighting Scheme for Text Categorization, Journal Expert Systems with Applications, Volume 38, Issue 10, pp. 12708–12716, 2011.
[11] Tatsunori Mori Miwa Kikuchi Kazufumi Yoshida, Term Weighting Method Based on Information Gain Ratio for Summarizing Documents retrieved by IR systems, Journal of Natural Language Processing, Volume 9, No.4, 2001.
[12] Ying Liu, Han Tong Loh, Aixin Sun, Imbalanced text classification: A term weighting approach, Expert Systems with Applications, Volume 36 Issue 1, pp. 690-701, 2009.
[13] Akiko Aizawa, An information-theoretic perspective of tf–idf measures, Information Processing and Management, Volume 39 Issue1, Pages 45 – 65, 2003.
[14] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu, Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 4, pp.721-735, 2009.
[15] Ariel Gómez, Jorge Ropero, Carlos León, Alejandro Carrasco, A Novel Term Weighting Scheme for a Fuzzy Logic Based Intelligent Web Agent, Department of Electronic Technology, University of Seville, Seville, Spain.
[16] S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil, Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution, Proceedings of World Academy of Science, Engineering and Technology, Volume 13,pp. 144-147, 2006.
[17]Mohammad Syafrullah and Naomie Salim, Improving Term Extraction Using Particle Swarm Optimization Techniques, Journal of Computing, Volume 2, Issue 2, pp.116-120, 2010.