Publication pour la présente candidature LAFPT

Titre	Term weighting schemes alternative to TF-IDF based on the Vector Space Model (VSM): state-of-the-art.
Auteurs	Demba Kandé; Fodé Camara; **Samba Ndiaye**
Référence	5th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)
Editeur	IEEE
Pages	en cours
Année	2018
DOI	en cours
URL	http://icetas.etssm.org/wp-content/uploads/2018/11/booklet-V10.pdf
Index	en cours
ISBN	978-1-5386-7966-1
Encadreur	Oui
Extrait d'une thèse	Oui

Term weighting schemes alternative to TF-IDF based on the Vector Space Model (VSM): state-of-the-art.

1st Demba Kandé
Departement of mathematics, Cheikh Anta Diop University
Dakar, Senegal
demba4.kande@ucad.edu.sn

2nd Fodé Camara
Department of mathematics, Alioune Diop University
Bambey, Senegal
fode.camara@uadb.edu.sn

3rd Samba Ndiaye
Departement of mathematics, Cheikh Anta Diop University
Dakar, Senegal
samba.ndiaye@ucad.edu.sn

***Abstract*—The explosion of textual data is leading to more and more important problems in text mining applications. The main feature of these applications is the need to have to search non-deterministically within each data item. To take advantage of these data, the researchers proposed several solutions called term weighting schemes. These ones constitute a fundamental problem in the exploitation of textual data. Among which we have the methods named Model Space Vector and Semantic, which are the two major domains of weighting at term used in the information retrieval (IR) and the texts classification (TC). However, the aim of this paper is to have an insight of the most recent term weighting algorithms. First, we will try to study primary objectives in this domain, and then we look at the most recent algorithms. A qualitative analysis of these algorithms will be the next steps. Only the term weighting methods based on the Vector Space Model (VSM) are described here. VSM-based methods are traditional or unsupervised methods, feature selection methods, and supervised or statistical methods. This paper could serve as a reference for a study related to the problems of methods and techniques associated with the terms weighting. This will allow both researchers and professionals in the domain to offer more sophisticated solutions than those existing in the task of automatic management of textual information.**

Keywords— Vector space model, classification, text mining, term weighting scheme.

I. INTRODUCTION

Term weighting is a problem of assignment of weight in the most appropriate way to a term in the corpus. It is fundamental in the tasks of the information retrieval (relevance of the answer of a request), and the texts classification (directly affects the precision of the classifier). Now , it is a complex process that resort to not only an analysis on the relevance of a term TF-IDF [3], LSA [16], ESA [17], Word2Vec [18], but also many other procedures such as determining the discriminating power of a term in a category TF-IGM [9], TF-ICD [20], TF-BCD [7], and the features selection to reduce the dimensionality IG [8], MI [21], CHI [22]. Term weighting remains an important and topical search domain, using a variety of complex techniques and combinations. In addition, researchers are developing new weighting schemes to improve existing ones to obtain better results in queries and texts classification algorithms.

In the literature, the research on term weighting is evolving considerably. This is due to the importance of the problems addressed. However, existing studies may be considered relevant because they address crucial problems in term weighting [11]. However, some of these works are slightly out of date because the latest studies are not taken into account. In addition, their explanations of term weighting is limited. For example, they use only on the known information of class labels for calculating the global factor TF-IDF-ICF [22]. The explanations also ignored some information in the corpus, or put emphasis on a specify domain of research [14]. These works remain still references of the research domain of term weighting. Thus, faced with the speed at which textual information is generated, an automatic organization of the most novice information is needed to better understand the facts.

In this paper, we will try to explain the limits listed above, then set out a solution to the term weighting problem. Indeed, this study summarizes the term weighting methods based on the vector-space model. We will focus our efforts on the most recent but also oldest works of the term weighting taking into account the patterns of supervised, unsupervised weighting.

The document is structured as follows: Section II explains the principle of the VSM method, limitations and alternatives proposed by the researchers. Next, Section III describes the unsupervised, feature selection, and supervised term weighting schemes in the literature and their limitations through qualitative analysis. Finally, Section IV concludes our research study.

II. TEXT REPRESENTATION BY VSM

VSM is the most popular method in the literature of the domain. The space vector model represents a document as a vector where the elements of the vector indicate the appearance of a word in a document. This translates into a high dimensional space. Typically, each distinct character string that appears in the document is a dimension.

The principle of the space vector model considers a document d_i as being a set of terms called vocabulary denoted V. Each term t is associated with a unique index.

Thus, we obtain a vector v, of dimension equal to the size of V and an element v_i of v constitutes the weight associated with the index term t_j. The component of v then represents the weight of the term t_j in the document d_i. So we have a matrix representation matrix t_{ij} where i (row of the matrix) designates the number of documents; j (column of the

matrix) is number of columns; and the intersection between *i* and *j* is the weight of term *j* in document *i*.

The lines and columns are very sparse, a document containing a small part of the total set of terms (words) and a term (word) being generally present in few documents. Any similarity between documents is explained by the presence of many common terms (words) between documents, any similarity between terms (words) is explained by their common presence in a large number of documents.

The space vector model makes an implicit hypothesis called bag-of-words hypothesis [24] that the word order in the document is not important. This seems like a great hypothesis, since a document must be read in a specific order to understand it.

Although this hypothesis works for many tasks like clustering and clustering, it is not a universal solution. Because in the search for information and the processing of natural language, the order of words is essential.

Another challenge of using the space vector model is the problems related to the terms used. For example, two synonyms (different terms but having the same meaning, such as "car" and "automobile") are associated with two different dimensions (or components) of the vectors that represent the documents; if two documents each use one of these terms, the comparison of their vectors will not show any similarity due to the common meaning of synonyms. Another example is a rare term, which corresponds to a stylistic particularity of the writing and has little relation to the meaning of the documents, will have a high weight TF-IDF [3] whereas for the comparison of documents or their classification TF-ICD [20], it can be considered as "noise". We also have the case of homonymies, i.e. used a term in several documents but having the same meaning. To provide an answer to these shortcomings, semantic analysis or indexing has been proposed. These seeks to identify the similarities between the occurrences of several terms of a corpus [16], or to determine the weight of a term by its semantic similarity to a specific TF-SW class [25]. Indeed, the Semantic methods [29] did not show better performance on statistical schemes.

Therefore, this study only describes the term weighting schemes based on statistical methods otherwise known as VSM. They includes two families namely supervised schemas which taking into account information known to categories, and unsupervised schemas which ignoring.

III. Term Weighting Methods Based on VSM Text Representationi

In the literature, several solutions have been proposed to resolve a term weighting problem. This section addresses the limitations of the most recent and most old term weighting schemes show in Table I. They are widely used for term weighting. We will explore them through some examples.

TABLE I. THE ELEMENTS OF TERM WEIGHTING SCHEMAS BASED ON VSM.

Weighting schemas	Denomination	Description / Variants
Unsupervised	Binary	$d = 1,1,0,0,1,0$ (1)
	Term-Frequency	$TF = f_{t,d}$ (2)
	Term-Frequency-Inverse Document Frequency	$tf_{ij} \times log\left(\frac{N}{df_j}\right);$ (3)
Supervised	Term-Frequency- Relevance of frequency	$tf_{ij} \times \log(2 + \frac{A}{\max(B,1)})$ (4)
	Relevance Term Frequency	$tf_{ij} \times Nu * \mid log\frac{(M1u+1)}{(M0u+1)} + log\frac{(M0+p)}{(M1+p)} \mid$ (5)
	Balanced Distributional Concentration	$tf_{ij} \times 1 + \frac{\sum_{k=1}^{\lvert c\rvert}\frac{P(t_j \mid c_k)}{\sum_{k=1}^{\lvert c\rvert}P(t_j \mid c_k)} log\frac{P(t_j \mid c_k)}{\sum_{k=1}^{\lvert c\rvert}P(t_j \mid c_k)}}{log\,(\lvert C\rvert)}$ (6)
	Term-Frequency-Inverse Gravity Moment	$tf_{ij} \times \left(1 + \lambda \frac{f_{j1}}{\sum_{r=1}^{m} f_{jr} \times r}\right)$ (7)
Feature selection	Term-Frequency- chi-square or $\chi 2$ statistic	$tf_{ij} \times \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)}$ (8)

	Term-Frequency- Information Gain	$tf_{ij} \times \sum_{c \in \{c_k\ \bar{c}_k\}} \sum_{t \in \{t_j\ \bar{t}_j\}} P(t_j, c_k) \log \frac{P(t_j, c_k)}{P(c_k)P(t_j)}$ (9)
	Term-Frequency- Mutual Information	$tf_{ij} \times \log \frac{P(t_j, c_k)}{P(c_k)P(t_j)})$ (10)
	Term-Frequency-Odds Ratio	$tf_{ij} \times \frac{P(t_j \mid c_k)(1 - P(t_j \mid \bar{c}_k))}{(1 - P(t_j \mid c_k))\, P(t_j \mid \bar{c}_k)}$ (11)
	Term-Frequency- Measure Relevance Distinction	$tf_{ij} \times \frac{A}{B} \times \frac{A}{C} \left(\frac{A}{B} \times \frac{A}{C} - \frac{B}{A} \times \frac{B}{C}\right)$ (12)

and others.

A. *Unsupervised methods*

The first term weighting in the literature are a Boolean (or binary), Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) term weighting methods for TC are generally derived from the IR and belong to the family of unsupervised term weighting schemes. In the literature, various unsupervised weighting schemes have been proposed and thus to optimize the results. We have focused on the limitations of these weighting schemes. These schemes are the most used and efficient unsupervised term weighting schemes. We can explore the literature, through a simple example.

L et's consider the following corpus, denoted *d*:

TABLE II. A SIMPLE EXAMPLE OF CORPUS D

Id doc	Document contain	Class
d_1	"the sky is blue"	negative
d_2	"the sun is bright today"	positive
d_3	"the sun in the sky is bright"	positive
d_4	"we can see the shining sun, the bright sun"	positive

Then, its dictionary is {'blue', 'sky', 'bright', 'sun', 'today', 'can', 'see', 'shining'}.

The Boolean Term weighting Method is the first and most simple of document representation [1]. In this approach, a document is represented as a binary vector. See equation (1) in Table I. This method transforms the document d into a vector which elements indicate the presence (value equal to 1) or the absence (value equal to 0) of a term in the document : *d = (1,1,0,1,0,1,0,0,1,0)*. This representation is still widely used because of its simplicity, its low processing time and its results are not bad. It is called "keyword" representation.

The Boolean term weighting is very limited and uninformative, because it ignores information about the occurrences of a term in the document that may be important for the classification operation, or the length of the text.

By considering the deficiencies in Boolean approach, researchers have proposed a new term weighting scheme denoted TF (Term Frequency) [2]. That is to say weighting a term using the known information about the corpus. See equation (2) in Table I.

In TF term weighting a document is a vector of which components correspond to the number of occurrences of terms in the document. It informs us about not only the presence or the absence of a term like the Boolean method but also about the number of time it appears in a document. A weight is assigned to each of the terms belonging to the document.

So we have three disadvantages in this method. Firstly, the non-support of the interaction between terms which translates in an independence of the latter's. Secondly, the syntactic restructuring of the document caused by the fact that the model does not keep the order of the words. Lastly, TF assigns large weights to the keywords that have a low discriminating power.

Let's consider the corpus of the Table. II:

The term frequency (i.e., TF) for "sun" in d_4 is 2, 4 in the corpus and appears in (d_2, d_3, d_4). So "sun" is a keyword with a higher weight but not more important than the others for example "blue" which a low weight is more discriminating for, which appears only in d1 (identity word in d_1). To correct for this weight difference, some normalizations have been proposed.

This, we have the frequency normalizations defined by:

$$TF = \frac{f_{t,d}}{\sum f_{t,d}} \tag{13}$$

to balance the difference of weight between the terms frequency and the rate term. Other variants of TF exist in the literature such as *log(TF+1)*, which makes it possible to minimize the influence of long documents on texts.

Nevertheless, they have not been up to a definitive solution of this deficit. A new weighting metric has been proposed to correct this weight difference.

To minimise this deficience, a global factor is combined with TF approach. See equation (3) in Table I. Where $f_{t,d}$ denotes the frequency of term t_j in document d_i and N is the total number of documents and $Df_{t,d}$ is the number of documents that contains the term t.

The weight is composed of two factors: the local factor TF (for Term Frequency) metric that calculates the number of times a word appears in a document; and the global factor IDF (Inverse Document Frequency) term is computed as the logarithm of the number of the documents in the corpus divided by the number of documents that are specific to the term. The basic idea of TF-IDF is to determine term weight that are frequent in the document (using the TF metric), but infrequent in the corpus (using the IDF metric).

Let's consider the corpus of the Table II:

The TF for *"sun"* in d_4 is *2* and *"blue"* is 1 in d_1. The word *"sun"* appears in tree documents. Then, the inverse document frequency (i.e., IDF) is calculated as *log(4/3) = 0.1249*. Thus, the TF-IDF weight is the product of these quantities: *2* x *0.1249 = 0.2498* and for "blue" the weight is *1* x *log(4/1) = 0.602*. From the above example, the TF-IDF weight of a term depends on its relevance in the corpus.

In spite of the successes obtained in the classification and extraction of textual data, this method of weighting is not the most efficient because it has shortcomings.

In addition to the first two limits listed in TF method, IDF ignores on the one hand the similarity between the terms. For example the total number of document in the corpora d of Table II is 4, and "sun" appears in 3 documents (d_4, d_3, d_4) in the same way as "bright ".

The global factor IDF assigns the same weight to the terms "sun" and "bright": IDF (sun) = IDF (bright) = *log(4/3)* = 0.1249, while they are not similar in all documents.

On the other hand, like TF, IDF assigns a high weight to the most relevant terms, which seems normal but a significant difference of weight is noted. To balance its results variants have been proposed.

The TFC is the frequency normalization defined as follow:

$$TFC = \frac{TF - IDF}{\sqrt{\sum(TF - IDF)^2}} \quad (14)$$

normalizes the TF-IDF according to all the terms in corpus so as not to favor the longer documents. LTC applies the logarithm to TFC to reduce the effects of frequency differences.

The BM25 [8] variant is based on an analysis of the behavior of the full elasticity model under different values of the global parameters K and k_1 which are in general unknown, but may be tuned on the basis of evaluation data.

$$BM25 = log(\frac{N}{Df_{tf_i}}) \frac{(k_1 + 1)tf_i}{K + tf_i} \quad (15)$$

In general, frequency normalization is the most used to avoid problems related to text lengths.

There may be other variants, which we do not cover in this study because they share the same idea of TF-IDF and their formats are basically similar to each other.

Since the traditional term weighting schemas is not fully effective. Several variants of TF-IDF based on supervised methods have been proposed in the literature. These variants introduce a new statistic and feature selection methods to evaluate the discriminating power (or relevance) of a term in a class label.

B. Feature-selection methods

In taking into account the limitation of the traditional term weighting, new approaches called supervised term weighting have been proposed [28]. These methods weight a term by distributing the knowledge to two classes. These information on the distribution shown in Table III. This table is named a contingency table. It contains the distribution of term t_j with respect to class c_k in the training corpus.

TABLE III. THE CONTINGENCE TABLE INFORMATION.

CLASS / TERM	C_K	$\bar{c}_k$		
	C_1	c_2	….	c_m
t_j	*A*	*B*		
$\bar{t}_j$	*C*	*D*		

A is the number of documents in c_k category that contains t_j term; *B* is the number of documents in $\bar{c}_k$ category that contains t_j term; *C* is the number of documents in c_k category that does not contain t_j term; *D* is the number of documents in c_k category that does not contain t_j term. The table III informs us that:

- if a t_j term is specific for the c_k, then it is a reference for this category, its discriminating power *A/B* is important, because the number of documents in *A* is higher for this t_j term compared to *B*;
- if the *A/C* ratio is important, then the category c_k contains more indexing documents than documents that do not index it;
- if a term t_j is more relevant for $\bar{c}_k$, this indicates that it is a reference term for $\bar{c}_k$, then the ratio *B/A* is high because the number of document indexing t_j is higher in $\bar{c}_k$ than the one in c_k;
- if the ratio *B/D* is important, then the number of documents referring to in $\bar{c}_k$ is higher than those not indexing t_j in $\bar{c}_k$;
- for the c_k category, the discriminating power of a t_j term is determined by (*A/B* x *A/C)* quantity. Whereas, the product of the *B/A* and *B/C* ratio, indicates the relevance of a t_j term in $\bar{c}_k$ category.

In general, these supervised methods use the feature selection metrics [29] as the global factor. Among them the Chi-deux statistic CHI [30], information gain IG [31], gain ratio GR [28], correlation coefficient CC [32], mutual information MI [33], odds ratio OR [31], and so on.

The intuition of these approach is if a term is related to the c_k class, then its relevance is high for this class.

Consequently, these methods associate this relevance with the traditional scheme TF for term weighting [4].

Thus, the χ2 Statistic is combined with the local factor TF to take into account the document information and class label knowledge in the determination of terms weight [34]. This CHI value is defined as follow:

$$\chi^2 = \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)} \quad (16)$$

Then TF-CHI weight of term t_j is calculated with the help of following equation.

$$tf_{ij} \times \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)} \quad (17)$$

Another metric is the Measure of Relevance and Distinction with the AD [5] is the another method of a term weighting based on the feature selection approach. It is based on the notion of relevance of feature from the distribution of terms in the category c_k. The more a term contributes to the distinction of the category c_k, the higher its relevance is for c_k , and then the weight assigned is high. AD of the term t_j for the category c_k can be defined in Table I equation (12).

In general, these methods calculate the weight of the terms by grouping set of class $C = \{c_1, c_2 \ldots c_k\}$ into two class c_k and its complement $\bar{c}_k$ [8], [13]. Therefore, for the multi-class text classification problem, $\bar{c}_k$ is considered as an aggregation of several class. See Table III.

However, this leads to a loss of information on the distribution of terms in the different categories. In this table the terms t_j are more numerous in $\bar{c}_k$ because it is a combination of several classes. This leads to inappropriate weights because *D* and *B* are skewed with respect to A and *C* which reflect the exact distributions of $\bar{t}_j$ and $\bar{c}_k$ in c_k [8].

Thus, the term weighting methods based on the selection of characteristics are not optimal for the problem of classification of multi-class texts. Because the reflection of the discriminating power of terms belonging to $\bar{c}_k$ is false.

C. Supervised methods

By considering the deficiencies of TF-IDF and features selections schemes, several supervised term weighting schemes have been proposed [4]. Otherwise, weighting a term by using an information known by the classes.

To avoid bias in D and B for term weighting, an approach called the frequency of relevance (RF) is proposed in [8]. This approach calculates the discriminating power of terms using the ratio of A and C. Because these reflect the exact distributions of terms in c_k. However, like feature selection methods, RF weights a term based only on knowledge of the contingency table, despite the limitations listed above. So it's a non-optimal term weighting method for the multi-class case.

Thus, taking into account the defiance's enumerated above, new schemes that consider the distribution of the term in the different classes are proposed. Let us consider equation (6) in Table I, the relevance of a term in the category of the value of entropy. More the entropy is high, it appears in several categories, and less discriminating it is. However, the concentration of the feature is more important than its discriminating power is. Conversely, a term in the different categories has often-higher entropy.

With $p(t_j,c_k)=f(t_j,c_k)/f(c_k)$, where $f(t_j,c_k)$ denotes the frequency of term t_j in category c_k and $f(c_k)$ denotes the frequency sum of all terms in category c_k.

Example: *in Table II, the term "sky" has an entropy more higher than the term "sun", but "sun" has a higher discriminating power because it is specific to the category "positive".*

In order to overcome the shortcomings of the bi-class schemes, Chen and al. propose in table I equation (7), Inverse Gravity Moment TF-IGM [9] in order to explore both the contribution of terms in the classification and the provision of information in corpus.

Where

$$1 + \lambda \frac{f_{j1}}{\sum_{r=1}^{m} f_{jr} \times r} \quad (18)$$

denotes the *igm* based global weighting factor of term t_j in document d_i and $\lambda \in [5, 9]$ is an adjustable coefficient for keeping the relative balance between the global and the local factors in the weight of a term. Theis $igm(t_j)$ defined as follows:

$$\frac{f_{j1}}{\sum_{r=1}^{m} f_{jr} \times r} \quad (19)$$

Where the frequency f_{rj} ($r=1,2,..,m$) usually refers to the class-specific document frequency of the term and f_{j1} the maximal frequency of the term of the class m (sort in descending order). TF-IGM is a supervised term weighting approach. Because the global factor IGM depends only on known class information.

In addition to the schemes listed in Table I, others exist in the literature as TF-IDF-ICSDF [34] and others.

Like all weighting schemes of supervised terms discussed in this article, only known information about the distribution of terms in the different class categories is used to determine the global factor. By ignoring the semantic structure of the corpus, information on the contribution of terms such as similarity between term and between documents is lost in the determination of the global weight of terms.

The global factor of a term must be calculate taking into account the labels of the classes and the similarity of the terms in the documents that contain them. Because any similarity between documents is explained by the presence of many common terms between the documents, and any similarity between terms is explained by their common presence in a large number of documents.

IV. Conclusion

In this study, we examined the works on term weighting based on the vector space model (VSM). Consequently, two objectives were set in this study: the primary objective was to study the principle of the functioning of the vector-space model and its limits; and the secondary objective of qualitative analysis of term weighting methods to identify their boundaries.

The first allowed us to identify the different prerequisites for document representation. These prerequisites include (1) the construction of space, (2) the matrix representation of documents, (3) the hypothesis of "bag of words" in machine learning algorithms, (4) the limits of model. In addition, we have been able to identify the advantages of the VSM representation as well as its limitations in the domain of term weighting with the semantic methods proposed as an alternative by the researchers.

The second allowed us to explore the literature by inspecting for each VSM-based term weighting method its operating principle and these limits. Among the studies encountered, we have unsupervised schemas [3], supervised schemes based on functionality selection [5], and statistical methods [20]. The unsupervised schemes ignore class labels, whereas those supervised takes them into account. Their goal is to improve the performance of text classification algorithms. However, features selection schemes are limited in the multi-class classification (section III B).

Thus, we can conclude by stating that important problems are being addressed in the domain of term weighting research. And the proposed solutions are very advanced despite some limitations noticed. However, new solutions can still be proposed and applied in other domain. A number of issues have not been fully addressed in this paper. For example the problems related to the "semantic gap" i.e. the difference between the interpretation that a computer can automatically obtain from a text and the meaning of this same text for a human (category targeted by the text), the semantic weighting methods [15]; [16] [17]; [26], Word2Vec [18]; [19], and others.

Finally, in this study, various term weighting problems were discussed. The solutions to these problems have been examined namely the advantages and disadvantages. This allowed us to offer a new intuition of how a term should be weighted. By combining all of these elements and the limitations of our study, we hope that this paper will serve the domain community in their search for new ideas.

References

[1] X. D. Maosong, "Chinese Text Categorization Based on the Binary Weighting Model with Non-binary Smoothing," European Conference on Information Retrieval ECIR Advances in Information Retrieval pp 408-419, 2003.

[2] G. Salton, and C. Buckley, "Term weighting approaches in automatic text retrieval," Information Processing & Management, 24 (5), 513–523. 1988.

[3] Z. Deng, S. Tang, D. Yang, M. Zhang, L. Li and K. Xie, "A comparative study on feature weight in text categorization," Springer Berlin Heidelberg, Hangzhou, China, 2004.

[4] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," Proceedings of the 2003 ACM symposium on Applied computing. (pp 784-788).

[5] J. Yang, J. Wang, Z. Liu, Z. Qu, "A Term Weighting Scheme Based on the Measure of Relevance and Distinction for Text Categorization," International Conference on Advanced Computing Technologies and Applications (ICACTA-2015) (pp 13-22), 2015.

[6] G. Feng, H. Wang, T. Sun, and L. Zhang, "A Term Frequency Based Weighting Scheme Using Naïve Bayes for Text Classification," Journal of Computational and Theoretical Nanoscience (pp 319-326)· 2016.

[7] T. Wang, Y. Cai, H. Leung, Z. Cai and H. Min. Entropy-based Term Weighting Schemes for Text Categorization in VSM. 2015 IEEE 27th International Conference on Tools with Artificial Intelligence.

[8] T. Mori, M. Kikuchi, K. Yoshida. Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems. Journal of Natural Language Processing, 9(4):3—32. September 2001

[9] K. Chen, Z. Zhang, J. Long, H. Zhang, "Turning from TF-IDF to TF-IGM term weighting in text classification," In journal Expert Systems with Applications, vol. 66 Issue C, December, pp. 245-260. 2016.

[10] G. V. Cormack, G. Hidalgo, J. M., and P. Sánz, "Spam filtering for short messages," International Conference on Advanced Language Processing and Web Information Technology, IEEE Xplore, 2008.

[11] J. Geng, Y. Lu, W. Chen, Z. Qin, "An improved text categorization algorithm based on VSM," IEEE 17th International Conference on Computational Science and Engineering, 2014.

[12] H. Wu, X. Gu, Y. Gu. "Balancing between over-weighting and under-weighting in supervised term weighting," Information Processing and Management 53 (2017) 547–557, 2017.

[13] P. Karisani, M. Rahgozar, F. Oroumchian. "A query term re-weighting approach using document similarity," Information Processing & Management Volume 52, Issue 3, May, Pages 478-489. 2016.

[14] M. Haddoud, A. Mokhtari, T. Lecroq S. Abdeddaïm, "Combining supervised term-weighting metrics for SVM TC with extended term representation," Springer-Verlag London 2016 Knowl Inf Syst

[15] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proceedings of the ACL 2004.

[16] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," JASIS, 41(6):391–407, 1990.

[17] E. Gabrilovich, and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," 20th International Joint Conference on Artifical Intelligence, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space," CoRR, abs/1301.3781, 2013.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013.

[20] D. Kande, F.Camara, R. M. Marone, and S. Ndiaye, "Vector Space Model of Text Classification based on Inertia Contribution of Document, "4th International Conference on Frontiers of Educational Technologies, Moscow, Russian Federation —ACM New York, NY, USA, June 25 - 27, 2018.

[21] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min Redundancy. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, pp.1226-1238, Mar. 2005.

[22] H. Schutze, J. O. Pedersen and D. A. Hull, "A Comparison of Classifiers and Document Representations for the Routing Problem,"

[23] F. Fen, F and M. G. ohrab, "Class-indexing-based term weighting for automatic text classification," Information Sciences, 236, 109–125, 2013.

[24] G. Adeva, P. Atxa, U. Carrillo and A; Zengotitabengoa, "Automatic text classification to support systematic reviews in

medicine," Expert Systems with Applications, 41(4), 1498–1508, 2014.
[25] McTear, Michael and al, "The Conversational Interface," Springer International Publishing, 2016.
[26] Q. Luo, E. Chen and H. Xiong, "A semantic term weighting scheme for text categorization," Expert Systems with Applications, 2011.
[27] F. Debole, and F. Sebastiani, "Supervised term weighting for automated text categorization," In Proceedings of the ACM symposium on applied computing (pp. 784–788), 2003.
[28] Y. ang, , and J. O. Pedersen, "A comparative study on feature selection in TC," In Proceedings of the fourteenth international conference on machine learning (ICML'97) (pp. 412–420), 1997.
[29] Z. Deng, S. Tang, D. Yang, M. Zhang, L. Li and K. Xie, "A comparative study on feature weight in TC," In Advanced web technologies and applications, lecture notes in computer Science: vol. 3007 (pp. 588–597).Springer Berlin Heidelberg. Proceeding. 2004.
[30] M. LAN, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for ATC," IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (4), 721–735, 2009.
[31] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for ATC," Information Sciences, 236, 109–125, 2013.
[32] H. Altinçay, and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," Pattern Recognition Letters, 31 (11), 1310–1323, 2010.
[33] D. Wang, and H. Zhang, "Inverse-category-frequency based supervised term weighting schemes for text categorization. Journal of Information Science and Engineering, 29 (2), 209–225, 2013.
[34] Q. Luo, E. Chena and H. Xiong, "A semantic term weighting scheme for text categorization," Expert Systems with Applications 38 12708–12716, 2011.