Semantic Sensitive TF-IDF to Determine Word Relevance in Documents

Jalilifard A.; Caridá V.F.; Mansano A.F.; Cristo R.S.
Data Science Team - Digital Customer Service
Itaú Unibanco, São Paulo, Brazil
amir.jalilifard; vinicius.carida; alex.mansano; rogers.cristo
@itau-unibanco.com.br

***Abstract*—Keyword extraction has received an increasing attention as an important research topic which can lead to have advancements in diverse applications such as document context categorization, text indexing and document classification. In this paper we propose STF-IDF, a novel semantic method based on TF-IDF, for scoring word importance of informal documents in a corpus. A set of nearly four million documents from health-care social media was collected and was trained in order to draw semantic model and to find the word embeddings. Then, the features of semantic space were utilized to rearrange the original TF-IDF scores through an iterative solution so as to improve the moderate performance of this algorithm on informal texts. After testing the proposed method with 200 randomly chosen documents, our method managed to decrease the TF-IDF mean error rate by a factor of 50% and reaching the mean error of 13.7%, as opposed to 27.2% of the original TF-IDF.**

***Index Terms*—Semantic sensitive TF-IDF, Keyword extraction, word relevance, semantic similarity**

I. Introduction

In the information era when huge number of digital documents are gathered in a daily basis, going through documents and extracting the most relevant information, understanding the general concept and finding the other related documents is more necessary than ever. Keywords are several relevant words that provide a rich semantic information about a text for many natural language processing applications. Thereby, many researches have been carried out in order to extract the most relevant words from a text.

Some made use of the already-known supervised classification methods such as Support Vector Machine (SVM) and Naive Bayes [1] [2]. Although these supervised approaches methods provided good results, the need for training data, which often needs involving human resources, still remain a problem. Moreover, the word relevance score provided by a method like SVM may or may not be directly proportional to the importance of terms in a particular document. Therefore, an unsupervised method which provide local weights considering a class of documents is desirable.

TF-IDF is a numerical statistics that, by scoring the words in a text, indicates how important a word is in a document considering the corpus that document belongs to. This method was studied in several researches for keyword and word relevance extraction. Ramos and his colleagues [3] examined the result of applying TF-IDF in determining the word relevance in document queries and concluded that this simple method efficiently classifies relevant words. Li et al. [4] applied TF-IDF for keyword extraction in Chinese texts based on analyzing linguistic characteristics of documents and providing several strategies including uni-, bi- and tri-gram extraction, new word finding and refinement. Chung et al. [5] proposed a probabilistic model based on TF-IDF which makes local relevance decisions for each location in a document and combines these local relevant decisions into a document wide relevance decision. Lee and colleagues [6] presented several variants of conventional TF-IDF for a more effective keyword extraction and topic summarization. They used cross-domain comparison for removing meaningless or irrelevant words.

Although the aforementioned methods improved the performance of conventional TF-IDF by providing probabilistic solutions or the use of multi-strategies, they are highly dependent on the original TF-IDF idea which is giving more weight to the words with high local and less global probability. This consideration specially fails when it comes to finding word relevance in informal documents , which are important sources of information in the era of social networks [7]. As an example, if a text contains informal words related to a specific ethnic communities, or words that have been used in a specific period of time, but not in the whole corpus, due to some changes in cultural expressions, both conventional TF-IDF and the related methods that attempt to improve its performance fail to find relevant words with high accuracy. Another example is informal conversations regarding formal topics like medical communities that provide rich information for users. Censuring the general semantic context, TF-IDF fails to detect the context-sensitive content which plays an important role in informal texts [8].

In this paper, we propose STF-IDF, a novel semantic sensitive method based on the conventional TF-IDF. The key idea is readjusting the conventional TF-IDF scores based on the semantic representation of most relevant words. Thereby, we assume that if a set of terms is considered important by TF-IDF, all the semantically similar words related to this set should be considered more important than those ones with less semantic relevance to the context. The next section explains the theoretical basis of the proposed method. The results and discussion are presented in the last section.

II. Materials and method

In this section we explain the materials and the mathematical definition of our proposed method. We start with the data acquisition and then we explain how our algorithm tries to improve the conventional TF-IDF through a finite numbers of iterations.

Data of nearly four million pages of online medical communities was gathered and after being pre-processed (i.e removing stopwords, punctuation, etc.), they were fed to word2vec [9] in order to learn the semantic space and words' distribution. Having the semantic distribution of terms of the corpus, our algorithm generates the word relevance score as following:

$$S_{wj}^{(k)} = S_{wj}^{(k-1)} \times \frac{1}{1 + \|e(w_j)\| \left\|\overline{e(w)}\right\| cos(\Theta)} \tag{1}$$

where $S_{wj}^{(k)}$ is the vector of word scores in Kth iteration and the initial scores are calculated using the conventional TF-IDF:

$$S_{wj}^{(0)} = P(W_j)_d * Log(P(W_j)_c) = TFIDF_{Wj} \tag{2}$$

and $\overline{e(w)}$ is the weighted expected value of the first $\lfloor\sqrt{n}\rfloor$ most relevant words in the previous iteration and is calculated as follow:

$$\overline{e(w)} = \frac{1}{\lfloor\sqrt{n}\rfloor} \sum_{i=1}^{\lfloor\sqrt{n}\rfloor} \left(\frac{1}{1 - \frac{S_{wj}^{(k-1)}}{\sum_{j=1}^{n} S_{wj}^{(k-1)}}} \times e(w_j)\right) \tag{3}$$

The algorithm is initiated with conventional TF-IDF scores. In each iteration, first, the mean embedding of $\lfloor\sqrt{n}\rfloor$ most relevant words from previous iteration are selected and the weighted mean embedding of them is calculated. The idea is that words with higher scores represent the context more than those which are less relevant. In order to calculate the weighted mean embedding, for each word its score in the previous iteration is divided by all the scores in order to get a number in the range [0,1]. This weight then is converted to a number greater than 1 and is considered the weight by which each word pushes the mean embedding toward itself. Afterwards, the previous scores are recalculated by being multiplied on the cosine distance of the word and the mean embedding. The idea is to repeatedly replace the words that have poor representation of general text context with those that are more related to the document context. The new scores are then passed to the next iteration and the scores are rearranged so that the words with better representation are moved toward the top of ranking.

By considering the embedding of the first $\lfloor\sqrt{n}\rfloor$ most relevant words as a multivariate unknown probability distribution, the less-relevant words are those instances which increase the variance of the distribution. Therefore, the goal is to decrease the variance by constantly replacing the outliers with those words that are semantically more related to the most important words and move the expected value of embeddings toward the value that perfectly matches the context. First, we prove that in each iteration the variance of the distribution from the mean in the previous iteration is decreased. Then, it is shown that in each iteration the mean value of set moves toward the mean of ideal distribution until it converges.

We define the mean embedding of the observed data in each iteration and the unknown distribution that best fits the context with $\mu^{(i)}$, and μ, respectively, as following:

$$\mu^{(i)} = \frac{X_1^{(i)} + X_2^{(i)} + X_3^{(i)} + ... + X_{m-1}^{(i)} + X_m^{(i)}}{m} \tag{4}$$

where $m = \lfloor\sqrt{n}\rfloor$ and $X_m^{(i)}$ represents the embedding of mth word of the set in the ith iteration and $\mu^{(0)}$ as the mean of the initial distribution generated by conventional TF-IDF, and $X_m^{(i)} \neq X_m^{(i-1)}$ if a word is substituted.

In each iteration either the set maintains the current members or some of them are replaced with words that have a less cosine distance. Multiplying the word score from previous iteration $S_{wj}^{(k-1)}$ to the inverse of cosine distance guarantees that a word with higher cosine distance from the expected value of embeddings is decreased more than a word with less distance. A factor of 1 is added to the equation in order to eliminate the reverse effect of distances lesser than 1.

In the simplest case, we assume that in each iteration one word is replaced with another. As it was explained, the equation (1) guarantees that the new word has less cosine distance to the expected value of embeddings than the replaced word. As a result:

$$\frac{1}{m}\sum_{k=1}^{m}(X_k^{(i)} - \mu^{(i-1)})^2 < \frac{1}{m}\sum_{k=1}^{m}(X_k^{(i-1)} - \mu^{(i-1)})^2$$
$$\Rightarrow Var(X^{(i)}) < Var(X^{(i-1)})$$

$$\Rightarrow \lim_{i\to\infty} \frac{Var(X^{(i-1)})}{Var(X^{(i)})} = 1 \tag{5}$$

Let's assume that the real context mean is bigger than the initially estimated mean by TF-IDF and one distant member per iteration is replaced with a closer one, say X_m in ith iteration, Then:

$$X_m^{(i)} > X_m^{(i-1)} \Rightarrow \frac{X_1^{(i-1)} + X_2^{(i-1)} + X_3^{(i-1)} + ... + X_{m-1}^{(i-1)} + X_m^{(i)}}{m} > \frac{X_1^{(i-1)} + X_2^{(i-1)} + X_3^{(i-1)} + ... + X_{m-1}^{(i-1)} + X_m^{(i-1)}}{m}$$

and consequently:

$$\mu^{(i)} > \mu^{(i-1)} > ... > \mu^{(1)} > \mu^{(0)} \Rightarrow (\mu - \mu^{(i)}) < (\mu - \mu^{(i-1)}) < ...$$
$$< (\mu - \mu^{(1)}) < (\mu - \mu^{(0)})$$

approaching the correct context's expected embedding μ in each iteration. In case of $\mu < \mu^{(0)}$:

$$\mu^{(i)} < \mu^{(i-1)} < ... < \mu^{(1)} < \mu^{(0)}$$
$$\Rightarrow (\mu^{(i)} - \mu) < (\mu^{(i-1)} - \mu) < ... < (\mu^{(1)} - \mu) < (\mu^{(0)} - \mu)$$

Finally, from:

$$\lim_{i\to\infty} \frac{X_k^{(i-1)}}{X_k^{(i)}} = 1 \Rightarrow$$
$$\lim_{i\to\infty} \mid X_k^{(i)} - X_k^{(i-1)} \mid = 0 \Rightarrow \lim_{i\to\infty} \mid \mu^{(i)} - \mu^{(i-1)} \mid = 0$$

and the condition of almost surely convergence is met after enough number of iterations:

$$\left| \mu - \mu^{(i)} \right| \leq \varepsilon \tag{6}$$

There are two ways that the proposed method can fail in improving the word rank. First, if the mth and $(m+1)$th words have exactly same score. In this case, the choice of words may change the expected embedding value and consequently lead to a totally different approximation of the document context. In order to solve this problem, the algorithm may simply check the score of mth and the $(m+1)$th words and in case of encountering the same scores, it can enter the $(m+1)$th word into the set as well. Second, if before starting the refining process the conventional TF-IDF produces scores with very high error rate, STF-IDF fails to find the correct context. Nevertheless, our results show that such a high error rate is not a common case and a moderate performance of TF-IDF is enough for the current method to produce significantly good results.

III. Problem Statement

xxxxxxx

IV. Method

xxxxxxxxx

V. Results and Discussion

The algorithm was tested for 160 randomly chosen informal medical documents. For both conventional TF-IDF and STF-IDF, the scores were evaluated with human annotated labels. Since the precision of word importance can be subjective, here, we define and analyze the ranking error which measures the number of words were put between the first $\lfloor \sqrt{n} \rfloor$ most relevant terms while they have the least relevance based on human evaluation.

As it is seen in Fig. 1, by replacing better words in the ranking table, STF-IDF has less error rate in comparison with TF-IDF. Since STF-IDF is initially constructed upon the TF-IDF scores, in the rare cases (less than 5% of times) when TF-IDF has abnormally big error rate, STF-IDF performs worse than TF-IDF.

We measured the error rate of STF-IDF against the original TF-IDF. As it is show in Fig. 2, our method improved the

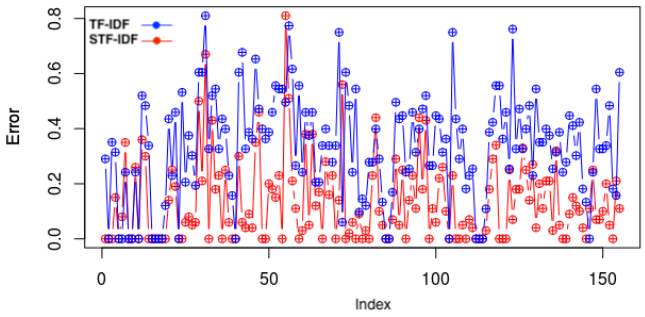

Figure 1. The error rate of STF-IDF in comparison with conventional TF-IDF for each document

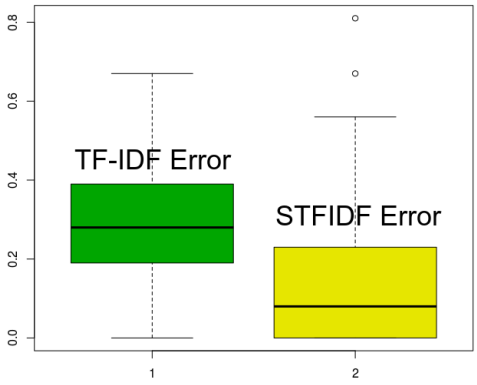

Figure 2. The boxplot of ranking error for STF-IDF and conventional TF-IDF

error rate by more than 50%, decreasing the error rate of TF-IDF from 27.68% to 13%. Among the rankings generated for 160 documents, in 50% of times the error rate of STF-IDF is significantly less ranging from 0% to 13% as opposed to high error of conventional TF-IDF ranging from 20% to 30% for 50% of tested documents.

VI. Conclusion

in this study we proposed a novel method based on semantically weighted TF-IDF scores for finding word relevance between a collection of documents. Textual data of nearly 4 million online medical communities were gathered and preprocessed. Afterwards, the corpus was fed to word2vec algorithm in order to generate word embedding. Initially, the words were ranked by conventional TF-IDF algorithm. Then these scores were repeatedly modified based on a semantic weight of each word proportional to the cosine distance of the world and the expected value of embedding of a set of most relevant words in each iteration. The algorithm stops when it reaches a predefined threshold which is a measure of dislocation of mean embedding distribution. Our results show

a significant decrease in error rate when STF-IDF is utilized. The future works will be focused on the convergence proof of the algorithm as well as replacing automatic tests with human-involved evaluation.

VII. Conflict of interest

The current method was proposed and tested by a group of data scientists from Itaú Unibanco. Any opinions, findings, and conclusions expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policy or position of Itaú Unibanco.

References

[1] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *International Conference on Web-Age Information Management*. Springer, 2006, pp. 85–96.

[2] Y. Uzun, "Keyword extraction using naive bayes," in *Bilkent University, Department of Computer Science, Turkey www. cs. bilkent. edu. tr/˜ guvenir/courses/CS550/Workshop/Yasin_Uzun. pdf*, 2005.

[3] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.

[4] J. Li, K. Zhang *et al.*, "Keyword extraction based on tf/idf for chinese news document," *Wuhan University Journal of Natural Sciences*, vol. 12, no. 5, pp. 917–921, 2007.

[5] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 13, 2008.

[6] S. Lee and H.-j. Kim, "News keyword extraction for topic tracking," in *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, vol. 2. IEEE, 2008, pp. 554–559.

[7] M. R. Morris, J. Teevan, and K. Panovich, "A comparison of information seeking using search engines and social networks." *ICWSM*, vol. 10, pp. 23–26, 2010.

[8] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional lstm networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.