PERFORMANCE EVALUATION OF INFORMATION RETRIEVAL SYSTEM USING VECTOR SPACE MODEL: A COMPARATIVE ANALYSIS

Omar Al-rassam [1]
[1] *Department of Mathematics,*
Faculty of Science and Health
Koya University, Koya KOY45,
Kurdistan Region - F.R. Iraq
omar.alrassam@koyauniversity.org

Miran Hama Saeed Mohammed Amin [2]
[2] Department of Software Engineering,
Faculty of Engineering
Koya University, Koya KOY45, Kurdistan
Region - F.R. Iraq
miran.hamahsaeed@koyauniversity.org

Zhenar Shaho Faeq[3]
[3] *Department of Software Engineering,*
Faculty of Engineering
Koya University, Koya KOY45,
Kurdistan Region - F.R. Iraq
Zhenar.shaho@koyauniversity.org

***Abstract* - The increasing use of the internet has created a vast amount of digital information and it is expanding extremely fast. Therefore, Information retrieval becomes a challenging task to fetch relevant information for users. The aim of this paper was to examine and evaluate the performance of the Information retrieval system through eight experiments to test all the features that can be used in a vector space model. These experiments were compared to show the best and the worst implemented features. The features are represented by applying (tf.idf, stop words, stemming), (tf.idf, No- stop words, stemming), (tf.idf, No- stop words, No-stemming), (tf.idf, stop words, No-stemming), (tf, stop words, stemming), (tf, No- stop words, stemming), (tf, No- stop words, No-stemming), (tf, stop words, No-stemming). Results showed that using stop words, stemming approach, and tf.idf improve the performance of the system. However, when tf was used without using stop words and stemming approaches the performance of the system is declined. In addition, results showed that stop words have a significant effect on the system while the stemming approach has no noticeable effect particularly with tf.**

Index Terms - Information Retrieval, Vector space model, inverse document frequency, Term frequency, stemming.

I. INTRODUCTION

Several algorithms and techniques have been developed for data mining and information retrieval. These techniques and algorithms are evaluated based on accuracy and performance. Ranking the most relevant documents at the top of the system output is one of the challenges of almost all Information retrieval systems [1]. The most classical used technique in this regard is Vector Space Model (VSM) [2]. Different procedures need to be applied to the data before applying VSM such as stop words, stemming and term weighting.

In order to evaluate the performance of Information retrieval systems, [3] have proposed visual and scalar evaluations methods to evaluate the overall performance of an information retrieval system. These evaluations are performed by precision, recall and F measure parameters. Furthermore, [4] claims that user satisfaction should be used as a criterion for measuring information system effectiveness besides other factors such as precision and recall. On the other hand, [5] compared the performance of Vector Space Model (VSM) and Latent Semantic Indexing (LSI), results showed that LSI performs better in most cases compared to VSM and it retrieves more relevant documents.

In addition, [6][7] argues that traditional measures such as precision and recall, are only based on binary relevance and cannot distinguish between different levels of relevance while the normalized distance performance measure provides the best performance in terms of document ranking and relevance of documents.

The aim of this paper is to examine and evaluate the performance of an Information retrieval system through eight experiments to test all the features that can be used in a vector space model. These experiments are compared to show the best and the worst implemented features. The features are represented by applying (tf.idf, stop words, stemming), (tf.idf, No- stop words, stemming), (tf.idf, No- stop words, No-stemming), (tf.idf, stop words, No-stemming), (tf, stop words, stemming), (tf, No- stop words, stemming), (tf, No- stop words, No-stemming), (tf, stop words, No-stemming).

II. RELATED STUDIES

Due to the dramatic increase of global internet usage, an enormous amount of text materials is produced in almost every field. Therefore, the task of text retrieval and classification remains a big challenge. One of the main application areas in information retrieval and text mining is Text classification or document categorization. Various machine learning algorithms and models have been applied in document classification, information retrieval and other text mining applications. Naive Bayes classifier and vector Space Model are among the oldest approaches for information retrieval and text classification.

The major problems of all Information Retrieval systems are fetching some irrelevant information together with the relevant one and not be capable of retrieving all relevant documents.

Boolean, Vector Space and Probabilistic models are the most widely used information retrieval models [8][9][10]. Vector space model can be considered as the most effective, influential

Doi: 10.25195/ijci.v47i2.332

and simple model for information retrieval systems and gives nearly exact matching results [10] [11].

Similarity values between queries and documents can be calculated using three different approaches in the Vector space model. Tf.idf and the normalization model provide better results compared to tf. Considering long documents, all the approaches can be used since it contains more appearance of the query terms [12]. Among the available frameworks, [8] argues that considering term-frequency, inverse document frequency measures, the vector space model can be used to achieve the ultimate relevancy in retrieving documents in information retrieval.

Moreover, [13] have proposed a new similarity measure that calculates the dissimilarity coefficient between the two instances and then the similarity value is obtained. This will improve the classification performance and accuracy of classifiers based on vector space models that are used in text categorization and information retrieval.

[9] implemented a vector space model for document retrieval system using open source technology and their results were adequate. It is suggested that lemmatization, spell-checking, and synonym expansion can be used to improve recall and precision.

Information retrieval system has been developed for languages other than English. [14] used a vector space model to develop a text retrieval system for Afaan Oromo language that can handle indexing and searching. It is argued that the performance the system can be increased if a stemming algorithm is improved.

Due to the effectiveness of the vector space model, it has been applied in different applications of information retrieval systems such as question answering systems [15], document summarization [16][17] and Image retrieval [18][19]

III. METHODOLOGY

A. *Pre-processing and preparing data*

Preprocessing data is a very important phase to prepare the data set before applying a vector space model. It is represented by loading data, Stopwords, tokenization, normalization, stemming, index data structure (term weighting), storing the indexed data into a file. The tested data file includes a collection of documents that has almost 3500 records of publications.

For tokenizing the input a simple regular expression was used to find the maximum alphabetic sequences and then normalizing the text to a form of lowercase. Stopwords are used to exclude less useful words.

Moreover, the stemming technique is used to reduce inflected words; in this Paper Porter stemmer is used to minimize the number of terms that might affect the performance of the system through stripping words of any derivations, for example, studying, studied, study’s, will become study which is the basic form of this term.

Now words are ready to be given weight. Two techniques were used for giving weight to the words; the first one is to count the number of a certain word that occurred in the document, in this case, the high weight is mapped to a word that has a high frequency. While the second technique is the inverse word frequency which gives a large weight to rare words that occur in documents.

The first technique does not differentiate among the importance of words, for example, articles normally appear in documents frequently while these articles do not provide real information such as (a, an, the,...etc.). On the other hand, the second technique gives importance to a seldom occurring word in the document and these words might have a good indicator to a certain topic, for instance, democracy, dictatorship indicate political topics. Finally, the revised information is stored in a file to be used later.

B. *Vector Space Model*

Vector space model (VSM) is used for information retrieval by representing documents as “bags of words”. Words in the documents are dots in multi-dimensional vector space. Not only documents are represented in vector space, but also queries are represented in vector space and each query is considered as a point in high-dimensional vector space. Each term (in documents/ queries) forms a dimension. Now the similarity between the vectors of the query and document is measured. The highest score of similarity the document has is the most relevant to the query. Figure (1) illustrates an example of how vector space works.

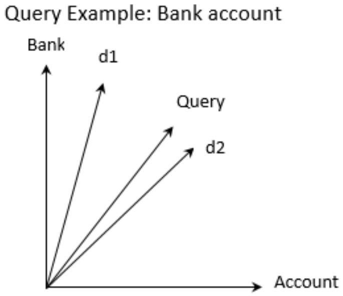


Fig. 1 shows the relationship between the query and the document in vector space

In the above example, the most similar document to the query is d2. The similarity between vectors is measured by calculating the cosine of the angle between two vectors. The following equation is used for calculating the cosine between angles of two vectors

$$\vec{x}\ and\ \vec{y}:\ \ cos\, cos\left(\vec{x},\vec{y}\right) = \frac{\vec{x}.\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} \square x_i y_i}{\sqrt{\sum_{i=1}^{n} \square x_i^2}\sqrt{\sum_{i=1}^{n} \square y_i^2}}$$

The output of this equation ranges from (-1, 0, 1):
1: the vectors in the same directions
0: orthogonal vectors
-1: the vectors in opposite directions

The following block diagram the process of the system:

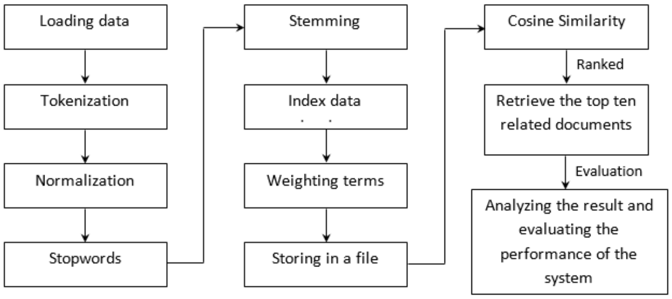


Fig. 2 block diagram of the system

IV. System Evaluation

The system was evaluated according to eight experiments as following

1. (tf.idf, stop words, stemming)
2. (tf.idf, No- stop words, stemming)
3. (tf.idf, No- stop words, No-stemming)
4. (tf.idf, stop words, No-stemming)
5. (tf, stop words, stemming)
6. (tf, No- stop words, stemming)
7. (tf, No- stop words, No-stemming)
8. (tf, stop words, No-stemming)

Figure 3 presents the performance of the system's results under different configurations such as using/or not using stop list, applying porter algorithm (stemming)/ or not, and using tf or tf.idf. The performance of the system is determined by the value of precision, recall and F-measure.

In general, figure 3 shows that the performance of the system was improved when the term weighting technique (tf.idf) was applied. In particular, the best performance of the system was shown when (tf.idf, stop words, stemming) were applied, with precision 0.27, recall 0.22 and f-measure 0.24. However, the performance was negatively affected by depending on other features such as using/ not using stop words and stemming. For example, the result of (tf.idf, No- stop words, No-stemming) experiment showed a decline in system performance because stop words and stemming were not used, with precision 0.17, recall 0.14 and f-measure 0.15.

In addition, when the second technique for term weighting (tf) was used in the experiments, the best result was roughly equal to the bad performance of the system when (tf.idf, No-stop words, No-stemming) was used, with precision 0.18, recall 0.15, and f-measure 0.16. Moreover, the worst system's performance among all the experiments was recorded when (tf, No- stop words, No-stemming) was implemented with precision 0.04, recall 0.03, and f-measure 0.04.

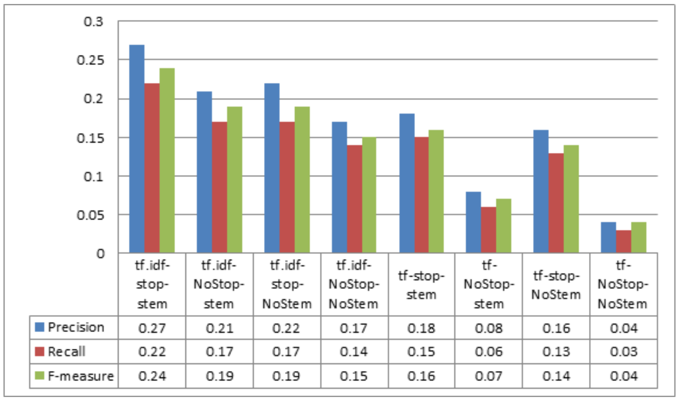
Furthermore, the feature of not using stop words in both experiments (tf, No- stop words, stemming) and (tf, No- stop words, No-stemming) had a negative effect on the performance of the system remarkably. However, the stemming approach did not have a huge effect on the performance. Finally, in both experiments (tf, stop words, stemming) and (tf, stop words, No-stemming) the value of precision, recall and f-measure were almost the same with (0.18, 0.15, 0.14) respectively.

The accuracy of the system is computed through applying precision, recall and f-measure. As well known, the closer of these metrics to one indicates the higher accuracy, in case the objective is to improve the accuracy of the system.

However, in this paper the main objective is to determine the mentioned features that might affect the performance of the system. These features were examined under eight experimental conditions, in each experiment only the top ten related results to a query were considered. Therefore, the values of the precision, recall, f-measure were shown in figure (3) close to 0.3. The values of these metrics will be close to one if all the results to a query are taken (not only the top ten), and the reason for taking only the top ten results was to make the system work fast and to avoid more calculations, particularly the system work under eight experiments.

Moreover, the difference among the outputs of the experiments are the same in either cases whether the top ten results to a query is taken or all.

Considering the limitations of the study, it is suggested that a bigger sample size can be used to test the performance of the system. In addition, this system can be applied to another language and compare the results to the current system.



	tf.idf-stop-stem	tf.idf-NoStop-stem	tf.idf-stop-NoStem	tf.idf-NoStop-NoStem	tf-stop-stem	tf-NoStop-stem	tf-stop-NoStem	tf-NoStop-NoStem
🟦 Precision	0.27	0.21	0.22	0.17	0.18	0.08	0.16	0.04
🟥 Recall	0.22	0.17	0.17	0.14	0.15	0.06	0.13	0.03
🟩 F-measure	0.24	0.19	0.19	0.15	0.16	0.07	0.14	0.04

Fig. 3 Result of the Experiments

Conclusion and Future work

Information retrieval is one of the most important topics in the field of computational linguistics. Therefore huge numbers of researches have been conducted in order to improve the accuracy and performance of the system. Vector space model is one of the most popular algorithms for information retrieval that have been used by many researchers. The aim of this paper was to examine the performance of Vector space model under different conditions. The result showed that using a stop-list collection and stemming raise the performance of the information retrieval system. In addition, using the idf approach raises the performance of the information retrieval system compared to the tf approach. The result of this paper can be further optimized in the future by taking 2- gram (bigram) of the input data set to be represented in one vector to compare the result with the current work. In addition, using a different data set (huge data set or different language) and comparing the result with the output of this article.

REFERENCES

[1] Clough, P. Sanderson, M. (2013) Evaluating the performance of information retrieval systems using test collections. Information Research, 18 (2). ISSN 1368-1613

[2] Ogheneovo E,E. Japheth R. B. (2016) Application of Vector Space Model to Query Ranking and Information Retrieval. International Journal of Advanced Research in Computer Science and Software Engineering 6(5), pp. 42-47H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[3] Zuva, K, Zuva, T. (2012) EVALUATION OF INFORMATION RETRIEVAL SYSTEMS. International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 3

[4] Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. Journal of the American Society for Information Science and Technology., 61, 859-868.

[5] Chawla, I., Singh, S. K. (2013). Performance evaluation of VSM and LSI models to determine bug reports similarity. 2013 Sixth International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2013.6612223

[6] Zhou B., Yao Y. (2008) Evaluating Information Retrieval System Performance Based on Multi-grade Relevance. In: An A., Matwin S., Raś Z.W., Ślęzak D. (eds) Foundations of Intelligent Systems. ISMIS 2008. Lecture Notes in Computer Science, vol 4994. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68123-6_46

[7] Mizzaro, S. (2001) A New Measure of Retrieval Effectiveness (Or: What's Wrong with Precision and Recalls). In: International Workshop on Information Retrieval, pp. 43–52

[8] Manwar, A., Mahalle, H.S., Chinchkhede, K.D., Chavan, V., & Porwal, S. (2012). A VECTOR SPACE MODEL FOR INFORMATION RETRIEVAL: A MATLAB APPROACH.Indian Journal of Computer Science and Engineering. vol. 3, no. 2, pp. 222–229.

[9] N. Jamil, N. A. Jamaludin, N. A. Rahman and N. Sabari, 2011, "Implementation of vector-space online document retrieval system using open source technology," 2011 IEEE Conference on Open Systems, pp. 395-399, doi: 10.1109/ICOS.2011.6079228.

[10] Yogish, D., & Hegadi, R. (2019). Variants of Term Frequency and Inverse Document Frequency of Vector Space Model for Effective Document Ranking In Information Retrieval.International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-7

[11] Zhao Y., Shi X. (2012) The Application of Vector Space Model in the Information Retrieval System. In: Zhang W. (eds) Software Engineering and Knowledge Engineering: Theory and Practice. Advances in Intelligent and Soft Computing, vol 162. Springer, Berlin, Heidelberg.https://doi.org/10.1007/978-3-642-29455-6_6

[12] Singh, J., & Dwivedi, S. (2012). Analysis of Vector Space Model in Information Retrieval.National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012 Proceedings published by International Journal of Computer Applications® (IJCA)

[13] Eminagaoglu, M., 2020. A new similarity measure for vector space models in text classification and information retrieval. Journal of Information Science, p.016555152096805.

[14] Bakala, N., 2019, Information Retrieval System By Using Vector Space Model, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH. VOLUME 8, ISSUE 10

[15] Jovita, Linda, Hartawan, A., Suhartono, D. (2015). Using Vector Space Model in Question Answering System. Procedia Computer Science, 59, pp.305-311.DOI: https://doi.org/10.1016/j.procs.2015.07.570

[16] Gupta, N., Saxena, P. , Gupta, J. (2013). Document summarisation based on sentence ranking using vector space model. International Journal of Data Mining, Modelling and Management, 5(4), p.380. DOI: 10.1504/IJDMMM.2013.057680

[17] Belwal, R. C., Rai, S., Gupta, A. (2021). Text summarization using topic-based vector space model and semantic measure. Information Processing & Management, 58(3), 102536. https://doi.org/10.1016/j.ipm.2021.102536

[18] Martinet, J., Chiaramella, Y., Mulhem, P. (2011). A relational vector space model using an advanced weighting scheme for image retrieval. Information Processing & Management, 47(3), 391–414. https://doi.org/10.1016/j.ipm.2010.10.003

[19] Karamti, H., Tmar, M., Gargouri, F. (2017). A new vector space model for image retrieval. Procedia Computer Science, 112, 771–779. https://doi.org/10.1016/j.procs.2017.08.202