Journal of Engineering and Applied Sciences 14 (10): 3378-3382, 2019
ISSN: 1816-949X

A Comparative Analysis of TF-IDF, LSI and LDA in Semantic Information Retrieval Approach for Paper-Reviewer Assignment

[1, 2]A. Ayodele Adebiyi, [2]Olawole Ogunleye, [1, 2]O. Marion Adebiyi and [3]J. Olatunji Okesola
[1]Department of Computer Science, Landmark University, Omu-Aran, Nigeria
[2]Department of Computer and Information Science, Covenant University, Ota, Nigeria
[3]Department of Computational Sciences, First Technical University, Ibadan, Nigeria

Abstract: The intelligent task of semantically assigning a paper to a reviewer with respect to his knowledge domain remains a challenging task in academic conferences. From literature, a number of automated reviewer assignment systems have been presented which are based on distributional semantic models such as Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) have been used to capture semantics. Thus, this study presents the comparative study of the three models based on their derived suitability scores between a paper meant for review and a reviewer's representation papers. From the experimental results obtained, it shows that TF-IDF outperformed the accuracy level of the other two models by a substantial margin.

Key words: Reviewer-assignment, latent semantic indexing, latent dirichlet allocation, term frequency-inverse document frequency, semantically, academic conferences

INTRODUCTION

Yearly academic conferences receive a lot of paper submissions which must be reviewed by an expert before it can be accepted for publication. To achieve the objective of a fair and accurate review of papers, the papers should be assigned with respect to the area of expertise of the experts. The automation of this intelligent task saves management time that could have been otherwise spent on manually sorting out the suitability of a paper being reviewed by a reviewer.

The approach to automating Paper Reviewer Assignment (PRA) can be undertaken as an information retrieval problem where a submitted paper is used as the query and each reviewer's set of published works typifies the document representations to be matched (Long *et al*., 2013). The information retrieval model seeks to compute a matching score which represents the relevance between the paper and the reviewer which in turn determines the suitability of the reviewer to review the queried paper (Li and Hou, 2016).

Drawing from literature, the state of the art information retrieval models used in PRA includes: Term Frequency-Inverse Document Frequency (TF-IDF), Latent Semantic Indexing (LSI), Latent Dirichlet Analysis (LDA) (Charlin and Zemel, 2013).

TF-IDF is a model for document representation that is often used in information retrieval. It is a model that evaluates how important a word is to a document. It weights the important words increasingly based on how frequently they appear in the document but decreases the weight proportionally as it occurs in other documents. TF-IDF can represent a document well by removing stop words from the documents. It is being used for text summarization and text categorization.

Latent Semantic Indexing (LSI) is a popular information retrieval method that uses linear algebraic indexing method to produce low dimensional representations by word co-occurrence. LSI uses a vector model to build a matrix of word co-occurrences. It identifies the position on a vector space where each term and a document in a collection are positioned. It is based on the hypothesis that words that are semantically similar will cluster together. It utilizes the Singular Value Decomposition (SVD) algorithm to create a denser matrix that approximately models the original document.

Latent Dirichlet Analysis (LDA) is a probabilistic topic model that generates topics based on word occurrences from a corpus or set of documents (Bengio *et al*., 2003). It assumes documents are a blend of several topics and that each word in the document can be grouped under the document's topics. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set. LDA is an unsupervised language model that transforms words from bag of words counts into continuous representative matrix.

A common fact about the mentioned approaches is that they are topic models. Bag of Words (BOW) Model

Corresponding Author: A. Ayodele Adebiyi, Department of Computer Science, Landmark University, Omu-Aran, Nigeria

view a document such as the paper and reviewer's expertise representation as a set of terms where the frequency of each term is significant but the ordering of each of the terms is ignored (Manning *et al*., 2008).

This study presents the results of a comparative study between TF-IDF, LSI and LDA which are the state of the art models used in information retrieval.

MATERIALS AND METHODS

This study describes the approach for paper-reviewer assignment using TF-IDF, LSI and LDA Models. It presents the different phases contained in our methodology in determining the semantic similarity scores. Finally, the scores from each of the models are compared.

Data preparation: The dataset used for the paper-reviewer assignment was curated from a typical large conference, Neural Information Processing Systems (NIPS) to test our model. The dataset includes published papers from 1985-2015 from their website to provide domain knowledge. The 2016 set of papers from the NIPS dataset was used as the query papers for assignment which were about 562.The 2016 reviewer's list on the NIPS website which includes about 100 area chairs was used as the reviewers. In modeling the reviewer's knowledge domain, we concatenated each of the publication of a reviewer to form a corpus that would represent the reviewer. The downloaded the paper in pdf format from their Google scholar profile page and then used native tools such as pdftotext to extract the text from the downloaded files and concatenated them to form a single corpus.

Data pre-processing: The dataset were obtained and pre-processed which includes the removal of unwanted characters, tokenization, elimination of numbers and punctuation, removal of stopwords of the datasets and lemmatization were performed. This produced a transformed list of documents from the pre-processing stage serves as input in the next phase of computing the similarities by the models.

Similarity computation: After building the reviewer's corpus, the 2016 paper set and the domain papers needed for training to develop the domain space, the text was vectorized for each of the reviewers and for the papers. Then the cosine distance was determined which is used as the semantic similarity between the reviewer and the paper. This was done for the 89 reviewers to the 562 papers for each of the Models: TF-IDF, LSI and LDA:

$$\text{Cos}\theta = \frac{\sum_{k=1}^{n} u_k v_k}{\sqrt{\sum_{k=1}^{n} u_k^2}\sqrt{\sum_{k=1}^{n} v_k^2}} \quad (1)$$

Assignment optimization: In optimizing the assignment process of submitted papers to the most suitable reviewers, integer linear programming formulation was used as presented by Taylor (2008) with an objective of maximizing the overall sum of suitability scores globally. This is subject to constraints that each submitted paper should be assigned to no more than a certain number of reviewers and no reviewer should be assigned more than a maximum workload of submitted papers using the mathematical model as:

$$\max \sum_{r}\sum_{p} S_{rp}\alpha_{rp} \quad (2)$$

s.t.

$$\sum_{p} \alpha_{rp} \leq W_r^{Max}, \forall_r \quad (3)$$

$$\sum_{p} \alpha_{rp} \geq W_r^{Min}, \forall_r \quad (4)$$

$$\sum_{p} \alpha_{rp} = R_p^{Min}, \forall_p \quad (5)$$

$$\sum_{p} \alpha_{rp} \in \{0, 1\}. \forall_r \ \forall_p \quad (6)$$

Evaluation: Firstly, evaluation was performed to check for the model that performed better in terms of accuracy in predicting the best reviewer of a paper. To achieve this, the accuracy of the results derived from analyzing the Models: TF-IDF, LSI and LDA. We conducted two experiments as follows:

Experiment 1: The evaluation is based on using 5 reviewers and 5 randomly selected papers from NIPS 2016 papers. The test was set up in the following way as used by Young etc. in 2012.

Each reviewer has at least 30 files to represent his expertise. For each reviewer, we removed 5 papers from the dataset to create a test set meant for review which will be divided into 5 different tests. The results were evaluated based on observation that in real life an author cannot review their own paper but theoretically should be the best-qualified reviewer. Therefore, if the paper was assigned its researcher as a reviewer the assignment was considered correct. The Accuracy, $A(x, p_i)$, of the model is tested such that: if the researcher of paper, p_i appears in the top x reviewers then $A(x, p_i) = 1$. If the researcher of

paper, p_i is not in the top x reviewers $A(x, p_i) = 0$. The graphs following shows each model's Average $A(x, p_i)$ for each of the p_i's averaged over the five tests on the y-axis. The x values 1-5 are shown on the x-axis.

Experiment 2: In a second test, we used a dataset containing 51 submitted papers and 5 reviewers to analyze the mean rankings of the models over time as documented by Li and Hou (2016).

RESULTS AND DISCUSSION

From the evaluation we have been able to obtain a satisfactory comparison of the different models. From each model the average accuracy in ranking values were obtained which is presented in Table 1. Table 1 presents the total scores obtained from the evaluation are displayed as. Below is the line plot depicting the average accuracy ranking values of each of the models (Fig. 1).

Also from the second evaluations, the following shows each of the scatter plots of the suitability scores computed for each of the models using a pilot pool of 51 submitted papers and five reviewers (Fig. 2-5).

From Table 2 it shows that TF-IDF slightly outperformed LSI and LDA in the prediction results as setup by Young *et al.* It is obvious that, the model that produced the most divergent scores is LDA with a standard deviation of 0.144, followed by LSI at 0.059 and TF-IDF at 0.027. This shows that LDA had clearer cut opinions about certain papers than LSI and TF-IDF as shown in Fig. 5. On plotting the mean values, it is apparent that the models produced the same signal shape of curve but at different scale which indicates a relative closeness in judgment.

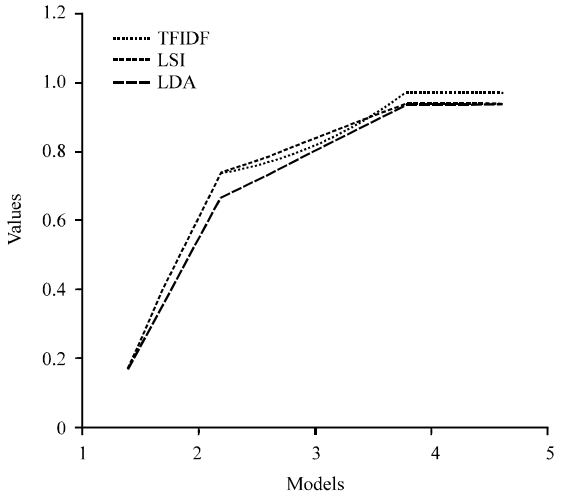

Fig. 1: Line plot of the average accuracy rankings of the models

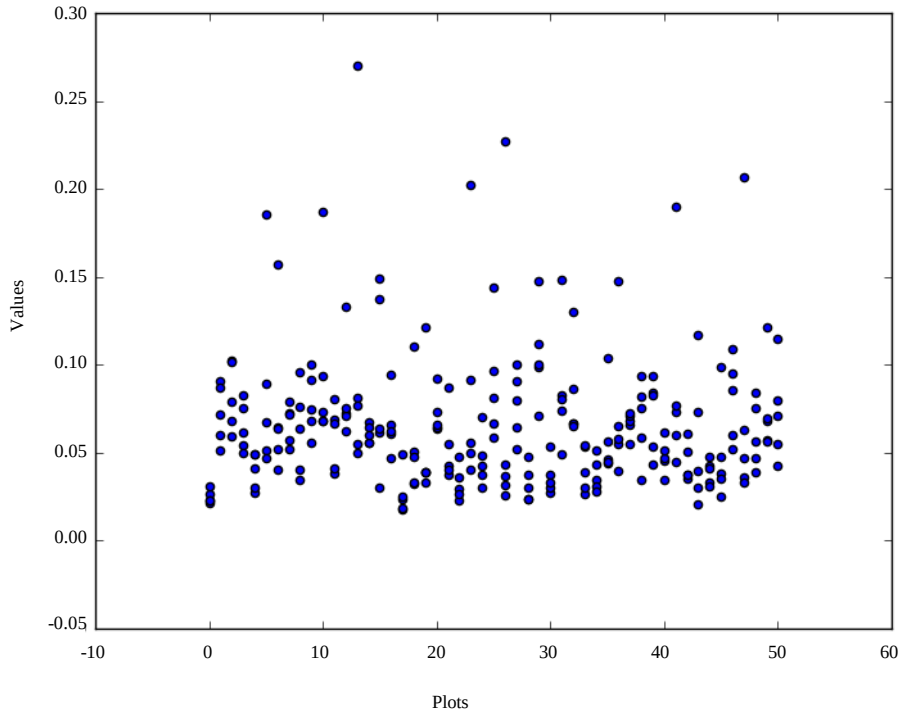

Fig. 2: TFIDF scatter plot

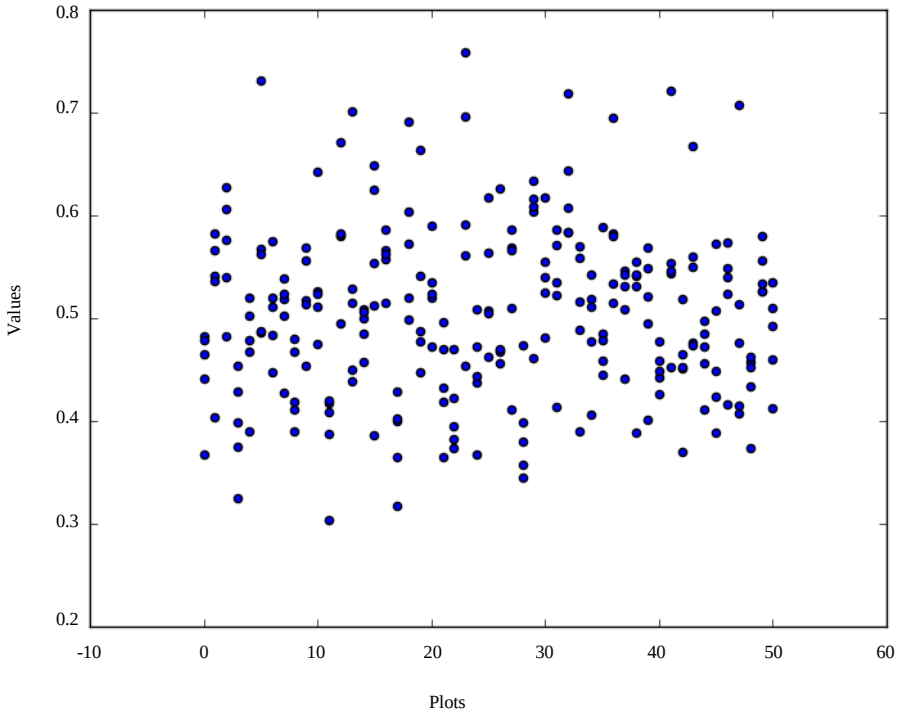

Fig. 3: LSI scatter plot

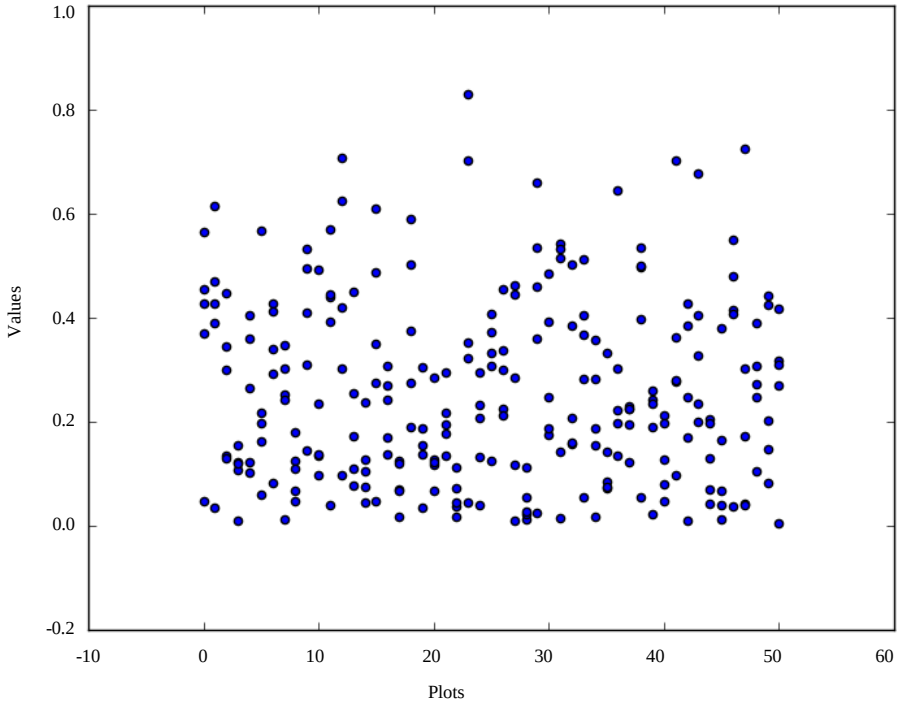

Fig. 4: LDA scatter plot

In this research work, we conducted two experiments using the dataset curated from NIPS 1985-2016 papers to check the performance in accuracy in determining the most suitable reviewer for a paper. To our surprise, from

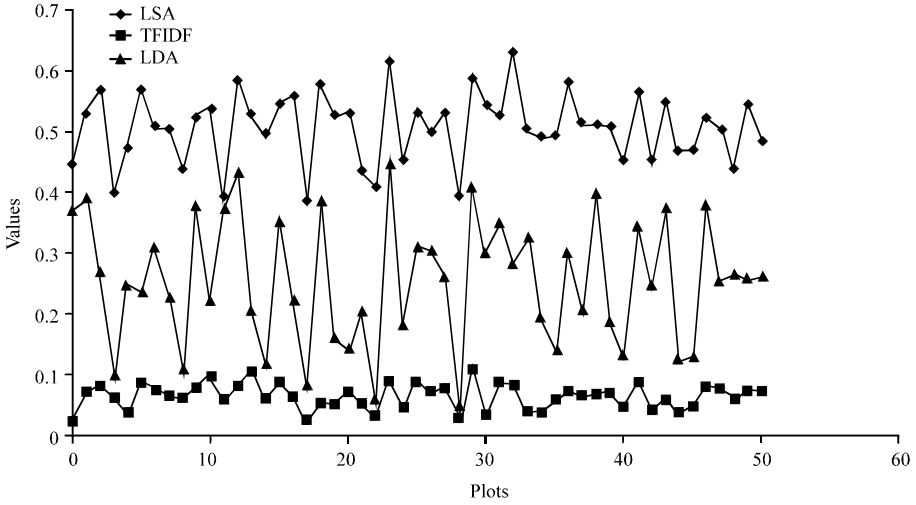

Fig. 5: Marked line scatter plot of the mean of the suitability scores

Table 1: Listing the total average scores for average rankings, A(x, p) for 5 times

Top (X) ranks	TFIDF	LSI	LDA
1	0.166534	0.166666667	0.166666667
2	0.733333333	0.733333333	0.666666667
3	0.8	0.833333333	0.8
4	0.96666	0.933333333	0.933333333
5	0.96666	0.933333333	0.933333333
	0.726637467	0.72	0.7

Table 2: Summary of mean and standard deviation of results

Models	Mean	SD
The LSI-computed suitability scores	0.504820382	0.059976297
The TFIDF-computed suitability scores	0.065378477	0.027717038
LDA-computed suitability scores	0.25592905	0.144661568

the results it showed that TF-IDF outperformed LSI and LDA in predicting the researcher as in the experiment 1 setup. Also, TF-IDF has the lowest standard deviation as compared to LSI and LDA. As in another study that compared LSI and LDA (Li and Hou, 2016), LSI seems to be better than LDA. In concurrence with Zhang *et al.* (2011), TF-IDF has semantic qualities that could make it useful for identifying the most suitable reviewer for a paper.

CONCLUSION

In this study, we present some experimental evaluations of distributional semantic models that has been used for paper-reviewer assignment systems. We used the NIPS 1985-2016 papers as our dataset. In our first experiment, our criterion for performance was the ability of the model to accurately predict the original researcher. The second experiment, we used the standard deviation of the mean results of the suitability scores generated by the models. And we found TF-IDF to be a clear winner for each of the experiments.

REFERENCES

Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin, 2003. A neural probabilistic language model. J. Mach. Learn. Res., 3: 1137-1155.

Charlin, L. and R.S. Zemel, 2013. The Toronto paper matching system: An automated paper-reviewer assignment system. Proceedings of the 30th International Conference on Machine Learning ICML'13 Vol. 28, June 16-21, 2013, Atlanta, Georgia, USA., pp: 1-9.

Li, B. and Y.T. Hou, 2016. The new automated IEEE INFOCOM review assignment system. IEEE. Network, 30: 18-24.

Long, C., R.C.W. Wong, Y. Peng and L. Ye, 2013. On good and fair paper-reviewer assignment. Proceedings of the 13th International Conference on Data Mining, December 7-10, 2013, IEEE, Dallas, Texas, USA, ISBN:978-0-7695-5108-1, pp: 1145-1150.

Manning, C.D., P. Raghavan and H. Schutze, 2008. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK., ISBN-13: 9780521865715, pp: 482.

Taylor, C.J., 2008. On the optimal assignment of conference papers to reviewers. MSc Thesis, University of Pennsylvania, USA.

Zhang, W., T. Yoshida and X. Tang, 2011. A comparative study of TF*IDF, LSI and multi-words for text classification Expert Syst. Applic., 38: 2758-2765.