

# Role of Weighting on TDM in Improvising Performance of LSA on Text data

**Sudarsun S<sup>1</sup> Venkatesh Prabhu G<sup>2</sup> Sathish Kumar V<sup>3</sup>**

R & D Team, Checktronix India Pvt Ltd, Chennai 600034

sudar@burning-glass.com

## Abstract

Latent Semantic Analysis (LSA) is a popular statistical technique, which describes underlying structure of texts. Since LSA represent terms and documents in Term – Document space, it is considered as Vector Space Information Retrieval model. LSA employs Singular Value Decomposition (SVD), a multivariate data reduction technique that approximates high dimensional dataset to low dimensional dataset still containing the significant information in the original dataset. Superiority and efficiency of LSA very much rely on Weighting Algorithms applied. Weighting functions increase the efficiency of Information Retrieval. These Weighting algorithms allocate weights to the attributes (Keywords) based on their occurrences in the corpus. The weights imply the relative importance of the attributes in the document collection.

In this paper, we will be concentrating on the fundamentals of LSA-SVD, Keyword Relevancy and Weightings. Effects of different weighting algorithms will be the central point of this paper. We experimented various weighting algorithms and they were evaluated to study their effects. Precision and Recall values are calculated to compare the performances. While building LSA IR models, a corpus of documents in represented a Term x Document Matrix (TDM) with keywords as the attributes of the corpus and documents being the observations of the corpus. Keyword queries can be projected on the LSA model thus obtained to find closely correlated keywords or a document (keyword collection).

Our experiments include weighting function application on TDM (Pre-Weighting) in order to increase or decrease the relative importance of words based on their occurrence. On the other hand, Weighting functions were applied when query is projected (Post-Weighting). We have experimented with various weighting functions like Inverse Word Frequency (IWF), Inverse Document Frequency (IDF), Normal Weighting, Normalized Document Vector (NDV), Weighted Inverse Document Frequency (WIDF) and also combinations of weightings like IWF+NDV, IDF+NDV. In general, IWF and IDF are used as global weighting functions while Normal weighting, NDV are used as local weighting functions.

We would be presenting a detailed analysis on the effect of the above algorithms. We did our analysis based on Precision-Recall estimates. We have developed a prototype IR query projection tool which projects keyword queries on the LSA model to retrieve relevant keywords with a floating-point score. Typically IWF+NDV pre-weighting produced 94% accuracy when compared with other algorithms like IWF, NDV and Normal weighting producing 88%, 91% and 90% respectively. In another experiment, out of IDF, IDF+NDV and IWF+NDV, IDF+NDV algorithm is found to be more effective in retrieving contextual relevant keywords. Accuracies of IDF, IDF+NDV and IWF+NDV are 95%, 97% and 91% respectively.

**Keywords:** Information Retrieval, SVD, LSA, TDM, Precision, Recall, Weighting Functions, IDF, IWF, WIDF, and NDV

## ABOUT THE AUTHORS

<sup>1</sup> **Sudarsun S** (IEEE Member) is the Director – R & D at Checktronix India Pvt Ltd, Chennai. He is an M.Tech Computer Science from IIT Madras. He got his B.E degree in Electronics and Instrumentation Engineering from Madras University with a Gold Medal.

<sup>2</sup> **Venkatesh Prabhu G** is a Research Associate at Checktronix India Pvt Ltd, Chennai. He is a M.E. Computer Science from Anna University, Chennai. He completed his B.E. Computer Science from MKU, Madurai.

<sup>3</sup> **Sathish Kumar V** is a Research Associate at Checktronix India Pvt Ltd, Chennai. He completed his M.Sc and B.Sc. Statistics from Loyola, Chennai.