

Unsupervised Contextual Keyword Relevance Learning and Measurement using PLSA

Sudarsun S¹ Dalou Kalaivendhan² Venkateswarlu M³

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

sudar@burning-glass.com

ABSTRACT

The primary goal of Contextual Keyword Relevance Learning is to find the related keywords for a given query terms in their context. Standard approaches, such as SVD, Document Clustering and Association Rules, do not generally provide the ability to automatically characterize or quantify the unobservable factors that lead to common patterns. Although SVD provides unsupervised solution to build association between keywords by addressing synonymy problem, it suffers the problem of polysemy in text data.

Probabilistic Latent Semantic Analysis (PLSA) is particularly useful in this context, since it can uncover latent semantic associations among keywords based on the contextual co-occurrence patterns of these keywords in documents. In general, any text can be seen as a distribution of words bound by some context or topic(s), which could be explicit or implicit. PLSA assumes that there are hidden topics, which is the cause for the word distribution found in a text and unearths the same. So we can say that a document text is composed of topics and topics lead to word distribution (Aspect model). If we consider the reverse of it words falling into the same topic are more relevant than words falling in different topics.

In this paper, we develop a probabilistic approach using PLSA for the discovery and analysis of contextual Keyword Relevance based on a words distribution from a training text corpus. We show the flexibility of this approach in classifying keywords in their correct domains. Since these relationships are measured in terms of probabilities, we are able to use probabilistic inference to perform a variety of analysis tasks such as Adaptive Document segmentation, Keywords classification, as well as predictive tasks such as collaborative recommendations.

The PLSA model building procedure spawns into three primary steps: 1) Building a Term-Document Matrix of the target text corpus 2) Computing the aspects model parameters 3) Keyword Query projection and estimation or relevancy. We would be presenting the various issues and their solutions while handling huge sparse floating-point matrices (TDM) in terms of computation time and memory requirements. Typically a 10,000 document x 40,000 keywords TDM would consume over 1.5 Gbytes of memory.

The aspect model parameters $P(w|z)$, $P(z|d)$, $P(z)$, β are computed by EM Iterations starting from random base values. As the frequency matrix is huge, the iterations generally take very long hours to complete. We will be presenting solutions to this by using alternate sparse matrix representation format so as to reduce the computation time & physical memory.

We have developed a prototype system that would allow us project keyword queries on the loaded PLSA model and gives back keywords that are closely correlated. The keyword query is vectorized using the PLSA model in the reduce aspect space and correlation is found by calculating dot product.

We also discuss about the parameters that control the PLSA performance viz. a) Number of aspects, b) Number of EM Iterations c) Weighting functions on TDM (Pre-Weighting) and their role in the quality of Relevancy Estimates. We have estimated the quality by Precision-Recall scores. We have performed various experiments on PLSA models build over varying corpus sizes varying and number of document text domains.

Finally we present a step-by-step procedure to build and tune a PLSA model from a scratch.

Keywords: SVD, Synonymy, Polysemy, Unsupervised Clustering, PLSA, Aspect model, Keyword Relevance

ABOUT THE AUTHORS

¹ **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.

² **Dalou Kalaivendhan** is a Research Associate at Checktronix India Pvt Ltd, Chennai. He is an M.Tech Distributed Computing from Pondicherry University, Pondicherry. He completed his B.Tech Computer Science from Pondicherry University, Pondicherry.

³ **Venkateswarlu M** is a Research Associate at Checktronix India Pvt Ltd, Chennai. He completed his B.Tech Computer Science from Pondicherry University, Pondicherry.