BMC Bioinformatics

BioMed Central

Meeting abstract

Open Access

# A systematic study on latent semantic analysis model parameters for mining biomedical literature

Mohammed Yeasin[1], Haritha Malempati*[1], Ramin Homayouni[2] and Mohammad Shahed Sorower[1]

Address: [1]Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN 38111, USA and [2]Bioinformatics Program, University of Memphis, Memphis, TN 38111, USA

Email: Haritha Malempati* - hmalempt@memphis.edu

* Corresponding author

*from* UT-ORNL-KBRIN Bioinformatics Summit 2009
Pikeville, TN, USA. 20–22 March 2009

Published: 25 June 2009

*BMC Bioinformatics* 2009, **10**(Suppl 7):A6 doi:10.1186/1471-2105-10-S7-A6

This abstract is available from: http://www.biomedcentral.com/1471-2105/10/S7/A6



## Background and rationale

Latent semantic analysis (LSA) is considered to be an efficient text mining technique [1] but most approaches developed on this paradigm are based on adhoc principles. A systematic study on the parameters affecting the performance of LSA is expected to provide guidelines to objectively select the LSA model parameters in a way that is consistent with the data and the application. In this study, empirical analyses were conducted using a previously published 50 gene data set [2] to examine the effects of the following parameters (outlined in Figure 1): Parameters are: (i) stemming, stop-words and word counts (to discard abstract with not enough information), (ii) corpus content (e.g., abstracts with and without titles), (iii) inclusion or exclusion of the dc component or 1st Eigen vector (that adds bias to the model), (iv) objective criteria to choose the number of factors (Eigen vectors) to create the model, (v) information theoretic criteria to select features (words in the corpus) instead of considering complete set of features.

## Methodology

Two datasets, one with titles and abstracts and the other with only abstracts were used to conduct empirical analyses. Preprocessing steps included stemming, stop word removal, as well as removal of documents with less than 100 terms. The term frequency-inverse document frequency (TF-IDF) matrix of size *8714*50* was constructed using the dataset. Singular value decomposition (SVD) on the TF-IDF matrix was used to compute the encoding of the dataset and only $k$ components were retained based on the following objective criteria:

1. Top 25 Eigen vectors

2. $\frac{\Sigma\sigma_p^2}{\Sigma\sigma_n^2} \sim 97\%$ ,

$\Sigma\sigma_p^2$ : energy content within p Eigen vectors,

$\Sigma\sigma_n^2$ : Energy content with n (all) Eigen vectors

3. $\sigma_k \sim \frac{0.7}{n}$ ,

n: number of documents, k: indices of Eigen vector, S: singular value

In addition, the effect of bias was studied by excluding the 1st Eigen vector (dc component).

Different combinations of these parameters were studied and the performance of various LSA models was evaluated by determining the average precision, recall values. The

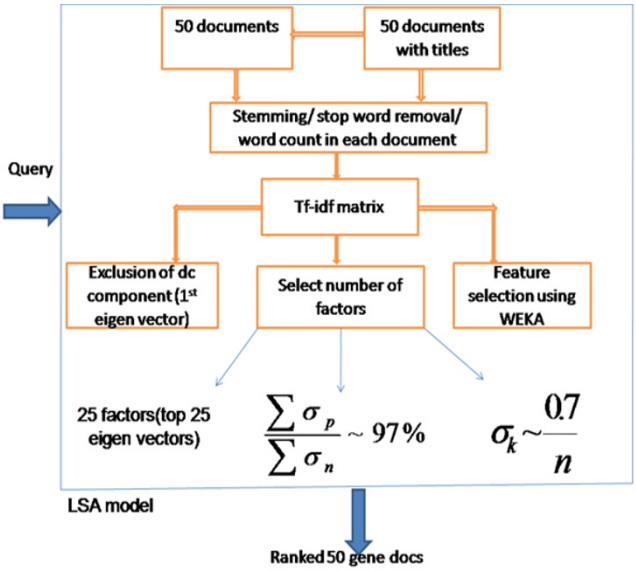


**Figure 1**
**Illustration of the methodology**.

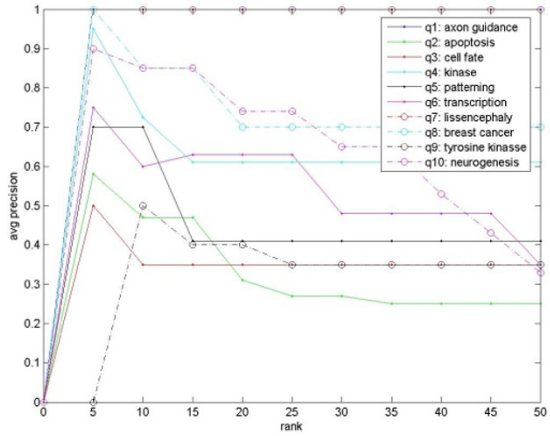


**Figure 2**
**Average precision vs. rank curve**. A combination of three parameters was used: inclusion of titles, exclusion of 1st Eigen vector and 0.7/n objective criterion for factor selection. This combination provides better performance than with individual parameters.

**Table 1: Average precision values of queries (rows) across different parameters (columns)**

| | Abstracts | Titles & Abstracts | 25 Eigen vectors | 97% energy criterion | 0.7/n criterion | Feature selection | No 1st Eigen vector |
|---|---|---|---|---|---|---|---|
| axon guidance | 0.33 | 1 | 1 | 1 | 1 | 1 | 1 |
| apoptosis | 0.2 | 0.20 | 0.18 | 0.19 | 0.23 | 0.3 | 0.27 |
| cell fate | 0.24 | 0.42 | 0.3 | 0.38 | 0.4 | 0.37 | 0.4 |
| kinase | 0.6 | 0.78 | 0.6 | 0.62 | 0.75 | 0.79 | 0.65 |
| patterning | 0.35 | 0.46 | 0.4 | 0.42 | 0.5 | 0.52 | 0.46 |
| transcription | 0.3 | 0.38 | 0.32 | 0.37 | 0.47 | 0.4 | 0.39 |
| lissencephaly | 0.4 | 1 | 1 | 1 | 1 | 1 | 1 |
| breast cancer | 0.25 | 0.74 | 0.37 | 0.5 | 0.75 | 0.67 | 0.7 |
| tyrosine kinase | 0.1 | 0.3 | 0.12 | 0.15 | 0.3 | 0.28 | 0.28 |
| neurogenesis | 0.18 | 0.32 | 0.13 | 0.18 | 0.32 | 0.3 | 0.3 |

best model is defined as the one with relatively high average precision across a set of varied queries.

## Results and conclusion

Performance analysis (average precision-recall curves, F-measure etc.) using Gene Ontology classifications corresponding to the 50 gene collection show that not all parameters significantly affect the performance of LSA model (Table 1). In general, adding titles in addition to the abstracts substantially increased the average precision. In addition, using 0.7/n criteria produced better results than using 25 Eigen vectors or the 97% criteria. It was found that the best performance was achieved by combining 3 parameters: inclusion of title in abstracts in the corpus, exclusion of the dc component, and selection of Eigen vectors based on objective criterion (Figure 2). This work provides a framework for determining the best parameters in using LSA for ranking genes with respect to queries. Future work will focus on evaluating this framework using different gene document collections.

## References

1. Vanteru BC, Shaik JS, Yeasin M: **Semantically linking and browsing PubMed abstracts with gene ontology.** *BMC Genomics* 2008, **9:**S10.
2. Homayouni R, Heinrich K, Wei L, Berry M: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21(1):**104.