Sheng-Ting Lin <slin@bcgsc.com>
UBC Bioinformatics Master's Student

Master's Thesis Progress Report

New technologies continue to accelerate the pace of biomedical research, causing an exponential increase in the number of scientific articles published every year. At the end of 2015, there were approximately 25.6 million records in the PubMed database. This astounding number makes it difficult for scientists to find publications relevant to their work. Hence, improving current technology for retrieving and recommending scientific articles is becoming increasingly important.

Many frameworks that retrieve related articles combine several methods to make recommendations. For instance, PubMed identifies a number of related articles for each record by performing pairwise analysis of shared terms between every record [1]. Scienstein uses a hybrid method consisting of citation analysis, author analysis and user activities to find similar articles [2]. CiteSeer uses an co-citation analysis and paper metadata in their calculation [3]. Most of these methods require information that is not always available, such as full text, keywords and citations.

While many have implemented paper recommendation systems, latent semantic analysis (LSA) has not yet been applied to retrieving related biomedical papers. LSA is the application of singular value decomposition (SVD) for finding relationships among a set of documents and terms [4]. In turn, SVD is a matrix factorization method that decomposes a matrix of two correlated dimensions into a set of uncorrelated components. These decomposed matrices expose the relationships from the original data and identify dimensions on which the data exhibits the most variance [4]. It has been widely applied as a collaborative filtering technique in implementing recommendation systems for commercial purposes, such as recommending movies or products to users based on their viewing and/or purchasing history [5]. In the context of mining biomedical literature, many have also applied LSA to discover knowledge and link information, such as constructing protein-protein interaction networks, annotating gene names, and finding previously unknown gene-disease relationships [6,7]. In other words, using LSA to find term-term relationships has been widely explored, but not with the objective of finding document relationships.

My project focuses on the application of LSA for identifying related biomedical articles using title and abstract text. The first step is to investigate the effect of various parameters on the recall and precision of the results, including the number of singular values in SVD, the similarity function and the type of matrix. A list of biomedical terms obtained from the Unified Medical Language System will be used to create a document-term matrix as the input for SVD. Document similarity will be measured by comparing document vectors in the decomposed matrix. A preliminary experiment was done using the TREC2005 genomics track data consisting of ~1,500 abstracts among 10 topics [8]. It showed that using cosine distance as the similarity function on TF-IDF (term-frequency and inverse document-frequency) matrix achieves the best precision score when using 15 singular values. Due to the inconsistency and relatively small size of the TREC2005 dataset, a new training dataset has been generated using the NCBI Entrez API: the most related papers from 100 randomly selected PubMed records. The same experiment will be performed on this new training set to compare the LSA method to the approach used by PubMed.

My project could also explore some of the following topics. First, finding the most similar papers for each article by measuring the distances between all possible document vectors is very computationally intensive, especially when the number of documents is large. This creates a need for a more efficient algorithm. Second, matrix factorization using SVD is also computationally expensive and thus redoing it whenever new papers are added to the matrix is intractable. Some have explored methods of incrementally building an SVD model and revising it when data is added or removed [9,10]. Applying these methods to my project can address the issue of scalability. Third, combining dissimilar document vectors might help bridge different research interests by uncovering relevant interdisciplinary papers. Finally, incorporating known associations of words into the document-term matrix, described as document padding or sprinkling in previous studies [11,12], could provide a way of performing partially supervised learning in SVD.

References

1. PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK3830/
2. Gipp B. Bell J. Hentschel C. Scienstein: A Research Paper Recommender System. ICETiC'09. 2009.
3. Giles CL, et al. CiteSeer: An Automatic Citation Indexing System. - Third ACM Conference on Digital Libraries. 1998
4. Chen H, et al. Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. Front. Physiol., 2013 January 30.
5. Ekstrand MD, et al. Collaborative Filtering Recommender Systems. Foundations and Trends in Human–Computer Interaction. 2010
6. Rodriguez-Esteban R. Biomedical Text Mining and Its Applications. PLoS Computational Biology. 2009 December
7. Zhu F, et al. Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics. 2012 Nov 15
8. Hersh W, et al. TREC 2005 Genomics Track Overview. The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings. 2005
9. Sarwar B, et al. Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems. Fifth International Conference on Computer and Information Science. 2002.
10. Brand M. Fast online SVD revisions for lightweight recommender systems. Proceedings of SIAM 3rd International Conference. 2003.
11. Abate F, et al. Improving Latent Semantic Analysis of Biomedical Literature Integrating UMLS Metathesaurus and Biomedical Pathways Databases. BIOSTEC. 2003.
12. Yang H, King I. Sprinkled Latent Semantic Indexing for Text Classification with Background Knowledge. Springer-Verlag Berlin Heidelberg. 2006