FYP Logbook

Ruohan Gao
gr013@ie.cuhk.edu.hk

Qiming Zhang
zq113@ie.cuhk.edu.hk

October 27, 2014

1 Expected Outcome

The expected result of this project is a system, which is able to find relevant & important papers given a query (which may be a few keywords or a paper itself). The project is targeting the arXiv (http://arxiv.org) database, one of the most active paper repository in computer science. The system may comprise several components: a data backend, a topic analysis algorithm that can determine how relevant two papers are, as well as a UI for query.

2 Paper Reading

2.1 Latent Dirichlet Allocation [1]

2.2 Introduction to Probabilistic Topic Models [2]

2.3 Replicated Softmax: an Undirected Topic Model [3]

2.4 Algorithms for Non-negative Matrix Factorization [4]

2.5 Probabilistic Latent Semantic Indexing [5]

2.6 Document Clustering Based On Non-negative Matrix Factorization [6]

2.6.1 Summary

In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a par- ticular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. NMF differs from the latent semantic indexing method based on the singular vector decomposition (SVD) and the related spectral clustering methods in that the latent semantic space derived by NMF does not need to be orthogonal, and that each document is guaranteed to take only non-negative values in all the latent semantic directions. These two differences bring about an important benefit that each axis in the space derived by the NMF has a much more straightforward correspondence with each document cluster than in the space derived by the SVD, and thereby document clustering results can be directly derived without additional clustering operations.

2.6.2 Future Pointers

1. Two problems addressed:
   a. topics can overlap, not necessarily orthogonal
   b. linear combination coefficients should all take non-negative values
2. Weighted term-frequency vector to represent each document (TF-IDF)
3. Steps of NMF document clustering algorithms

2.6.3 Link

http://web.stanford.edu/class/ee378b/papers/xu-liu-gong-document.pdf

2.7 Research Paper Recommender System Evaluation [7]

2.8 Projected Gradient Methods for Non-negative Matrix Factorization [8]

2.8.1 Summary

Non-negative matrix factorization (NMF) can be formulated as a minimization problem with bound constraints. This paper proposes two projected gradient methods for NMF, both of which exhibit strong optimization properties. The one solving least square sub-problems leads to faster convergence than the popular multiplicative update method.

2.8.2 Future Pointers

1. Conventional Approach: $V \approx WH$ where V is an $n \times m$ data matrix, $W \in R^{n \times r}$ and $H \in R^{r \times m}$. The objective is to find W and H which minimize the difference between V and WH:

$$\min_{W,H} f(W,H) \equiv \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} (V_{ij} - (WH)_{ij})^2$$

2. Two projected gradient methods for NMF

   a. Alternating Non-negative Least Squares Using Projected Gradient Methods

   b. Directly Applying Projected Gradients to NMF

3. Experiments and evaluations are done systematically

2.8.3 Link

http://www.csie.ntu.edu.tw/ cjlin/papers/pgradnmf.pdf

3 Dataset Description

3.1 NIPS

The NIPS data set contains papers from the NIPS conferences between 1988 and 2003. The conference is characterized by contributions from a number of different research communities in the general area of learning algorithms.

3.2 arXiv

The arXiv is a repository of electronic preprints, known as e-prints, of scientific papers in the fields of mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance, which can be accessed online.

4 Implementation

5 Evaluation

6 Future Work

References

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[2] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[3] Hinton G E, Salakhutdinov R. Replicated softmax: an undirected topic model[C]//Advances in neural information processing systems. 2009: 1607-1614.

[4] Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances in neural information processing systems. 2001.

[5] Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

[6] Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003.

[7] Beel, Joeran, et al. "Research paper recommender system evaluation: a quantitative literature survey." Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation. ACM, 2013.

[8] Lin C J. Projected gradient methods for nonnegative matrix factorization[J]. Neural computation, 2007, 19(10): 2756-2779.