Bipartite Graph for Topic Extraction*

Thiago de Paulo Faleiros and Alneu de Andrade Lopes
University of São Paulo
São Carlos, Brazil
{thiagopf, alneu}@icmc.usp.br

Abstract

This article presents a bipartite graph propagation method to be applied to different tasks in the machine learning unsupervised domain, such as topic extraction and clustering. We introduce the objectives and hypothesis that motivate the use of graph based method, and we give the intuition of the proposed Bipartite Graph Propagation Algorithm. The contribution of this study is the development of new method that allows the use of heuristic knowledge to discover topics in textual data easier than it is possible in the traditional mathematical formalism based on Latent Dirichlet Allocation (LDA). Initial experiments demonstrate that our Bipartite Graph Propagation algorithm return good results in a static context (offline algorithm). Now, our research is focusing on big amount of data and dynamic context (online algorithm).

1 Introduction

A huge amount of data stored today are in semi-structured and unstructured forms. Most of these data sets are texts. In fact, the most likely form of storing information is text. Automatic techniques to help organize, manage, and extract knowledge from such textual data are worthwhile research topics for the machine learning and data mining communities.

In an unsupervised context, a common knowledge extraction task is text clustering, *i.e.* the grouping of objects that represent documents or words in such a way that objects in the same group are more similar to each other than to those in other groups. Dimensionality reduction can be considered a subtype of clustering; these include a well-know technique Latent Semantic Analysis (LSA) based on Singular Value Decomposition (SVD). Although this matrix decomposition technique have been successfully applied to different domains, it has drawbacks, such as expensive storage requirements and computing time. As an alternative, Hofmann [Hofmann, 1999] proposed the Probabilistic Latent Semantic Analysis (PLSA). However, it has problems in generalize the inferred model leading to overfitting [Blei *et al.*, 2003]. To overcome this problem, Blei [Blei *et al.*, 2003] proposed Latent Dirichlet Allocation (LDA), a fully Bayesian Model with a consistent generative model. LDA has influenced a huge amount of work and have become a mainstay in modern statistical machine learning. LDA based models have a rigorous mathematical treatment of decomposed operations that discover the latent groups (topics). From the practitioner's perspective, creating a new model and deriving it to an effective and implementable inference algorithm are hard and tiresome tasks [Rajesh *et al.*, 2014]. Moreover, the mathematical rigour hampers a rapid exploration of new assumptions, heuristics, or adaptations that could be useful in many real scenarios.

A simple way to describe operations that distil knowledge or infer patterns is to use a graph or network to represent relations among the entities of a problem. Describing algorithms over a graph representation enables an easy incorporation of heuristic methods. Network representations offer several advantages: (1) they avoid sparsity and ensure low memory consumption; (2) they enable operations for the inclusion and exclusion of new vertices, edges and subnetworks to be easily described; (3) they provide an optimal description of the topological structure of a dataset; and (4) they provide local and global statistics of a datasets structure to be combined.

The usual strategies to tackle the knowledge extraction task is to represent the text collection as a document-term matrix. Nevertheless, more expressive representations, such as homogeneous or heterogeneous [Rossi *et al.*, 2012] networks may be employed. In a bipartite heterogeneous network, only links between objects of different types are allowed (document-term relations). By assign edge-weights equal term frequencies, we can capture the strength of this links. In such a scenario, text mining is a complex task and demands specific processing algorithms.

Our aim is to investigate and develop techniques that combine the expressive network representation made possible by the complex networks theory with data streaming techniques for dealing with problem of topic extraction. In addiction, to bring the advantages associated with bipartite network representation to the unsupervised learning process. The hypothesis is that the combination of networks (or graph) representation and the data streaming techniques can generate effective and efficient models for topic extraction.

An undirected bipartite graph is a triple $G = (\mathcal{D}, \mathcal{W}, \mathcal{E})$

*This research is supported by São Paulo Research Foundation (FAPESP), proj. number 2011/23689-9

where $D = \{d_1, \ldots, d_m\}$ and $\mathcal{W} = \{w_1, \ldots, w_n\}$ are sets of vertices, and $\mathcal{E}$ is the set of edges $\{e_{j,i} = (d_j, w_i) | d_j \in \mathcal{D}, w_i \in \mathcal{V}\}$ each with edge weight f_{ji} (see Figure 1).

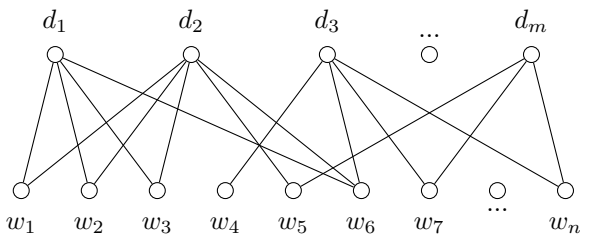

Figure 1: Heterogeneous Graph G for document representation

For the representation of a corpus of documents, $\mathcal{D}$ is the set of documents and $\mathcal{W}$ is the set of words. An edge $e_{j,i}$ exists if word w_i occurs in document d_j – an edge signifies an association between a document and a word. Edge-weight $f_{j,i}$ is equal word frequencies w_i in document d_j.

Our proposed method is based on exploration of an effective unsupervised learning algorithm based on labels propagation. However, unlikely traditional label propagation techniques, we represent the data as a bipartite heterogeneous graph. Moreover, each vertex is associated to many labels, as a soft-cluster schema.

During the propagation procedures, the labels propagate to all nodes through the edges. Higher edge weights allow labels to "travel" through more easily. The rationale behind our method (see Algorithm 1) is to locally spread out the influence of each word of a document d_j, and globally propagate the cluster proportions of the entire documents collection back to a word w_i. These local and global labels propagations (lines 6 and 9) are performed until convergence. In the end, the labels can be used in tasks as clustering or topic extraction.

Algorithm 1: Bipartite Graph Propagation Algorithm

Input: bipartite graph G,
Output: labels A_j for each document d_j and labels B_i for each word w_i

```
1  begin
2      Initialize labels A_j's and B_i's;
3      while convergence do
4          foreach d_j ∈ D do
5              repeat
6                  A_j ← localPropag(G,d_j, A_j, B);
7              until A_j convergence ;
8          end
9          B ← globalPropag(G,A,B);
10     end
11 end
```

The bipartite graph structure can be easily adjusted with the insertion of new vertices. Moreover, our propagation method can be parallelizable and work in a dynamic context of stream of documents.

2 Final Considerations

Our proposed method is formally based on aspects of the Nonnegative Matrix Factorization (NMF) and Variational Bayesian (VB) Inference algorithm for LDA. We have shown the equivalence between these methods. In addiction, we did experiments to evaluate the performance of our bipartite graph propagation algorithm compared with LDA in the task of topic extraction and document representativeness. We used two evaluation metrics: *Pointwise Mutual Information*, and *Classification Accuracy*. The initial experiments was performed in a static context. The results indicates that our algorithm is a promising method to explore in topic extraction task. Mainly when the problem is easily represented as a graph and has heuristics knowledge that could be incorporated to improve the topic extraction process.

Modern text mining algorithms require computation with large volume of documents. For some applications, documents are being published at a rate high enough to make its long-term storage cost outweighs its benefits. When the dynamic context is taking into account, such factors as unknown vocabulary size, memory consumption, processing time, and concept drifts are concerns that bring new research challenges and a promising exploratory area for data mining algorithms. Our research goals is to create a novel approximation stream handling algorithm that incrementally analyse arriving documents in one pass and extract hidden knowledge.

The graph representation is a simple structure that can be incorporated in dynamic methods, it allows to enable operations for the inclusion and exclusion of new vertices, edges and subgraphs. Bipartite Graph Propagation method fits well in unsupervised tasks, it can be easy parallelized, and its propagation procedures works in an online fashion.

References

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57. ACM, 1999.

[Rajesh *et al.*, 2014] R. Rajesh, G. Sean, and D. M. Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 33 of *JMLR Proceedings*, pages 814–822, 2014.

[Rossi *et al.*, 2012] R. G. Rossi, T. P. Faleiros, A. A. Lopes, and S. O. Rezende. Inductive model generation for text categorization using a bipartite heterogeneous network. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 1086 –1091, 2012.