A FRAMEWORK FOR USER GUIDED DOCUMENT CLUSTERING

Josan Koruthu, Manjeet Rege, and ***Reynold Bailey***

ABSTRACT

A major proportion of business data being generated today is in unstructured text format. To discover valuable information from the growing text corpora, document clustering is a typical approach employed in business analytics. However, many of the document clustering methods are completely unsupervised, and hence are unable to incorporate any available domain knowledge. We present an approach for integrating user provided constraints in document clustering. Specifically, the user expresses the domain knowledge in the form of must-link and cannot-link constraints. These constraints are then utilized by the framework for guiding the clustering process.

Keywords: Text, Analytics, Documents, Clustering

MOTIVATION

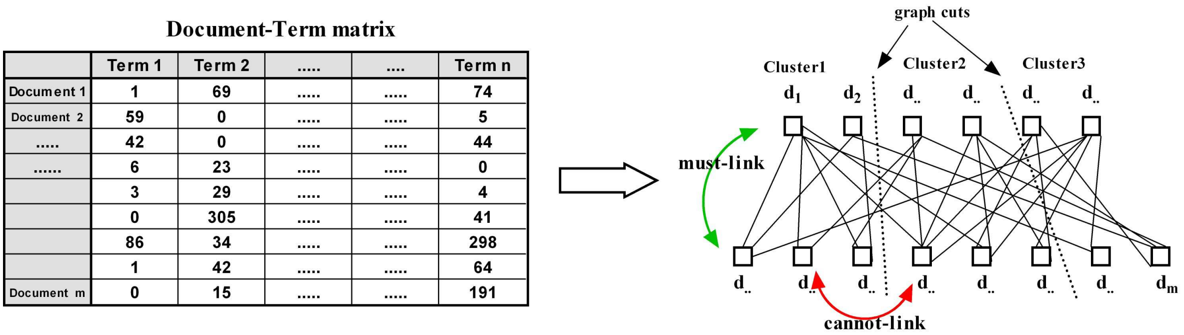
There has been an exponential growth in the rate at which data is being generated today in diverse application domains. For example, large corporations develop profiles of customers based on heterogeneous data collected such as online blogs, transcripts of customer service calls, written communications, etc. Much of this text data is in an unstructured format. From text analytics perspective, a logical step is to employ clustering techniques to discover natural groupings of data in order to create data summaries. Consequently, document clustering has received significant attention in business analytics recently. Generally, the data is represented using the vector model (or dictionary model) in which a set of m documents with n unique terms is represented as an m x n document-term matrix. An entry i,j in this matrix is usually the frequency of word j appearing in document i (*see* Figure 1). An application of a clustering algorithm to this matrix yields clusters of documents, such that documents belonging to the same cluster are similar together, and dissimilar from those grouped in a different cluster. While very effective in generating grouping of documents, many of the algorithms implemented in analytics tools are completely unsupervised. In many applications, there might be limited domain information available from the user that could be embedded into the document clustering process. A publishing company for instance, interested in clustering of books would like to provide expert domain information indicating that there are certain books (e.g. for mature audience) that should always be clustered together, and that they should never be clustered along with certain other books (e.g. children's books). An unsupervised document clustering algorithm might cluster these two kinds of books together based on the similarity amongst them, i.e. commonly occurring words according to the vector model. A semi-supervised approach, on the other hand, can integrate the user provided knowledge to guide the clustering process. Note that this is different from document classification (or supervised learning) where a class label is available for all documents, which is then learned by a classification algorithm to predict document classes for new documents.

PROPOSED FRAMEWORK

We propose a framework in which a user is able to provide pairwise constraints, viz., *must-link* and *cannot-link* constraints on documents (*see* Figure 1). An existence of a must-link constraint between two documents indicates that the two corresponding documents must be clustered together, irrespective of their dissimilarity. Similarly, a *cannot-link* constraint signifies that the two documents should never be clustered together, irrespective of their similarity. Although, semi-supervision in the form of constraints has been proposed before [4,5], the novelty of our work lies in the underlying clustering algorithm, viz., the Isoperimetric Graph Partitioning (IGP) algorithm [1,3], which integrates the constraints. Having IGP at the heart of the framework to compute the clusters brings us the following advantages:

(1) **Applicability to large datasets**: IGP quickly solves a sparse system of linear equations to get the clusters, as compared to an iterative approach adopted by non-negative matrix factorization [5], or eigenvalue decomposition in spectral clustering [2]. As a result, the framework is applicable to large datasets and scales extremely well as the size of the document set grows.

(2) **Performance in the presence of noise**: It is common to have “noise” in real-world datasets. The proportion of the noise present varies based on various factors such as the data source, cleaning methodology, or the data capturing process. It has been shown that the performance of other clustering algorithms (such as spectral clustering [2]), steadily deteriorates as the amount of noise in the data increases. On the other hand, in the presence of various kinds and amounts of noise, IGP is known to obtain stable clusters.



Document-Term matrix

	Term 1	Term 2	.....	....	Term n
Document 1	1	69	.....	.....	74
Document 2	59	0	.....	.....	5
.....	42	0	.....	.....	44
......	6	23	.....	.....	0
	3	29	.....	.....	4
	0	305	.....	.....	41
	86	34	.....	.....	298
	1	42	.....	.....	64
Document m	0	15	.....	.....	191

Figure 1: The initial document-term matrix is converted into a document similarity graph matrix. The user provided constraints guide the graph partitioning process in obtaining document clusters.

EVALUATION

For evaluating the framework, we have used some of the publicly available text datasets, and conducted two sets of experimental studies. First, we have studied the clustering results as the amount of domain knowledge available to the framework increases. We have seen that as the percentage of constraints on documents gradually increase (1% to 5%), the clustering accuracy correspondingly improves as well. This shows that the framework is successfully able to integrate the available domain expertise into the clustering process. Second, we have compared the performance of IGP document clustering with spectral clustering, in the absence of any supervision. These set of experiments are primarily to demonstrate the performance if there is no constraint knowledge available. The results in both the experimental studies are extremely promising.

CONCLUSIONS

We present a text analytics framework, where user provided domain knowledge can be integrated into the document clustering process. The *must-link* and *cannot-link* constraints provided guide the document clustering process. By representing document similarity using a weighted graph, we treat document clustering as a graph partitioning problem. The Isoperimetric Graph Partitioning algorithm is capable of dealing with large sparse and noisy datasets leading to optimal document clusters.

REFERENCES

1. Grady, L., & Schwartz, E. L. (2006), Isoperimetric Partitioning: A new algorithm for graph partitioning, SIAM Journal on Scientific Computing, 27(6), 1844-1866.
2. Chung, F. R. K. (1997). *Spectral Graph Theory*. American Mathematical Society.
3. Grady, L., & Schwartz, E. L. (2006), Isoperimetric Graph Partitioning for Image Segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(3), 469-475.
4. Ji, X., & Xu, W. (2006), Document clustering with prior knowledge, Proceedings of ACM SIGIR Conference, 405–412.
5. Chen, Y., Rege, M., Dong, M., & Hua, J.(2007), Incorporating user provided constraints into document clustering, Proceeding of IEEE ICDM Conference, 103 –112.