A Neighborhood-Based Approach for Clustering of Linked Document Collections

Ralitsa Angelova
Max Planck Institute for Informatics
66123, Saarbrücken, Germany
angelova@mpi-inf.mpg.de

Stefan Siersdorfer
Max Planck Institute for Informatics
66123, Saarbrücken, Germany
stesi@mpi-inf.mpg.de

ABSTRACT

This paper addresses the problem of automatically structuring linked document collections by using clustering. In contrast to traditional clustering, we study the clustering problem in the light of available link structure information for the data set (e.g., hyperlinks among web documents or coauthorship among bibliographic data entries). Our approach is based on iterative relaxation of cluster assignments, and can be built on top of any clustering algorithm. This technique results in higher cluster purity, better overall accuracy, and make self-organization more robust.

Categories and Subject Descriptors: H.3.3 Information Systems: Information Search and Retrieval I.5.3 Pattern Recognition: Clustering

General Terms: Theory, Algorithms, Reliability

Keywords: Clustering, Exploiting Link Structure

1. INTRODUCTION

The issue of automatically structuring heterogeneous document collections into thematically coherent subsets is relevant for a variety of applications, such as organizing large personal email folders, dividing topics in large web directories into subtopics, structuring large amounts of company and intranet data, etc.

Graph-based clustering is a well established problem in the literature. A detailed overview of existing methods is presented in [6]. Typically, the underlying graph G is constructed by representing each data point as a node in G and each edge, connecting any two data points, by a weight, indicating the distance (dissimilarity) between its end points.

Our approach is orthogonal to the approaches discussed in [6] as we use statistical knowledge about the cluster assignments of the nodes in the formed neighborhoods in G. Furthermore, the assignment of the edge weights, and thus the type of graphs used by the above approaches, are based on node-node similarity, and it is not clear how to carry this forward to a hyperlinked environment. Closest to the approach presented in this paper is our own recent work on neighborhood-based classification [2].

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

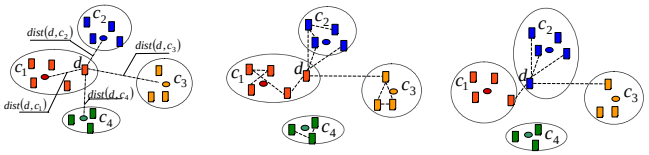
2. A PROBABILISTIC FRAMEWORK FOR GRAPH-BASED CLUSTERING

We propose two approaches that consider the link structure in the test graph.

2.1 Content Combination

An intuitive way of combining the content of a document d with the content of its neighbors $d' \in N(d)$ is to assign term weights $w'(t_i, d)$ to all terms $t_i \in d$ while considering in a linear way the term weights t_i in d's neighbors $d' \in N(d)$. The impact of the neighborhood content on the final term weights in document d is controlled by a parameter α. The correspondingly adjusted feature vectors can be used as an input to all vector-based clustering algorithms, e.g., the k-means algorithm.

2.2 Graph-based Clustering



(a) Content-Based Clustering (b) Link Structure (c) Final Clustering

Figure 1: Neighborhood-Based Clustering

The approach, we propose here, adopts a probabilistic formulation of the clustering problem and is based on the so called relaxation labeling technique [2]. We aim to cluster a set of documents $\mathcal{D}$, where each document $d \in \mathcal{D}$ corresponds to a vertex in the graph G and each link between two documents in $\mathcal{D}$ corresponds to an edge in G. The clustering algorithm requires as an input the the text of each document d and information about which documents of G constitute its neighborhood, $N(d)$. Let $c(d)$ denote the cluster of node d whose validity can be associated with a probability. The content of document d is represented as a set of terms that occur in d and denoted by $\tau(d)$. The output of the algorithm should be an assignment of clusters to the graph nodes such that each document $d \in G$ belongs to its maximally likely cluster i, selected from a finite set of clusters $[1..m]$.

The intuition behind our approach is sketched in Figure 1. Figure 1 (a) shows the clustering entirely based on the content information about each document. Here document d is assigned to its nearest (most similar) cluster. The similarity of document d is measured with respect to the cluster centroids, (e.g. produced by a content-based clustering algorithm such as k-means). Figure 1 (b) shows the link structure among the documents. In the content-only world, such link information is completely ignored. Our attempt to make use of it starts with the observation that document d

is linked to a significantly higher number of documents from cluster c_2 than from cluster c_1. Furthermore, the graph structure suggests that documents from cluster c_2 typically link to documents that belong to this very same cluster - c_2. Thus, with high probability a document can be clustered in c_2 if it is linked to many documents that belong to cluster c_2. In our toy example, this leads to reassigning document d to cluster c_2. The final clustering after taking both content and link information into account is shown in Figure 1 (c).

Formally, taking into account the underlying link structure and document d's content-based feature vector, the probability of a document d to be assigned to cluster i is:

$$\Pr\left[c\left(d\right)=i \mid \tau, G\right]=\Pr\left[c\left(d\right)=i \mid \tau(d), c\left(d_1\right), \cdots, c\left(d_l\right)\right]$$

where d_1 through d_l are the documents in $\mathcal{D}$.

For tractability, it makes sense to focus on the strongest dependencies among immediate neighbors. Such a model is called a first-order Markov Random Field or MRF [4, 5]. Computing the parameters of an MRF such that the likelihood of the observed training labels is maximized is a difficult problem that cannot be solved in closed analytic form and is typically addressed by an iteration technique known as relaxation labeling (RL). Our approach builds on this mathematical technique.

In the spirit of emphasizing the influence of the immediate neighbors for each document, $N\left(d\right)$, we obtain $\Pr\left[c\left(d\right)=i \mid \tau(d), G\right]=\Pr\left[c\left(d\right)=i \mid \tau(d), N\left(d\right)\right]$ and denote it by $\Phi_{i,d}$. This reflects the MRF assumption that the label of a node is conditionally independent of the labels of other nodes in the graph given the labels of its immediate neighbors. We abbreviate $\Pr\left[c\left(d\right)=i \mid \tau(d)\right]$, the graph-unaware probability based only on d's local content, by $\sigma_{i,d}$. Let $c\left(N\left(d\right)\right)$ denotes the assignment of clusters to the group of neighbors of d, $N\left(d\right)$.

Then, $\Phi_{i,d}$ can be computed in an iterative RL manner as follows:

$$\Phi_{i,d}^{(r)}=\sigma_{i,d}\cdot\sum_{c(N(d))}\left(\prod_{d'\in N(d)}\Pr\left[c\left(d\right)=i\wedge c\left(d'\right)=j\right]\right)^{(r-1)} \quad (1)$$

where $r>1$ and $i,j\in[1..m]$ are cluster assignments.

To avoid the potential increase in the level of noise, we can consider only a subset of good neighbors. These neighbors should be similar enough to the document in question, d. For this purpose, we introduce a similarity threshold, which can be computed as the cosine-similarity between the pair of neighboring documents and which selectively determines the neighborhood of each document d.

However, calculating the sum over all possible cluster assignments in Equation 1 is hard as we have $m^{|N(d)|}$ summands, where m is the number of distinct clusters. To solve this problem we employ two major methods. We approximate the sum over all possible cluster assignments of the neighborhood to either its most significant summand, treating it as if it were the true set of clusters, or the most significant summands and their associated probabilities. The algorithm efficiently re-computes and updates the probabilities of particular cluster assignments to the neighborhood $N\left(d\right)$ after *each* RL iteration.

To improve the robustness of the algorithm we propose two beneficial extensions. We aim to ignore the unnecessary and most probably noisy information behind all irrelevant links in a neighborhood by assigning to each edge e a weight w_e equal to the cosine similarity between the feature vectors of the documents connected by the edge. We also explore further the hypothesis that neighboring documents should receive similar cluster assignments. We introduce a metric over the set of clusters where thematically close clusters are separated by a shorter distance and therefore impose smaller cost for assigning neighbors to similar clusters.

3. EXPERIMENTS

We conducted experiments on three sets of data obtained from the database of scientific papers DBLP, the movie database IMDB, and the online encyclopedia Wikipedia. Full detail of the experiments can be found in [1]. All datasets are available at www.mpi-inf.mpg.de/∼angelova.

3.1 Results

We compared the content-based method k-Means [3] - **k-Means**, and the Content Combination - **CComb**[α], with the graph-based clustering method - **GC**, and its enhanced variant using the edge weighting scheme and the cluster similarity metric - **wmGC**. All graph-based methods use the result of the simple content based k-Means in the initialization step. We tested all graph-based methods with different influence of the neighborhood described by α which are correspondingly shown in squared brackets after the method abbreviation.

We also tested an MST-based graph-cut clustering algorithm [7], computing the edge weights as weighted sum of hyperlink based neighborhood and content similarity of the documents, and pruning edges in the corresponding spanning tree. However in our preliminary experiments, the k-means algorithm (despite of its simplicity) showed superior performance on our data sets. Hence, we did not consider building our algorithm on top of a graph-cut approach.

The outcome of the comparison among the above methods along with the 95% confidence intervals is shown in Table 1.

Table 1: Comparison of Clustering Methods

	DBLP	**IMDB**	**Wikipedia**
	Accuracy	*Accuracy*	*Accuracy*
$kMeans$	0.4245±0.0074	0.3569±0.0145	0.5054±0.0133
GC	0.4609±0.0075	0.3809±0.0147	0.5497±0.0133
$wmGC[H]$	0.4689±0.0075	0.3790±0.0147	0.5938±0.0131
$CComb[1.0]$	0.5218±0.0075	0.4024±0.0149	0.6391±0.0128
$GC[1.0]$	0.5914±0.0074	0.4338±0.0150	0.6254±0.0129
$wmGC[H][1.0]$	0.6108±0.0074	0.4540±0.0147	0.6394±0.0128

The graph-based approach significantly outperforms all pure content-based methods. Our experiments show improvements of up to 9% over the k-Means algorithm as well as significant gains close to 10% over the content combination approach. The performance of the graph-based clustering is even better if the content combination technique is used as initialization step for the graph-based methods. Including the cluster distance metric in the computations improves the graph-based clustering by gently imposing constraints on the pair of cluster assignments for each two neighboring documents resulting in up to 6% gain in accuracy in some cases.

The newly proposed graph-based clustering method, especially applied on top of the content combination technique, is very robust and outperforms the previously known state-of-the-art algorithms by a significant margin.

4. REFERENCES

[1] R. Angelova, S. Siersdorfer, and G. Weikum. A neighborhoodbased approach for clustering of linked document collections. Research Report MPI-I-2006-5-005, 2006.

[2] R. Angelova and G. Weikum. Graph-based text classification: Learn from your neighbors. In *ACM SIGIR '06*, 2006.

[3] J. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics, 28:100-108*, 1979.

[4] S. Z. Li. *Markov random field modeling in image analysis.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[5] L. Pelkowitz. A continuous relaxation labeling algorithm for markov random fields. 20:709–715, 1990.

[6] A. Schenker, H. Bunke, M. Last, and A. Kandel. Graph-theoretic techniques for web content mining. *Series in Machine Perception and Artificial Intelligence*, 62, 2005.

[7] C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comp.*, C20:68–86, 1971.