# Discovering Communities in Linked Data by Multi-View Clustering

Isabel Drost, Steffen Bickel, and Tobias Scheffer

Humboldt-Universität zu Berlin, Institut für Informatik
Unter den Linden 6, 10099 Berlin, Germany
{drost, bickel, scheffer}@informatik.hu-berlin.de

**Abstract.** We consider the problem of finding communities in large linked networks such as web structures or citation networks. We review similarity measures for linked objects and discuss the k-Means and EM algorithms, based on text similarity, bibliographic coupling, and co-citation strength. We study the utilization of the principle of multi-view learning to combine these similarity measures. We explore the clustering algorithms experimentally using web pages and the CiteSeer repository of research papers and find that multi-view clustering effectively combines link-based and intrinsic similarity.

## 1 Introduction

Citation Analysis has originally been carried out manually (Garfield, 1972), but many discovery tasks in this problem area can be automated. Finding communities in linked networks is a sub-problem of citation analysis. The task here is to find clusters of thematically related papers or web pages (White & McCain, 1989; Kautz et al., 1997; Getoor, 2003) where objects within clusters are similar and dissimilar between clusters.

When clustering publications or web pages it seems appropriate to make use of the similarity of their textual content. Yet also the inbound and outbound links can be used to define the similarity of two documents. The $k$-means algorithm already has been applied to citation analysis (Hopcroft et al., 2003). The EM algorithm (Dempster et al., 1977), and the recently developed multi-view clustering method (Bickel & Scheffer, 2004), appear to be suitable. But it is not clear how these approaches differ in terms of cluster quality.

We discuss how partitioning cluster algorithms can be applied to linked data. We review vector space representations of linked documents and their correspondence to the bibliographic coupling and co-citation similarity measures. We study appropriate distributional models that can be used to instantiate EM. When having different measures of similarty at hand the natural question is whether algorithms can use a combination. We develop an undirected graph model and use multi-view clustering algorithms. A comparative analysis of the resulting clustering methods leads us to results on their cluster quality. We obtain results on the benefit of the co-citation, bibliographic

Conference of the German Classification Society, 2005.

coupling, the undirected, and the multi-view model. Additionally we compare link based clustering to clustering based on the textual content of papers or web pages.

The rest of the paper is organized as follows. Section 2 reviews related work, we describe the problem setting in Section 3. In Section 4, we discuss clustering algorithms and their application for citation analysis. Section 5 presents empirical results, and Section 6 concludes.

## 2 Related Work

Citation analysis dates back to Garfield (1972) who proposed the impact factor as a performance measure for journals. White and McCain (1989) coined the term *bibliometrics* for automated analysis of citation data. Bibliometrics focuses on two graphs: the *co-citation graph* (White & McCain, 1989) relates papers by the proportion of jointly cited work. The *collaboration graph* (White, 2003), by contrast, relates papers by jointly authored research papers (the mathematician Pál Erdős is believed to be the node with highest degree, having more than 500 co-authors).

It is known that many properties (such as the degree of the nodes) of naturally grown graphs, such as citation or social networks, follow power laws (Redner, 1998). This distinguishes them from random graphs (Liljeros et al., 2001; Alberich et al., 2002). Small-world properties are typical for such compounds (Watts & Strogatz, 1998). In this respect, the web exhibits the same properties as a citation network and the same algorithms can be applied to analyze its cluster structure (Gibson et al., 1998; Getoor, 2003).

The problem of clustering web search results has been addressed using modified versions of $k$-means (Modha & Spangler, 2000; Wang & Kitsuregawa, 2001) as well as a spectral clustering algorithm (He et al., 2001); here, the instances are represented using a combination of document content, inbound, and outbound links. The multi-view EM and multi-view $k$-means clustering methods can be applied when each instance has a representation in two distinct vector spaces. In our problem area, those spaces can be inbound links, outbound links, and text content. Multi-view clustering appears interesting for citation analysis because, if this requirement is met, then it often outperforms the regular EM substantially (Bickel & Scheffer, 2004).

## 3 Problem Setting

We consider the problem of clustering linked objects. More precisely, we assume that each document has an unknown "true" class membership. This true class label is not visible to the clustering algorithm, but we use the labels to evaluate the quality of the resulting clusters as the homogeneity of true class memberships within the returned clusters. The homogeneity measure is the entropy of the true classes within the generated clusters (Equation 1). $C$

is a partitioning of the instances $X$ into clusters $c_i$, and $L$ is the (manual) partitioning into true classes $l_j$. Hence, $p(l_j|c_i)$ is the fraction of instances in $c_i$ that have true class label $l_j$. Intuitively, the entropy is the average number of bits needed to encode the true class label of an instance, given its cluster membership. Since the true class memberships are not visible, no algorithm can directly optimize this criterion.

$$E_{C,L} = \sum_{c_i \in C} \frac{|c_i|}{|X|} \left( - \sum_{l_j \in L} p(l_j|c_i) \log p(l_j|c_i) \right) \quad (1)$$

The $k$-means and EM algorithms require instances to be represented in a vector space. Let $V = \{1, \ldots, n\}$ be a universe of documents of which we wish to cluster a subset $X \subseteq V$. Let $E \subseteq V \times V$ be the citation graph; $(x_j, x_k) \in E$ if $x_j$ cites $x_k$. For every $x_j \in X$, we define a vector $\boldsymbol{x}_j^{in}$ of inbound links: $x_{jk}^{in} = 1$ if document $x_j$ *is cited by* $x_k$, and 0 otherwise. The outbound vector $x_j^{out}$ is defined analogously: $x_{jk}^{out} = 1$ if $x_j$ *cites* $x_k$. In addition, we consider the intrinsic, text-based representation $x_j^{txt}$. In the context of $k$-means, $x_j^{txt}$ is a normalized tfidf vector; in the context of multinomial EM, it is a vector that counts, for every word in the dictionary, the number of occurrences in document $x_j$.

Let us review common concepts of similarity for linked documents. Intuitively, the *bibliographic coupling* measures the number of common citations in two papers whereas the *co-citation* is a measure of how frequently two papers are being cited together. That is, the bibliographic coupling of two papers is the correspondence of their sets of documents connected by outbound links whereas the co-citation strength of two papers equals the similarity of their sets of documents connected by inbound links.

The general EM algorithm is instantiated with a model-specific likelihood function. Based on the *bibliographic coupling* this likelihood has to quantify how well the vector of outbound links $x_j^{out}$ of a document $x_j$ corresponds to some cluster; based on *co-citation*, the vector of inbound links $x_j^{in}$ has to be considered. The $k$-means algorithm requires a similarity measure. A natural similarity function based on the *bibliographic coupling* is the cosine between two vectors of outbound links $bc(x_j, x_k) = \frac{\langle x_j^{out}, x_k^{out} \rangle}{|x_j^{out}||x_k^{out}|}$; the *co-citation* similarity $cc(x_j, x_k)$ is defined as the cosine similarity of $x_j^{in}$ and $x_k^{in}$. In the textual view, text similarity $ts(x_j, x_k)$ can naturally be calculated as the cosine between document vectors $x_j^{txt}$ and $x_k^{txt}$.

In addition to the concepts of co-citation and bibliographic coupling, we will also study an undirected model, $x_j^{undir} = x_j^{in} + x_j^{out}$.

## 4 Clustering Algorithms for Citation Analysis

In this section, we discuss how $k$-means and EM clustering can be applied to citation analysis.

### 4.1 Clustering by k-Means

The well known $k$-means algorithm starts with $k$ random mean vectors and then, in turns, assigns each instance to the cluster with nearest mean vector and re-calculates the means by averaging over the assigned instances as long as there is a change in the cluster assignments.

### 4.2 EM for Citation Analysis

The Expectation Maximization algorithm (Dempster et al., 1977) can be used for maximum likelihood estimation of mixture model parameters. Applied to citation analysis, the mixture components are the clusters of related papers that we wish to identify. We get cluster assignments from the estimated mixture model by assigning each instance $x_j$ to the cluster of highest *a posteriori* probability $\text{argmax}_i P(c_i|x_j)$.

We introduce the multinomial citation model for clustering linked data. In this model, a paper has a certain number $n$ of links, where $n$ is a random variable governed by $P(n)$. Each of these $n$ links is a random variable that can take $|V|$ distinct values, it is governed by a cluster-specific distribution $\theta_i(x_k)$. References are drawn without replacement as there can be at most one link between each pair of papers. The distribution of $n$ random variables with $|V|$ values, drawn without replacement, is governed by the multi-hypergeometric distribution.

The multi-hypergeometric distribution is the generalization of the hypergeometric distribution for non-binary variables. Unfortunately, it is computationally infeasible because calculation of probabilities requires summation over a huge trellis and even a lookup-table is impractically large. Since the number of links in a paper is much smaller than the number of papers in $V$, it can be approximated by the multinomial distribution. This corresponds to drawing citations with replacement. The likelihood in the multinomial citation model is given in Equation 2. The "$n!$" term reflects that there are $n!$ ways of drawing any given set of $n$ citations in distinct orderings.

$$P_\Theta(x_j|c_i) = \prod_{x_k \in V} P(n) n! \theta_i(x_k)^{x_{jk}} \tag{2}$$

Again, $x_j = x_j^{in}$ for co-citation and $x_j = x_j^{out}$ for bibliographic coupling. The E and M steps for the multinomial model are given in Equations 3, 5, and 6 (posterior and maximum likelihood estimator for the multinomial distribution are well-known). As we see in Equation 4, it is not necessary to know $P(n)$ if only the posterior $P_\Theta(c_i|x_j)$ is of interest. We can apply Laplace smoothing by adding one to all frequency counts.

$$\text{E step: } P_\Theta(c_i|x_j) = \frac{\pi_i P_\Theta(x_j|c_i)}{\sum_k \pi_k P_\Theta(x_j|c_k)} = \frac{\pi_i \prod_{x_l \in V} P(n) n! \theta_i(x_l)^{x_{jl}}}{\sum_k \pi_k \prod_{x_l \in V} P(n) n! \theta_k(x_l)^{x_{jl}}} \tag{3}$$

$$= \frac{\pi_i \prod_{x_l \in V} \theta_i(x_l)^{x_{jl}}}{\sum_k \pi_k \prod_{x_l \in V} \theta_k(x_l)^{x_{jl}}} \tag{4}$$

$$\text{M step: } \theta_i(x_k) = \frac{\sum_{x_l \in X} x_{lk} P(c_i|x_l, \Theta)}{\sum_{j \in V} \sum_{x_l \in X} x_{lj} P(c_i|x_l, \Theta)} \tag{5}$$

$$\pi_i = \frac{1}{|X|} \sum_{x_k \in X} P_\Theta(c_i|x_k) \tag{6}$$

The multinomial distribution is also frequently used as a model for text. In the multinomial text model, words are drawn with replacement according to a cluster-specific distribution $\theta_i(x_k)$. The likelihood of a document $x_j = x_j^{txt}$ in cluster $c_i$; can be characterized analogously to Equation 2; the E and M steps for the multinomial text model follow Equations 4 and 6, respectively (with $x = x^{txt}$).

### 4.3 Combining Text Similarity, Co-Citation and Bibliographic Coupling

The methods that we studied so far can be applied using text similarity, co-citation, or bibliographic coupling as similarity metric. It is natural to ask for the most effective way of combining these measures. A baseline for the combination of inbound and outbound links that we consider is the *undirected model* (Section 3) in which inbound and outbound links are treated alike.

We study the multi-view clustering model (Bickel & Scheffer, 2004). Multi-view clustering can be applied when instances are represented in two distinct (ideally independent) views. Here, distinct views naturally are $x^{in}$, $x^{out}$, and $x^{txt}$. Two interleaving EM algorithms then learn the parameters of distinct models, each model clusters the data in one of the views. The parameters are estimated such that they maximize the likelihood plus an additional term that quantifies the consensus between the two models.

This approach is motivated by a result of Dasgupta et al. (2002) who show that the probability of a disagreement of two independent hypotheses is an upper bound on the probability of an error of either hypothesis. Table 1 briefly summarizes the multi-view clustering algorithm (Bickel & Scheffer, 2004). In our experiments, we study multi-view $k$-means and multi-view EM with multinomials.

The multi-view clustering algorithm returns two parameter sets $\Theta^{(1)}$ and $\Theta^{(2)}$ and two clustering hypotheses, one in each view. A unified cluster assignment can be obtained by using the argmax of a combined posterior applying bayes rule and a conditional independence assumption (Equation 7). Equation 7 needs the definition of a combined prior $\pi_i$, we use $\pi_i = \frac{1}{2}(\pi_i^{(1)} + \pi_i^{(2)})$.

$$P_\Theta(c_i|x_j) = \frac{\pi_i P_\Theta(x_j|c_i)}{\sum_k \pi_k P_\Theta(x_j|c_k)} = \frac{\pi_i P_{\Theta^{(1)}}(x_j^{(1)}|c_i) P_{\Theta^{(2)}}(x_j^{(2)}|c_i)}{\sum_k \pi_k P_{\Theta^{(1)}}(x_j^{in}|c_k) P_{\Theta^{(2)}}(x_j^{(2)}|c_k)} \tag{7}$$

In the multi-view $k$-means algorithm, we assign an example $x_j$ to the cluster with $\text{argmax}_i \frac{\langle x_j^{(1)}, m_i^{(1)} \rangle}{|x_j^{(1)}||m_i^{(1)}|} \frac{\langle x_j^{(2)}, m_i^{(2)} \rangle}{|x_j^{(2)}||m_i^{(2)}|}$, where $m_i^{(1)}$ and $m_i^{(2)}$ are the mean vectors of the $i$-th cluster in the respective view.

**Table 1.** Multi-view Clustering.

Input: instances $\{(x_1^{(1)}, x_1^{(2)}), \ldots, (x_m^{(1)}, x_m^{(2)})\}$.

1. Randomly initialize parameters $\Theta^{(2)}$ in view (2).
2. E step in view (2): compute posterior $P(c_i|x_j^{(2)}, \Theta^{(2)})$ of cluster membership given the model parameters in view (2).
3. **Until** convergence:
   (a) **For** $v \in \{(1), (2)\}$:
      i. M step in view $v$: Find model parameters $\Theta^v$ that maximize the likelihood given the posterior $P(c_i|x_j^{\bar{v}}, \Theta^{\bar{v}})$ computed in the last step.
      ii. E step in view $v$: compute posterior $P(c_i|x_j^v, \Theta^v)$ of cluster membership given the model parameters in the current view.
   (b) **End For**.
4. Return combined model $\Theta = \Theta^{(1)} \cup \Theta^{(2)}$.

## 5 Comparative Analysis

In this section, we will investigate the relative benefit of the different algorithms and representations in terms of cluster quality and regarding different applications (scientific publications or web pages). In order to measure the cluster quality as the average entropy (Equation 1) we use manually defined labels that are hidden to the clustering algorithms. For our experiments we use the CiteSeer data set (3,312 scientific publications, six classes) (Lu & Getoor, 2003) and the well known WebKB collection (8,318 university web pages, six classes).

Let us first study how the different clustering methods compare in terms of cluster quality for only link-based representations. Fig. 1 shows the averaged cluster quality over ten runs of multinomial EM and $k$-means for both data sets. Error bars indicate standard error (in most cases unperceivably small).

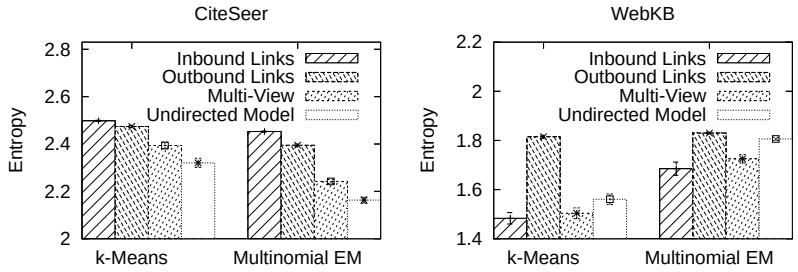


**Fig. 1.** Cluster entropy for link-based clustering.

The multinomial model fits the CiteSeer data best. Simple $k$-means clustering gives the best performance for WebKB. For this problem, the inbound links (co-citation) contain the most relevant information and lead to the best results. For the CiteSeer data, the undirected model works best.

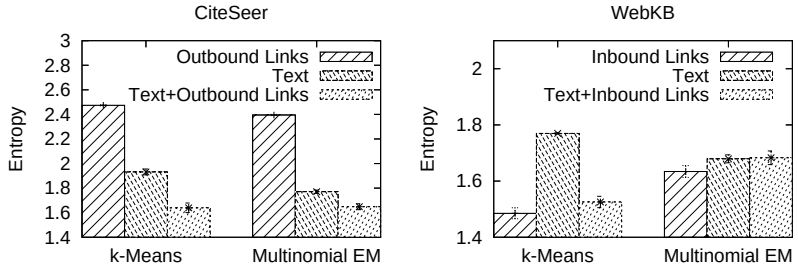


**Fig. 2.** Cluster entropy for link- and text-based clustering.

In Figure 2 we want to answer the question whether the usage of textual content has a positive impact on cluster quality. For CiteSeer, we combine outbound link information and text because outbound links lead to a better clustering results; for WebKB we combine inbound links information and text for the same reason. For CiteSeer combining textual content and link information by multi-view EM works better than each of the single approaches. For the WebKB data, combining link and text information did not lead to an improvement in clustering quality. It is remarkable, that for WebKB data the inlinks seem to contain far more valuable information for clustering than the textual content of the web pages. We also ran experiments with concatenated text and link vectors. Yet for all datasets and algorithms, clustering quality was significantly worse in comparison to multi-view clustering.

## 6 Conclusion

We analyzed how partitioning clustering algorithms can be applied to the problem of finding communities in linked data using similarity metrics based on co-citation, bibliographic coupling, and textual similarity as well as a combinations of them. For the combination of different similarity metrics we considered an undirected and a multi-view model. We motivated and discussed the multinomial distributional model for citation data that can be used to instantiate general EM.

Experiments show that for publication citation analysis (CiteSeer data) the combination of different measures always improves the clustering performance. The best performance is achieved with the multi-view model based on outlink and textual data. By contrast, for web citation analysis (WebKB data) the inbound links are most informative and combining this measure with others (outbound links or text) deteriorates the performance.

## Acknowledgment

This work was supported by the German Science Foundation DFG under grant SCHE 540/10-1. We thank Lise Getoor for kindly providing us with the CiteSeer data set.

# Bibliography


Alberich, R., Miro-Julia, J., & Rosselló, F. (2002). *Marvel universe looks almost like a real social network*Preprint). arXiv id 0202174.

Bickel, S., & Scheffer, T. (2004). Multi-view clustering. *IEEE International Conference on Data Mining*.

Dasgupta, S., Littman, M. L., & McAllester, D. (2002). Pac generalization bounds for co-training. *Advances in Neural Information Processing Systems 14* (pp. 375–382). Cambridge, MA: MIT Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journ. of Royal Stat. Soc. B*, *39*.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*, 471–479.

Getoor, L. (2003). Link mining: A new data mining challenge. *SIGKDD Exploration 5*.

Gibson, D., Kleinberg, J. M., & Raghavan, P. (1998). Inferring web communities from link topology. *UK Conference on Hypertext* (pp. 225–234).

He, X., Ding, C. H. Q., Zha, H., & Simon, H. D. (2001). Automatic topic identification using webpage clustering. *ICDM* (pp. 195–202).

Hopcroft, J., Khan, O., & Selman, B. (2003). Tracking evolving communities in large linked networks. *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kautz, H., Selman, B., & Shah, M. (1997). The hidden web. *AI Magazine*, *18*, 27–36.

Liljeros, F., Edling, C., Amaral, L., Stanley, H., & Aberg, Y. (2001). The web of human sexual contacts. *Nature*, *411*, 907–908.

Lu, Q., & Getoor, L. (2003). Link-based text classification. *IJCAI Workshop on Text Mining and Link Analysis, Acapulco, MX*.

Modha, D. S., & Spangler, W. S. (2000). Clustering hypertext with applications to web searching. *ACM Conference on Hypertext* (pp. 143–152).

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, *4*, 131–134.

Wang, Y., & Kitsuregawa, M. (2001). Link based clustering of Web search results. *Lecture Notes in Computer Science*, *2118*.

Watts, D., & Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.

White, H. (2003). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, *54*, 423–434.

White, H., & McCain, K. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, *24*, 119–186.