Costco: Robust Content and Structure Constrained Clustering of Networked Documents

Su Yan[†], Dongwon Lee[‡], and Alex Hai Wang[♯]

[†]IBM Almaden Research Center
San Jose, CA 95120, USA
[‡♯] The Pennsylvania State University
[‡]University Park, PA 16802, USA
[♯]Dumore, PA 18512, USA
syan@us.ibm.com, {dongwon, hwang}@psu.edu

Abstract. Connectivity analysis of networked documents provides high quality link structure information, which is usually lost upon a content-based learning system. It is well known that combining links and content has the potential to improve text analysis. However, exploiting link structure is non-trivial because links are often noisy and sparse. Besides, it is difficult to balance the term-based content analysis and the link-based structure analysis to reap the benefit of both. We introduce a novel networked document clustering technique that integrates the content and link information in a unified optimization framework. Under this framework, a novel dimensionality reduction method called COntent & STructure COnstrained (Costco) Feature Projection is developed. In order to extract robust link information from sparse and noisy link graphs, two link analysis methods are introduced. Experiments on benchmark data and diverse real-world text corpora validate the effectiveness of proposed methods.

Key words: link analysis, dimensionality reduction, clustering

1 Introduction

With the proliferation of the World Wide Web and Digital Libraries, analyzing "networked" documents has increasing challenge and opportunity. In addition to text content attributes, networked documents are correlated by links (e.g., hyperlinks between Web pages, citations between scientific publications etc.). These links are useful for text processing because they convey rich semantics that are usually independent of word statistics of documents [8].

Exploiting link information of networked documents to enhance text classification has been studied extensively in the research community [3, 4, 6, 14]. It is found that, although both content attributes and links can independently form reasonable text classifiers, an algorithm that exploits both information sources has the potential to improve the classification [2, 10]. Similar conclusion has been drawn for text clustering by a growing number of works [1, 2, 7, 11, 13, 20]. However, the fundamental question/challenge still remains

How to effectively couple the content and link information to get the most of both sources?

Existing work either relies on heuristic combination of content and links, or assumes a link graph to be dense or noise-free, whereas link graphs of real-world data are usually sparse and noisy. To this end, we propose a novel clustering approach for networked documents based on the *COntent and STructure COnstrained (Costco) feature projection*, and cluster networked documents from a *dimension reduction* perspective. Compared to existing work, Costco has the following advantages

1. Couples content and link structure in a unified objective function, and hence avoids heuristic combination of the two information sources;
2. Alleviates the curse-of-dimensionality problem by constrained dimensionality reduction;
3. Does not rely on dense link structure and is robust to noisy links, which suits the method well for real-world networked data;
4. Is very simple to implement, so can be used for exploratory data analysis before any complicated in-depth analysis.

2 Related Work

The techniques for analyzing networked documents can be broadly categorized as content-based, link-based, and combined approaches. As more and more work confirm the effectiveness of using link structure to enhance text analysis, novel approaches to merge content and link information attract increasing interest in the text mining domain.

[6] proposes generative probabilistic models for document content and links. [4] uses factorized model to combine the content model and the link model. [14] tackles the problem by using the relaxation labeling technique. Besides the vast amount of work on link-enhanced text classification, there are increasing number of work focusing on link-enhanced clustering. [1] extends the relaxation labeling method to text clustering. The cluster assignment for each document is not only determined by content attributes, but is also influenced by the assignments of neighborhood documents on the link graph. [2] focuses on clustering scientific literature, and weights words based on link information. [11] extends the term-based feature space with in-link and out-link features. [7] treats networked document clustering as a spectral graph partitioning problem. [13] shares a similar idea of adopting graph-partitioning techniques, but merges content and links by weighting the link graph with a content similarity metric. Our technique is orthogonal to all the existing work by clustering networked documents from a dimension reduction perspective and is robust to sparse and noisy link graphs.

3 Main Proposal

3.1 Problem Statement

Text data, usually represented by the bag-of-words model, have extremely high-dimensional feature space (1000+). A feature projection approach can greatly reduce the feature space dimensionality while still preserve discriminative information. In the networked environment, semantically related documents tend to cite each other. If the link structure is noise-free and dense enough, then link-based clustering augmented by textual content [1, 2], will generally yield well separated clusters . However, the link structure is often noisy and sparse. For instance, many links in Web pages are for navigational purpose and therefore not indicators of semantic relations [15]. We introduce an algorithm to bridget the disconnect between text and link structure from a feature projection perspective.

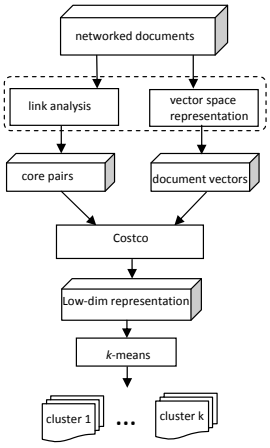


Fig. 1. Framework of Costco-based networked document clustering

The overall clustering framework is outlined in Figure 1. Given networked documents, two preprocessing steps are performed. On the one hand, link analysis is performed to extract *core pairs*, which are pairs of documents strongly correlated with each other according to the link structure. On the other hand, the vector space model is employed to convert documents into high-dimensional vectors. Each dimension is a word after preprocessing (stopping, stemming etc.). Core pairs and document vectors are then input into the feature projection module Costco. The generated low-dimensional data are partitioned by the traditional k-means clustering method into k clusters, where k is the desired number of clusters provided by users.

3.2 Local Link Analysis

The link graphs of real-world networked documents are usually sparse and noisy. Instead of naively assuming a pair of connected documents being similar in topic, we need schemes to extract more robust link information from the graph. A local link analysis scheme is introduced in this section.

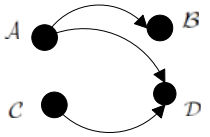


Fig. 2. Cociting vs. Cocited

We model a link graph as *directed and unweighted*, denoted by $G(\mathbb{V}, \mathbb{E})$, where $\mathbb{V}$ is the set of the vertices/documents, and $\mathbb{E}$ is the set of edges/links between vertices. If document d_i links to/cites document d_j, then there is an edge of unit weight starting from d_i and pointing to d_j. Let matrix $L \in \mathbb{R}^{n \times n}$, where

n is the number of documents, be the corresponding *link matrix* defined as

$$L_{i,j} = \begin{cases} 1 \; d_i \;\; cites \;\; d_j \\ 0 \;\; otherwise. \end{cases} \tag{1}$$

L embodies two types of document concurrences: *cociting* and *cocited*, as illustrated in Figure 2. For example, both $\mathcal{A}$ and $\mathcal{C}$ cites $\mathcal{D}$, and $\mathcal{B}$ and $\mathcal{D}$ are being cocited by $\mathcal{A}$.

In order to capture the concurrences, two adjacency matrices $X \in \mathbb{R}^{n \times n}$ and $Y \in \mathbb{R}^{n \times n}$ are calculated

$$X_{i,j} = \frac{|L_{i*} \cap L_{j*}|}{|L_{i*} \cup L_{j*}|}, \;\; 0 \leq X_{i,j} \leq 1 \tag{2}$$

$$Y_{i,j} = \frac{|L_{*i} \cap L_{*j}|}{|L_{*i} \cup L_{*j}|}, \;\; 0 \leq Y_{i,j} \leq 1 \tag{3}$$

where L_{i*} and L_{*i} represent the i-th row vector and column vector of L respectively. $X_{i,j}$ measures the Jaccard similarity of two documents d_i and d_j in terms of the cociting pattern, and $Y_{i,j}$ measures the similarity of the cocited pattern. Combining the two concurrences patterns, we have

$$Z = \alpha X + (1 - \alpha) Y \tag{4}$$

where $\alpha \in [0, 1]$ is the parameter that controls the contribution of each individual link pattern to the overall structure-based similarity. Given Z, the set $\mathbb{C}$ of core pairs is then defined as

$$\mathbb{C} = \{(d_i, d_j) | Z_{i,j} > \theta\} \tag{5}$$

where θ is a threshold that controls the reliability of link-based similarities.

3.3 Global Link Analysis

The link analysis scheme introduced in the previous section is a "local" method in the sense that for any query vertex/document in the graph, only the links between the query vertex and its direct neighbors are considered. Local analysis can miss some informative document pairs. For example in Figure 3, the relations among $\mathcal{A}$, $\mathcal{B}$, $\mathcal{D}$ and $\mathcal{E}$ are lost.

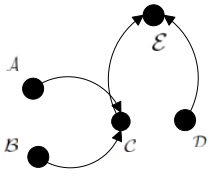


Fig. 3. Local method misses informative pairs

In the global scheme, we define a Markov random walk on the link graph. The link graph is modeled as *undirected and weighted*, denoted as $\tilde{G} = (\tilde{\mathbb{V}}, \tilde{\mathbb{E}})$. If there is a link between two documents d_i and d_j, we consider a relation (thus an edge) exits between them, no matter who starts the link. The edge is further weighted by the pairwise similarity $\mathfrak{D}(d_i, d_j)$ of the two documents. Let matrix $W \in \mathbb{R}^{n \times n}$, where $w_{i,j} = \mathfrak{D}(d_i, d_j)$, be the weight matrix. The one-step transition probabilities p_{ik}, which are the probabilities of jumping from any state (vertex) i to one of its adjacent state k, are obtained

directly from these weights $p_{ik} = W_{ik} / \sum_j W_{ij}$. We can organize the one step transition probabilities as a matrix P whose i,k-th entry is p_{ik}.

Due to the sparseness of a link graph, two documents that are strongly correlated in topics may not be linked together. For example, a scientific article can not cite all the related work, and several Web pages with similar topics may scatter in the Web without any link among them. To remedy this problem, for each vertex whose degree is below the average, we add artificial links between the vertex and its s nearest neighbors where s is a small number.

For the augmented link graph, the transition matrix P has the property that $Pe = e$, i.e., P is stochastic, where e is the vector with all 1 elements. We can now naturally define the Markov random walk on the undirected graph $\tilde{G}$ associated with P. The relation between two documents is evaluated by an important quantity in Markov chain theory, the *expected hitting time* $h(j|i)$, which is the expected number of steps for a random walk started at state i to enter state j for the first time. Formally, $h(j|i)$ is defined as

$$\begin{cases} h(i|i) = 0 \\ h(j|i) = 1 + \sum_{k=1}^{n} p_{ik} h(j|k) \; i \neq j \end{cases} \tag{6}$$

The choice of using expected hitting time to evaluate the correlation between two documents is justified by the desired property that the hitting time from state i to state j decreases when the number of paths from i to j increases and the lengths of the paths decrease. The core pairs can be naturally defined as

$$\mathbb{C} = \{(d_i, d_j) | (h(j|i) + h(i|j))/2 < \gamma\} \tag{7}$$

for some threshold γ.

3.4 Content & Structure Constrained Feature Projection (Costco)

Let matrix $D \in \mathbb{R}^{f \times n}$ be the document-term matrix where each column d_i is a vector in the f-dimensional space. Let $\{(d_{j,1}, d_{j,2})\}_{j=1...m}$ be the set of m document pairs that have been identified as core pairs at the link analysis step. Since these pairs of documents are strongly connected according to the link structure, there is a high probability that a core pair of documents are also semantically similar. We then desire a projection direction, such that any two documents of a core pair will be *more similar* to each other after being projected along the direction. To achieve this goal, we can minimize the variance between a pair of documents. Let us define the covariance matrix V to encode the pooled variances for all the core pairs

$$V = \frac{1}{m} \sum_{\{(d_{j,1}, d_{j,2})\} \in \mathbb{C}} (d_{j,1} - d_{j,2})(d_{j,1} - d_{j,2})^T \tag{8}$$

Then the desired projection is

$$S^* = \arg\min_S Tr(S^T V S) \tag{9}$$

Algorithm 1: Networked Document Clustering Based on Costco.

Input : A set of n networked documents
Desired # clusters k
Desired # dimensionality r
Output: a set of clusters
begin link analysis
  Extract *core pairs* $\mathbb{C}$ by local link analysis (Eq. 5)
  or global link analysis (Eq. 7)
begin content analysis
  Represent n documents using vector space model to get $D \in \mathbb{R}^{f \times n}$;
Construct covariance matrix U (Eq. 10);
Construct covariance matrix V (Eq. 8);
Solve Eq. 11 to get low-dimensional data as $\widehat{D} = S^T D$;
Clustering low-dimensional data: k-means($\widehat{D}$, k);
return a set of clusters;

where $S \in \mathbb{R}^{f \times r}$ denotes the optimal transformation matrix, r is the desired subspace dimensionality provided by users, and $Tr(\cdot)$ is the *trace* of a square matrix, defined as the summation of the diagonal elements.

Directly minimizing Eq. 9 leads to trivial solutions. For example, if the entire data set is projected to one point, then the covariance between core pair documents is minimized. To avoid trivial solution, we can put constrains on the variance of the entire data set to prevent all the data points huddle together. The covariance matrix of the entire data set is defined as

$$U = \frac{1}{n} \sum_{i=1}^{n} (d_i - \mu)(d_i - \mu)^T \tag{10}$$

where $\mu = \sum_{i=1}^{n} d_i$ is the global mean. Accordingly, we define the following objective

$$\begin{aligned} S^* &= \arg\max_S Tr \frac{S^T U S}{S^T V S} \\ &= \arg\max_S Tr((S^T V S)^{-1}(S^T U S)) \end{aligned} \tag{11}$$

The objective function defines a linear feature projection direction that both maximally preserves the variations of the entire data set and minimizes the total variances of core pairs. Simply put, after being projected along the optimal projection direction, the documents that are strongly connected (according to link structure) will be more similar to each other, while the rest documents are still well separated.

After the transformation matrix S is solved, the high-dimensional (f-dim) data can be optimally represented in the r-dim subspace as $\widehat{D} = S^T D$, where $\widehat{D} \in \mathbb{R}^{r \times n}$, $r \ll f$. The optimization problem of Eq. 11 is a general eigenvector problem. Usually a regularization term is added to solve an ill-posed problem or to prevent overfitting [12]. We skip detailed discussion about it due to space limit. The overall clustering scheme is outlined in Algorithm 1.

Table 1. UCI data sets

Datasets	# classes	# instances	# features
balance	3	625	4
vehicle	4	846	18
breast-cancer	2	569	30
sonar	2	208	60
ionoshpere	2	351	34
soybean	4	47	35

Table 2. 20-Newsgroups data sets

Datasets	topics	# features
difficult	comp.windows.x, comp.os.ms-windows.mis, comp.graphics	3,570
mediocre	talk-politicis.misc, talk.politics. guns, talk.politics.mideast	4,457
easy	alt.atheism, sci.space, rec.sprot.baseball	4,038

Table 3. Reuters data sets

Datasets	# classes	# instances	# features
reu4	4	400	2,537
reu5	5	500	2,257
reu6	6	600	2,626

Table 4. WebKB and Cora Data sets

Datasets	# classes	# instances	# features	# links
WebKB	5	877	1,703	1,608
Cora	7	2,708	1,433	5,429

4 Performance Evaluations

4.1 Set-up

The proposed networked document clustering framework has been evaluated on 6 UCI benchmark data sets [1], 3 data sets generated from the 20-Newsgroups document corpus [2], 3 data sets generated from the Reuters document corpus [3], the WebKB data sets [4] of hypertext, and the Cora data set[4] of scientific publications. Statistics of the data sets are listed in Table 1 to Table 4.

For the 20-Newsgroups document corpus, 3 data sets are generated, each of which is a balanced combination of documents about 3 topics. Depending on the similarities in the topics, the 3 data sets show various levels of clustering difficulties. To generate the Reuters data sets, for a given number of topics b, firstly, b topics are randomly sampled, and then about 100 documents of each topic are randomly sampled and mixed together. Table 3 shows the average statistics of 5 sets of independently generated data sets.

Spherical k-means [5] the Normalized Cut (NC) [19] [5] are chosen as baseline clustering methods. Both techniques have shown success in clustering text data [9]. Costoco and nr-Costco are our proposals with and without regularization respectively. For competing dimensionality reduction techniques, we compare to two well-known unsupervised dimensionality reduction methods, the principal component analysis (PCA)[16] which is a linear method and the locally linear embedding (LLE)[17][6] which is a non-linear method. For competing techniques that couple content and link information, we implement *Augmented*[11] and *L-Comb* [7, 13]. *Augmented* augments the content-based vector space model with link features and applies k-means to the augmented document vectors. *L-Comb*

[1] `http://archive.ics.uci.edu/ml/`

[2] `http://people.csail.mit.edu/jrennie/20Newsgroups/`

[3] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[4] `http://www.cs.umd.edu/~sen/lbc-proj/LBC.html`

[5] original authors' implementation is used. `http://www.cis.upenn.edu/~jshi/software/`

[6] original authors' implementation is used `http://www.cs.toronto.edu/~roweis/lle/`

linearly combines content similarity with link similarities and uses NC as the underlying clustering scheme. The method *Links* is a k-means clustering based on link similarity only.

To avoid biased accuracy results using a single metric, we used three widely-adopted clustering evaluation metrics: 1) *Normalized Mutual Information* (NMI), 2) *Rand Index* (RI), and 3) *F-measure*.

4.2 Controlled Experiments

In controlled experiment, given a data set, artificial links are generated and inserted between data points. In this way, we can control the density of a link graph as well as the error rate of links, and evaluate a method with various settings. Every method that uses link information will take use of all the available links instead of pruning out some links with preprocessing steps. With controlled experiments, clustering schemes can be evaluated in a fair setting without being influenced by preprocessing.

Table 5. Performance on UCI data sets measured by RI and F (noise-free) (best results are bold-faced)

Datasets	# of links	FF(kmeans)	PCA	LLE	Augmented	FF(NC)	L-Comb(NC)	**Costco**	**nr-Costco**
balance		0.1806	0.6177	0.5730	0.5911	0.6706	0.6772	**0.7151**	0.7132
vehicle		0.6462	0.6408	0.6507	0.6431	0.6709	0.6761	**0.7404**	0.7180
breast-cancer	400	0.7504	0.7504	0.6356	0.7504	0.7554	0.7541	**0.8008**	0.7486
sonar	RI	0.5032	0.5032	0.5031	0.5041	0.5043	0.5046	**0.6700**	0.5749
ionosphere		0.5889	0.5889	0.5933	0.5889	0.5841	0.5841	**0.6509**	0.6196
soybean		0.8283	0.8291	0.7761	0.9065	0.8372	0.8372	**1.0000**	**1.0000**
balance		0.4629	0.5010	0.4506	0.4658	0.5686	0.5771	**0.6290**	0.6270
vehicle		0.3616	0.3650	0.3597	0.3635	0.3594	0.3730	**0.5365**	0.4785
breast-cancer	400	0.7878	0.7878	0.6520	0.7878	0.7914	0.7905	**0.8330**	0.7866
sonar	F	0.5028	0.5028	0.6042	0.5064	0.5041	0.5048	**0.6828**	0.5945
ionosphere		0.6049	0.6049	0.6580	0.6049	0.5997	0.5997	**0.7346**	0.7188
soybean		0.6761	0.6805	0.5485	0.8282	0.6716	0.6716	**1.0000**	**1.0000**

To generate artificial links, we sample the cluster membership relation of pairs of documents and uniformly pick x pairs to add links in. Given an error rate e of links, we control the samples such that $\lceil x * e \rceil$ pairs of documents belong to different topic, which means these links are noise.

Coupling Content and Links. We first fix the error rate of links to be zero $e = 0$, and vary graph density by introducing x =100 to 800 links between documents. This experiment measures the performance of a method in the noise-free setting with various levels of graph density. Figure 4, 5 and 6 show the clustering performance measured by NMI for the UCI, 20-Newsgroups, and Reuters data sets respectively. Table 5, 6 and 7 show the same result measured by RI and F score, with fixed 400 pairs of links. For all the data sets and different graph density levels, Costco consistently and significantly outperforms other competing methods. Notice that, L-Comb and Augmented improve clustering accuracy for some data sets i.e., *vehicle*, *balance*, *easy*, but do not consistently perform well for all the data sets.

Table 6. Performance on 20-Newsgroup data sets measured by RI and F (noise-free) (best results are bold-faced)

Datasets	# links	FF(kmeans)	PCA	Augmented	(FF)NC	L-Comb(NC)	Costco	nr-Costco
difficult		0.5231	0.3910	0.4111	0.4493	0.4506	**0.7868**	0.5543
mediocre	400	0.5865	0.4579	0.4674	0.7105	0.7499	**0.9375**	0.6488
easy	RI	0.6858	0.2350	0.1610	0.9251	**0.9431**	0.9256	0.5565
difficult		0.4424	0.4792	0.4786	0.4681	0.4660	**0.7157**	0.5444
mediocre	400	0.5299	0.4926	0.5088	0.6686	0.7072	**0.9064**	0.5978
easy	F	0.8375	0.4725	0.4725	0.9781	**0.9833**	0.9746	0.6370

Table 7. Performance on Reuters data sets measured by RI and F (noise-free) (best results are bold-faced)

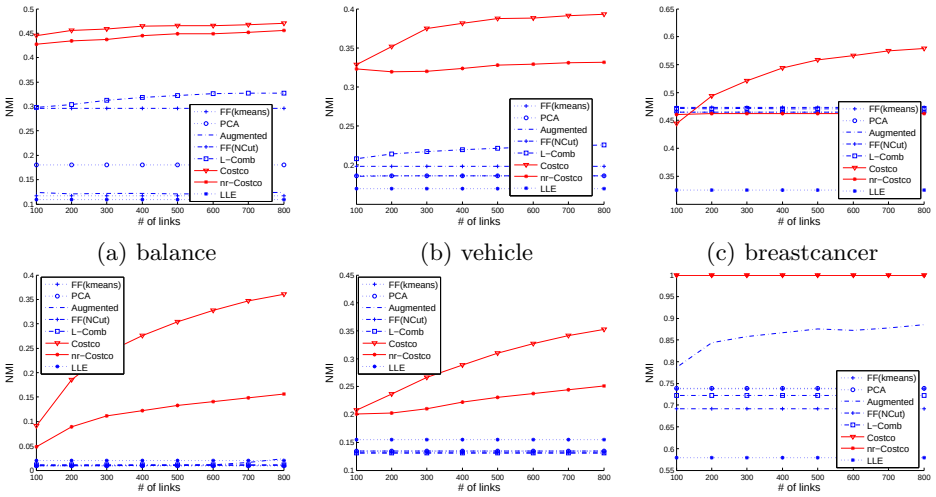
Datasets	# links	FF(kmeans)	PCA	Augmented	(FF)NC	L-Comb(NC)	Costco	nr-Costco
Reu4		0.6422	0.6694	0.6227	0.8141	0.8241	**0.9891**	0.8996
Reu5	400	0.8172	0.7484	0.6626	0.8358	0.8405	**0.9781**	0.8973
Reu6	RI	0.8563	0.6127	0.5433	0.9046	0.8791	**0.9888**	0.8809
Reu4		0.4932	0.5125	0.5297	0.6842	0.6977	**0.9779**	0.8323
Reu5	400	0.6084	0.5285	0.4921	0.642	0.6493	**0.9442**	0.7761
Reu6	F	0.6092	0.3966	0.3596	0.7240	0.6882	**0.9657**	0.6974

Robustness to link errors. Follow a similar setting of the previous experiment, we now fix the density of link graphs to have $x = 400$ pairs of links, but vary the error rate e of links from 0 to 1. Figure 7 shows the behavior of Costco for 3 representative data sets (results on other data sets show similar patterns and thus omitted). As long as most of the links are informative (i.e., the percentage of noisy links is below 50%), without any link-pruning preprocessing steps, regularized Costco always improve clustering accuracy. These results indicate the robustness of Costco to noisy link graphs.

Dimensionality Reduction. In our experiments, for UCI data sets which have relatively low-dimensional features, the reduced dimensionality r is fixed to be the half of the original dimensionality. For text data sets, the reduced dimensionality is set to 40 (this number does not change the relative performance comparison among competing methods). As reported results show, Costco always outperforms the other two unsupervised dimensionality reduction method, PCA and LLE. The performance gain is due to the use of link information. PCA and LLE, however, can not exploit link information even when available. We observed that LLE does not perform well for text data sets, thus did not report its result on text data. This observation is due to the fact that LLE fails to handle sparse and weakly connected data such as text [18].

Table 8. Perormance on Cora and WebKB data sets (best results are bold-faced)

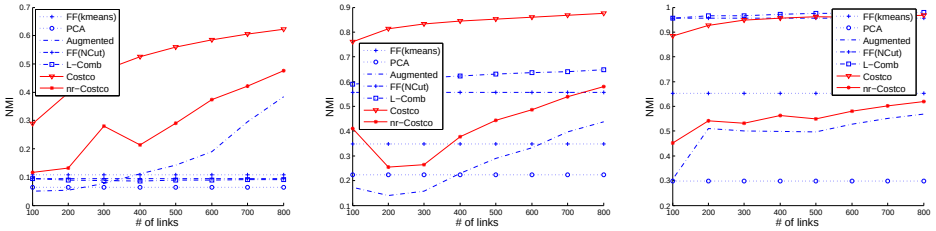
Datasets	*k*means	PCA	Costco	nr-Costco	Links	Augmented	L-Comb(NC)
Cornell	0.2163	0.3058	**0.3809**	0.2054	0.1365	0.2105	0.3544
Texas	0.2276	0.3291	0.3755	0.2163	0.1643	0.3149	**0.4121**
Wisconsin	0.3977	0.4067	**0.4846**	0.2609	0.0977	0.3982	0.4592
Washington	0.3469	0.3352	**0.3885**	0.1599	0.1991	0.3221	0.3404
Cora	0.1361	0.1592	**0.3712**	0.1631	0.0336	0.1496	0.1817



(a) balance (b) vehicle (c) breastcancer

(d) sonar (e) ionoshpere (f) soybean

Fig. 4. Clustering results on UCI data sets



(a) difficult (b) mediocre (c) easy

Fig. 5. Clustering results on 20 Newsgroups data sets

Local vs. Global Link Analysis. In this experiment, instead of using all the available links, Costco adopts the local and global link analyses to extract robust core pairs of documents and does dimensionality reduction accordingly. With fixed 400 links and an error rate of 0.5, Figure 8 shows the clustering results. In most cases, both link analysis methods can prune noise in links and improve clustering performance. Global link analysis usually outperforms local analysis as can be expected.

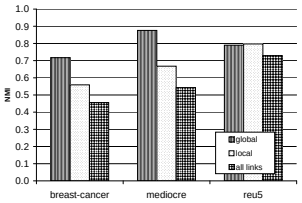
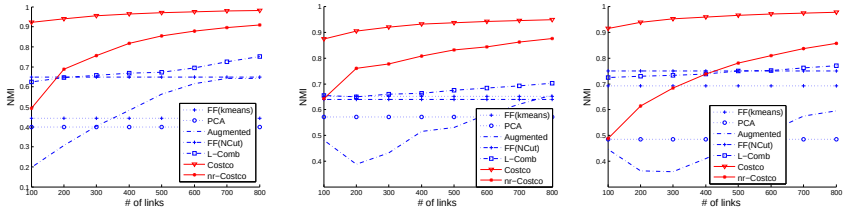


Fig. 8. Link analysis: global vs. local methods

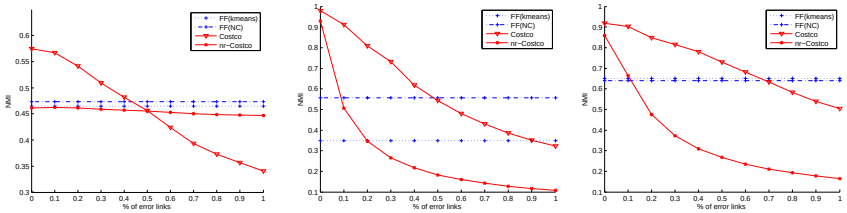
4.3 Unrestrained Experiments

We evaluate all the methods with real-world networked documents. Experimental results are are shown in Table 8. Basically, similar patterns to controlled



(a) Reu4 (b) Reu5 (c) Reu6

Fig. 6. Clustering results on Reuters data sets



(a) breast-cancer (b) mediocre (c) reu5

Fig. 7. Clustering results on Reuters data sets

experiments are observed. For example, in most cases, Costco outperforms competing clustering methods and dimensionality reduction methods. The regularization improves the robustness in clustering performance, and dimensionality reduction in general alleviates the curse-of-dimensionality problem related to text data and generates more accurate data partitions. Note that, because all our data sets have very sparse and noisy link structures, the clustering method *Links*, which entirely relies on link structures, has the worst performance. But when combining link structure with content information, all the three content and link coupling techniques improve clustering performance. This observation confirms the usability of link structure (can be sparse and noisy) in text analysis.

5 Conclusion

A novel clustering model for networked documents is proposed. The Costco feature projection method is designed to represent high dimensional text data in an optimal low-dimensional subspace, and adopts the traditional k-means clustering method to partition the reduce-dimension data. Instead of using a stiff weighted combination of content-based and link-based similarities, Costco explores the correlation between the link structure and the semantic correlations among documents, and constrains the search for the optimal subspace using both content and link information. Local and global link analysis methods are proposed to extract robust link information from noisy and sparse link graphs.

References

1. R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. In *CIKM*, pages 778–779, 2006.
2. L. Bolelli, S. Ertekin, and C. L. Giles. Clustering scientific literature using sparse citation graph analysis. In *PKDD*, pages 30–41, 2006.
3. S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *SIGMOD*, pages 307–318, 1998.
4. D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2000.
5. I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175, 2001.
6. L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *J. Mach. Learn. Res.*, 3:679–707, 2003.
7. X. He, H. Zha, C. H.Q. Ding, and H. D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, 2002.
8. M. Henzinger. Hyperlink analysis on the world wide web. In *HYPERTEXT*, pages 1–3, 2005.
9. X. Ji and W. Xu. Document clustering with prior knowledge. pages 405–412, 2006.
10. F. Menczer. Lexical and semantic clustering by web links. *JASIST*, 55(14):1261–1269, 2004.
11. D. S. Modha and W. S. Spangler. Clustering hypertext with applications to web searching. In *HYPERTEXT*, pages 143–152, 2000.
12. A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40:636–666, 1998.
13. J. Neville, M. Adler, and D. Jensen. Clustering relational data using attribute and link information. In *Proceedings of the IJCAI Text Mining and Link Analysis Workshop*, 2003.
14. H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext catergorization method using links and incrementally available class information. In *SIGIR*, pages 264–271, 2000.
15. H. W. Park and M. Thelwall. Hyperlink analyses of the world wide web: A review. *J. Computer-Mediated Communication*, 8(4), 2003.
16. K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
17. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
18. L. K. Saul, S. T. Roweis, and Y. Singer. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.
19. J. Shi and J. Malik. Normalized cuts and image segmentation. 2000.
20. Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *CIKM*, pages 499–506, 2002.