Web clustering based on the information of sibling pages

Caimei Lu Xiaodan Zhang Jung-ran Park Xiaohua Hu
College of Information Science and Technology, Drexel University
3141 Chestnut Street
Philadelphia, PA 19104, USA
Caimei.lu@drexel.edu, jung-ran.park@cis.drexel.edu {xzhang, thu}@ischool.drexel.edu

Abstract

This paper is dedicated to investigating the value of information from sibling pages for web page clustering. We use a link-based clustering algorithm to examine the usefulness of sibling links for improving clustering quality. The algorithm is extended by two types of edge weighting techniques. The results of the experiments conducted on WebKB4 dataset prove that: (1) using information from sibling pages can significantly improve clustering quality; (2) sibling pages are more useful than parent and child pages in enhancing clustering performance; (3) weighting and pruning sibling links can not improve the clustering quality. We also conducted an experiment on the citation dataset Cora7. The results indicate that sibling links are not more useful than the direct citation links when used to cluster collections of research papers.

1. Introduction

Traditional clustering algorithms, such as k-means, cluster documents based on their textual content. These algorithms achieve good results when used to cluster structured corpora, such as academic papers. Web page documents also contain useful on-page features, such as textual contents and HTML tags. Nevertheless, these features are sometimes missing, misleading and unrecognizable due to the lack of well-controlled authoring styles and other reasons [1]. Therefore, it is desirable to combine information from other sources to enhance clustering effectiveness. A central property of web documents is that they are connected through hyperlinks. The link structure of the web provides a rich information source for web clustering. A good way for improving clustering quality is to combine on-page features and features extracted from the neighboring pages when clustering a web page. However, a question when using features from neighbors is of which links or neighbors to select.

Figure 1 shows the six types of neighbors a page have at a distance of two: parent, child, sibling, spouse, grandparent and grandchild [1].

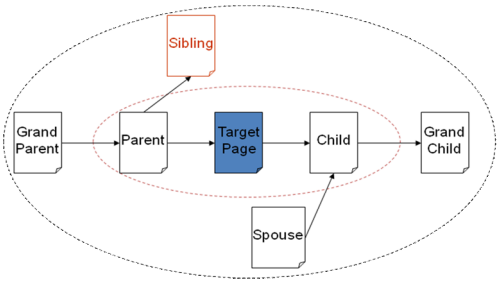

Figure 1. Six types of neighbors of a web page at a distance of two

Parent and child are most commonly used neighbors in existing research on web clustering or classification [2-10]. Grand parent, grand child, and other pages beyond the outer circle in Figure 1 are rarely used because the farther a page from the target page, the less likely that the page shares the same topic with the target page. Sibling and spouse are not directly linked to the target page, but they are also important neighbors. Especially, sibling pages have been empirically demonstrated to be even more useful than parents and children in classification tasks [2, 10]. Qi and Davison explains this phenomenon based on the process of hyperlink creation[1]. When creating out-links, authors tend to choose pages with related topics. But these pages may not share the same topic with the source page. However, the out links of a page tend to have the same class, especially when they appear adjacent to each other. As Qi and Davison explain, ñA page often acts as a bridge to connect its

outgoing links, which are likely to have common topics.ò Chakrabarti et al. also called the common in-links shared by two or more pages as ñbridgesò, which ñhint that two or more pages have the same class, while not committing what that class could beò [2].

Although the value of sibling pages for improving classification performance has been proved in previous research, information from sibling pages has rarely been used in clustering tasks. This paper aims to investigate whether information from sibling pages can be used to improve clustering quality, and whether sibling pages are more useful than parent and child pages for enhancing clustering performance. For this purpose, we choose the link-based clustering algorithm proposed in [9] to do the experiment. This algorithm uses the theory of Markov Random Fields to derive an iterative relaxation of clustering assignments. We adapt the algorithm to cluster web pages based on the sibling links. We also extend the algorithm with two kinds of edge weighting techniques. The algorithm are experimented on the hypertext dataset WebKB4 [11]. In order to investigate whether co-citation links between research papers have the same effect on clustering quality as sibling links between web pages do, we also conduct a experiment on the citation dataset Cora7 [12].

Our main findings include: (1) Using information from Sibling pages can significantly improve clustering quality; (2) Sibling pages are more useful than Parent and Child pages in improving clustering quality; (3) Weighting and thresholding can not improve the quality of sibling-based clustering; (4) sibling links are less effective than the direct citation links (parent and child links) when used to cluster collections of research papers.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 reviews the clustering algorithm and introduces two ways to weight the sibling link. Section 4 present and discusses the experimental results. Section 5 concludes the paper.

2. Related Work

The Sibling relationship between web pages is like co-citation relationship between academic papers, which is first exploited by Small and Griffith to measure the common intellectual interest between two documents [13]. In the context of web documents, the value of sibling link has also been exploited by text mining community.

Many methods have proposed to enhance web page classification by utilizing the information of sibling pages. Chakrabarti et al. compared the performance of a pure term-based classifier with a pure link-based classifier that uses the class labels of sibling pages in classifying a set of 849 Yahoo pages [2]. The results show that the link-based classifier, while using orders of magnitude fewer features, achieves much better performance. The link-based classifier is also refined by considering the position of out-links on the bridge page, which further improves the classification quality. However, the coverage of the pure link-based classifier is limited by the fact that any page in the dataset has to have some sibling pages.

Slattery and Mitchell try to use sibling relationships within the test set to help identify functional category regularities [5]. The category regularities learned from the test set based on the sibling relationships are combined with the training set regularities learned by a relational learner called FOIL to improve the classification quality.

Some methods also combine the information from sibling pages with information from other types of neighboring pages to enhance the classification quality. Lu and Getoor propose a link-based logistic classifier, which classify a page on the basis of both its textual content and the category distributions of its three types of neighboring pages: parents, children and siblings [8]. Comparison study shows that the link based classifier outperforms a simple content-only classifier.

Qi and Davison proposed a method to enhance the classification performance by combing the class and content information from four kinds of neighboring pages (parents, children, siblings and spouses) with the textual information of the target page [10]. Experimental results show that the enhanced classifier achieves higher classification accuracy than text-based classifier. Furthermore, the study of the contribution of different neighbor types revealed that while all neighbor types are useful, sibling pages are the most important neighbors to use.

The sibling relations have also been utilized in clustering tasks. However, in this case, the sibling link is mostly used to improve the similarity metrics [6, 14, 15]. Weiss et al. define a similarity function between two hypertext documents by taking into account both the content similarity of the two documents and their hyperlink structures [14]. The similarity function assigns higher similarity values to documents that have common ancestors and descendants. He et al. also propose a new approach combing textual information, hyperlink structure and sibling relations into a single similarity metric [6]. Sibling pages sharing more common parents are given higher similarity values. It

is found that clustering method based on the new similarity metric achieves higher results. Drost et al. propose a way of utilizing the principle of multi-view learning to combining three different similarity metrics: textual similarity, similarity based on sibling relation, and similarity based on spouse relation [15]. The three kinds of similarity metrics and their combinations are successfully applied in partitioning clustering algorithms for finding communities in large linked data.

The above studies of utilizing sibling relation to improve similarity metrics to some extent imply the usefulness of sibling relation in clustering tasks. However, they do not show whether the information of sibling pages, such as class labels, can be directly used to enhance the clustering performance. This study of this paper bridges this gap. We use the link-based clustering algorithm proposed in [9] to investigate the value of the information of sibling pages for clustering tasks. The reason of choosing this algorithm is that it takes into account of both textual content and the cluster label of the neighboring pages during the clustering process. The experimental results show that compared to content-based algorithm, this algorithm can significantly improve the clustering results by utilizing the information from parents and children pages [9]. In this paper, we examine whether it can achieve same or better improvements by utilizing the information of sibling pages. The algorithm is reviewed in next section.

3. Clustering Algorithm

The algorithm used [9] is adapted for clustering web pages documents based on their sibling information.

3.1 Basic Model

Let $D = \{d_i, i = 1,2,...,n\}$ be a document set, which is represented with an undirected graph *G*. Each document $d \in D$ corresponds to a vertex in the graph *G,* and each link between two documents in *D* corresponds to an edge in G. Here, we only consider sibling links between two documents. All the sibling neighbors of document D is denoted by *S*(*d*). Let *c*(*d*) stands for the cluster of document *d, t*(*d*) denotes the set of terms contained in *d*.

Then the cluster assignment of *d* is based on both *d*'s on-page features *t*(*d*) and the cluster distributions of all of its siblings *S*(*d*). If we use $\Phi_{i,d}$ to denote the probability of a document *d* to be assigned to cluster *i*, then:

$$\Phi_{i,d} = \Pr(c(d) = i \mid T, S(d))$$
$$= \Pr[c(d) = i \mid t(d), c(d_1), c(d_2), ..., c(d_k)]$$

where d_1 through d_k are the sibling pages of document *d* in *D*.

If assuming that there is no direct coupling between the texts of a document and the cluster labels of its siblings, then the above equation can be rewritten as:

$$\Phi_{i,d} = \Pr[c(d) = i \mid t(d)] \sum_{c(S(d))} \Pr[c(d) = i \mid c(S(d))] \Pr[c(S(d))]$$

If further assuming that the labels of a document's siblings are independent of each other, then we get the following formula:

$$\Phi_{i,d} = \sigma_{i,d} \sum_{c(S(d))} \left(\prod_{d' \in S(d)} \Pr[c(d) = i \mid c(S(d))] \Pr[c(S(d))] \right)$$

where $\sigma_{i,d} = \Pr[c(d) = i \mid t(d)]$.

The above equation considers all the combinations of cluster assignments of document *d*'s siblings. For simplicity, we adopts hard clustering approach, that is we only consider the most probable cluster assignment of document *d*'s each sibling. Then the above equation is simplified as:

$$\Phi_{i,d} = \sigma_{i,d} \prod_{d' \in S(d)} \Pr[c(d) = i \mid c(S(d))] \Pr[c(S(d))]$$

The equation is resolved through iterative Relaxation Labeling techniques. First, the class label of each document is initialized through a content-based clustering process. Then the cluster assignment of each document is re-estimated using the label assignments of its siblings and its own content. The re-estimation process based on the above equation iterates until the probability $\Phi_{i,d}$ for each document stabilizes or the changes drop below some threshold or the times of iterations reaches a certain number.

3.2 Weighting

In [9], the algorithm is extended through edge weighting. Each edge in graph *G* is weighted by the cosine similarity between the feature vectors of the two documents connected by the edge (See the $\boldsymbol{W}$ in the following formula).

$$\Phi_{i,d} = \sigma_{i,d} \prod_{d' \in S(d)} \Pr[c(d) = i \mid c(S(d))] \Pr[c(S(d))] * W$$

In our study, rather than weighting and pruning the edges based on the content similarity of the connected documents, we propose two types of weight based on the property of the sibling relationship.

For the first type is simple weight, W_S, which is equal to the number of common in-links or parent pages shared by the target page and its sibling page.

In order to avoid the bias caused by the different number of in-links each page accepted. We also introduce a normalized weight, with value falls in [0, 1]. It is calculated through the following formula:

$$W_N = \frac{P_l(d \cdot d_s)}{\sqrt{P_l(d) * P_l(d_s)}}$$

where d_s is one of document d's sibling pages. $P_l(d)$ is the probability that document d is linked by some page in document set D. It is calculated as the ratio of the number of in-links accepted by d to the total number of hyperlinks in D. $P_l(d \cdot d_s)$ refers to the probability that document d and its sibling d_s are jointly linked by some page in D. It is actually the ratio of the simple weight W_S to the total number of hyperlinks in D.

4. Experiment

4.1 Dataset

We choose the WebKB [11] dataset for experiment. It contains about 8300 web page documents about university computer science departments and around 11, 000 hyperlinks. These pages are divided into seven categories: student, faculty, staff, course, project, department and other. In our experiments, we only use the 4190 pages from four most populous categories: student, faculty, course and project.

In order to compare the effect of sibling linkage between hypertext to that of co-citation linkage between academic papers on clustering performance, we also conduct an experiment on the citation dataset Cora [12]. Cora is an online archive containing 30,000 computer science research papers, which are categorized into a Yahoo-like topic hierarchy. The papers are categorized according to text content. We only select a subset of 7 classes under the machine learning category from the Cora database.

4.2 Quality Metrics

The cluster quality is evaluated by three metrics: F-score [16], purity [17], and normalized mutual information (NMI) [18]. F-score combines the information of precision and recall which is extensively applied in information retrieval. Purity assumes that all samples of a cluster are the members of the actual dominant class for that cluster. NMI is defined as the mutual information between the cluster assignments and a pre-existing labeling of the dataset normalized by the arithmetic mean of the maximum possible entropies of the empirical marginals. A merit of NMI is that it does not necessarily increase when the number of clusters increases. All the three metrics range from 0 to 1, and the higher their value, the better the clustering quality is.

4.3 Clustering approaches under comparison

In the experiment, we compare the results of four kinds of clustering approaches. The first is K-means using TFIDF (term frequency * inverse document frequency) scheme. The other three are link-based clustering approaches using the algorithm proposed in [9]. Among these three approaches, one is based on both child and parent links (out-links and in-links), one is only based on child links (out-links), and the last one, which we are concerned with, is based on the sibling links. All the link-based clustering approaches use TFIDF-based K-means for initialization. Table 1 lists the notations for these four clustering approaches, which are used in the following result tables and figures.

Table 1. Clustering Approaches under Comparison

Notation	Explanation
Text_kmean	k-means using TFIDF scheme
PC_link_kmean	Link-based clustering using both Parent and Child links.
C_link_kmean	Link-based clustering only using Child links.
S_link_kmean	Link-based clustering using Sibling links.

4.3 Results

4.3.1 Comparison of different clustering approaches. The experimental results on WebKB4 dataset using sibling link-based clustering are compared to that of content-based K-means. Table 2 lists the results. The symbols * and ** indicates the change is significant according to the paired-sample T-test at the level of $p<0.05$ and $P<0.01$ respectively. These two symbols have the same meaning in all the following experimental result tables and figures.

Table 2. Sibling link-based clustering Vs. content-based clustering

	F-score	Purity	NMI
Text_kmean	0.432	0.663	0.328
S_link_kmean	**0.489**	**0.709**	**0.400**
change	13% *	7%**	22%**

From Table 2, we can see that the improvement of sibling link-based clustering over content-based clustering is significant according to all the three quality metrics. In order to prove the superiority of sibling links over other types of links, we also compare the results of link-based clustering using child links to the content-based clustering (see Table 3).

Table 3. Child link-based clustering Vs. content-based clustering

	F-score	Purity	NMI
Text_kmean	0.432	0.663	0.328
C_link_kmean	0.468	0.694	0.369
change	8%	5%*	12%**

Table 3 shows that the improvement by utilizing child link in clustering is not as great as that by using sibling links in clustering.

The results of sibling link-based clustering is directly compare to child link-based clustering (see Table 4) and clustering using the both parent and child links (see Table 5).

Table 4. Sibling link-based clustering Vs. Child link-based clustering

	F-score	Purity	NMI
C_link_kmean	0.468	0.694	0.369
S_link_kmean	**0.489**	**0.709**	**0.400**
change	4.5%*	2%*	8%**

Table 5. Sibling link-based clustering Vs. clustering using both Parent and Child links

	F-score	Purity	NMI
PC_link_kmean	0.470	0.693	0.367
S_link_kmean	**0.489**	**0.709**	**0.400**
change	4%*	2%**	9%**

From Table 4 and Table 5, we can see that link-based clustering using sibling links can significantly improve the clustering quality compare to the link-based clustering using out-links or using both out-links and in-links. This indicates that the information from sibling pages is more useful than the information from parent and child pages for clustering tasks.

4.3.2 The effect of Weighting and Thresholding. We apply two kinds of weight to the sibling edges to see if the clustering quality can be improved. Table 6 shows the results of using simple weight W_S, where W_S Ó 2 means two sibling pages share at least two parent pages, and all the sibling linkages with less than 2 common parents are pruned.

Table 6. The effect of the simple weight

	F-score	Purity	NMI
Without using Weight	0.489	0.709	0.400
W_S Ó*1*	0.486	0.705	0.393
W_S Ó*2*	0.471	0.693	0.367

Table 6 shows that using the simple weight does not improve the clustering performance. And as the threshold of weight increases to 2, the clustering quality decreased. This indicates that the value of sibling links is not proportional to the times it co-occurs with the target page. And pruning some sibling links based on the simple weight leads to the loss of useful information for clustering. We also test the effect of using normalized weight W_N (as defined in section 3.2) on clustering quality. Table 7 shows the clustering results with different threshold values.

Table 7. The effect of the normalized weight

	F-score	Purity	NMI
Without using Weight	0.489	0.709	0.400
W_N Ó0	0.486	0.705	0.393
W_N Ó0.1	0.484	0.705	0.398
W_N Ó0.3	0.484	0.705	0.395
W_N Ó0.5	0.483	0.704	0.394
W_N Ó0.7	0.480	0.700	0.384
W_N Ó0.9	0.475	0.695	0.379

It is clear to see that even using the normalized weight can not improve the clustering performance. And as the threshold increases, the value of all the three quality measures decline. This further indicates that thresholding only leads to information loss.

4.3.2 Sibling link in research papers. In order to check whether the sibling links or co-citation links have the same effect when used to cluster collections of research papers. We also do an experiment on the citation dataset Cora7. The result is shown in Table 8, where Text_kmean is the baseline clustering method for comparison.

Table 8. Clustering Results on Cora7

	F-score	Purity	NMI
Text_kmean	0.494	0.545	0.312
PC_link_kmean	0.560**	0.612**	0.404**
C_link_kmean	0.559**	0.609**	0.395**
S_link_kmean	**0.562****	**0.610****	**0.395****

]

According to Table 7, for the cora7 data set, clustering based on sibling links can still significantly improve the clustering quality compared to the content-based clustering. However, sibling link is not proved to be more useful than the direct hyperlinks (child and parent link). The reasons maybe that citations in research papers are more meaningful than hyperlinks in web pages. Therefore, in the collections of research papers, sibling or co-citation links do not have the superiority over citation links as they do in hypertext collections.

5. Conclusion

In this paper, we adopt a MRF-based clustering algorithm to examine the impacts of sibling linkage on the clustering results. Based on the experimental results, we argue that sibling linkage is more useful than parent and child linkage for clustering hypertext documents. We also investigate the impact of edge weighting and pruning on the performance of clustering based on sibling links. Two types of weights are adopted based on the properties of sibling linkage. The results indicate that weighting does not improve the clustering performance and pruning leads to the decline in clustering performance because of the loss of useful link information. The sibling link is also utilized for clustering collections of research papers. Although applying sibling link can still improve the clustering quality compared to content-based clustering method. However, the experimental results imply that sibling links are not more useful than the direct citation links when used to cluster the collection of research papers. In the future, we will experiment on more datasets and investigate the effect of other properties, such as the proximity of the sibling link to the target link in the parent pages, on the clustering performance.

References

[1] X. Qi and B. D. Davison, "Web Page Classification: Features and Algorithms," *Technical Report*, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 2007.

[2] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in *SIGMOD '98: Proceedings of the ACM SIGMOD International Conference on Management of Data* New York, NY: ACM Press, 1998, pp. 307-318.

[3] D. S. Modha and W. S. Spangler, "Clustering Hypertext with Applications to Web Searching," in *11th ACM on Hypertext and Hypermedia*, San Antonio, Texas, 2000.

[4] H.-J. Oh, S. H. Myaeng, and M.-H. Lee, "A Practical Hypertext Categorization Method using Links and Incrementally Available Class Information," in *SIGIR*, Athens, Greece, 2000, pp. 267-271.

[5] S. a. Slattery and T. Mitchell, "Discovering Test Set Regularities in Relational Domains," in *17th International Conference on Machine Learning (ICML)*, Morgan Kaufmann, 2000, pp. 895-902.

[6] X. He, H. Zha, C. H. Q. Ding, and H. D. Simonb, "Web document clustering using hyperlink structures," *Computational Statistics & Data Analysis,* vol. 41, pp. 19-45, 2002.

[7] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: OrganizingWeb document collections based on link semantics," *the VLDB Journal,* vol. 12, pp. 320-332, 2003.

[8] Q. Lu and L. Getoor, "Link-based Text Classification," in *the 20th International Conference on Machine Learning (ICML)*, Menlo Park, CA, 2003.

[9] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in *CIKM '06: the 15th ACM International Conference on Information and Knowledge Management*, New York, NY, 2006, pp. 778-779.

[10] X. Qi and B. D. Davison, "Knowing a Web Page by the Company It Keeps," in *CIKM '06: the 15th ACM International Conference on Information and Knowledge Management*, New York, NY, 2006, pp. 228-237.

[11] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. a. Slattery, "Learning to extract symbolic knowledge from the world wide web," in *AAAI-98*, 1998.

[12] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval,* vol. 3, pp. 127-163, 2000.

[13] H. Small and B. Griffith, "The structure of scientifc literatures, Identifying and graphing specialities," *Science Studies,* vol. 4, pp. 17-40, 1974.

[14] R. Weiss, B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," in *the seventh ACM conference on Hypertext*, Washington DC, 1996, pp. 180-193.

[15] I. Drost, S. Bickel, and T. Scheer, "Discovering Communities in Linked Data by Multi-View Clustering," in *29th Annual Conference of the German Classification Society, Studies in Classification, Data Analysis, and Knowledge Organization* Berlin, 2005, pp. 342-349.

[16] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Conference*

on Knowledge Discovery in Data, San Diego, California, 1999, pp. 16 - 22

[17] Y. Zhao and G. Karypis, "Criterion functions for document clustering: experiments and analysis," *Technical Report*, Department of Computer Science, Univ. of Minnesota 2001.

[18] A. Strehl and J. Ghosh, "Cluster ensemblesð a knowledge reuse framework for combining partitions," *Journal of Machine Learning Research,* vol. 3, pp. 583ï 617, 2002.