Enhancing Contents-Link Coupled Web Page Clustering and Its Evaluation

Yitong Wang and Masaru Kitsuregawa
Institute of Industrial Science, The University of Tokyo
{ytwang, kitsure@tkl.iis.u-tokyo.ac.jp}

Abstract

Web page clustering is a fundamental technique to offer a solution for data management, information locating and its interpretation of Web data and to facilitate users for navigation, discrimination and understanding. Most existing clustering algorithms cannot adapt well to Web clustering directly in terms of efficiency and effectiveness. Combining contents analysis and hyperlink structure analysis has been proven a better approach. However, how to effectively combine the two features with different nature in clustering to get satisfactory results remains an open problem and there is still little work on it. In this paper, we present an experimental study on enhancing coupling of links and contents analysis of Web pages for robust clustering. In particular, we introduce two techniques: in-link reinforcement and anchor window analysis to improve the adaptability of contents-link coupled clustering. Our detailed evaluation indicates those techniques can effectively improve the quality of Web pages clustering for a wide range of topics.

1. Introduction

Currently, there are more than 2 billion pages on the web without counting those so-called hidden Web pages that can be generated from the underneath databases. At the same time more than 100 million pages become obsolete every month. Locating truly needed Web pages and interpreting them appropriately is a big challenge faced by researchers in the fields of database, Information Retrieval (IR) and data mining. So, correctly clustering both the source Web pages and results of search engines is very important to help end users in navigation, discrimination, summarization and interpretation of the Web. Most existing and well-cited topic directories such as Yahoo! (www.yahoo.com) and open directory (www.dmoz.com) are mainly created and maintained manually by domain experts. Therefore those topic directories cover only a very small portion of the whole Web due to extremely low scalability of manual creating and maintenance. They are also more often outdated as the Web changes all the time. Some topics also have no corresponding sub-categories in Yahoo or open directory. Such unsatisfactory performance calls for the needs of semi-automatic or automatic clustering of Web pages that is expected to scale well and be able to follow the evolution of the Web well.

Document clustering has been well studied in the field of tradition IR. The most commonly used techniques are developed under the vector-space model. Under this model, documents are represented by vectors of terms extracted from the documents. During the clustering process, similarity between documents is used to determine whether two documents should be put into the same cluster or not. Although dozens of similarity functions have been developed, they are more or less built on top of the same hypothesis: more common terms two documents have, more similar the two documents are. Unfortunately, such hypothesis may not hold in the context of Web pages. The fundamental reason is that Web is a place where everyone can publish. Web pages are thus somehow created randomly by various authors and different authors may use different words to express similar ideas [18]. Furthermore, because of the ease of multimedia presentation, some Web pages in fact contain only small portion of concrete texts from which terms can be meaningfully extracted. What makes the problem more complex is that, there are many "junk" pages on the Web, which should be removed before the clustering process. All these unique features of Web pages make the traditional term-based clustering techniques work poorly in a lot of cases.

Compared to traditional text documents, one extra feature contained in Web pages is the hyperlink. Hyperlinks (abbreviate to link here after for simplicity) are in fact the authors view about the relationships among various Web pages, which should be naturally considered in the clustering process, in addition to terms. Recently, works have been reported in Web page clustering that attempted to combine link information with term analysis in the clustering process. The basic approach is to measure both the similarity of contents (represented by terms) between two Web pages and the links related to them. The initial results seem rather promising [16, 21]. However, due to the complex link structure of the Web and the diversity of Web pages, it is very difficult to control such clustering process with respect to the proper assignment of the contributions of the link and terms in similarity functions. As a result, most content-link coupled clustering approaches suffer from the problem of sensitivity to clustering parameter setting and non-uniform performance in clustering Web pages with wide range of topics.

The above observations motivated our work reported in this paper. Through a comprehensive performance study on a content-link coupled clustering technique, we provide some insights on the general problem of such

clustering approach. We proposed two techniques, *in-link reinforcement* and *anchor window analysis* to alleviate the problem and improve the effectiveness of content-link coupled clustering algorithms. The in-link reinforcement technique addresses the issues of orthogonality of link space (it is common that a Web page has thousands of links pointing to it) and complex link structure of Web. It uses the contents of the pages to which those links point to reduce the dimensionality of the link space and simplify the link structure. The technique of anchor window analysis makes use of another important feature that does not exist in traditional text documents, anchor text and text surround the links, to make the semantics of terms clearer so that the term-based similarity analysis could be more accurate. The results of our experimental study indicated that with the help of these two techniques, the clustering algorithm become more robust. That is, it can well adapt to different kinds of Web pages and cluster a wide range of topics accurately. Furthermore, the clustering results are less sensitive to parameter settings in the similarity functions.

Our main contribution in this paper can be summarized as follows. (1) We identified the common problems in contents-link coupled Web page clustering approach through the results of a comprehensive experimental study and provided explanations to the poor performance of the approach for certain Web pages; (2) We introduced two simple yet effective techniques to offer a solution to the problems pinpointed; and (3) We conducted a performance study to evaluate the effectiveness of two proposed techniques.

The rest of the paper is organized as follows. Related work on Web page clustering is briefly reviewed in Section2. The general approach of contents-link coupled clustering is discussed in Section 3. In Section 4, we described two proposed techniques and their effects. Section 5 presents the results of a performance study. Finally Section 6 concludes the paper.

2. Related Work

Clustering is a classical problem that attracted new interests in the recent surge of data mining research. There are lots of clustering algorithms have been developed, that can be classified into a number of categories, include center-based partitioning clustering (K-means and its variations), hierarchal clustering (HAC), density and connectivity-based clustering (DBSCAN), grid-based clustering and graph-based clustering etc. Orthogonal to the general clustering algorithms is the representation of data objects to be clustered. For example, in traditional text clustering, documents to be clustered are represented by terms extracted from those documents using TFIDF (term frequency and inverted document frequency) model.

Hyperlink is an important feature in the context Web pages, for identifying associative relationship among pages and can be used to obtain high quality search results, as indicated by PageRank [17] and HITS [9]. While the term-based algorithms cluster Web pages based solely on the contents of Web pages, the contents-link coupled approaches consider both the contents and link information of Web pages. Web-log based approaches cluster Web pages according to the page access patterns discovered from Web logs. We omit the discussion of web-log based clustering approaches here.

T.H.Haveliwala *et. al.* proposed a technique LSH (Local – Sensitive – Hash) for clustering the entire Web, which emphasizes more on the scalability of clustering. Snippet-based clustering is well studied in [14, 20]. Shingle method, which is often used for duplicates removal is proposed in [1] to measure the similarity between pages for clustering. Applying the technique of association rule mining to term vectors is another clustering approach proposed in [3]. It can automatically produce groups of pages without defining the similarity between pages. These approaches differ with each other on clustering method and are all based on common terms shared among web pages.

Many works [5,9,10,14,17] tried to explore link analysis to improve quality of Web information retrieval or to mine useful knowledge from Web. Kleinberg suggested that there are two kinds of pages on the web for a specific topic: “hub” pages (include many useful links) and “authority” pages (include relevant contents to a topic and are citied by many hub pages) and they reinforce each other [9]. Gibson, *et. al.* proposed a hierarchical network search engine that clusters hypertext documents based on the contents as well as the link structure of each hypertext document to structure a given information space to support various services like browsing and querying [16]. Clustering hypertext documents by *co-citation analysis* is explored in [7]. By applying HITS algorithm [9] to the vicinity graph around a seed URL, the approach proposed in [10] could find similar pages to the seed URL in a more narrow way. However it focuses more on finding similar pages rather than clustering Web pages.

We proposed a contents-link coupled clustering algorithm that combines content and link analysis to cluster web search results [19]. The basic clustering method used is an extended K-means. By varying weighting factor of contents analysis and link analysis, we also compared the performance of term-based, link-based and content-link coupled approaches and gave some empirical recommendations for weighting factors.

3. Contents-Link Coupled Clustering

3.1 Link Analysis

Hyperlinks are helpful since they demonstrate objective opinions of the authors of other web pages to the pages they point to. **Co-citation** [6] and bibliographic **coupling**

[12] are two fundamental measures to be used to characterize the similarity between two documents. **Co-citation** measures the *number of citations* (*out-links*) *in common* between two documents and **coupling** measures the number of documents (*in-links*) that cite both of two documents under consideration (for a hyperlink q->r, q is called in-link page of r and in reverse, r is called out-link page of q). Since both "hub page" and "authority page" are possible to be included in search results, both co-citation and coupling are considered in link analysis.

3.2 Contents Analysis

For contents analysis, some works use all terms in a page (the whole document) and others may only use certain part(s) of the page. Possible choices include snippet, title, meta-contents or even anchor window of in-link pages of the page. Snippet of page *u* means sentences attached with URL *u* appeared in search results. The anchor window of in-link page *v* of *u* means the anchor text and text around the hyperlink *v->u* in the source page *v.* It might include concise and important terms to describe the main topic of the page that the link points to. Meta-content is an optional tag for most web pages and gives the summary of the page by its author. By applying stemming to the contents collected above, it is possible for us to extract all distinct terms for all pages.

3.3 Contents-Link Coupled Clustering

When we say *C*ontents-*L*ink *C*oupled (*CLC*) clustering, we mean that in clustering process both link and contents analyses are considered for measuring the similarity between pages regardless of clustering algorithm used.

By *CLC*, each web page q in data set D is represented as 3 vectors: q_{Out} q_{In} and q_{Kword} with *M, N and L* as the vector dimension respectively. The *ith* item of vector q_{Out} indicates whether q has the corresponding out-link as the *ith* one in M out-links. If yes, the *ith* item is 1, else 0. q_{In} is identically defined. The *kth* item of vector q_{Kword} indicates the frequency of the corresponding *kth* term of L appeared in page q. We have several notations: *n, m, M, N, L* are positive integers, D is dataset, the set of specified number of web pages for clustering. Pages in D are called target pages. We use n to denote the size of D. We use m to denote specified number of out-link/ in-links extracted for each URL/page in D. *M, N, L* denotes total number of distinct in-links, out-links as well as terms after applying link and contents analysis for all n pages in D respectively.

So, the similarity of two pages q and r is the linear combination of three parts:

$$p_{out} S(q_{out}, r_{out}) + p_{in} S(q_{in}, r_{in}) + p_{term} S(q_{term}, r_{term})$$
$$p_{out} + p_{in} + p_{term} = 1 \quad (1)$$

$S(q_{out}, r_{out})$ is defined as *Cosine* of two out-link vectors. So, total similarity of two pages is the linear combination of corresponding out-link similarity, in-link similarity and term similarity. Centroid (center point) is used to represent the cluster when calculating the similarity of a page with a cluster. By varying weighting factors in formula (1), it is possible to study the effects of out-links, in-link and terms on clustering process. As stated in [21] that from statistical point of view, terms mainly affect noise pages removal and percentages of pages that could be clustered; in-link pages affect size of big cluster and number of clusters produced; out-link pages affect the percentage of page clustered. Here, we would like to give a more detailed and intuitive explanation of contribution of terms and links from semantic point of view. We have following findings:

1. According to the experimental results, we found that the results of term-based clustering is rather coarse and usually includes very general groups, which are totally different each other from semantic point of view. E.g. for topic "jaguar", "car" group and "animal" group are two very general groups with very different semantic topics; while "car driver club" and "racing car" are finer classification. So, term-based clustering could only roughly separate pages into general semantic groups and failed to handle the finer case, like "racing car" and "car driver club" since both pages may include some terms like "car, model etc. The main reasons of poor "purity" of clusters produced by term-based clustering are: i) noise pages are included into clusters instead of removing since noise pages share some unimportant terms with other pages; ii) pages that on different finer topics (but the same general topic) are mixed together. If the query topic itself is very single-idea (a special case we study is "HIV"), most pages will be grouped in a cluster.
2. Since hyperlinks represent the authors' view of the relationship among Web pages, hyperlink-based clustering expresses "association" of pages. Therefore, we could say that clusters produced by link-based clustering are in finer granularity. The problem of link-based clustering is that some similar pages (e.g. new created pages) may not have enough co-citation/ citation to be grouped together. That is to say, recall is some low.
3. Since hyperlink and terms are features with very different nature and in turn link analysis and contents analysis may in different "scale". By "scale", we mean due to different dimension and sparseness of feature spaces (link and term), average link similarity and average term similarity among all pages are in different scales.
4. Terms in anchor window of a hyperlink have proved to be a precious contents summary for the page it points to as stated in many literatures [4][17]. The

reason is that pages talking about "automobile" may not include the word "automobile" in its source. However, it is very likely to include these terms in the anchor window of in-link pages. We should deal with terms in the anchor window of its in-link pages separately.

In order to have an in-depth understanding of contents-link coupled clustering, we have implemented an extension of K-means as clustering method. Combing with CLC analysis, we call it CLCK clustering as depicted in Fig. 1.

1. Define the similarity threshold
1. Filter irrelevant pages and only relevant pages join clustering process
2. Assign each relevant pages to the Top *C existing cluster(s)* based on the similarities (that above the similarity threshold) between the page and the centroids
3. The page will be one cluster itself if no *existing cluster* meet step 3
4. Recompute the centroids of the cluster
5. Repeat Step 2 until 5 until all relevant pages are assigned and the centroids do not change

Fig. 1 CLCK Clustering method

In the clustering procedure, we use similarity threshold to control the clustering process instead of using pre-defined K value and K centroid as in standard k-means. It is possible to apply HAC (hierarchical agglomerative clustering) on the base clusters produced by Fig. 1 to make the final results more concise and easy-to-interpret semantically.

4. Improving the Adaptability

Experimental results have demonstrated that contents-link coupled clustering could improve the clustering quality significantly by utilizing the merits of link analysis and terms analysis. It also handles noise page more gracefully. Yet, we noticed that CLCK suffers limitations in following two aspects:

1) Quality of main/biggest cluster is still unsatisfactory;
2) Cannot adapt it well to some topics

Quality measurement of a cluster is mainly based on entropy, that is, a) whether noise pages are included in the cluster; 2) whether pages in the cluster are tightly related and focused on the same sub-topics.

In this part, we first analyze experimental results in more detail to pinpoint the underlying reasons of limitations. That is why contents-link coupled clustering works poor for some topics and quality of the biggest cluster produced is unsatisfactory. We then introduce two techniques to alleviate the problems so as to improve the adaptability of contents-link coupled clustering to a wide range of topics. We choose not to use standard IR collections, as we are interested in the performance of real data on the web. We download search results of Search engine for various topics as dataset and use "T", "L" and "CLC" to denote terms–based (with p_{out}, p_{in} and p_{Kword} as 0, 0, 1), link-based (with p_{out}, p_{in} and p_{Kword} as 0.5, 0.5, 0) and contents-link coupled (with p_{out}, p_{in} and p_{Kword} as 0.2, 0.3, 0.5) clustering approaches respectively.

4.1 Effects of Contents Analysis and Link Analysis on Clustering Process

Feature (term, links) distribution for different topics on the web is various. Based on the quality of link and contents analysis on clustering process, various topics could be summed up to four kinds of cases. We demonstrate them by presenting an example for each case as shown in Table 1 to Table 4. The results shown in Table 1-4 are produced by applying algorithm depicted in Fig. 1 with similarity threshold 0.1 and different weighting factors for "T", "L" and "CLC". Similarity threshold and weighting factors are chosen according to empirical evaluations stated in [21]. The label of each cluster shown in Table 1-4 is identified automatically by *term* vector of centroid for each cluster. We modified the label a little to make it more natural and easy to understand.

Table 1. Clustering results for Topic "Jaguar"

	Main /Distinct Sub-topics
T (0.37, 0. 78)	6 (Car, Club, Game, Big cat, Parts, Racing car)
L (0.29, 0.63)	8 (Car, Club, Game, Big Cat, Atari Emulate, Touring place, Online dealership, research project)
CLC (0.26, 0.75)	10 (Club, Car model, Game, Big cat, Atari Emulate, Parts, Racing car, Touring place, Online dealership, Research Project)

Table 2. Clustering results for Topic "Salsa"

	Main /Distinct Sub-topics
T (0.48, 0.82)	2 (Latin dancing, hot sauce)
L (0.31, 0.67)	7 (Latin music, dancing, hot sauce, recipes, club, food, salsa in Germany)
CLC (0.29 0.79)	6 (Latin music, Dancing, Hot sauce, Recipes, Club, Food)

Table 3. Clustering results for Topic "Abduction"

	Main /Distinct Sub-topics
T (0.39, 0.80)	3 (Abuse/ court, Child abduction, Peircean Abduction)
L (0.35, 0.51)	4 (Alien abduction, child abuse, Peircean Abduction, parents divorce)
CLC (0.37, 0.71)	3 (Abuse/court, Child, Peircean Abduction)

Table. 4 Clustering results for Topic "HIV"

	Main /Distinct Sub-topics
C (0.76, 0.76)	1 (AIDS/ therapy)
L (0.38, 0.31)	5(Aids, prevention, Treatment, Research/Fund)
CLC (0.59, 0.55)	1 (AIDS/ therapy)

The two numbers in the left entries of each table are values for two evaluation metrics: average entropy and percentage of page clustered (it is calculated by deducting number of singleton clusters produced from total number of pages in dataset). Detailed definition of entropy is given in Section 5. Here, we only need to sate that low entropy value denotes high "purity"/"quality" of the cluster. The right entries in Table 1-4 are the number of main distinct clusters and their corresponding labels.

We use "distinct clusters" since there might be more than one cluster on the same topic. We choose clusters with clear semantic meaning. E.g. based on Table 3, we could know that with term-based clustering, only two topics "Latin dancing" and "hot sauce" are roughly identified and separated. Let us give a detailed look on the four examples shown above.

1) Jaguar

For this kind of topic, both term-based clustering and link-based clustering produced some semantically meaningful groups as depicted in Table 1. Two evaluation metrics: average entropy and percentage of page clustered indicate that link similarity and term similarity are in the same scale. The main effect of link analysis in CLC for this kind of topic is to "purify" each cluster by removing some noise page and forming more detailed/narrowed clusters as "Game" with "Atari Emulate" and "car model" with "online dealership" instead of identifying new semantic clusters.

2) Salsa

In this kind of topic, term-based clustering works poor as entropy value 0.48 indicates. It only separates two general groups "Latin dancing" and "hot sauce". Link-based clustering works well as depicted in Table 2. It is obvious that link similarity and term similarity are on the same scale. By combining link and contents analysis, CLC gives significant improvements both in "purity" and "percentage of page clustered". In this case, link analysis is very effective in noises handling, cluster purifying and new semantic clusters forming as clusters "recipes" and "club" indicated in Table 2.

3) Abduction

In this case, term-based clustering works reasonable and link-based clustering works poor since it is obvious that link space is rather sparse and about 50% of pages in target dataset cannot be grouped. Comparing metrics for "C" and "L", it could be concluded that link similarity and term similarity are not at the same scale. The effect of link analysis was too weak to have any practical improvements for CLC as shown in Table 3.

4) HIV

The last kind of topics is that both term-based clustering and link-based clustering works rather poor, especially term-based clustering. The topic itself is a medical terminology and has very fixed meaning in different contexts. Its link space is very sparse and there is big difference between link similarity and contents similarity. The effect of link analysis is too weak and totally was washed out by contents analysis. The last kind of topic usually is very tough for automatic clustering since both terms and links are not effective due to their low discrimination. For this kind of topic, the classification is more associational than semantic.

Another drawback of sparseness of feature space is that there is no big gap between similarity of similar pages and dissimilar pages and this leads to low similarity threshold in CLCK. So, the bigger the cluster, the drawback of low similarity threshold is more evident.

In summary, we need to improve the discernibleness of both link analysis and contents analysis, especially link analysis so as to improve similarity threshold accordingly.

4.2 In-Link Reinforcement

From the above discussion, we could see that in order to improve clustering, we need to combine link analysis and contents analysis and to make link similarity and term similarity in the same scale. Comparing with contents analysis, link analysis is more susceptible to orthogonality since when authors put citations/hyperlinks in their pages, they usually have certain preferences. Some newly-created high quality pages may not have many pages link to them. Another point need to mention is that by vector space model, when we map one page to a point in a high-dimensional feature space, we have the assumption that each dimension is orthogonal with other dimensions. While in traditional database clustering, it is possible to guarantee this; for web page clustering, it is difficult to do so. Some work has tried to utilize WordNet to deduce relationship between words. While WordNet is mainly a kind of abstract-concrete relation (like snake-serpent), clustering results are sensitive to the degree of abstractness or concreteness. In this paper, we would like to try to find the relationship between in-link pages. (Since in-link pages are more influential than out-link pages in web search results clustering, here, we only consider in-link reinforcement) Some dimension-reduction approaches can also be applied to filter some minor links so as to improve the situation. However, our object is not only dimension reduction but also finding the relationship between in-links, which in turn will improve the coupling analysis. There is still little work on finding relationship between hyperlinks. We first cluster similar in-link pages into groups and map each group to a dimension in in-link vector space. By doing so, it is possible to reduce dimension of in-link vector space greatly and also improve link analysis.

Mutual reinforcement that applying knowledge of one kind of objects to another kind of objects to get relationship between objects of latter kind and then doing reverse is not very new and has been proposed and checked in some literatures. However, in this paper, we

are not focusing on spirally mutual reinforcement but on how in-link reinforcement will affect the final contents-link coupled clustering.

Method 1: In-Link Reinforcement

1. Executing term-based clustering and cluster n target pages into x groups.
2. Each in-link page p is then represented as an x-dimensional vector and its kth item indicates that when the in-link p has a out-link to pages in *kth* cluster of x groups.
3. Cluster in-link pages into y group based on vector similarity
4. Map each in-link page to the cluster it belongs to and then for dataset, in-link space is reduced to y-dimensional space
5. Execute CLCK approach with renewed in-link vectors

Fig. 2 In-Link Reinforcement

It is possible to cluster in-link pages based on its snippet (when extracting in-link pages for a given page based by service of search engine, snippets are attached with each in-link page) and out-link information. Here, we consider clustering in-link pages based on term- based clustering results as depicted in Fig. 2.

4.3 Anchor window analysis

As mentioned in Section 3.2, when an author put a citation/hyperlink in his page, he usually writes some text around hyperlink to describe the contents of destination page that the hyperlink points to. Since it is a kind of objective description, it usually reveals the topic of destination page more substantially and includes little noisy information. E.g. for a homepage of a search engine, keyword "search engine" may not appear. However, in the anchor window of pages that cite the "search engine" page, it is very like to include the keyword "search engine", which is very essential for clustering. Many works have suggested the effectiveness of anchor window. However, most clustering works treat terms in anchor window equally with those terms appeared in target page. Just as described in section 3.2 that all terms are extracted and stemmed and equally measured. Term vector are formed based on total distinct terms. Nevertheless, as we have analyzed in section 3.3 and section 4.1 that in-link pages and target pages have very different contributions in clustering, so terms in anchor window of in-link pages should be paid much more emphasis. By detailed analysis, we could find that distinct terms appeared in all anchor windows are rather focused and much fewer than terms appeared in target pages. We divide term vector into two parts: anchor window term vector and target page term vector. We could find that comparing with using two vectors, when two target pages share some common terms in the anchor window of their in-link pages (the two target pages are very likely similar), using only one term vector as we previous do actually decreases the term similarity. ($\frac{1}{2} * S(q_{AWterm}, r_{AWterm}) + \frac{1}{2} * S(q_{TPterm}, r_{TPterm}) > S(q_{term}, r_{term})$)

In reverse case, if two pages do not share or only share few terms (two pages are likely dissimilar) in anchor window, using one merged term vector actually increases the term similarity. So, anchor window analysis could help a lot to discriminate similar pages from dissimilar pages based on terms sharing.

Method 2: Anchor Window Analysis

We use four vectors to represent page in target set and similarity measurement of formula (1) is revised as:

$$p_{out} S(q_{out}, r_{out}) + p_{in} S(q_{in}, r_{in}) + \frac{p_{term}}{2} * S(q_{AWterm}, r_{AWterm}) + \frac{p_{term}}{2} * S(q_{TPterm}, r_{TPterm})$$

$$p_{out} + p_{in} + p_{term} = 1 \qquad (2)$$

To make it simple, we adopt above similarity function. Of course, we can use different weighting factors for anchor window term vector similarity and target page term similarity.

5. A Performance Study

We present the results of a performance study that investigates the effectiveness of proposed techniques.

5.1 Data Preparation and Evaluation Methods

In order to have an understanding of web page clustering of real data, we have tested search results from search engine directly for more than 40 topics that cover a wide range. We downloaded resulting pages from search engine to form the target dataset of clustering. For each page in target dataset, we extract 100 in-link pages and out-link pages.

We use following objective metrics to evaluate the quality of clustering,

Average entropy

Entropy is used to measure the "purity" of a cluster, which is defined as: $E(j) = -\sum_i p_{ij} \log(p_{ij})$

where p_{ij} is the "probability" that a page of cluster j belongs to the given class i, approximated by the number of pages that belong to class i divide by the total number of pages in cluster j . It is obvious that $E(j)$ become 0 if the cluster is pure, that is, all pages in a cluster belong to the same class. In other words, entropy measures "*whether pages in the same group are truly focusing on the same topic*" by comparing the groups produced by clustering algorithm to known *classes*. Low entropy means high quality of the cluster due to high intra-cohesiveness while high entropy means that the cluster members are not tightly related. High entropy usually indicates two possibilities: there are some noise pages in the cluster, or the pages in the cluster cover different topics under the general query topic.

In order to measure the overall performance of a clustering scheme CS, overall entropy is defined as:

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E(j)}{n}$$

where $E(j)$ is the entropy for cluster j, n_j is the size of cluster j and n is the total number of data points in dataset. Since clustering is meant to group similar ones together, average entropy is more appropriate when evaluating the quality of a clustering algorithm. When the entropy of a cluster is around 0.2 (0.2-0.25), it means that around 80% pages of the cluster is on the same topic and the rest 20% pages are on another topic or other topics.

***F*- measure**

F measure combines two metrics, *precision* and *recall*, used in information retrieval to evaluate whether clustering algorithm can remove noise pages and cluster high quality pages as much as possible. To apply this measure, we treat each cluster as the results of a query and each *class* as the desired set of results for a query. We calculate the *precision* and *recall a*s:

$P(i, j) = N(i, j)/n_j$, $R(i, j) = N(i, j)/g_i$ n_j, g_i are the size of cluster j and *class i* respectively and $N(i, j)$is the number of pages of *class i* in cluster j. The *F measure* of cluster j and *class i* is then given by (n is the total number of data points in dataset):

$$F(i, j) = \frac{2(P(i, j) * R(i, j))}{(P(i, j) + R(i, j)}, \quad F = \sum_i \frac{g_i}{n} \max_j \{F(i, j)\}$$

It is obvious that the higher the F-measure value, the better the quality of clustering results. *F*-measure gives an overall view of clustering results based on quality and percentage of page clustered.

5.2 Basic Performance

In the previous section, we used four topics to analyze the limitations of existing content-link coupled approach. In this subsection, we use the same four topics to demonstrate the effectiveness of two techniques introduced in Section 4. The name convention is as follows. CLCK denotes the traditional content-link coupled approach depicted in Fig 1. CLCK-I denotes the algorithm that integrates in-link reinforcement in CLCK as depicted in Fig 2. While CLCK-A denotes the algorithm that integrates anchor window analysis in CLCK, CLCK-AI denotes the algorithm that integrates both techniques. We use the same weighting factors as mentioned above for all the four approaches. As we analyzed in [19] that both contents analysis and link analysis have their own feature and contribution, pro and con. So, when contents and link analysis are at the same "scale", the recommended weighting factors still work.

Table 5. The basic performance

	Jaguar	Salsa	Abduction	HIV
CLCK	10/0.26	6/0.29	3/0.37	1/0.59
CLCK-I	10/0.22	6/0.24	7/0.26	5/0.39
CLCK-A	10/0.19	6/0.21	5/0.33	4/0.43
CLCK-AI	10/0.20	6/0.22	7/0.25	6/0.36

Comparison of the performance of four different algorithms is given is Table 5. There are two numbers in each entry, representing the distinct semantic clusters identified by the specific clustering approach and the average entropy, of the clustering results, respectively. During experiments, we use higher similarity threshold (e.g. 0.2 for the four topics).

The results in Table 5 indicate clearly the effectiveness of the two techniques proposed in Section 4. While each individual technique, in-link reinforcement or anchor window analysis may improve the clustering quality in certain types of clusters, only the combination of two techniques can give uniform better results in all cases. Furthermore, the final entropy of the algorithm that integrates both techniques is very close to the best one, if not lower.

For topic *jaguar*, CLCK has already obtained reasonably good results and there seems no much room for improvements as we discussed in the previous sections. Yet, both techniques improved the clustering quality to certain extent. By analyze the results, we found that the two techniques mainly improve the quality of big/main cluster(s) produced, which reduces the average entropy significantly. For topic *salsa*, since the effects of in-link reinforcement more or less depend on the quality of term-based clustering at the first place, anchor window analysis contributes more to the improvements. For the latter two cases *abduction* and *HIV*, the original in-link information considered in CLCK seems not sufficient. In-link reinforcements improved the clustering quality significantly in terms of both the entropy and number of clusters produced. For topic *abduction*, CLCK-I has produced 7 distinct semantic clusters while original CLCK could only get roughly three clusters. While terms in anchor window of in-link pages are more summarized and general, it is not so effective in identifying new clusters/topics as in-link analyses.

5.3 Performance Comparison Based on Objective Metrics

The performance of four algorithms in terms of average entropy and F-measure is depicted in Fig. 3 and Fig. 4 respectively. As we can see from the Fig. 4, for different topics, in-link reinforcement and anchor window analysis demonstrated different contributions. Not only "purify" clusters is improved but also semantic new clusters can be identified. For polysemous topics like "jaguar" and "salsa", which have very different meaning under different context, anchor window analysis contribute most by improving contents analysis. For non-polysemous topics, like "abduction" and "HIV", in-link reinforcement gives better results by identifying more semantic clusters. In-Link reinforcement also depends on the quality of term-based clustering to certain extent. For the worst case like topic *HIV*, combining two methods gives biggest improvements.

CLCK-I works poor in terms of *F*-measure for topic

like *jaguar* and *salsa*s as shown in Fig. 4. The reason is that CLCK-I split some cluster on a topic into several small clusters. According to definition of *F*-measure, it gets low recall for the topics. Since during web page clustering, entropy and precision are more important and influential given acceptable percentage of page clustered, we could say that in-link reinforcement is useful and effective in web page clustering.

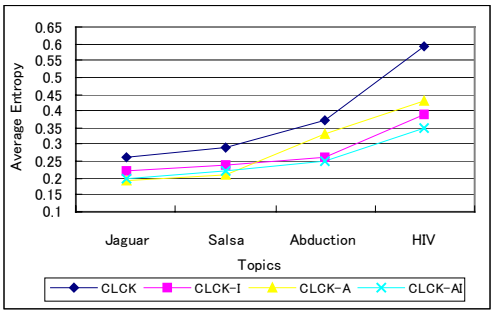

Fig. 3 Evaluation of two techniques based on average entropy

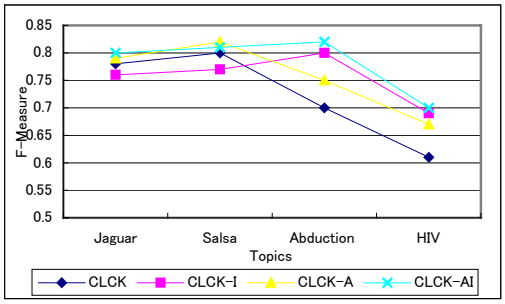

Fig.4 Evaluation of two techniques based on *F*-measure

6. Conclusion

Web page clustering is one of the most important techniques for discrimination, navigation, summarization and understanding web pages. However, traditional clustering approaches cannot offer a satisfactory solution for web page clustering. In particular, clustering algorithms based solely on contents of Web pages works poorly. While hyperlinks provide extra information for Web page clustering, early content-link coupled clustering algorithms cannot make full use of such information and work well only for certain types of Web pages because of the difficulties caused by the huge link space. In this paper, we analyzed the reasons why those algorithms fail to provide good performance to all types of Web pages based on our experimental results, and proposed two techniques, in-link reinforcement and anchor window analysis. By integrating these two techniques into the content-link coupled clustering algorithms, the quality of clustering can be improved significantly even for very tough topics. For the point view of clustering, we would like to develop even more robust algorithms that rely little on parameter setting, and are adaptive to various types of Web pages. With the success of clustering topical Web pages, the next step is to extend the technique to cluster a corpus of web pages (even entire web).

Reference

[1] C. Glassman, M. S. Manasse, and G. Zweig. 97 Syntactic clustering of the Web. WWW6, CA, USA, 1997, pages 587-595

[2] D. R. Cutting, D. R. Karger, et. al Scatter/gather: A Cluster-based approach to browsing large document collections. *ACM SIGIR'92*,Copenhagen, Denmark, pages 318-329

[3] Daniel Boley et. al. Partitioning-based Clustering for web document Categorization , it can be found at www.enterpriseware.net/ EWRoot/Files/ Boley1999a.pdf

[4] E. Amitay *Using common hypertext links to identify the best phrasal description of target web documents*, in Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.

[5] G. Kleinberg and Raghavan. *Inferring Web communities from link topology.* ACM Hypertext'98,*USA,* pages 225-234

[6] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents, *J. American Soc. Info. Sci.*, 24(1973), pages 265-269.

[7] J. Pitkow et. al. Life, Death and lawfulness on the Electronic Frontier. *SIGCHI'97*, USA, pages: 383 - 390

[8] J. Dean and M. Henzinger *Finding related page in the World Wide Web*. WWW8, Toronto, 1999

[9] Kleinberg Authoritative sources in a hyperlinked environment. *SODA*, January 1998.

[10] L Kaufman et. al.. *Finding groups in Data: an introduction to cluster analysis*. Wiley-Interscience Publication, 1990.

[11] M. Ester, H. Peter et. al. *A Density-based Algorithm for Discovering Cluster in Large Spatial Database with Noise*, SIGKDD'96, USA, pages 226-231

[12] M.M. Kessler, Bibliographic coupling between scientific papers American Documentation, 14(1963), pages 10-25

[13] O. Zamir and O. Etzioni Grouper: A Dynamic Clustering Interface to Web Search Results, WWW8, Toronto

[14] O. Zamir and O.Etzioni. *Web document clustering: A feasibility demonstration* SIGIR'98, Melbourne, Australia, pages 46-54

[15] R. Kumar, P Raghavan, S. Rajagopalan and A. Tomkins *Trawling the Web for emerging cyber-communities* WWW8, Toronto, Canada, 1999

[16] R. Weiss, B. Vélez and M. A. Sheldon *Hypursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering* Hypertext'96, Washington D.C, USA, pages 180-193

[17] S.Brin and L.Page *The anatomy of a large scale hypertextual web search engine*. *WWW7*, Brisbane Australia, 1998 pages 379-388

[18] T.H. Haveliwala, A. Gionis, and P. Indyk. *Scalable techniques for clustering the web*. In Proc. of the WebDB Workshop, USA, 2000

[19] Y.Wang and M. Kitsuregawa *Evaluation of Contents-Link Coupled Clustering for Web Search Results*, CIKM'02, USA, 2002, pages 499-507

[20] Z. Jiang et. al. Retriever: Improving Web Search Engine Results Using Clustering

[21] Z. Su, Qiang Yang et.al *Correlation-based Document Clustering using Web Logs* 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 5