データベースシステム 128－41
（２００２．７．１８）

リンク・コンテンツ統合ウェブページクラスタリング手法の効果の検証

王軼トン　　喜連川優　{ytwang, kitsure@tkl.iis.u-tokyo.ac.jp}
東京生産技術研究所

近年クラスタリングはウェブ上における莫大な量の情報を処理（例えば、リソース探索や情報解釈）するための最も重要な手法の一つとなっている。本論文では、リンク情報とコンテンツを統合することによって、クエリートピックにおける検索結果をクラスタリングすることを可能にした新しい手法であるリンク・コンテンツ統合クラスタリング手法を提案し、本手法の質を検証した。種々の実験を行った結果、本手法によって、検索結果が返す莫大な量のウェブページを簡潔な階層構造による高クオリティーでセマンティックに意味のあるグループに分類し、また、そのグループに関するトピック名と共に提示できることを確認した。本論文では、これらの実験を通して得られた結果を提示し本手法が非常に効果的で有望であるということを示す。

Examining the Quality of Link-Contents Coupled Clustering for Web Pages

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

Abstract Clustering is currently one of the most crucial techniques for dealing (e.g. resources locating, information interpreting) with massive amount of heterogeneous information on the web. In this paper, we present a unifying clustering algorithm to cluster web search results for a specific query topic by combining link and contents information. In particular, we examine the quality of the proposed link-contents coupled clustering approach. The proposed approach automatically clusters the web search results into high quality, semantically meaningful groups in a concise, easy-to-interpret hierarchy with tagging terms. We conduct experiments and comparisons and the experimental results show that the proposed approach is effective and promising. Keywords: co-citation, coupling, anchor window, snippet

1. Introduction

The web creates new challenges for research in the fields of database, IR and data mining. The quality (recall and precision) and correspondent interpretation of search results for current search engines are far from satisfying due to various reasons like huge volume of information; users differ on requirements for search results; users may be just interested in “most qualified” information or one peculiar part of information etc. Especially, synonymity (different terms have similar meaning) and polysemy (same word has different meanings) make things more complicated.

Many works [1,2,3,16,25] tried to explore link analysis to improve the quality of web search results or mine useful knowledge on the web. Kleinberg proposed HITS algorithm in [1] to locate the “most authoritative” (authority) pages for a query topic and suggested that there are two kinds of pages in search results: “hub” and “authority ” and they reinforce each other. However, sometimes one’s “most authoritative” pages are not useful for other people and further investigations on the above challenges are in high demand. The goal of our work is to cluster high-quality pages in web search results into more detailed, semantically meaningful groups with tagging terms to facilitate user’s searching and interpretation. *Web search results /search results* is used to denote web pages returned from web search engine on a specific topic. We use URLs or pages interchangeably when referring to search results.

Clustering approaches could be classified in two broad categories: term-based clustering [7, 8, 12, 14, 21, 24] and link-based clustering [9,11, 20,25]. Term-based clustering that is based on common terms shared among documents does not adapt well to web environment since it ignores the availability of hyperlinks between web pages and is susceptible to spam. Hyperlinks could provide valuable information to determine the related page since they give objective opinions for the topic of the pages they point to. Moreover, web search results are also different from a corpus of text documents in words distribution [23]. It is pointed out in [1] that many “authority” pages contain very little text. All these facts present difficulties in using term-based approach for web page clustering. In [20], we proposed a link-based clustering algorithm by co-citation and coupling analysis. According to preliminary experimental results, link-based clustering could produce some medium size but high quality clusters of web search results. However, it suffers from the facts that pages without sufficient in-links (out-links) could not be clustered, which means the recall is low. So it is very natural to investigate how to combine link and contents information in clustering algorithm to overcome the above problems. Unlike clustering in other fields, web page clustering should separate irrelevant ones from relevant pages and only cluster relevant pages into meaningful groups.

The paper is organized as follows. Section 2 is an assessment of related work of clustering in web domain. In Section 3, we describe the link-contents coupled clustering algorithm. Subsequently in Section 4, we report experimental results and evaluations. We present conclusion and future work in Section 5.

2. Related Work

Related work can be classified into following categories: one is clustering hypertext documents in a certain information space and the other one is clustering web search results. It is in [9] that a hierarchical network search engine is proposed to cluster hypertext documents to structure a given information space for supporting various services like browsing and querying based on the contents as well as the link structure of each hypertext document. In [21], a technique called LSH (Local–Sensitive–Hash) is proposed for web clustering. It plays more emphasis on the scalability of clustering. Snippet-based clustering is well studied in [7,8]. Shingle method, which is often used for duplicates removal is proposed in [14] to measure the similarity between pages for clustering. Applying the technique of association rule mining to term vectors is another clustering approach proposed in [24]. 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.

Clustering hypertext documents by *co-citation analysis* is explored in [11]. By applying HITS algorithm [1] to the vicinity graph around a seed URL, the approach proposed in [25] could find similar pages to the seed URL in a more narrow way, which is more focusing on finding similar pages than clustering web pages.

3. Link-Contents Coupled Clustering

Hyperlinks are helpful since they demonstrate objective opinions of the authors of other web pages to the pages they point to. **Co-citation** [19] and bibliographic **coupling** [18] are two more 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. Both co-citation and coupling are considered in the proposed approach.

The anchor text or snippet of page *u* means anchor text or snippet attached with the hyperlink that points to *u in* search results. Anchor window of a hyperlink includes anchor text as well as text that surrounds the hyperlink, which might include concise and important terms to describe the main topic of the page that the link points to. We consider four parts of text in our contents analysis for each URL/page *u* in search results: snippet, anchor text, meta-content and anchor window of the in-link *v* of *u*. Meta-content is an optional tag for most web pages and gives the summary of the page by the author. We "glue" the four parts for each page *u* in search results and apply stemming processing to it to extract terms.

By combining contents and link analysis, the proposed approach clusters search results based on common terms, in-links and out-links shared among them. We have several notations: *n, m, M, N, L* are positive integers, *R* is the set of specified number of search results for a topic. We use *n* to denote specified number of search results used for clustering, *m* to denote specified number of in-links extracted for each URL/page in *R*. *M, N, L* denote total number of distinct in-links, out-links as well as terms after applying link and contents analysis for all *n* pages in *R* respectively. We describe the clustering algorithm in more detail:

1) Representation of each page *P in R*

Each web page *P* in *R* is represented as three vectors: P_{Out} (*N*-dimension), P_{In} (*M*-dimension) and P_{KWord} (*L* – dimension). The *ith* item of vector P_{Out} indicates whether *P* has the correspondent out-link as the *ith* one in *N* out-links. If yes, the *ith* item is 1, else 0. P_{In} is identically defined. The *kth* item of vector P_{KWord} indicates the frequency of the corresponding *kth* term of *L* appeared in page *P*.

2) Centroid-based similarity measurement

We adopt traditional *Cosine* similarity measurement and the similarity of two pages *P, Q* includes three parts: out-link similarity *OLS (P, Q)*, in-link similarity *ILS (P, Q)* and contents similarity *CS (P,Q)*, which are defined as follows:

$$OLS(P,Q) = (P_{Out} \bullet Q_{Out}) / (\| P_{Out} \| \| Q_{Out} \|)$$

$$ILS(P,Q) = (P_{In} \bullet Q_{In}) / (\| P_{In} \| \| Q_{In} \|)$$

$$CS(P,Q) = (P_{KWord} \bullet Q_{KWord}) / (\| P_{KWord} \| \| Q_{KWord} \|$$

|| || is length of vector. Centroid or center point *C* is used to represent the cluster *S* when calculating the similarity of page *P* with cluster *S*, ***Sim(P, S)***. Centroid is usually just a logical point, which also includes three vectors. ***Sim(P, S)=Cosine(P, C)=***

P1* *OLS (P,C)*+P2* *ILS(P,C)*+P3* *CS(P,C), where P1+P2+P3=1* (1)

Centroid C is defined as:

$$C_{Out} = \frac{1}{|S|} \sum_{P_i \in S} P_{iOut} \quad C_{In} = \frac{1}{|S|} \sum_{P_i \in S} P_{iIn} \quad C_{Kword} = \frac{1}{|S|} \sum_{P_i \in S} P_{iKWord}$$

|S| is number of pages in cluster S. By varying the value of *P1, P2 and P3*, we could get an in-depth understanding of the role of out-link, in-link as well as term in clustering process.

3) Clustering method

We make some extensions to standard K-means and the clustering method is as follows:

- Filter irrelevant pages
- Define similarity threshold

Since similarity is meant to capture the common links and terms shared by different pages the similarity threshold could be easily defined and adjusted.

- Assign each page to clusters iteratively

Each page *is assigned top C existing clusters* according to *similarity threshold*. If none of current existing clusters meet the demand, the page under consideration becomes a new cluster itself. We limit *C* to top *3* clusters based on similarity values. All pages that join clustering procedure are processed sequentially and the whole process converged when centroids of all clusters are no longer changed.

- Merge two base clusters

Two base clusters produced by previous steps are merged if they share majority members. ***Merge threshold*** is used. Merging process is also iteratively executed until no clusters share more members.

Experimental results show that final clustering results are insensitive to the processing order; however, further investigation about this point is needed, which is not discussed here. The convergence of the approach is guaranteed by K-means itself since our extension does not affect this aspect.

4) Introducing some heuristic rules

- Differentiating among links

We would like to differentiate among links by weighting them. It is very common for a page *u* that many of its in-link pages are from the same website. E.g. for URL/page www.jaguar.com, more than 20 in-link pages are from website www.fort.com.

Rule For an URL/page *u*, if more than one in-link (out-link) page of it is from the same website, we would replace these in-link (out-link) pages (e.g. the number is *K*) that from the same website with one website page with weight K1 (1<K1<K).

The value of K1 is determined according to the value of K. In our experimentation, we set K1 as 1, 2 or 3 when K with the value intervals as K=1, (1<K <20) or K>20. For the above example, we replace all in-link pages that from the website www.fort.com with one in-link page http://www.fort.com/ with weight 3.

- Hierarchical Clustering

We apply hierarchical clustering on previous clustering results to make the final clustering result into a concise, easy to interpret hierarchy. Another HR-merging threshold is used as the halt condition. Similarity between two clusters is identically calculated as defined in formula (1).

(i) Compute the similarity for every possible pair of clusters;

(ii) For all pair of clusters that similarity is bigger than HR-merging threshold, we preserve them for further processing. We select one pair, say *(a, b)* and then merge them into a higher-level cluster *A*. Other cluster pairs that share one member with *A*, say *(a, c) or (b, c)* will be add into *A*, which result in *a*, *b*, *c* are in *A*. If there is no such cluster pair, select another cluster pair to process. The selection order is descendent based on the similarity values.

(iii) Repeat step (i), (ii) until the similarity of all possible pairs of clusters are smaller than HR-similarity threshold.

5) Tagging each cluster

We present tagging terms for each cluster since it is important for users to have a flavor of the main topic of the cluster by a glance of the tagging terms. Say for cluster *S*, *C* is its centroid, from C_{Kword} , it is easy to know terms that have higher values and are most shared by the members of Cluster *S*, which might convey the main topic of the cluster.

	C/0.1	L/0.1	MA/0.1
1	Car , type (6/ 87))	Car, type, part (3/ 67)	Car, type, part, restore, race (4/ 68)
2	Club, support (3/ 57)	Club (1/ 23)	Club (1/37)
3	Game, Atari (3/28)	Game, Atari (2/ 17)	Game, atari (2/ 32)
4	Cat, onca (3/ 15)	Cat, onca (1/ 8)	Cat, wildlife, onca (2/ 13)
5	**	Book, magazine (1/ 6)	Book, jag, magazine (3/10)
6		Tour, reef (1/ 4)	Reef, tour (1/ 5)

Table 1. Final clustering results for topic "Jaguar "

	C/0.1	L/0.1	MA/0.1
1	New, York, City (4/ 98)	New, York, city (3/ 54)	New, York, City (2/ 76)
2	Theater, circus (6/ 41)	Circus (1/12)	Theater, Broadway, ticket (3/ 17)
3	Classic, Sybase golf (1/ 11)	Game, user, group (1/ 9)	Circus, trapeze (2/14)
4	Company, offer (1/ 18)	Sports (1/ 9)	Game, user, group (2/ 11)
5			Sports, company, product (2/ /14)
6		Classic, Sybase golf (1/ 3)	Classic, Sybase, golf (1/ 3)

Table 2. Final clustering results for topic "big apple"

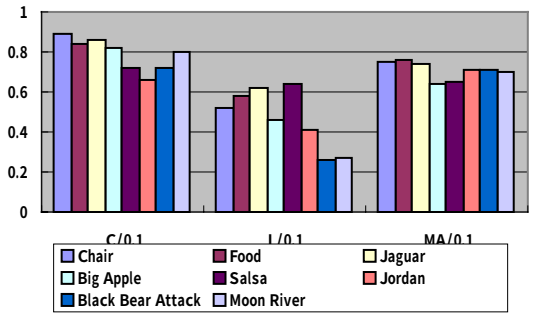

Fig.1 Percentage of page clustered (see Section 4 for definitions of C, L, MA)

Example: Noise web pages in search results of topic "jaguar" that are clustered by term-based clustering
http://www.folkart.com/~latitude/folktale/tale_3.htm
http://www.crica.com/hotels/jaguar.html
http://centralamerica.com/cr/hotel/jaguar.htm
http://www.jaglair.com/rain/jag-rain.htm (Cat)
http://www.jaguarpc.com (Car saloon)

http://jaguar.online.fr (Car saloon)
http://www.jindal.com/jaguar (Car saloon)
http://www.yerbamate.com (Car dealer)
http://www.rogers16.freeserve.co.uk (Club)
http://www.raf.mod.uk/airpower/jaguar.htm (Car service support)
http://www.audiovisualizers.com/toolshak/vidsynth/jag_vlm/jag_vlm.htm (Atari Game)
http://www.theatlantic.com/issues/2000/07/johnson.htm (book)

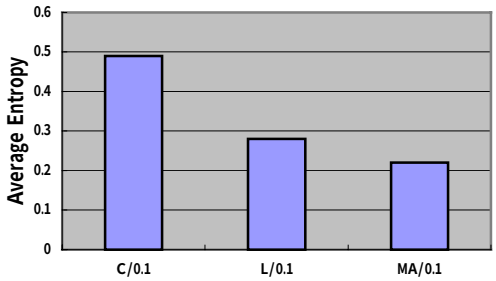

Fig2 Average comparisons based on entropy for eight topics (see Section 4 for definitions of C, L, MA)

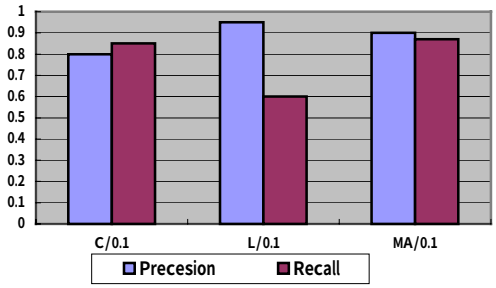

Fig. 3. Average comparisons based on precision and recall for eight topics

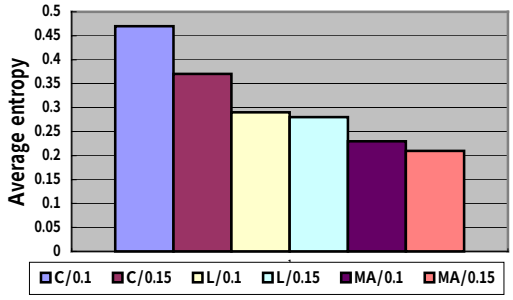

Fig. 4. Comparisons for topic "jaguar" with different similarity thresholds

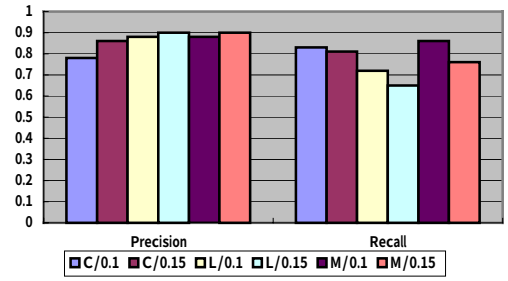

Fig.5 Comparison for topic "jaguar" with different similarity thresholds

4. Experiments and Evaluations

We arbitrarily select eight topics for experimenting, which include rather general ones like "food" and "chair"; relatively specific ones like "black bear attack" and "moon river"; as well as other topics like "jaguar", "big apple", "salsa", "jordan". Especially, we choose topic "jaguar" for detailed comparison. We test 200 URLs for each topic and extract 100 in-links for each URL in search results. All results are obtained by Google search engine. By varying the parameters in formula (1), it is possible to examine the effect of link and contents analysis on the clustering process. Term-based clustering is denoted as "C" (0, 0, 1 for P1, P2, P3); link-based clustering is denoted as "L" (0.5, 0.5, 0 for P1, P2, P3); Link-contents coupled clustering is denoted as "MA" (0.2, 0.3, 0.5 for P1, P2, P3). The choice of parameter values for clustering approach "MA" is based on empirical evaluation. Similarity threshold 0.1 and merging threshold 0.75 is used in our experimentation as recommended in [20]. So "C/0.1" means term-based clustering with similarity threshold 0.1. Another HR-merging threshold is introduced in the hierarchical clustering process. We deliberately choose a relatively strict one 0.4 for it since we would like to make sure that only clusters that are similar enough will be merged into one higher-level cluster. The anchor window we tried in our experimentation is 4, which include two word to the left and two words to the right of the anchor text.

4.1 Experimental Results

As final clustering results reveal, one page could belong to more than one cluster or belong to singleton cluster, which means that it cannot be grouped with other pages. In the rest of discussion, "pages/URLs clustered" means pages or URLs that appear in final clusters whose size are no less than 3. The size of a cluster is the number of pages in the cluster. We ignore singleton clusters or very small clusters. In Table 1 and Table 2, we give the final clustering results after hierarchical clustering process for topic "jaguar" and "big apple", which could give a flavor of clustering results from the semantic point of view. Each entry in the tables is the main tagging terms we get according to part 5 of Section 3. The two numbers in the parenthesis of each entry in the two tables are: a) the number of sub-clusters included in this cluster to indicate whether the cluster is a higher-level cluster; b) the number of distinct pages /URLs clustered in this cluster. E.g. for the first entry of term-based clustering "C/0.1" in Table1, the tagging words are "car, type" and the two numbers are 6, 87. It means that the cluster is a higher-level cluster composed of 6 sub-clusters and there are totally 87 distinct URLs/ pages are grouped in this cluster. Its main topic is about parts of Jaguar cars.

From the two tables, we get impression that for term-based clustering, it could only identify the most

popular ideas around the topic and fail to separate pages if they are differ slightly in topics. From link-based clustering L/0.1, we know that it could identify some medium size, tightly related meaningful clusters. The main disadvantages of link-based clustering are low recall and the quality of big clusters is not good. For combining links and contents in clustering as MA/0.1, it is clearly that it could "pull" some pages with the same topic but missing common links into the cluster.

4.2 Evaluation of Clustering Results

We would like to use three metrics *precisions*, *recall* and *average entropy* to evaluate the quality of final clusters. In our initiative evaluations, we manually check 200 URLs for each topic and then give our judgments. Each page is given two estimates: relevant or not (to the query topic), its main topics and then create *classes* manually. Although this is time-consuming and it could lead to bias in our evaluations, it is possible to carry out user experiment to counteract potential bias. Of all 200 pages for each topic, around 75% are marked "relevant" on the average.

4.2.1 Evaluation Metrics

(A) *Precision* and *recall*

We use *A* to denote the number of URLs clustered and *B* to denote the number of URLs that marked 'relevant'; then we redefine *precision* and *recall a*s follows:

Precision= $|A \cap B| / |A|$

Recall= $|A \cap B| / |B|$

Precision and *recall* are two global metrics that used to measure: for all pages in search results, whether noise pages are removed from being clustered and high quality pages are clustered respectively.

(B) *Average Entropy*

In order to get a clear understanding for each cluster, we use "entropy" to measure the "goodness" or "purity" for un-nested clusters by comparing the groups produced by the clustering technique to known classes. Low entropy means high quality of the cluster because of high intra-cohesiveness while high entropy means that the cluster members are not tightly related but cover different sub-topics under the general query topic. Since clustering is meant to group similar ones together, we think average entropy is more influential when evaluating the quality of a clustering algorithm. We adopt the computing of entropy introduced in [10]: Let CS is a cluster solution and *E (j)* is the entropy for cluster *j*. The average entropy for a set of clusters is calculated as the sum of entropy of each cluster weighted by its size. The definitions are as follows:

$E(j) = -\sum_i p_{ij} \log(p_{ij})$, $E_{CS} = \sum_{j=1}^{m} \frac{n_j * E(j)}{n}$, p_{ij} is the "probability" that a member of cluster *j* belongs to the given class *i*. n_j is the size of cluster *j*, m is the number of clusters and n is the total number of page clustered.

4.2.2 Comparisons among Different Clustering Approaches

By varying the value of parameters in formula (1) it is possible to compare different clustering approaches. From Fig.1, we could know that on the average, term-based clustering gives the highest ratio of page clustered and link-based clustering gives the lowest. However, the highest ratio of page clustered does not produce the highest recall, as shown in Fig.3, which is the average value of precision and recall for all eight topics. This means that term-based clustering fails to separate noise pages and group them into final clusters. In order to get an in-depth understanding of this, we give a detailed check for the clustering results of topic "jaguar" by term-based clustering. Some noise pages clustered by term-based clustering are presented in *Example*. They could be kicked off by combining links and contents information in clustering. These noise pages are clustered because they accidentally share some important terms with other high quality pages or they share some terms each other. E.g. the first three URLs listed in Example are in one cluster since they share terms like "hotel", "food", "little" etc, however, they are in no way similar. The text in the parenthesis of each URL in *Example* is the topic of the cluster it belongs to. The fourth URL is clustered with other pages on topic "Jaguar Cat" since it includes some terms like "wildlife", "endanger" etc, while actually it is just a video clip on rainforest. Since the noise pages share no links between each other or with other pages, they could be prevented from being clustered by combining link and contents analysis in clustering process.

Based on the evaluation metrics introduced in section 4.2.1, we compare the quality of clustering results among the three clustering approaches as demonstrated in Fig.2 and Fig.3. The average entropy is calculated according to clustering results before applying hierarchical clustering. In general, the average entropy for term-based clustering ("C/0.1") is rather high, which means that the clusters obtained by this way are very coarse, pages in one cluster actually cover different subtopics. Link-based clustering ("L/0.1") could improve a lot for this but with low recall since it could produce some medium but tightly related clusters. Link-Contents coupled clustering ("MA/0.1") could complement this without sacrificing the "purity" but at a little cost of precision, which is clearly conveyed in Fig.2 and Fig. 3 since snippets and anchor windows usually bring noises. We also try different similarity thresholds. When increasing the similarity threshold, the average entropy decreases, which gives better "purity" as shown in Fig.4. The precision of clustering results increases while recall decreases, as shown in Fig.5.

Since clustering web search results is meant to give clear classified information to facilitate user's locating and interpretation, the proposed link-contents coupled clustering is effective in separating noise pages from high quality ones and clusters high quality pages into meaningful groups. In general, it works much better than current term-based clustering and link-based clustering as well.

5. Conclusion

We present a unifying clustering approach in the paper by combining link and contents information that appeared in anchor text, snippet, meta-content as well as anchor window of the in-links, which might give a reasonable summary for the topic of the page under consideration. In particular, we investigate the effect of link and contents analysis on clustering process. We conducted experiments and evaluations on eight topics, which include rather general ones like "chair" and rather specific ones like "black bear attack" as well as several other topics like "jaguar", "big apple" etc. According to preliminary experimental results, contents analysis is useful to identify the general idea since term-based clustering produces "coarse" clusters and fail to relate pages in a more narrow way. Link-based clustering could identify tightly related, medium size but meaningful groups by link analysis. However, it suffers from the problems that pages with few/insufficient in-links or out-links will not be clustered and the "purity" of big-size clusters is also not so good (high entropy). Combining contents and links provide much help for the mentioned problems. The final clustering results of the proposed approach are presented in a concise, easy to interpret form of hierarchy. Our evaluation is based on three metrics: average entropy, precision and recall, which we think that average entropy is more influential when evaluating a clustering algorithm. The experimental results suggest that the proposed link-contents coupled clustering gives improvements over term-based and link-based clustering approach in following several ways: 1) improve the recall by "pulling" more high quality pages into the cluster with same topic and "removing" some noise pages; 2) balance the clustering process to give reasonable clusters; 3) improve the average entropy as a whole.

While our preliminary experimentation on the proposed approach gives positive results, we still need to conduct detailed analysis and interpretation of the experimental results. Further investigations and improvements like more extensive check on other topics as well as the effects of parameters introduced in similarity measurement are also among our next step works.

References

1. **Kleinberg 98** *Authoritative sources in a hyperlinked environment.* SODA, January 1998.
2. **Ravi Kumar *et. al.* 99** Trawling the Web for emerging cyber-communities WWW8, 1999
3. **Brin and Page 98** *The anatomy of a large scale hypertextual web search engine*. WWW7, Australia
4. **Oren Zamir and Oren Etzioni 99** *Grouper: A Dynamic Clustering Interface to Web Search Results*, WWW8, Toronto Canada.
5. **Richard C. Dubes and Anil K.Jain,** *Algorithms for Clustering Data*, **Prentice Hall, 1988**
6. **Oren Zamir and Oren Etzioni 97** *Fast and Intuitive clustering of Web documents*, KDD'97,
7. **Oren Zamir *et. al.* 98** *Web document clustering: A feasibility demonstration* SIGIR'98, Australia.
8. **Zhihua Jiang *et. al.*** *Retriever: Improving Web Search Engine Results Using Clustering*
9. **Ron Weiss *et. al.* 96** *Hypursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering* Hypertext'96 Washington USA
10. **Michael Steinbach, *et. al.*** *A Comparison of Document Clustering techniques* KDD'2000.
11. **James Pitkow *et. al.*** *Life, Death and lawfulness on the Electronic Frontier*. SIGCHI'97
12. **Cutting, D.R. *et. al.92*** *Scatter/gather: A Cluster-based approach to browsing large document collections*. ACM SIGIR'92, pp 318-329, 1997
13. **A.V. Leouski *et. al.* 96** *An evaluation of techniques for clustering search results.* Technical Report, University of Massachusetts, Amherst,
14. **Broder *et. al.* 97** *Syntactic clustering of the Web.* WWW6,
15. **Lenoard Kaufman and Peter J. Rousseeuw.** *Finding groups in Data: an introduction to cluster analysis* Wiley, 1990
16. **Gibson, Kleinberg *et. al* 98** *Inferring Web communities from link topology.* Hypertext'98.
17. **Agrawal and Srikant 94** *Fast Algorithms for mining Association rules* VLDB'94, Chile.
18. **M.M. Kessler,** *Bibliographic coupling between scientific papers* American Documentation, 14(1963), pp 10-25
19. **H. Small,** *Co-citation in the scientific literature: A new measure of the relationship between two documents*, J. American Soc. Info. Sci., 24(1973), pp 265-269
20. **Yitong Wang and Masaru Kitsuregawa,** *Use Link-based clustering to improve web search results*, WISE'01, 2001
21. **Taher H.Haveliwa *et. al. 99*** *Scalable techniques for Clustering the Web.*
22. **Taher H.Haveliwa *et. al.*** *Similarity Search on the Web: Evaluation and Scalability Considerations* Extended Technical Report, 2000
23. **Einat Amitay** *Using common hypertext links to identify the best phrasal description of target web documents*, SIGIR'98 workshop for Hypertext IR for the web
24. **Daniel Boley *et. al.*** *Partitioning-based Clustering for web document Categorization* , www.enterpriseware.net/EWRoot/Files/Boley1999a.pdf
25. **J. Dean and M. Henzinger 99** *Finding related page in the World Wide Web*. WWW8