D-3 An Link-Contents Coupled Clustering for Web Search Results

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

1. Introduction

There is a big gap about the interpretability and accessibility of web search results of current search engine for a specific topic to meet users' demands. The reason might be various like too many returns, users differ with requirements and expectations for search results; sometimes a search request cannot be expressed clearly with few keywords etc. Especially, synonymity (different terms have similar meaning) and polysemy (same word has different meanings) make things more complicated. All these facts as well features of web data have created challenges for many research fields like database, IR and data mining.

Kleinberg argued in [1] that links between web pages could provide valuable information to determine related pages (with query topic). So, many works [1,2] try to explore link analysis to improve quality of web search process or mine useful knowledge on the web. We think clustering high quality web pages in search results into meaningful groups could help a lot for the above challenges. By doing so, it is much helpful for users to identify the main ideas around the topic on the web since users could have an overview or just select the interested group to view. We use *URLs* or *pages* interchangeably when referring to search results, which usually include list of ranked URLs point to correspondent web pages.

We propose link-contents coupled web page clustering approach by combining link (co-citation and coupling) and contents analysis. Especially we study their contributions in clustering process. According to preliminary experimental results, the proposed approach could produce reasonable clusters and works much better than current term-based or link-based clustering approaches.

2. Clustering Approach

Link analysis we considered here includes co-citation (capture common out-links between pages) and coupling (capture common in-links between pages). The contents analysis we considered is meant to capture common terms shared by pages in their *snippets*, *anchor text*, *meta-contents* and *anchor window of in-link pages*. By "gluing "the four parts of text and applying stemming process, we could obtain meaningful and useful terms to identify the main idea of the page under consideration.

Based on link and contents analysis, our approach clusters search results based on common in-links, out-links and terms they shared. In the rest of paper, M, N, L denote total number of distinct out-links, in-links extracted and terms after stemming processing for all *n* pages in Search Results *R* respectively.

1) Representation of each page *P in R*

Each web page *P* in *R* is represented as 3 vectors: P_{Out}, P_{In}, P_{KWord} with M, N and L dimension respectively. The *kth* item of each vector the frequency of the corresponding item (link or term) appeared in page *P*.

2) Centroid-based similarity measurement

The similarity of two pages includes three parts: out-link similarity *OLS(P,Q)*, in-link similarity *ILS(P,Q)* and term similarity *CS(P,Q)*, which are defined as:

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

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

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

$\| \|$ 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) is defined as***

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

Centroid C is defined as:

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

$$C_{KWord} = \frac{1}{|S|} \sum_{P_i \in S} P_{iKWord} \qquad (2)$$

|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 contributions of out-link, in-link and terms to clustering process.

3) Clustering method

We extend the standard K-means to overcome its disadvantages and our clustering method is as follows:

1. Define the similarity threshold
2. Filter irrelevant pages (associate with few links or terms)
3. 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 correspondent centroids
4. The page will be one cluster itself if no *existing cluster* meet step 3
5. Recompute the centroids of the clusters if its cluster members are changed
6. Repeat Step 2 until 5 until all relevant pages are assigned and all centroids do not change any more
7. Merge two base clusters produced by step 6 if they share most members based on merge threshold

The convergence of the approach is guaranteed by standard K-means itself since our extension does not affect this aspect. The two parameters introduced here are *similarity threshold* and *merge threshold.*

4) Introducing some heuristic rules

- Differentiating among links
- Hierarchical Clustering

While link analysis usually suffers from the problem of orthogonality, we would like to alleviate it by weighting links to differentiate among them. We merge out-links (in-links) from the same domain into a domain page with weight. We also apply hierarchical clustering to clusters produced by previous steps to make the final clustering results more concise and easy to interpret. Another HR-merging threshold is used as the halt condition. Similarity between two clusters is identically calculated as in formula (1).

5) Tagging each cluster

We attach tagging terms with each cluster since it is important for users to have a flavor of the main topic of each cluster by a glance of its tagging terms, which is based on C_{KWord} vector with

C as its centroid.

3. Experiments and Evaluations

We arbitrarily select eight topics for experimentation, which include rather general ones like "chair" and "food"; some topics with more than one meaning under different context like "jaguar", "big apple", "Jordan", "salsa" as well as relatively specific ones like "black bear attack" and "moon river". This is also the order of topics in x-axles of Fgure1. In Table 1 and Table 2, we give clustering results with tagging words for two topics.

By varying the parameters in formula (1), it is possible to try different clustering approaches for a specific topic. Link-based clustering is denoted as "L" (with P1, P2, P3 as 0.5, 0.5, 0); term-based clustering is denoted as "C" (with P1, P2, P3 as 0, 0, 1); combining links and contents analysis is denoted as "M" (with P1, P2, P3 as 0.2, 0.3, 0.5). Similarity threshold 0.1 and merge threshold 0.75 is used in our experimentation as recommended in [3]. Another HR-merging threshold is introduced in the hierarchical clustering process. We deliberately choose a relatively strict one 0.4 for it. 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.

The two numbers in the parenthesis of each entry in the 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. "**" in the tables indicates the corresponding cluster is not interpretable.

	C	L	M
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	L	M
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"

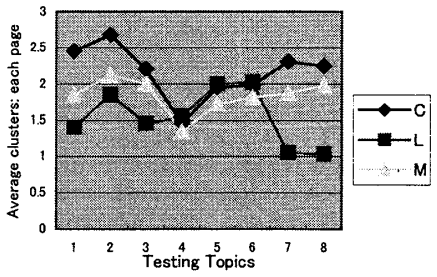

Fig.1 Average clusters each URL belongs for each topic based on three clustering approaches

Experimental results (Table 1 and table 2) suggested that term-based clustering are rather "coarse", could only clearly identify the most popular ideas around the topic but fail to separate pages if they are differ slightly in topics. The main disadvantages of link-based clustering are low recall and the quality of big clusters is not good although it could identify some medium but meaningful groups.

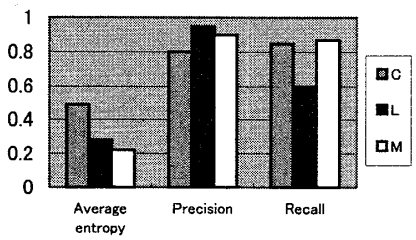

Fig2. Evaluation of three clustering approaches based on three metrics (see section3 for definitions of C, L, M)

Our evaluations are based on three metrics: average entropy, precision and recall. We adopt the computing of entropy introduced in [3], which provides a measure of "goodness" or "purity" for un-nested clusters by comparing the groups produced by the clustering technique to known classes, which are done manually in our experiments 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 includes some noises or covers more than one sub-topic under the general query topic. We use A to denote the number of URLs clustered and B to denote the number of URLs that marked 'relevant', then *precision* and *recall* are redefined as follows: $Precision = |A \cap B| / |A|$; $Recall = |A \cap B| / |B|$

Term-based clustering usually has the highest overlap while link-based clustering gives the lowest as shown in Fig.1. According to Fig. 2, the average entropy for term-based clustering is rather high, which means that the clusters obtained by this way are very coarse, pages in one cluster actually covers different subtopics. Link-based clustering could improve a lot for this but with low recall since the clustering results for L are some medium but tightly related, meaningful clusters. Combining link and contents analysis will compensate this without sacrificing "purity" but at a little cost of precision as shown in Figure 2 since snippets usually bring some noises.

4. Conclusion

In this paper, we extend the previous work on link-based clustering by combining link and contents analysis. Our goal is to cluster high quality pages (by filtering some irrelevant pages) in web search results for a specific query topic into semantically meaningful groups with useful tagging keywords to facilitate users' locating and interpretation. We also extend standard K-means algorithm to overcome its disadvantages to make it more natural to handle noises. Experimental results suggested term-based clustering is too "coarse" and link-based clustering could identify tightly related, meaningful group with low recall and high entropy for big-size clusters. Experimental results and evaluations suggest that the proposed approach gives significant improvements over term-based and link-based approach in several ways: 1) improve the recall without loss of quality (entropy) by "pulling" more pages into the cluster with same topic; 2) balance the clustering process to give reasonable clusters; 3) improve the average entropy as a whole.

Reference

1. Kleinberg 98 Jon Kleinberg. *Authoritative sources in a hyperlinked environment.* SODA, January 1998.

2. Ravi Kumar *et. al.* 99 Trawling the Web for emerging cyber-communities WWW8, Toronto, Canada, 1999

3. Yitong Wang and Masaru Kitsuregawa, *Use Link-based clustering to improve web search results*, WISE'01, pp. 119-128, 2001