Improve Information Interpretation by Clustering Web Search Results

１Ｘ－０２

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

1. Introduction

While web search engine could retrieve information on the Web for a specific topic, users have to step a long ordered list in order to locate the needed information, which is often tedious and frustrating due to various reasons like huge volume of information; users differ with requirements and expectations for search results; sometimes a search request cannot be expressed clearly with few keywords etc. Especially, synonym (different terms have similar meaning) and homonym (same word has different meanings) make things more complicated. In general, the resources locating (recall and precision) and accordingly interpretation of search results of current search engines are far from satisfying and has created big challenges for many research fields.

Kleinberg argued in [1] that links between web pages could provide valuable information to determine related pages (with query topic). So, many works [2,3] try to explore link analysis to improve quality of web search process or mine useful knowledge on the web. We think clustering web search results could help a lot. The goal of our work is to cluster high-quality pages in web search results into more detailed, semantically meaningful groups with tagging keywords to facilitate user's searching and information interpretation. 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.

We propose a clustering approach combining link (co-citation and coupling) and contents analysis. Especially we study their contributions in clustering process. According to preliminary experimental results, the approach could give reasonable results.

2. Clustering Approach

The links analysis we considered here includes co-citation (capture common out-links between pages) and coupling (capture common in-links between pages). The contents analysis is meant to capture common terms shared by pages in their snippets and anchor text. The anchor text and snippet for a page is the text appeared as the link and the text attached with the link in search results returned by search engine for a specific topic.

By link (co-citation and coupling) and contents analysis, our approach clusters search results based on common terms, in-links and out-links they shared. In the rest of paper, M, N, L denote total number of distinct out-links, in-links extracted and terms appeared in snippet and anchor text (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) Similarity measure

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

$$(P_{Out} \bullet Q_{Out})/(\| P_{Out} \| \| Q_{Out} \|) \qquad (1)$$

$\| P_{Out} \|^2 = (\sum_1^N P_{Out\,i}^{\;2})$ (Total number of out-links of *P*),

$\| Q_{Out} \|^2 = (\sum_1^N Q_{Out\,i}^{\;2})$ (Total number of out-links of Q),

ILS(P,Q) and *CS(P,Q) are defined identically with corresponding vectors*. Dot product is to capture the common out-links, in-links and terms shared by *P* and *Q*. $\| \ \|$is length of vector.

Centroid *C* of cluster *S* is used when calculating the similarity of page *P* with cluster *S*. Centroid is usually C_{out}just a logical point. It is defined as:

$$= 1/|S| \sum_{P_i \in S} P_{iOut}, \quad C_{In} = 1/|S| \sum_{P_i \in S} P_{i\,In} \quad C_{Kword} = 1/|S| \sum_{P_i \in S} P_{i\,KWord}$$

|S| is the number of pages in cluster S. So the similarity of pages *P* and clusters *S*, ***Sim(P, S)*** is defined as:

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

P1, P2 and P3 are parameters and can be adjusted to different clustering

3) Clustering method

We extend standard K-means to meet requirements for clustering of web search results as well as to overcome disadvantages of K-means. Our clustering method is:

a) Filter irrelevant pages
b) Define similarity threshold
c) Assign each page to clusters

Each page *is assigned to existing clusters* when the similarity between the page and the correspondent cluster is above the *similarity threshold*. If none of current existing clusters meet the demand, the page becomes a new cluster itself. Centroid vector is *incrementally* recalculated when new members are

Main Keywords (Cluster Size>3)	C/ 0.18	C/0.24	C/0.3	L/0.1	CL/0.1
1	Car (1.7), Club (1.4)	Car (1.8), Club (1.2)	Car (1.9), Club (1.2)	Car (2.1), Club (0.5)	Car (2.2), club (0.5)
2	Club, (1.7) Car (1.2)	Club (1.9), part (1.1), Car (0.7)	Club (1.6)	Club (1.5), frame (0.6)	Club (1.7)
3	Game (1.4), Atari (1.1)	Game (1.7)	Game (1.2)	Game (1.7)	Game (2)
4	**	**	**	Magazine (1.4)	Atari Emulate (1.9)
5	**	**	**	Cat (1.6)	Cat (1.8)
6	**	Magazine (1.3)	**	Atari Emulate (1.5)	Magazine (1.5)
7	**	**	**	Part (1.1), type (0.7)	Part (1.2), Type (0.9)
8	**	**		Race (0.9)	Race (0.8)
9	**			Tour (1.1)	Tour (1.1)
10					Link (1.2)
11					Frame** (0.9)
12					Support (1)

Table 1. Main tagging keywords of clusters for different clustering patterns

introduced to the cluster. All pages that join clustering procedures are processed sequentially and the whole process is iteratively executed until it converges (centroids of all clusters are no longer changed). Experimental results for topic "Jaguar" as well as the main tagging keywords and average weight for each cluster are shown in Table1. Different clustering patterns are tried by varying parameters in formula (1) with different similarity threshold. "C", "L" and "CL" represent "contents only", "link only" and "combining link and contents" respectively. Experimental results suggested term-based clustering is too coarse to identify some medium but meaningful groups as shown in Table1. "**" in Table 1 means that the corresponding cluster is not interpretable.

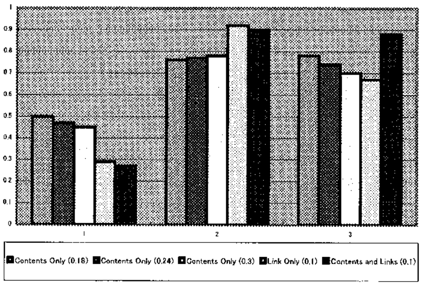

Figure 1 **Comparing of different clustering patterns based on average entropy, precision and recall**

3. Evaluation

Our evaluations are based on three metrics: average entropy, precision and recall and this is also the order of x-axles in Fgure1. We adopt the computing of entropy introduced in [3] and the other two as:

Precision=number of URLs that are both clustered and 'relevant' marked / number of URLs clustered

Recall=number of URLs that are both clustered and 'relevant' marked /number of 'relevant' marked URLs

In general, 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/0.1 are some medium but tightly related, meaningful clusters. Combining link and contents will compensate this without sacrificing "purity" but at a little cost of precision as shown in Figure 1 since snippets usually bring some noises.

4. Conclusion

In this paper, we proposed an approach to improve information interpretation by clustering web search results. Our goal is to cluster high quality pages (by filtering some irrelevant pages) in search results returned from web search engine for a specific query topic into semantically meaningful groups with useful tagging keywords to facilitate users' locating and information 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. Combining links and contents could solve this and generate reasonable clustering results with useful tagging keywords to help accordingly interpretation.

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. **Ron Weiss *et. al.* 96** *Hypursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering* Hypertext'96 Washington USA