Finding Pertinent Page-Pairs from Web Search Results

Takayuki Yumoto and Katsumi Tanaka

Dept. of Social Informatics, Graduate School of Informatics, Kyoto University,
Yoshida Honmachi Sakyo-ku Kyoto 606-8501, Japan
{yumoto, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. Conventional Web search engines evaluate each single page as a ranking unit. When the information a user wishes to have is distributed on multiple Web pages, it is difficult to find pertinent search results with these conventional engines. Furthermore, search result lists are hard to check and they do not tell us anything about the relationships between the searched Web pages. We often have to collect Web pages that reflect different viewpoints. Here, a collection of pages may be more pertinent as a search result item than a single Web page. In this paper, we propose the idea to realize the notion of "multiple viewpoint retrieval" in Web searches. Multiple viewpoint retrieval means searching Web pages that have been described from different viewpoints for a specific topic, gathering multiple collections of Web pages, ranking each collection as a search result and returning them as results. In this paper, we consider the case of page-pairs. We describe a feature-vector based approach to finding pertinent page-pairs. We also analyze the characteristics of page-pairs.

1 Introduction

Web search engines can find pertinent pages, and lead us to them. However, there are some cases when they cannot find pertinent answers. We consider two of them here. The first case is where information a user wishes to have is distributed on multiple Web pages. Conventional search engines do not suggest the misleading results but they do not tell us which pages include which part of the information we want. The second case is where we have to collect Web pages that reflect different viewpoints. For example, suppose that we wish to obtain information about "wind power generation" and "nuclear power generation". Some pages are described from the viewpoint of "wind power generation" and others are described from the viewpoint of "nuclear power generation". A single page with one viewpoint will not provide enough answers. Also, a conventional search engine will not tell us anything about the relationships between searched Web pages.

This is due to the same reason, i.e. conventional Web search engines evaluate each single page as a ranking unit. In both cases, a collection of pages may be more pertinent as an item for a search result than a single Web page.

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 301–310, 2005.

In this paper, we propose the new concept of “multiple viewpoint retrieval”, which means searching Web pages described from different viewpoints for a specific topic, gathering multiple collections of Web pages, ranking each collection as a search result and returning them as results. We also describe a simple approach to achieve multiple viewpoint retrieval and analyze the characteristics of page-pairs.

This paper is organized follows. Section 2 explains our motivation and the concept behind multiple viewpoint retrieval. Section 3 describes our approach to achieve multiple viewpoint retrieval, which we evaluate in Section 4. Section 5 is the conclusion and discusses future work.

2 Multiple Viewpoint Retrieval

2.1 Motivation

Although Web search engines can find pertinent pages, there are two cases conventional search engines cannot find these. This is where

– Information, the user wishes to have is distributed on multiple Web pages and where
– We have to collect web pages that reflect different viewpoints.

These cases have common problems. There are that conventional search engines do not reflect on the relationships between search results and search result lists output by conventional search engines give us no information about the relationships between Web pages.

2.2 Concept

To solve these problems, we propose “multiple viewpoint retrieval”, which means searching Web pages described from different viewpoints for a specific topic, gathering multiple collections of Web pages, ranking each collection as a search result, and returning them as results. When pages described from different viewpoints include the same topics, their content is different and the points they focus on are also different. To achieve multiple viewpoint retrieval, we need to establish the following:

1. Collecting Web pages: What Web pages should be collected?
2. Gathering multiple collection: What Web pages should compose each collection and what relationships they satisfy?
3. Ranking the each collection: What collection is pertinent?

2.3 Our Approach

We focused on gathering multiple collections and ranking each collection and took the approach re-ranking search results with conventional search engines[1].

[1] In this paper, we used Google[1].

This was because conventional search engines can find good results as a single page. To achieve "multiple viewpoint retrieval" simply, we considered page-pairs as ranking units.

The multiple viewpoint retrieval was executed in three steps:

1. Submit a query to a conventional search engine, and collect the Web pages,
2. Collect page-pairs taking the relationship between pages into consideration, and
3. Calculate the evaluation function for page-pairs and rank them.

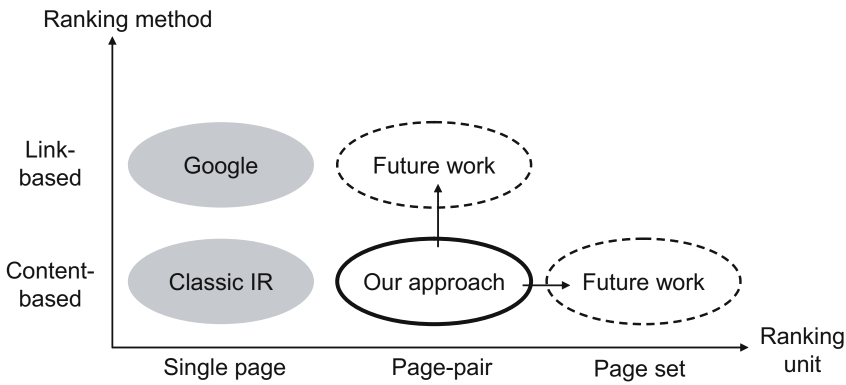

Fig. 1. The relationship between our approach and other research

Conventional Web search engines compute ranking scores for searched pages by the content-analysis approach (computation of page similarity to a query) or the link-analysis approach (such as Google's PageRank). In this paper, we also use the content-analysis approach like classic information retrieval. The major differences of our work from conventional work is that the information unit for ranking is not a single page, but a page-pair. Extensions of our approach to the link-analysis method and to the arbitrary collection of pages are remained as future work.

2.4 Related Work

Retrieval with Clustering. Cutting *et al.* proposed document clustering for efficient browsing [2], and some search engines take this approach. They prepare clusters from search results and display each page[3]. These are different approaches to ours. Clusters are collections that consist of similar pages. Our "multiple viewpoint search" prepares a collection from pages that are similar but have some different parts. Web pages in different clusters, which are prepared by search engines with clustering, are sometimes described from different

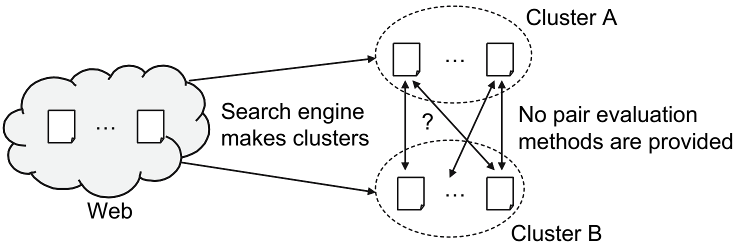

Fig. 2. Search engine with clustering

viewpoints. However, search engines with clustering do not provide us with information on how to choose pages from each cluster to make pertinent page-pairs. (See Figure 2.)

Summarization Using Multiple Documents. Summarization using multiple documents is used to summarize news [4, 5]. This approach prepares clusters from news articles, matches each sentence for each article in clusters, and makes a summary. It is important to detect the similar articles or sentences with this approach. Our goal was more challenging in that it was more important to detect the differences than the similarities. We attempted to detect the different viewpoints.

3 Multiple Viewpoint Retrieval for Page-Pairs

3.1 Model

We used a vector-space model to describe Web pages, page-pairs, and queries. Term Frequency Inverse Document Frequency (TFIDF)[6] word weight was used for the feature-vector. TF is the number of times words appeared in each document, and IDF of keyword kw was calculate as follows:

$$IDF(kw) = log\frac{N}{df(kw)} + 1 \tag{1}$$

N is the number of searched results and $df(k)$ is the number of documents with keyword "kw". IDF scores were calculated from collections of search results and also page-pairs. (p_1, p_2) denotes page-pair consisting of pages p_1 and p_2. Even if $p_1 \neq p_2$, $(p_1, p_2) = (p_2, p_1)$. In the feature-vector of page-pairs, the TF values are the summation of the TF values of p_1 and p_2, and the IDF values are calculated from all of page-pairs.

The feature-vector of query v_q is :

$$v_q = (v_q^{(1)}, v_q^{(2)}, \cdots, v_q^{(n)}) \tag{2}$$

$$v_q^{(i)} = \begin{cases} 1 \text{ if term } t_i \text{ in query} \\ 0 \text{ if otherwise.} \end{cases}$$

3.2 Feature Values

We defined three feature values to analyze characteristics of page-pairs:

– *Inter-page similarity* : $sim(v_{p_1}, v_{p_2})$,
– *Page-pair relevance* : $sim(v_{(p_1,p_2)}, v_q)$, and
– *Page relevance* : $sim(v_p, v_q)$

We adopted a cosine correlation value for similarity. Similarity in feature vector v_1 and v_2 was calculated as:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|} \tag{3}$$

Inter-page similarity indicates how much duplication there is between pages composing page-pairs. Page-pair relevance indicates how pertinent a page-pair is for given query. Higher values are best. We adopted it as evalua tion value for page-pairs. Figure 3 shows the relationship between the feature vectors of pages p1 and p2 (denoted as v_{p1}, v_{p2}), and page-pair (p1,p2) (denoted as $v_{(p1,p2)}$). v_q is feature vector of query q. Each bar corresponds to each element of a feature vector. Three bars from the left-most one correspond to the keyword included in query (k1,k2,k3). If $sim(v_{(p1,p2)}, v_q)$ has a high value, the following conditions are required:

– The values of elements, which corresponds to a query, complement each other in $v_{(p1,p2)}$ and reach a high value.
– The values of other elements are set off against each other in $v_{(p1,p2)}$ and stay low.

When pages are described from the different viewpoints, the above conditions are satisfied. (Duplication in query terms occurs many times, but occurs little in other terms.)

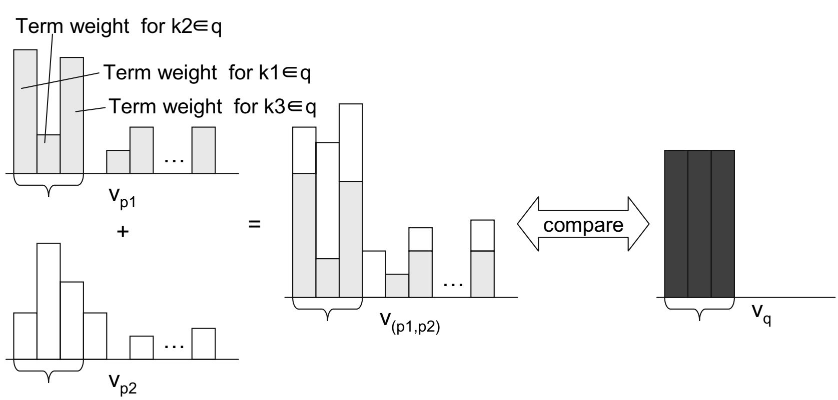

Fig. 3. Feature vector for pertinent page-pair

Page relevance indicates how pertinent a single page is for given query.

We defined valuable page-pair as page-pair which has higher page-pair relevance than page relevance of both pages consisting of it. In other words, valuable page-pairs satisfy following equation.

$$sim(v_{(v_{p1},v_{p2})}, v_q) > max(sim(v_{p1}, v_q), sim(v_{p2}, v_q)) \quad (4)$$

The valuable page-pair is more pertinent than the single pages which compose it.

4 Analysis for Page-Pairs

We analyzed characteristics of page-pairs. We first analyzed the relationship between page-pair relevance and page relevance. We then analyzed the relationship between page-pair relevance and inter-page similarity. We also analyzed the relationship between page-pair relevance and Google's ranking. We used following four queries in Table 1. We obtained 100 URLs for each query by Google[1], and made page-pairs from the available pages.

Table 1. Queries used for the experiments

Query name	Query terms	# of page-pairs
Q_A	"wind power generation", "nuclear power generation"	4656
Q_B	"America", "Iraq"	4753
Q_C	"Nobunaga Oda", "Mitsuhide Akechi" (They were Japanese feudal warlords in the 16th century.)	4656
Q_D	"Hong Kong", "gourmet"	4095

4.1 Page-Pair Relevance and Page Relevance

We analyzed the relationships between page-pair relevance and page relevance. Table 2 lists the number of page-pairs and valuable page-pairs. 30–50% of page-pairs are valuable page-pairs. It also lists the maximum of page-pair relevance and page relevance. In all the cases, the maximum of page-pair relevance is higher than the maximum of page-relevance.

In Figure 4, valuable page-pairs in the case of query Q_A are plotted on the graph, where the horizontal axis corresponds to higher page relevance and the

Table 2. The numbers of valuable page-pairs

Query name	# of valuable page-pairs	# of page-pairs	Max. of page-pair relevance	Max. of page relevance
Q_A	1599	4656	0.631579	0.624543
Q_B	2042	4753	0.428426	0.378591
Q_C	2102	4656	0.467308	0.436177
Q_D	2002	4095	0.443854	0.407625

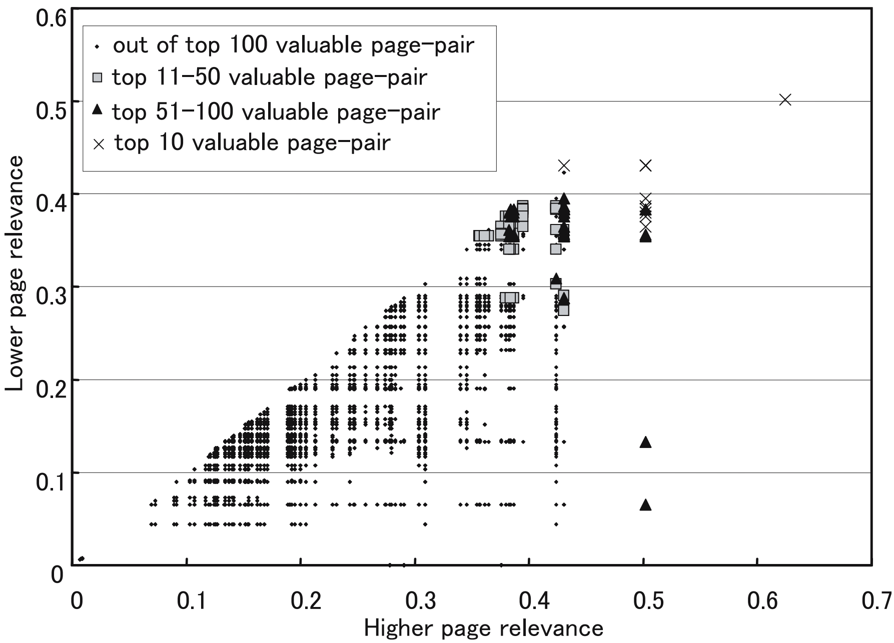

Fig. 4. The distribution of valuable page-pairs

vertical axis corresponds to lower page relevance for each page-pair. In this graph, the shapes of points are classified by the page-pair relevance ranking of valuable page-pairs. We found that many of highly-ranked valuable page-pairs appear in upper-right corner of the graph. It should be noted that some valuable page-pairs, having a page whose page relevance is low, has a high rank score of page-pair relevance. It means that there are the pages which have low page relevance but are valuable as the members of page-pair.

4.2 Page-Pair Relevance and Inter-page Similarity

We analyzed the relationship between page-pair relevance and inter-page similarity. Figure 5 shows the relationship between page-pair relevance and inter-page similarity in the case of query Q_A. Each point in the graph corresponds to a valuable page-pair or other page-pair, where the horizontal axis corresponds to the page-pair relevance and the vertical axis corresponds to the inter-page similarity.

Page-pair A and B in Figure 5 are valuable page-pairs and have the same page-pair relevance. Their inter-page similarity values are different. Page-pair A has a high inter-page similarity, and page-pair B has a low inter-page similarity. The both pages which compose Page-pair A describe about electric power circumstance, including both of “wind power generation” and “nuclear power generation”. On the other hand, page-pair B consists of the page which mainly describes “wind power generation” and the other which mainly describes “nuclear power generation”.

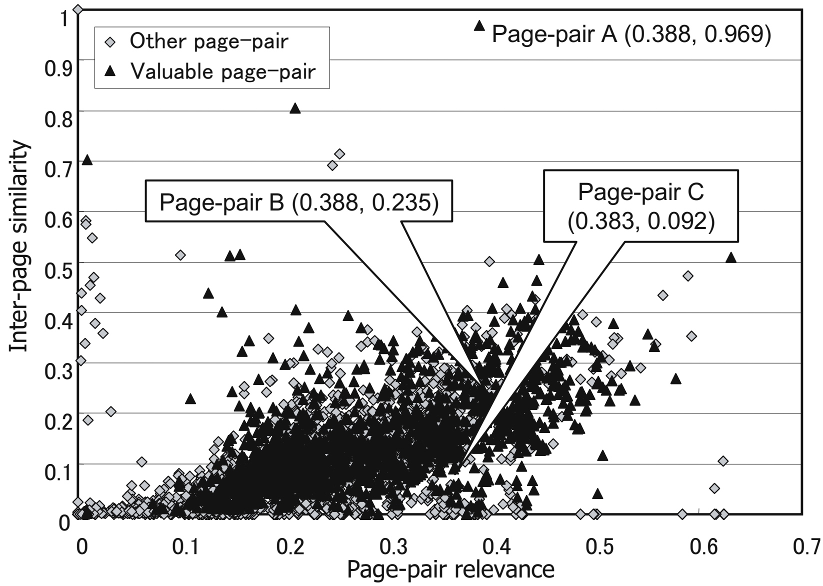

Fig. 5. The relationship between page-pair relevance and inter-page similarity

Page-pair C in Figure 5 is also a valuable page-pair but has very low inter-page similarity. It includes the pages which are much larger than the other. In such page-pairs, the characteristics of smaller pages are ignored. They are regarded as 'noise'.

Considering this, we can say,

- When inter-page similarity is too high, two pages are described from the same viewpoints, and
- When inter-page similarity is too low, page-pair depends on only one page.

Therefore, pages are regarded to be described from different viewpoints when the inter-page similarity satisfies the following for appropriate thresholds θ_1 and θ_2,

$$\theta_1 < sim(v_{p_1}, v_{p_2}) < \theta_2 \tag{5}$$

4.3 Page-Pair Relevance and Google's Ranking

We analyzed the relationship between page-pair relevance and Google's ranking of pages composing page-pairs. We classified page-pairs into four groups, i.e.,

- Group A : Page-pairs composed by the pages in the top 20 for Google's ranking.
- Group B : Page-pairs composed by the pages in the top 50 for Google's ranking, which are not in group A.
- Group C : Page-pairs composed by the pages in the top 100 for Google's ranking and which were not in groups A or B or D.

Table 3. Distribution of top 10 and 50 pertinent page-pairs

Query name		# of page-pairs			
		Group A	Group B	Group C	Group D
Q_A	top 10	0	3	7	0
	top 50	3	26	21	0
	all	190	1035	2350	1081
Q_B	top 10	2	1	6	1
	top 50	6	12	24	8
	all	190	1035	2400	1128
Q_C	top 10	2	4	4	0
	top 50	19	13	18	0
	all	190	1035	2350	1081
Q_D	top 10	0	8	2	0
	top 50	2	28	20	0
	all	190	1035	2050	820

– Group D : Page-pairs composed by the pages from the top 50 to 100 for Google's ranking.

We prepared page-pairs from 100 search results by using several queries and ranked them with their page-pair relevance. Table 3 lists the distribution of the top 10 page-pairs and the top 50 of pertinent page-pairs. As a result, we found that:

1. At least about 60% of top ranking page-pairs were in groups B and C,
2. At most only about 40% were in gourp A, and
3. There were very few in group D.

When we browsed Web pages with Google's ranking, we noticed page-pairs in group A. However, there are few good pertinent page-pairs in group A. Considering 1 and 3 above, most good pertinent page-pairs consists of pages with a high and a low Google's ranking. When we browsed Web pages with Google's ranking, such page-pairs were difficult to find. Therefore, our approach was better than browsing Web pages with Google's ranking.

5 Conclusions

We proposed the new concept, multiple viewpoint retrieval and explained our simple approach to achieve it. We analyzed the characteristics of page-pairs. We found that

– There are the pages which have low page relevance but are valuable as the members of page-pair.
– Page-pairs consisting of the pages which are described from different viewpoints has a high page-pair relavance and a low inter-page.

– Pertinent page-pairs are difficult to find by browsing with Google's ranking but multiple viewpoint retrieval can find them easily.

Future work is as follows:

– The development of the algorithm for finding pertinent page-pairs quickly, and
– The extensions to the link-analysis method and arbitary collection of pages.

Acknowledgements

This work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology under a Grant-in-Aid for Software Technologies for Search and Integration across Heterogeneous-Media Archives, and the Informatics Research Center for Development of Knowledge Society Infrastructure (COE program by Japan's Ministry of Education, Culture, Sports, Science and Technology).

References

1. Google, http://www.google.com/.
2. D. R. Cutting, J. O. Pedersen, D. Karger and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.318–329, 1992.
3. Clusty the Clustering Engine, http://clusty.com/.
4. NewsInEssence, http://lada.si.umich.edu:8080/clair/nie1/nie.cgi .
5. Columbia NewsBlaster, http://www1.cs.columbia.edu/nlp/newsblaster/ .
6. G. Salton. Developments in automatic text retrieval. Science, (253):pp.974–979, 1991.