Retrieving Multimedia Web Objects Based on PageRank Algorithm

Christopher C. Yang
Department of Systems Engineering
and Engineering Management
The Chinese University of Hong Kong
(+852) 2609 8239
yang@se.cuhk.edu.hk

K. Y. Chan
Department of Systems Engineering
and Engineering Management
The Chinese University of Hong Kong

ABSTRACT
Hyperlink analysis has been widely investigated to support the retrieval of Web documents in Internet search engines. It has been proven that the hyperlink analysis significantly improves the relevance of the search results and these techniques have been adopted in many commercial search engines, e.g. Google. However, hyperlink analysis is mostly utilized in the ranking mechanism of Web pages only but not including other multimedia objects, such as images and video. In this project, we propose a modified Multimedia PageRank algorithm to support the searching of multimedia objects in the Web.

Categories & Subject Descriptors:
H.3.3 [**Information Systems**]: Information Search and Retrieval – search process, retrieval models

General Terms: Algorithms, Experimentation

Keywords: Hyperlink analysis, multimedia retrieval, Web search engines, HITS, PageRank, content based retrieval.

1. INTRODUCTION
Information retrieval is a multidisciplinary research area that involves computer scientists, information scientists, library scientists and information system scientists. Before the advent of the Web, research in information retrieval is primarily focused on the indexing, retrieval, and categorization of text and multimedia objects. Individual text document and multimedia object is treated independently although some work in library science study the citation of documents. Given the advance of hyperlinks and markup languages in the Web, the retrieval of information will no longer consider the indexed features of the multimedia objects alone but also the hyperlinks that represent the relationships among the linked objects. In this project, we develop a modified Multimedia PageRank algorithm for multimedia retrieval in the Web that make uses of the hyperlink and embedded links among the multimedia objects available in the Web.

1.1 Content Based Retrieval
Content based retrieval is the most popular techniques for multimedia retrieval, especially in content based image retrieval (CBIR) [4]. In content based retrieval, low level features, such as color, texture, shape, motion, etc. are extracted from each multimedia object, and then relevant objects are retrieved based on the similarity of the features. However, feature extraction and similarity measurement among a large number of multimedia objects are usually a time consuming process.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

1.2 Hyperlink Analysis
Hyperlink analysis has been studied in the last ten years to support the Web search engines, in particular, in the ranking of the retrieved Web pages [1]. The prominent techniques are HITS [2] and PageRank [3], developed by Kleinberg and Page, respectively. A hyperlink from Web page A to Web page B may corresponds to a recommendation of Web page B by the author of Web page A or the same topic being presented in both Web Page A and Web page B. The hyperlink analysis supports the evaluation on relevance of pages to a searching query or a topic.

In PageRank algorithm, the PageRank value is computed by weighting each hyperlink to the Web page proportionally to the quality of the Web page containing the hyperlink. The PageRank values are computed recursively with arbitrary initial values.

In HITS algorithm, Web pages are categorized into authorities and hubs. Authorities provide good content on the topic but hubs are like directory type of Web pages. Hub and authority scores are computed for each Web page recursively.

Both PageRank and HITS and most of the other modified algorithms are applied on crawling and ranking of Web pages. However, the Web contains multimedia objects, such as images, videos, and audio, in addition to Web pages in HTML format. Although there are search engines for searching multimedia objects separately, hyperlink analysis has seldom been adopted. Yang et al. [5] has developed a system, named Octopus, that applied hyperlink analysis for multimedia retrieval on the Web. However, the direction of the hyperlink and whether a multimedia object is embedded in a Web page are not considered in order to make the analysis simpler.

2. Hyperlink Analysis for Retrieving Multimedia Objects
In this project, we propose to a new algorithm to adopt the hyperlink analysis for retrieving multimedia Web objects, including Web pages, images, and video. A directed graph representing the hyperlinks among the multimedia objects is first constructed as illustrated in Figure 1. A solid directed arc represents the hyperlink and a dash directed arc represents the embedding relationship. If the undirected graph in Octopus [5] is used, the graph as presented in Figure 2 (a) will be used instead.

The PageRank algorithm is then modified in order to rank the multimedia objects in our directed graph. The PageRank algorithm computes the PageRank values recursively based on the hyperlinks among Web pages. Our modified Multimedia PageRank algorithm computes the modified Multimedia PageRank values based on the

hyperlinks and embedded links among multimedia objects and the weightings for different types of links.

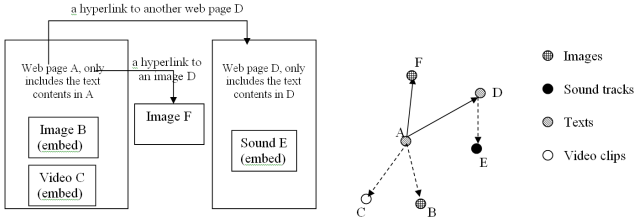

Figure 1 Directed graph representing the hyperlinks among multimedia objects.

(a) (b)

Figure 2 (a) Undirected graph representing the hyperlinks among multimedia objects in Octopus, (b) Directed graph representing the hyperlinks among Web pages using the PageRank algorithm.

Given the directed graph representing the hyperlinks among the Web pages (Figure 2 b), the formulation in Page Rank for computing the PageRank values is

$$R(u) = (1-d)\sum_{v \in B_u} \frac{R(v)}{N_v} + dE(u)$$

where *u, v* are Web pages
B_u is a set of web pages that point to *u*
N_u is the number of links from *u*
d is a constant
R(u): rank of web page *u*
E(u) is a population containing web pages corresponding to a source of rank with uniform probability distribution

In our modified Multimedia PageRank algorithm, given the directed graph representing the hyperlinks among the multimedia Web objects, the formulation for computing the Multimedia PageRank values is

$$R(u) = (1-d)\sum_{v \in B_u} \frac{w_{v,u}}{\sum_{a \in A_v} w_{v,a}} R(v) + dE(u)$$

where *u, v, a* are multimedia objects
B_u is a set of multimedia objects that point to *u*
A_v is a set of multimedia objects that pointed by *v*
N_u is the number of links from *u*
d is a constant
R(u) is the Multimedia PageRank value of the multimedia object *u*
E(u) is a population containing multimedia objects corresponding to a source of rank with uniform probability distribution
$w_{v,u}$ is the weight of the link from object *v* to object *u*.

3. Experiment

We have conducted an experiment to test the performance of the modified Multimedia PageRank algorithm in terms of the precision of retrieved multimedia objects and compared with the original PageRank algorithm in terms of the precision of retrieved Web pages. Table 1 presents the result of the experiment.

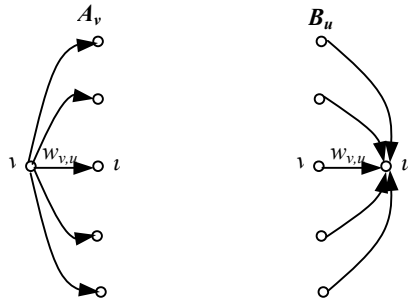

Figure 3 $w_{v,u}$ is the weighting of the hyperlink from Object *v* to Object *u*, A_v is the set of objects linked from Object *v*, and B_u is the set of objects linked to Object *u*.

Table 1 Experimental results

	Multimedia PageRank	Page Rank
Web pages	0.868	0.842
Images	0.633	
Videos	0.980	

A T-test has been conducted and it is found that the modified Multimedia Page Rank is significantly better than the original Page Rank in retrieving Web pages at the significant level of 0.1.

4. Conclusion

We proposed a modified Multimedia PageRank algorithm for retrieving multimedia objects in the Web. The hyperlinks and embedded links among multimedia objects are analyzed based on the PageRank algorithm. It is found that the modified Multimedia PageRank algorithm produce satisfactory performance in retrieving images and videos. In retrieving Web pages, the Multimedia PageRank algorithm also has significantly better performance than the original PageRank algorithm.

5. REFERENCES

[1] Monika R. Henzinger, Hyperlink Analysis for the Web, *IEEE Internet Computing*, January-February, 2001, pp.45-50.

[2] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Envoirnoment. *Journal of the ACM*, vol. 46, no. 5, September, 1999, pp.604-632.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Library Technologies Working Paper*, Standford University, Palo Alto, California, US, 1998.

[4] C. C. Yang, Content Based Image Retrieval: A Comparison between Query by Example and Image Browsing Map Approaches. *Journal of Information Science*, vol.30, no.3, 2004, pp.257-270.

[5] J. Yang, Q. Li, and Y. Zhuang, Octopus: Aggressive Search of Multi-Modality Data Using Multifaceted Knowledge Base. In *Proceedings of the International Conference World Wide Web Conference (WWW2002)*, Hawaii, May, 2002.