Internet Multimedia Information Retrieval based on Link Analysis

by

Chan Ka Yan

Supervised by

Prof. Christopher C. Yang

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of Master of Philosophy

in

Division of Systems Engineering and

Engineering Management

August 2004

Acknowledgement

I wish to gratefully acknowledge the major contribution made to this paper by Prof. Christopher C. Yang, my supervisor, for providing me with the idea to initiate the project as well as guiding me through the whole process; Prof. Wei Lam and Prof. Jeffrey X. Yu, my internal markers, and my external marker, for giving me invaluable advice during the oral examination to improve the project.

I would also like to thank my classmates and friends, Tony in particular, for their help and encouragement shown throughout the production of this paper. Finally, I have to thank my family members for their patience and consideration and for doing my share of the domestic chores while I worked on this paper.

Abstract

Ever since the invention of Internet World Wide Web (WWW), which can be regarded as an electronic library storing billions of information sets with different types of media, enhancing the efficiency in searching on WWW has been becoming the major challenge in the Internet world while different web search engines are the tools for obtaining the necessary materials through various information retrieval algorithms.

The current web search engines usually process queries according to user-given keywords or samples and try to satisfy the users' needs base exclusively on the analysis of the words in web pages. In most cases, this not only allows users to receive the information they need in only one medium, but also ignores some potentially useful reference in retrieving the web pages like the linkage between web pages.

Recently, some scientists proposed the use of hyperlinks in web pages to grab and rank different web pages in the WWW. If there exists a hyperlink between two web pages, it is most likely that they have similarities in their contents, or one of the web pages is the detailed description of another page. Moreover, using hyperlinks for information retrieval can gather information in different media, which means that we can retrieve multimedia information in a single query.

The objective in this research project is to retrieve multimedia information through link analysis. We compare the performance of Page Rank Algorithms with hypertext-induced topic selection (HITS) Algorithm in searching multimedia information in order to introduce a modified algorithm for multimedia searching.

We take into account several factors which may affect the ranking of multimedia links for the modification of the Page Rank and HITS Algorithm. Comparisons among the modified algorithms and the traditional ones will be provided for evaluation of our proposed ranking algorithms.

概要

自互聯網面世及普及化後，提高於萬維網中搜尋資訊之效率便成爲其中一個主要的互聯網科技研究範疇。萬維網可媲美一個儲存了上千萬各式各樣媒體資料的電子圖書館，而各種的網路搜尋器就是利用不同的資訊檢索計算方法爲用戶提供所需資訊。

現存的互聯網搜尋器一般都利用使用者的輸入的關鍵字或範例進行檢索。它們大都只是利用網頁內容中的文字內容來分析網頁事否符合用戶需要，但這不但令用戶每次檢索只能查詢一個媒介的資料，同時亦完全忽略了網頁中其他潛在並有利於檢索的參考資料，比如網頁間之超連結關係。

近年有學者提出利用網頁中的超連結來搜集萬維網網頁及爲它們排序。在某層面上，如果兩個網頁間有一條超連結把兩者聯繫起來，那兩者的內容必然是有共通點，或者其中一個網頁是另一個的詳細敘述。再者，利用超連結進行資訊檢索可獲取不同媒體的資料，從而令用戶在同一次檢索中能取得多種媒體資訊.

本研究習作的目的爲利用網頁中各類連結分析來獲取多種媒體資料。我們會比較網頁排名法(Page Rank Algorithm)及超連結引發主題選取法(HIT

Algorithm)在多媒體資料搜尋上的異同，以提供一種改良的多媒體資訊搜尋運算法。

我們會考慮各種可能會影響網頁排名法及超連結引發主題選取法爲多媒體連結排名的因素，而我們亦會比較改良後的運算法改良與傳統運算法，加以闡釋兩者各方面分別，讓讀者評估我們提出的改良法之優劣。

Table of Content

List of Figure

Chapter 2

Chapter 3

Chapter 4

List of Table

Chapter 3

Chapter 4

Appendix

Chapter 1. Introduction

1.1 Background

The WWW contains an enormous amount of information. There are currently over billions of Web pages in the WWW, which continues to grow in a phenomenal rate. It is undoubtedly difficult to locate the required information with high quality and relevancy in such a large corpus. To make it even complicated, the non-unifying structure, variability in authoring style and content render the harvest of the useful information impossible just with traditional techniques for database management and information retrieval.

In the past years, content-based search engines have been one of the primary tools in information retrieval inside the WWW. They usually have a large database which store and index the majority portion of WWW, and build giant indices for all web pages containing certain keywords. Whenever a query keyword is entered, the search engine returns with a list of web pages with the given query term.

As all the resulting web pages are ranked in advance, the searching process only takes several seconds. But there are certain drawbacks for the index-based search engines. One of them occurs when a broad query topic is used as search keyword – thousands to millions web pages with the keywords will be returned to users including those unrelated to the query. This seriously affects the quality and

relevance of the searching results. So, how can the search engine select the most representative and descriptive ones?

Some scholars, like Kleinberg[3], Page[2], etc., proposed that other than the text content of the web pages, we can also use the hyperlink structures in the pages for crawling and ranking the web pages so as to distill and downsize the result set for the users' query. This is known as the hyperlink analysis in information retrieval. Though researchers, such as Arsu et al.[28], have done much comparison in different searching strategies, there is still no conclusion for which searching methodology can be out-performed than others in all aspects.

1.2 Importance of hyperlink analysis

Hyperlink analysis algorithms allow search engines to deliver focused results to user queries. A hyperlink from page A to page B is usually a recommendation of page B by the author of page A [1]. On the other hand, if page A and page B are connected by a hyperlink, they may be on the same topic [1].

Before we discuss the details of different link analysis algorithms, we should first understand some definitions of keywords which are commonly used in describing the hyperlink analysis.

- Authority: measuring the number of objects pointing to an object, calculated by HITS
- Hub: measuring number of objects pointed by an object, calculated by

HITS

Remarks: Hubs and Authorities are mutually reinforcing relationship, which means for better Hub pages, they will always point to good Authority pages; for better Authority pages, they will always pointed by good Hub pages.

- Rank: scores of a page calculated by the Page Rank algorithm
- out-degree: number of pages a page links to
- in-degree: number of pages have links to a page
- out-links: links the page point out
- in-links: links point to the page

There are several applications in hyperlink analysis: crawling, ranking pages, mirrored hosts, web page categorization, and geographical scope [1]. In the following sections, we will only cover crawling and ranking, which are the two most vital topics in our research approach later on.

Apart from the link relationships widely used in hyperlink analysis, the anchor text, the link types, position of the link, etc., can also be used in the hyperlink analysis. Our research project also focused on investigating how these extra pieces of information affect the existing hyperlink analysis in ranking web pages.

Chapter 2. Related Work

As we stated in the previous chapter, information retrieval on WWW can be processed by content-based analysis and link-based analysis. Content-based analysis grabs and ranks web pages by considering the relevance of the contents of the HTML web pages to user's query. These contents include the text body of the pages, the title of the pages, the page keywords in the meta data of the pages, etc. Link-based analysis often ranks the web pages by the number of links connected to them. In most of the link-based algorithms, a rule is followed – the more the number of links pointing to a web page, the higher its rank. In this chapter, we will focus on describing various link-based analyses in information retrieval.

2.1 Crawling

For any search engine, there are two necessary functions: how to grab the set of relevant data specified by user's query and how to rank the data grabbed. Crawling is the process of collecting web pages.

2.1.1 Crawling method for HITS Algorithm

Kleinberg proposed that an ideal collection of web pages relevant to user's query should follow the 3 properties below [3]:

1 The set should be relatively small

2 The set should be rich in relevant pages

3 The set should contain most (or many) of the strongest Authorities

Most scientists followed these three properties when building their own searching algorithms so that less computer resources and time would be used. When they started to do crawling, they tried to grab as fewer pages as possible while these pages should be predicted to have higher relevance to the user's query.

Firstly, they collected t highest ranked pages for query from a text-based search engine as the Root Set of the resultant web page collection. The Root Set should satisfy the 3 properties of ideal collection of web pages. Then they expanded the Root Set by crawling the hyperlinks in the web pages inside the Root Set (Fig. 2.1).

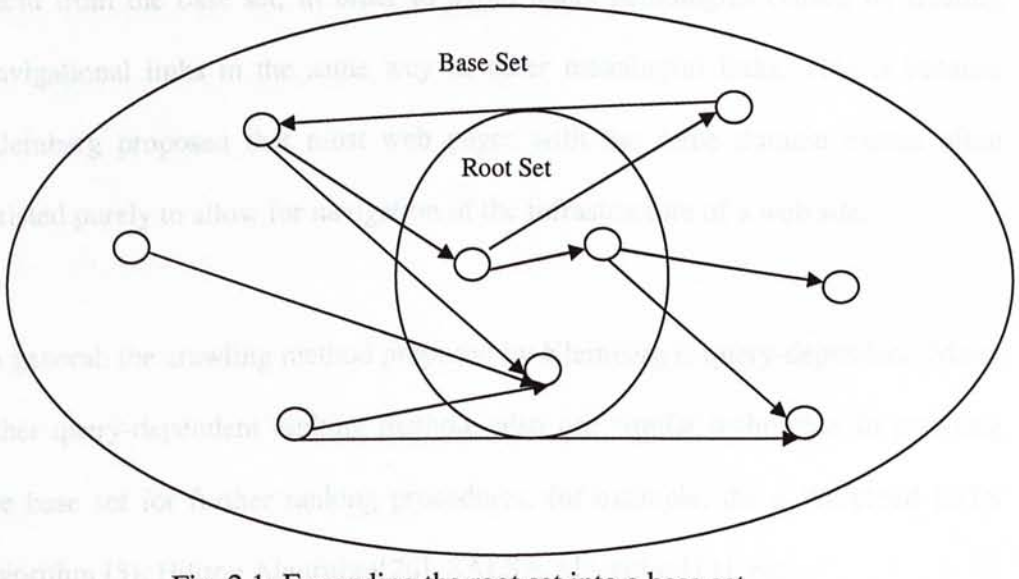

Fig. 2.1: Expanding the root set into a base set

The process repeated until a certain number of web pages were collected or no

more new pages could be added in the set. Except for restricting the number of web pages collected, the number of layers in crawling, that is, the web page distances, is commonly another constraint in gathering web pages. The expanded Root Set was called a Base Set, which was used for further ranking of the pages. The Base Set should also satisfy the above 3 properties. After building the Base Set, a linked network was formed and various ranking algorithms could be used for ordering the web pages.

The crawling techniques introduced by Kleinberg are usually used in the HITS Algorithm[1, 3, 5, 8, 10, 11, 13, 16, 18, 22, 27, 29]. In order to limit the size of base set, Kleinberg[3] suggested to restrict the number of in-links and include all the out-links of the root set links in the base set. Besides, he suggested using the domain name to decide which links are purely for navigational use and remove them from the base set, in order to avoid many pathologies caused by treating navigational links in the same way as other meaningful links. This is because Kleinberg proposed that most web pages with the same domain names often existed purely to allow for navigation of the infrastructure of a web site.

In general, the crawling method proposed by Kleinberg is query-dependent. Many other query-dependent ranking methods also use similar techniques in crawling the base set for further ranking procedures, for example, the a Weighted HITS algorithm [5], Hilltop Algorithm[26], SALSA Algorithm[11], etc.

2.1.2 Crawling method for Page Rank Algorithm

For Page Rank Algorithm[1, 2, 6, 7, 8, 12, 13, 14, 15, 18, 24, 27, 30, 31], as it considers the whole WWW as the base set of the ranking, the crawling strategy is to grab as many links as possible through the crawlers. Therefore, not much distillation rules are imposed on finding the base set links for carrying out Page Rank.

We can take the crawling system of the Google search engine [15, 17] as an example. In order to scale to hundreds of millions of web pages, Google has a fast distributed crawling system. A single URLserver serves list of URLs to a number of crawlers. Each crawler maintains its own DNS cache so that it needs not do a DNS lookup before crawling each document. The list of URLs can be obtained from the out-links of each web pages crawled or provided by some web page service provider and the author of the web pages, etc. Then, the URLs crawled are cached and indexed for further analysis.

2.2 Ranking

After collecting a set of relevant web pages, we arrange these pages in order and return them to the user. Ranking is the process of arranging the returned web pages in descending order of relevance. There are several common ranking methods, namely the Page Rank Algorithm [1, 2, 6, 7, 8, 12, 13, 14, 15, 18, 24, 27, 30, 31], the HITS Algorithm [1, 3, 5, 8, 10, 11, 13, 16, 18, 22, 27, 29], the Page Rank-HITS Algorithm [13], the Stochastic Approach for Link-Structure

Analysis (SALSA) [11], the Average and Sim Algorithm [8], the Netscape Approach [9] and the Cocitation Approach [9].

2.2.1 Page Rank Algorithm

Page Rank is a method for computing a ranking for every web page based on the graph of the web proposed by Brin and Page [2, 15]. It is a query-independent scheme which assigns ranking scores independent of a given query. The web pages which contain the query keywords will be extracted and then ranked based on the pre-calculated page rank scores. The collection of web pages can therefore be very large and not all web pages are relevant to user's query. A random surfer model [2, 15] is used in the calculation of Page Rank. This model suggests that user usually performs a random walk on the web graph. The random surfer simply keeps clicking on successive web page links at random. However, the surfer periodically get bored and jump to a random page chosen based on the distribution of $E(u)$, a population of web page corresponding to a source of rank.

In general, a highly linked page is more important than a page with few links. Page Rank provides a more complicated method than the citation counting. A web page is ranked higher if the sum of the ranks of its backlinks is high, which states that it has many backlinks or it has a few highly ranked backlinks. Backlinks are parent links of a web page pointing to it. A simple example of Page Rank algorithm is shown in Fig. 2.2.

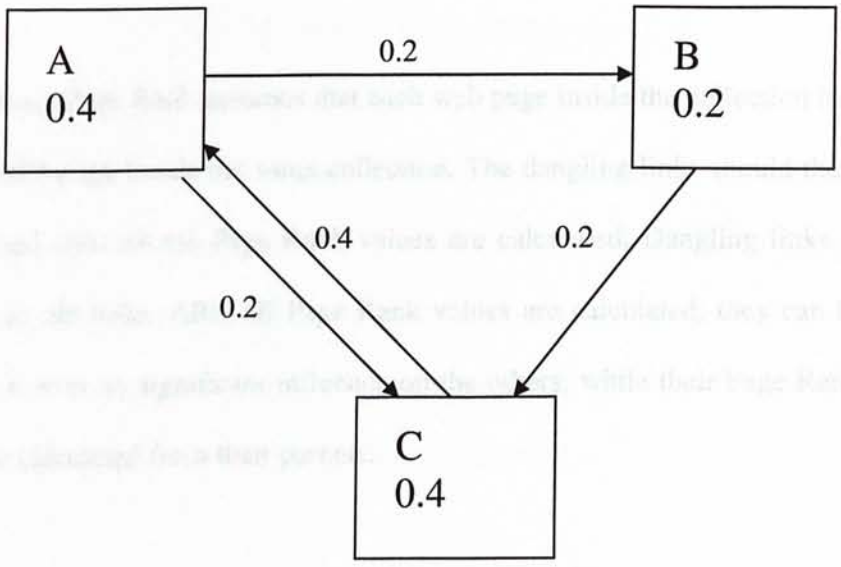

Fig. 2.2: Simplified Page Rank Calculation

This is the general formulation of Page Rank:

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

such that d is maximized and $\|R\|_1 = 1$ ($\|R\|_1$ denotes the L_1 norm of R)

where u: a web page in the collected web pages

B_u: set of web pages that point to u

N_u: number of links from u

d: probability of user does not jump to a page linked from the current page, but jumps to a random sample in a population of web pages $E(u)$ with certain probability distribution, $0<d<1$

$1-d$: probability of user jumps uniformly at random to one of the pages linked from the current page

$R(u)$: rank of web page u

$E(u)$: a population contains web pages corresponding to a source of rank

with uniform probability distribution

However, Page Rank assumes that each web page inside the collection has at least one child page inside the same collection. The dangling links should therefore be removed until all the Page Rank values are calculated. Dangling links are links with no out-links. After all Page Rank values are calculated, they can be added back in with no significant influence on the others, while their Page Rank values can be calculated from their parents.

Many scientists have made modifications to the Page Rank Algorithm. Brin and Page [15] suggested including the random surfer model to the Page Rank Algorithm. At each step, with some probability, the surfer teleports to a completely random web page, independent of out-links of the current page. Another scientist, Haveliwala, proposed a topic-sensitive Page Rank algorithm [6] to rank pages by their importance scores (ranking scores) as well as the classified topic of user query. Haveliwala has introduced several proposals for efficient computation of Page Rank [24], too. Diligenti, Gori and Maggini also formulated two algorithms (Focused Page Rank [13] and Double Focused Page Rank [13]) to compute a relative ranking of web pages when focusing on a specific topic. Ding, He, et al. [20] have proposed the implementation of mutually reinforcing calculations in the Page Rank like the Out-link normalized Rank, In-link normalized Rank and the Symmetric normalized rank.

One of the most famous applications of Page Rank Algorithm is the Google

Search Engine [15, 31] found by Brin and Page. Yet, Google's ranking system is not a purely link-based algorithm. It factors a web page in two rankings: the text score on account of the keyword hits in text content and the Page Rank score. A hit list for the web page is used for storing the counts of keyword hits in the content of the web page, the positions and fonts of the keyword in that page, etc. The information in the hit list affects the text score of that particular web page. Thus, we can say that Google's ranking algorithms is actually based on both content-based analysis and the link-based analysis.

2.2.2 HITS Algorithm

Kleinberg suggested the HITS Algorithm [3] to identify the most central web pages for broad search topics in the context of the WWW as a whole. HITS is a query-dependent ranking method. Unlike Page Rank, the initial collection of web pages for applying HITS algorithm is relatively small and more relevant to user's query. To perform the HITS algorithm, we should first collect a root set from the resultant set of a text-based search engine, and then expand it to a base set as stated before.

Then, we can compute the Hubs and Authorities by the following 2 equations iteratively:

$$x^{<p>} = \sum_{q:(q,p)\in E} y^{<q>} \quad \text{and} \quad y^{<q>} = \sum_{p:(q,p)\in E} x^{<p>}$$

where p, q: web pages in base set

$x^{<p>}$: Authority value of web page p

$y^{<q>}$: Hub value of web page p

(q, p): there exists a directed hyperlink from web page q to web page p

E: the set of directed hyperlinks in the base set collection

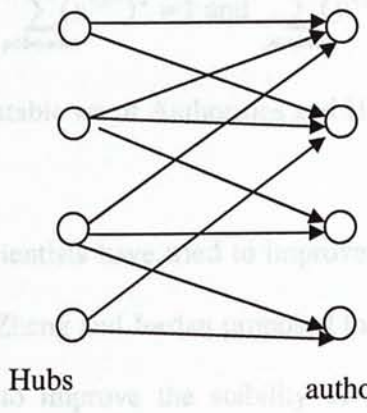

Fig. 2.3: An illustration of Hubs and authorities

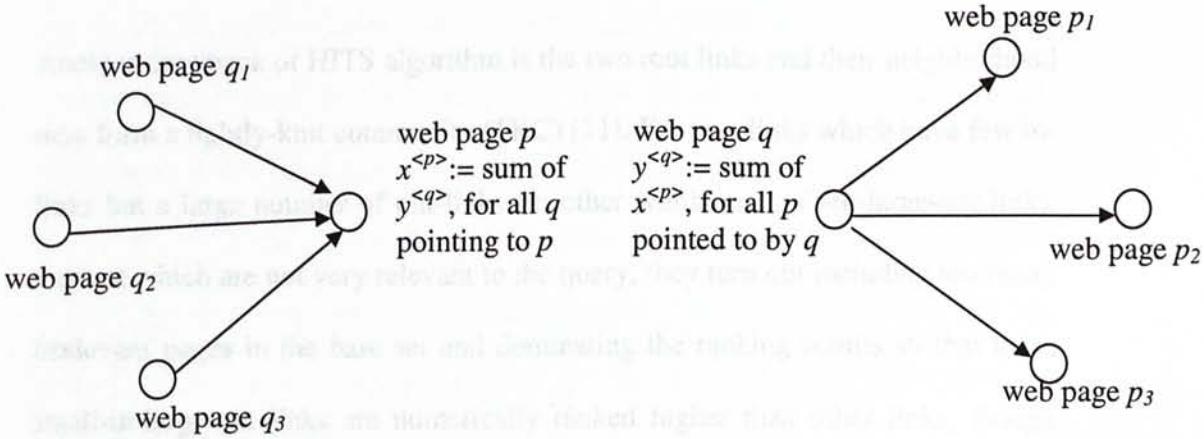

Fig. 2.4: The basic operation of HITS

Fig. 2.3 illustrates the meaning of Hub and Authority and Fig. 2.4 shows the basic

operation of HITS algorithm. First, the value of Authority for each web page is calculated, which is the I Operation. Based on the Authority calculated, we then calculated the Hub value, and this is the O operation. Note that the Authority values and the Hub values should be normalized in each iteration so that their squares sum to 1, i.e., $\sum_{p \in baseSet} (x^{<p>})^2 = 1$ and $\sum_{p \in baseSet} (y^{<p>})^2 = 1$. Several iterations are performed to obtain a stable set of Authorities and Hubs.

In past years, many scientists have tried to improve the performance of HITS by various methods. Ng, Zheng and Jordan proposed the Randomized HITS [14] and Subspace HITS [14] to improve the stability of HITS algorithm. Bharat and Henzinger [16] alleviated the mutually reinforcing relationships between hosts problem and topic drift problem by adding weights to web pages. The Companion algorithm [9] of Dean and Henzinger is derived from Kleinberg's HITS algorithm, which exploits not only links but also links' order on a web page.

Another drawback of HITS algorithm is the two root links and their neighborhood may form a tightly-knit community (TKC) [11]. For root links which have few in-links but a large number of out-links (in other words, a small-in-large-out link), most of which are not very relevant to the query, they turn out including too many irrelevant pages in the base set and dominating the ranking results so that these small-in-large-out links are numerically ranked higher than other links, though they in fact may not be more relevant to the query than other links. Usually, the average in-degree and out-degree of a root link are much smaller than the out-degree of a small-in-large-out link. Li, Shang and Zhang suggested a Weighted

HITS algorithm [5] to improve the TKC effect. By adding weights in both in- and out-links of a small-in-large-out link, the Hub and Authority values of this links become less dominate so that the result of HITS becomes more reliable. Li, Shang and Zhang also proposed combining HITS algorithms with some relevance scoring methods [5], for instance, Vector Space Model (VSM), Okapi Similarity Measurement (Okapi), Cover Density Ranking (CDR) and Three-Level Scoring Method (TLS).

Above all, Hub and Authority values can combine with other link properties such as the host so as to enhance the accuracy, as suggested by Bharat and Milhaila in their Hiltop Algorithm[26]. Gibson and Kleinberg have also demonstrated the use of anchor text in improving the HITS Algorithm [25]. Similar to anchor text, Yang [19] has also illustrated the possibility of fusion of the text-based retrieval method and the Authorities. Human factor can also be one of the possible components to be inserted in the HITS. Chen, Tao, et al. [23] have tried to make use of the DirectHit system, ranking of which are based on the click popularity and stickiness, with the Authority and Hub values in order to return more representative searching results to users. Other than different external factors, the results of HITS can be worked with other ranking models, too. The Hub-Averaging-Kleinberg Algorithm and the Bayesian Algorithm of Borodin, Roberts, et al. [17] are the combination of HITS and SALSA algorithms, and HITS and Bayesian statistical model respectively to reduce the drawback of pure HITS Algorithm.

2.2.3 PageRank-HITS Algorithm

There are many differences between Page Rank and HITS algorithm. Page Rank algorithm is a query-independent approach for ranking, while HITS algorithm is a query-dependent approach. Therefore, the query processing time for HITS is much longer than Page Rank as the ranking scores of Page Rank are pre-computed once only before the user input the query. In reverse, the Hub and the Authority values in HITS algorithm can only be calculated after the user's query.

Page Rank algorithm is more stable than HITS algorithm. Page Rank has a well-defined behavior because of its probabilistic interpretation and it can be applied to large collections without canceling the influence of the smallest web communities. HITS, on the other hand, magnifies the effect of the largest web communities, which restricts the HITS algorithm to be applied on large web page collection. Nevertheless, Page Rank is sometimes simplifying the complex relationships of web page citations, which is weaker than HITS in capturing the relationships among web pages.

To combine the advantage of Page Rank and HITS Algorithm in ranking web pages, Diligenti, Gori and Maggini introduced a Page Rank-HITS model [13]. They employed two surfers in the new model, each implementing a bi-directional Page Rank surfer. Surfer 1 either followed a back-link or jumped to a random page whereas surfer 2 either followed a forward link or jumped to a random page. Considering the interaction between the surfers in a matrix like in HITS, the ranking scores were calculated.

2.2.4 SALSA Algorithm

Similar to HITS Algorithm, SALSA Algorithms [11] also preserves the measure of Hub and Authority as the indicator of ranking relevant pages for user query. Considering a bipartite graph G, whose two parts correspond to Hubs and Authorities, an edge between Hub r (r_h) and Authority s (s_a) means that there is an informative link from page r to page s. Authorities and Hubs pertaining to the dominant topic of the sites in G should then be highly visible (reachable) from many sites in G. Thus, Moran and Lempel suggested identifying these sites by examining certain random sites more frequently than others, less connected sites.

SALSA is based upon the theory of Markov chains and relies on the stochastic properties of random walks performed on the collection of sites. It follows a meta-algorithm which is a version of the spectral filtering method and differs from Kleinberg's HITS Algorithm in which the association matrices are defined.

The meta-algorithm used for building up the association matrices is stated here. Given a topic t, construct a site collection C which should contain many t-Hubs and t-Authorities, but should not contain many Hubs or Authorities for any other topic t'. Let $n=|C|$. Then, deriving two $n \times n$ association matrices — a Hub matrix H and an Authority matrix A from C and the link structure induced by it. Association matrices are widely used in classification algorithms, and are used in the SALSA Algorithm in order to classify the web pages into communities of Hubs or Authorities.

SALSA combines the theory of random walks with the notion of the two distinct types of web pages, Hubs and Authorities, and in fact, analyzes two different Markov chains: A chain of Hubs and a chain of Authorities. It performs a random walk by alternately (a) going uniformly to one of the pages which links to the current page, and (b) going uniformly to one of the pages linked to by the current page. Unlike traditional random walks, state transitions in these chains are generated by transversing one forward link and one backward link in a row. A Hub score and an Authority score are obtained then from these chains.

The following two formulas are the stochastic matrices which are the transition matrices of the two Markov chains:

(a) The Hub-matrix $\tilde{H}$:

$$\tilde{h}_{i,j} = \sum_{\{k|(i_h,k_a),(j_h,k_a)\in\tilde{G}\}} \frac{1}{\deg(i_h)} \bullet \frac{1}{\deg(k_a)}$$

(b) The Authority-matrix $\tilde{A}$:

$$\tilde{a}_{i,j} = \sum_{\{k|(k_h,i_a),(k_h,j_a)\in\tilde{G}\}} \frac{1}{\deg(i_a)} \bullet \frac{1}{\deg(k_h)}$$

where $\tilde{a}_{i,j}$: a positive transition probability in the Authority-matrix used for calculating the Authority value, implies that a certain page k points to both pages i and j, and hence page j is reachable from page i by two steps: retracting along the link $k \rightarrow i$ and than following the link $k \rightarrow j$

$\tilde{h}_{i,j}$: a positive transition probability in the Hub-matrix used for calculating the Hub value, implies that a certain page k is pointed to by both pages

i and j, and hence page j is reachable from page i by two steps: retracting along the link $k \rightarrow i$ and than following the link $k \rightarrow j$

i_h: the Hub i in the bipartite undirected graph

i_a: the Authority i in the bipartite undirected graph

$deg(i_h)$: out-degree of Hub i

$deg(i_a)$: in-degree of Authority i

(i_h , j_a): an edge between Hub i (i_h) and Authority j (j_a) means that there is an informative link from page i to page j

The SALSA Algorithms, in some sense, can reduce the problem of TKC effect caused in the HITS Algorithm as it also includes the stochastic random walk model in the calculations. Moran and Lempel [11] has proved that SALSA Algorithm is less vulnerable to the TKC effect and can find meaningful Authorities in collections where Kleinberg's THIS Algorithm fails to retrieve.

2.2.5 Average and Sim

The idea of Average and Sim [8, 27] was proposed by Gevrey and Ruger in combining the similarity measures obtained by text-based search engine with linkage analysis. They thought that HITS and Page Rank algorithm made some loss in part of the information obtained via text analysis from the text-based search engine. Therefore, they tried to reuse the similarity measures obtained by the search engine along with link analysis. Both Average and Sim represented the Authority value of a page p. Average equaled to the average over similarity measures of all incoming links q, while Sim equaled to the sum of the similarity

measure of page p and the average over all similarity measures of incoming links q.

The following are the formulation of Average and Sim:

$$\text{Average: } authority(p) = \frac{1}{|\{q \mid q \rightarrow p\}|} \sum_{q \rightarrow p} similarity(q)$$

$$\text{Sim: } authority(p) = similarity(p) + \frac{1}{|\{q \mid q \rightarrow p\}|} \sum_{q \rightarrow p} similarity(q)$$

where *authority(p)*: the authority value of web page p

similarity(p): the similarity value of web page q calculated by text-based search engine

$q \rightarrow p$: web page q has a link to web page p

$|\{q|q \rightarrow p\}|$: number of web pages point to web page p

2.2.6 Netscape Approach

Netscape introduced a "What's Related?" feature in the version 4.06 of the Netscape Communicator browser [9]. This approach used the connectivity information, usage information and content analysis to determine the relationships. The details of this approach can be found in the What's Related FAQ page of Netscape.

2.2.7 Cocitation Approach

Cocitation Approach [9] was suggested by Dean and Henzinger to examine the siblings of a starting node u in the web. Two nodes are cocited if they have a

common parent. The number of common parents of two nodes is their degree of cocitation. The Cocitation approach ranks the web pages by the degree of cocitation in descending order.

On account that the majority of researchers have carried out many sounded reports in the HITS Algorithm and the Page Rank Algorithm, we may be able to gain more ideas and information from their experiences as well as the more mature techniques in reducing the side-effects during experiments. Therefore, in our research project, we only investigate into the application of the Page Rank algorithm and the HITS algorithm in multimedia information retrieval.

2.3 Multimedia Information Retrieval

When user search for image, sound track, or video clips, they can type their query in the search engines for image, audio or video databases and get the necessary information they need from the databases. Generally, all these databases use content-based analysis to index all the multimedia objects in advance. When the user initiates a query, the search engine returns the pre-ranked results to the users. This, however, restricts the users to retrieve the objects they desire from only the specific database rather than the whole WWW. It is obvious that no multimedia database can gather all the media objects in its own database. Therefore, the search engine will be unable to reach some of the useful objects not included in the database.

Besides, these search engines usually pay less attention to the relationship

between the web pages containing these multimedia objects and the objects' ranking, which can significantly affect the degree of relevancy from user's point of view. For instance, users tend to consider a picture of Mickey Mouse in a well-known homepage like Disney or IMDB more important than a similar one in a personal homepage.

The convenience in searching multimedia information is another concern for users. Most of the search engines can only report results in one medium for each user query. It will be convenient to users for enquiring in any one medium and receiving result in several media.

Yet, many scientists have developed various hyperlink analysis algorithms on web page searching. However, from my understanding, these algorithms have never been performed on searching objects other than text. Indeed, it should be more useful for users to input queries in only one medium to obtain relevant objects from different media in an instance. Yang, Li and Zhuang suggested an Octopus [4] system for multimedia information retrieval through user feedbacks, link analysis and object contents.

2.3.1 Octopus

The main idea of Octopus[4] is to divide the whole searching process into three different layers based on a Layered Graph Model (LGM). These 3 layers include the user layer, the structure layer and the content layer. All objects in these three layers, whether they are in the form of texts, image, etc., are called media objects.

Therefore, the relevance of two media objects can be evaluated by the following three aspects:

1. User's interpretation of the two media objects deduced by the user's interactions
2. Structural relationships between the two media objects (hyperlinks)
3. Similarity in the contents of the two media objects from the lower-level features

There is a diagram showing the structure of LGM in Fig. 2.5. The LGM stores knowledge as the links between media objects. The information retrieval can be restricted in a relatively small locality connected via links instead of in a whole database. The search space is therefore effectively reduced and more complicated algorithms can be applied on each layer.

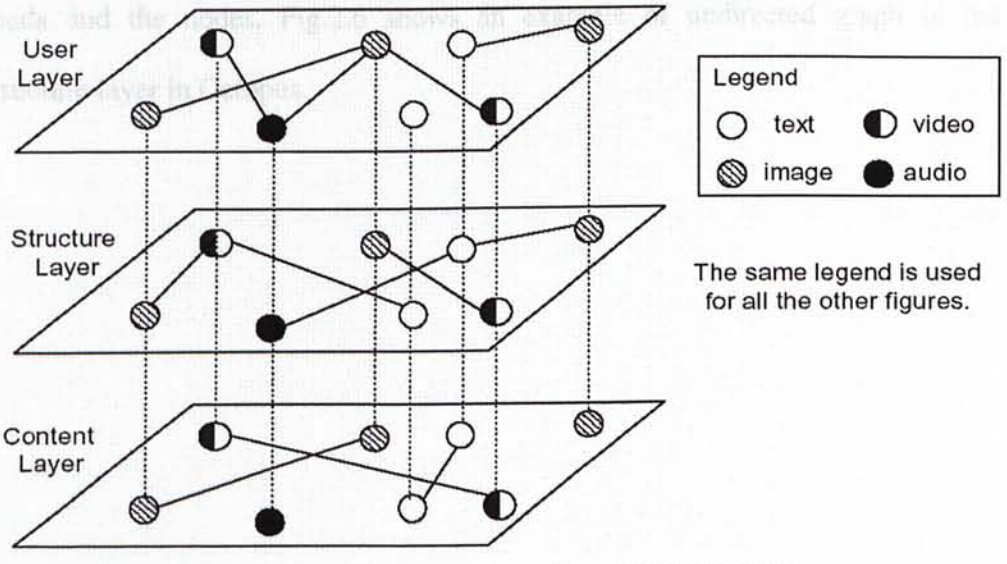

Fig. 2.5: The layered graph model (LGM) [4]

Moreover, we usually can only find similarities between objects in the same

media in the content layer. This means that if we need to retrieve multimedia objects, we also have to consider the structure layer and the user layer. Although the user layer is the most reliable measurement in ranking, it is less objective and requires user's feedback. Thus, in our research, we will only focus on the information retrieval algorithms on the structure layer.

Mechanism for link analysis of Octopus in the structure layer is to consider the connectivity of all media objects in an undirected graph for which each of its nodes represents one media object, without considering their object type and whether they are embedded in other media objects. If there is a link from one media object I to another media object II, there should be links connecting object I to object II and all the media objects inside object II. The undirected links are used due to consistency with links in the user layer and content layer in the LGM. Besides, all links are of the same weight, regardless of the distances between the seeds and the nodes. Fig.2.6 shows an example of undirected graph in the structure layer in Octopus.

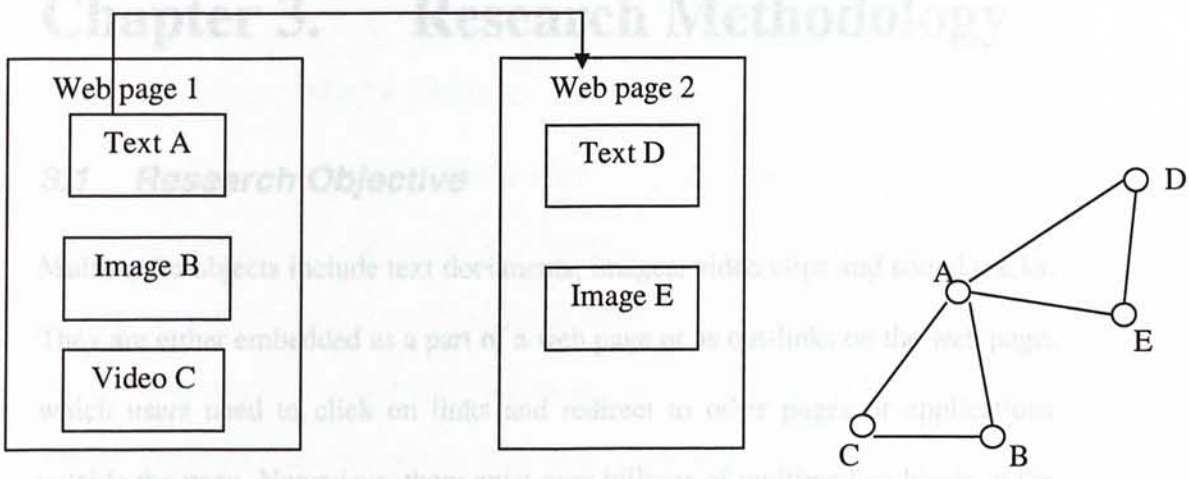

Fig. 2.6: Undirected graph showing connectivity between web pages

There are some obvious advantages in using link analysis for multimedia information retrieval. For example, hyperlinks connect objects with different media together to form the Base Set for ranking. Besides, we can use the distance between 2 media objects to calculate the relevance of them. Moreover, we can also use wordings in the filename and anchor text of hyperlinks to enhance the quality of object ranking.

In the following sections, we are going to discuss how we apply the above observations into our research work.

Chapter 3. Research Methodology

3.1 Research Objective

Multimedia objects include text documents, images, video clips and sound tracks. They are either embedded as a part of a web page or as out-links on the web page, which users need to click on links and redirect to other pages or applications outside the page. Nowadays, there exist over billions of multimedia objects in the whole WWW. The demand on Internet multimedia information retrieval techniques is therefore gradually increased.

As we discussed in above sections, most Internet link analysis algorithms can be applied to web pages or text documents only. In our project, we attempt to find out how well they perform in handling multimedia object links. We examine the Page Rank algorithm and the HITS algorithm only since they are the most typical ways in analyzing hyperlinks and plenty of improvements on these two algorithms have been announced by other scientists in the past.

Apart from the link analysis algorithms, we would like to identify factors affecting the ranking of media objects and their degrees of influence on Page Rank and HITS. We can then put weights on the formulation of Page Rank and HITS to represent the effects of these factors in a more concrete manner. Our target is to obtain the most suitable formulation for searching multimedia object through the link analysis.

3.2 Proposed Crawling Methodology

3.2.1 Collecting Media Objects

To begin with, the primary task of our project is to crawl the media objects' links necessary for ranking afterwards. Our approach is similar to the one used by Kleinberg's HITS algorithms [3] and our root set and base set follow the three principles of ideal collection [3]. We first send a textual query to Google, a well-known text-based search engine. The first ten web page links returned from Google are the seed components, and also the root set, of our media objects collection.

The second step is expanding the root set to a base set which our proposed ranking algorithms work on. We try to explore the HTML tag of the web pages and extract the out-links. If an out-link is not inside the root set or base set, we will add it into the base set. At the same time, we store for each object link the in-link and out-link relationships, the object type, at which layer does this link obtain from its parents and whether it is embedded or not in its parents, etc. We use the term "parents" to represent the in-links of an object link and "children" as the out-links of an object link, while the word "layer" refers to the distance between an object link and the root set object links.

We define an object A is embedded in another object B base on the following criteria:

(a) For text, the frames pages inside the frameset are embedded by the

frameset HTML page

(b) For images, if they can be viewed inside their parent objects, they are embedded in their parent object

(c) For videos, if they are broadcasted inside their parent objects, they are embedded in their parent object

(d) For audios, if they are heard by users inside their parent objects, they are embedded in their parent object

The expanding process repeats for all object links until the total number of object links in the base set equals to a certain amount. We restrict the number of the media object links so as to keep the size of the collection smaller and reduce the computational time and resources. We do not set the maximum number of layers so that more relevant media objects and more relationships between objects can be included in the base set.

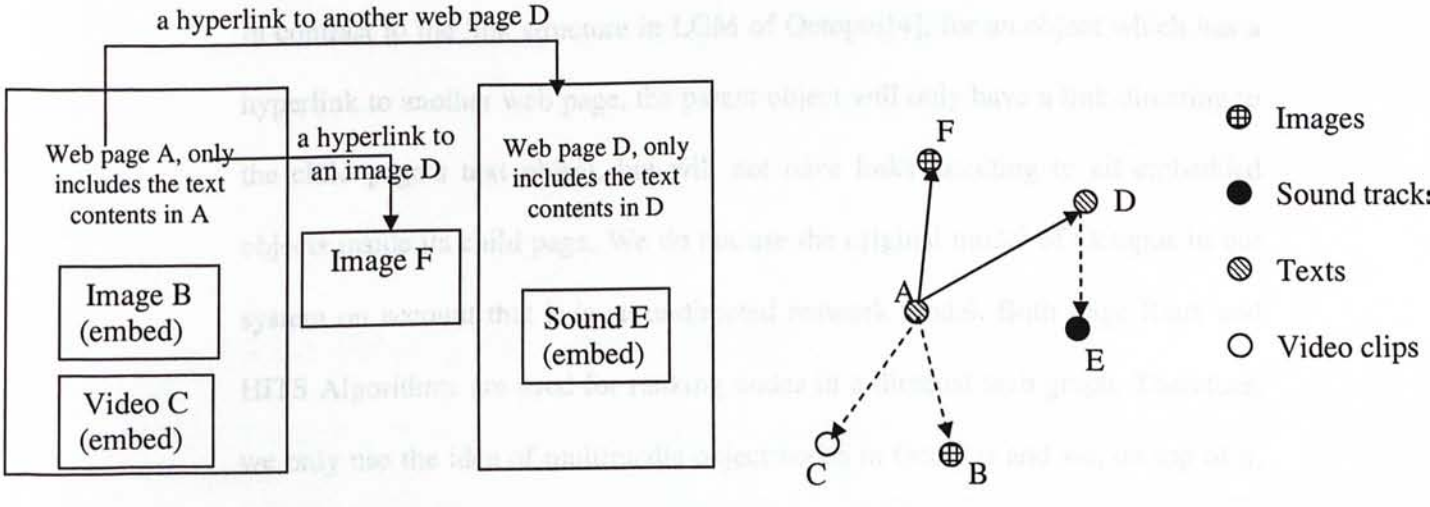

Fig. 3.1: Directed graph representing the connectivity relationships between media objects in our research

An example describing the directed graph formed by the media objects is shown in Fig. 3.1. We treat a web page and all the text contents in the web page as one whole text object, which means that text object *A* represents only the text contents of the web page. There is an image and a video clip inside the web page *A*, we define that all the objects other than text inside another object are embedded in another object. Image object *B* and video object *C* are therefore said as embedded links in text object *A*. The embedded links from object *A* to objects inside the same web page as *A* are drawn in dot lines in the figure. There are two hyperlinks in text object *A*, one is linked to an external image object *F*, the other one is linked to another web page, text object *D*. The hyperlinks from text object *A* linked to objects outside *A* are constructed in solid lines. There is a sound track embedded in text object *D*. Therefore, an embedded link is drawn from text object *D* to sound object *E*.

In contrast to the link structure in LGM of Octopus[4], for an object which has a hyperlink to another web page, the parent object will only have a link directing to the child page's text object, but will not have links directing to all embedded objects inside its child page. We do not use the original model of Octopus in our system on account that it is an undirected network model. Both Page Rank and HITS Algorithms are used for ranking nodes in a directed web graph. Therefore, we only use the idea of multimedia object nodes in Octopus and we, on top of it, have defined a new directed network structure to connect object nodes and some node attributes for storing the possible factors affecting the ranking of nodes,

which will be more suitable for adapting the Modified Page Rank and HITS algorithm later on. These factors will be discussed in the later sections.

Our system does not handle the links in various scripts as it is time-consuming and inefficient to extract links in the script programming. To be frank, the links in script are comparatively less important as much of them are advertisement links or fancy interface.

The time spent to build up and run the crawling system takes almost half of the whole research period. You may argue that why we need to build up our own system for crawling. Many current search engines, such as Google, Yahoo, and Lycos provide the services for finding the links which link to a particular page. However, these search engines could only help find URL links (single medium). Apart from single medium in-links, we also need to consider different factors that can potentially affect the ranking objects. By using our own crawling system, we find it easier to store necessary information, such as the embedded state of the children links, for further use. Therefore, we have tried to build up our own system for collecting the object links.

3.2.2 Filtering the collection of links

Inevitably, there should be some links for advertising or navigational purposes. In order to distill away these meaningless links from our base set, we have imposed several filtering rules in adding object links into the children set and parent set. These rules are usually based on the text, common words, repeated links, or

positions of the links.

3.2.2.1 HTML tag positions

For advertising links, they are usually in a certain position of a web page, for example, in the <Head> tag or at the end of the HTML file. With this, we can thus overlook the common positions for advertising links when we extract the object links out from a web page.

3.2.2.2 Keyword in the surrounding paragraph or table row

Surrounding letters or HTML coding are usually good descriptions of the details of an object link. Therefore, we use the surrounding words of an object link for filtration of less relevant links.

When an object link is inside a table in its parent object, we will check if there is the keyword searched in the same table row. In such case, we will include the object link in the base set. Otherwise, it will not be included in the base set. The checking can be done by detecting the <table> tag and the <tr> tag in the HTML codes.

Similarly, when an object is inside a paragraph or a list rather than a table in its parent object, we will check if there is the keyword searched between the previous two paragraphs and the following two paragraphs or the previous two list items and the following two list items before we add that object link into the base set.

3.2.2.3 Self-reference links

Self-reference links can be a source of steering in ranking. Self-reference means self-recommendation or self-citation of a media object. An object with many self-reference links can mislead the ranking algorithms using the in-link properties to deliver a high ranking to it, due to the reinforcing calculations of the object ranks in the algorithm. Therefore, we remove these self-reference links before we do further calculations in order to have a more objective and meaningful result.

3.2.2.4 Duplicated links

Duplicated links may attenuate the score of an object which should be originally rank higher during the ranking process, so that they may not be able to retrieve by the user. Therefore, all duplicated child links of the same parent link should be removed. Some links with alias, for example, www.yahoo.com and www.yahoo.com/index.html, are easily detected and thus can be removed in case of duplication. For some link alias, however, we can never notice that two object links are linking to the same content until these links are opened and read.

One of the examples is http://disney.go.com/disneyvideos/index.html and http://disneyvideos.disney.go.com/ . These two links link to the same web site, but their wordings in the link are much different. We will not treat these links as duplicates as it is a time-consuming process to find out all these kind of alias and at the same time bear a heavy load to the system.

3.2.2.5 Special links or words should be excluded

We can observe that some of the object links are purely for advertising and carry no meaning. For example, some graphics which indicate for update information, with the links named new.gif or new.jpeg, or some text object links with the anchor text "Terms and Policy" which commonly appear inside the text objects, we will remove them from our base set. This is because most of these links always appear in the page, no matter what the query terms are.

3.2.2.6 Maximum number of parents and children links

In avoidance of certain domain dominating the number of object links in the base set, children set and parent set, we try to restrict the number of parent links and the children links to not more than 30 and 75 respectively. Applying all the above filtering strategies, we can then effectively reduce the number of navigational links, advertising links and meaningless links in the base set without using the Kleinberg's Removal Strategy, which we will discuss in the next paragraph.

3.2.2.7 Kleinberg's Removal Strategy on Navigational links

Kleinberg suggested deleting all the intrinsic links from the base set [3]. In fact, all links in the world can be classified into transverse links and intrinsic links. Transverse links are links with different domain, while intrinsic links are links with same domain. In Kleinberg's idea, intrinsic links are generally for navigation of infrastructure of site only and it is less relevance with user's query. Therefore, Kleinberg proposed to remove them from the base set.

However, Kleinberg's removal strategy is not suitable in the crawling of

multimedia object links. Considering object links with same domain but different media, these objects are not for navigational use and they are actually useful and relevant to the users' query. As a result, we do not use Kleinberg's removal strategy in removing the navigational links.

Above all, not all the objects are filtered by the same restrictions. Certain flexibilities can be adjusted according to different kinds of object. For some objects with extremely few out-links, like some simple-design personal homepages, we will try to loose our filtering rules in order to include more possibly relevant links for the user's query. In contrast, for some objects with huge amount of out-links, like some message boards or pages like www.imdb.com or www.amazon.com, we will try to tighten our filtering rules to reduce the number of links that have higher chance to be irrelevant to user's query.

Take IMDB as an example. The "Page Flipper" feature in the IMDB web page is often appeared in all films' description page. The graphics involve in the Page Flipper feature is irrelevant to the film content themselves. Therefore, we can identify the characteristics of the HTML codes in making the Page Flipper and then try to remove the object links related to it. These can reduce the number of unrelated object links in the base set and increase the efficiency of information retrieval.

For those object links which are already in the base set, we will ignore the filtering rules and directly add them into the children set of their parents, unless

the maximum number of parent and children links are not reached.

3.3 Proposed Ranking Methodology

3.3.1 Identifying the factors affect ranking

Before we apply different ranking algorithms to the base set of media object links, we should first understand what components of the object links will potentially affect the ranking values of the links. Here we have identified seven factors which may influence the ranking results: object type, distance between objects, number of in-links (parents) and out-links (children), parents' and children's object type, whether the object is embedded or not embedded in parent object, the position of the hyperlink of the object in there parent object and finally the size of the object inside its parent object. We will describe these factors in details in the following paragraphs.

3.3.1.1 Object Type

We have defined 4 object types for our reference, text, image, video clips and sound tracks. There should be some differences in the importance and degree of relevance for different media. In general, people consider video clips most important and have the highest degree of relevance corresponding to user's query, especially when they are searching for some TV programmmes or films. Image come second while text documents come third, and the last one is sound tracks.

People conceptually receive graphical information better than text information,

this can be seen from our ascendants thousands of years ago. Therefore, image and video clips, which are generally using meaningful or representative symbols to describe different ideas, should be more relevant to users than text objects. Video clips are more informative than images as video is a combination of sound tracks and images. Sound track is less informative than the visual media objects providing that visual objects are usually more comprehensible to human beings than vocal objects. Other types of links, such as email links, newsgroup links and links to FTP sites are out of our consideration at this moment.

3.3.1.2 Distance between Objects

The distance between objects can be obtained from the number of layers of the objects relative to the root set objects. Intuitively, for two objects with longer distances, they are less relevant. We may think that the objects in layer three are the least relevant objects to users' query in most cases.

3.3.1.3 Number of In-links (Parents) and Out-links (Children)

These factors already exist in the formulation of Hubs, Authorities and Page Rank values. Number of parents of each object affects the number of in-links' Page Rank values to be summed up in the calculating process of Page Rank of an object and the number of Hub values to be added in that of HITS. The more the parents, the more the number of in-links' page rank value and Hub values to be added.

On the other hand, the number of out-links affects the Page Rank values for the in-links to distribute to their children objects and the number of Authority values

to be added in the HITS. The more the children, the smaller the Page Rank values each child received from its in-link and the more the number of Authority values to be added.

However, there may be some extra manipulation based on the number of in-links and out-links to reduce the drawback of Page Rank Algorithm and HITS, for instance, to reduce the TKC effect [11] by adding weights in the HITS algorithm.

3.3.1.4 Parents' and Children's Object Type

Similar to the reason for taking into account the object type of the media object itself, as the ranking scores of media objects' parents and children are iteratively used to calculate their own ranking scores, we should also consider the object type of the media object itself. There may be difference in importance between an object pointed to by an image and an objected pointed to by a text link. This information can be involved in the calculation of Page Rank and HIITS Algorithms for multimedia object ranking.

3.3.1.5 Embedded or Not Embedded

Whether an object is embedded in its parent object or user is redirected to other pages or application outside its parent object by clicking the hyperlink of the object inside its parent should also be considered in ranking media objects. Some people think that the object links not embedded in their parents should be ranked higher as those media objects are in some sense a detail description of their parent objects. In contrast, other people think that the object links embedded in their

parents should be ranked higher as objects inside the same web page must be more relevant to the page rather than the objects outside the web page.

3.3.1.6 Position of the Objects

The positions of media objects in their parent objects are also affecting its importance. Objects in the <Head> tag of HTML files are commonly less important than those inside the <Body> tag because media objects inside <Head> tag are mainly for advertising purpose. At the same time, links inside the <Script> tag are usually for advertisements or fancy interface. Our filtering rules can therefore make use of these properties to discard the potentially irrelevant links out from our Base Set.

3.3.1.7 Size of the Objects

The size of media objects inside their parent objects are also a good indicator of their importance – the larger the size of an object, the more important it is and therefore, it should be ranked higher.

3.3.2 Modified Ranking Algorithms

After crawling the base set of relevant media objects and identifying the factors potentially affecting the ranking of media objects, we have to modify and apply the ranking algorithms to the base set. We only focus on the Page Rank algorithm and HITS algorithm for further modifications because of the publicity and simplicity of their mathematical models. Many detailed studies have carried out for these algorithms and we can use more researchers' experience in defining the

new model for our multimedia searching purpose. Also, modifications of these algorithms are based on the fundamental model of Brin and Page [2]'s Page Rank and Kleinberg [3]'s HITS so as to demonstrate a clear difference in the effect of object model in these algorithms.

At this stage, we need to incorporate the factors we discussed in the above section to the Page Rank and the HITS algorithms. Our modifications mainly focus on adding a weight, representing the combined effects on the distance between objects and the root set, their parents' and children's object types, embedded states, position, size and object types of the objects themselves. Different combinations of these factors are used for calculating the Page Rank, Authority and Hub values. We are going to discuss in detail each modified algorithms in the following sections.

3.3.2.1 Modified Page Rank Algorithm

The original formulation of Page Rank is:

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

such that d is maximized and $\|R\|_1 = 1$ ($\|R\|_1$ denotes the L_1 norm of R)

where u: a web page in the collected web pages

B_u: set of web pages that point to u

N_u: number of links from u

d: probability of user not jumping to a page linked from the current page, but jumping to a random sample in a population of web pages $E(u)$ with

certain probability distribution, $0<d<1$

$1-d$: probability of user jumping uniformly at random to one of the pages linked from the current page

$R(u)$: rank of web page u

$E(u)$: a population containing web pages corresponding to a source of rank with uniform probability distribution

The definition of Page Rank value of a certain web page u is the summation of its parent's average page rank value per that particular parent's number of children. We can understand it in this way: a parent object is likely to distribute the same amount of its Page Rank value to its entire children objects. All the children objects are of the same weight of $\frac{1}{N_v}$ and the sum of all the weight of the parent link equals to one. This is an important guideline for our later modification. No matter how we set the weights initially, we must normalize the weights in the calculation of Page Rank value so that the total weight for each parent link is equal to one. Otherwise, the Page Rank value at the convergence will become zero or infinity for all objects.

Besides, the way of handling the root set objects and dangling links are important in the Page Rank Algorithm. The major assumption of the web graph of Page Rank is all objects in the world must have connection to each other, each link will have at least one in-link and one out-link. However, from our observation in the experiment, a large portion of object links is dangling links. And, the root set we gathered from Google does not promise to have their corresponding in-links

within the Base Set, on account of the limited size of our base set. In the original Page Rank Algorithm, Brin and Page assume that the algorithm is applied to a huge amount of web page in the whole WWW. The chance of web pages having no in-links is inevitably low or this even never happens. Therefore, they only mention how to handle dangling links or looping links, but not the in-link-absented root links.

We follow the way Brin and Page done towards dangling links. We remove the dangling links before we calculate the iterations in Page Rank Algorithm. After all the Page Rank values are calculated, they can be added in without affecting the Page Rank values of other links significantly. Several iterations are needed to carry out in order to calculate the Page Rank scores of these dangling links as well as normalization of all Page Rank values.

For the in-link-absented root links, we pretend them have self-references to themselves, which take 0.1 portion of their total weights corresponding to their children. You may doubt that the Page Rank value of the root object links become higher than what they should be, which is unfair to other object links. However, as they are the source of rank and origin of base set, it is reasonable to award slightly higher score for them. Moreover, from our observation in the experiment, we can always find that such little increase in the Rank value does not bring much impact on the ranking position of the return objects for the queries.

The following is the modified formula of Page Rank:

$$\text{Page Rank: } R(u) = (1-d)\sum_{v \in B_u} \frac{w_{v\_u}}{\sum_{a \in A_v} w_{v\_a}} R(v) + dE(u)$$

such that d is maximized and $\|R\|_1 = 1$ ($\|R\|_1$ denotes the L_1 norm of R)

where *u, v, a*: media object in the base set

B_u: set of media objects that point to *u*

A_v: set of media objects that pointed to by *v*

N_u: number of links from *u*

d: probability of user not jumping to a media object linked from the current object, but jumping to a random sample in a population of media objects *E(u)* with certain probability distribution, *0<d<1*

1-d: probability of user jumping uniformly at random to one of the media objects linked from the current object

R(u): rank of media object *u*

E(u): a population containing media objects corresponding to a source of rank with uniform probability distribution

w_{v_u}: combined weight of distance between objects, object types, embedded states, positions and sizes of the objects corresponding to media object *v*, $w_{v\_u} = w_{object\_type} \times w_{embedded} \times w_{position} \times w_{size} / w_{layer}$

where w_{object_type}: weight representing the object type corresponding to media object *u*

w_{layer}: weight representing the number of layers of object u relative to the root set objects when its parent is object *v*

$w_{embedded}$: weight showing whether the object *v* embeds the object *u*

$w_{position}$: weight representing the importance of position of object u in object v

w_{size}: weight reflecting the size of object u in object v

For any page rank vector R, where $R = \begin{bmatrix} R(u_1) \\ R(u_2) \\ \vdots \\ R(u_n) \end{bmatrix}$, the L_1 norm of X is the sum of the absolute value of all elements in X, i.e., $\|R\|_1 = \sum_{i=1}^{n} |R(u_i)|$, u_i stands for any media object i. As all the page rank values are positive, L_1 norm of R indicates that the sum of all page rank values must equals to 1.

Calculation of Page Rank value of a media object is the weighted sum its parents' Page Rank value, multiplied by the probability of the random surfer to choose the successive link of the parents and finally added by the chance of the surfer jumping to other media objects corresponding to the source of rank. We normalize the combined weight w_v so that the sum of all combined weight of each parent v of object u equals to one.

As what we have mentioned in previous sections, the distance of the object link, the size of the media object in its parent, the position where the media object located in its parent and whether the object link is embedded or not in its parent may have much effect on the Page Rank of an media object. In terms of the object type, our Modified Page Rank algorithm takes only the media object's own object

type but not it parents', because we think that the weight of object type should only affect the amount of share of Page Rank value the object media under calculation can obtain from its parent. And also, as we need to normalize the combined weight by the sum of combined weight of each parent, if we involve the weight of parent type in the combined weight, both the numerator and denominator will have the same factor—weight of parent type. Then the normalized combined weight will result in no net effect for the weight of parent type. Besides, the calculation of Page Rank is iterative. The effect of parent type of the media object on the Page Rank score will always appear in the later iteration. Therefore, we have decided not to involve the parent object type in the formulation of Modified Page Rank algorithm.

Fig. 3.2 describes the basic operation of Modified Page Rank Algorithm graphically. Objects v_1, v_2 and v_3 are the parent objects of object u, while objects p and q are the children of u. The Page Rank score of u is the product of object type weight of u and the weighted sum of Page Rank scores of its parents v_1, v_2 and v_3. The weight $w_{v_1\_u}$ is the combined weight of v_1 corresponding to u. It depends on the object type of u, embedded state, position and size of u in v_1 and layer of u with parent v_1. The combined weights $w_{v_2\_u}$ and $w_{v_3\_u}$ are calculated in the same way. The combined weight is normalized by sum of all the combined weights of the same parent. For example, in Fig. 3.2, both p and q have the parent u, the combined weights w_{u_p} is normalized by the sum of w_{u_p} and w_{u_q}.

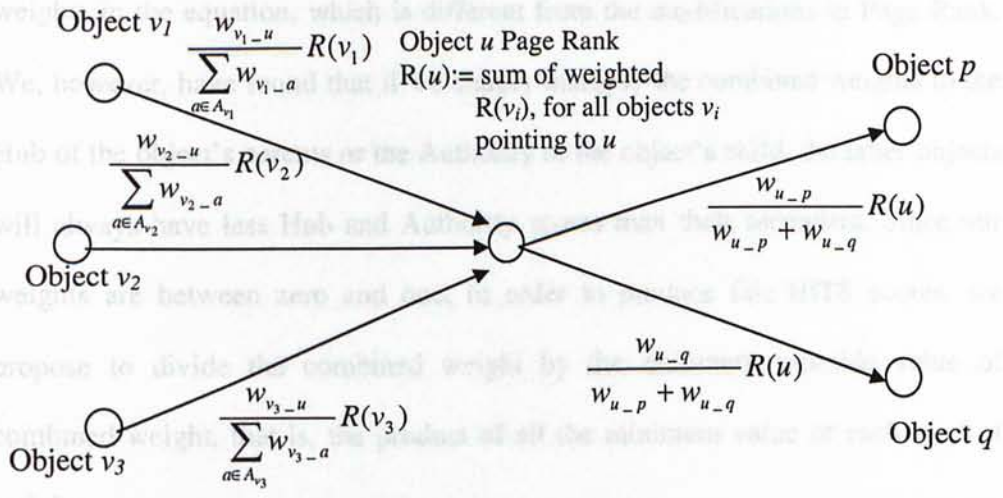

Fig. 3.2: Basic operation of Modified Page Rank calculation

3.3.2.2 Modified HITS Algorithm

Another algorithm we have modified is Kleinberg's HITS Algorithm [3]. The two fundamental equations for HITS are as follows:

$$x^{<p>} = \sum_{q:(q,p)\in E} y^{<q>} \quad \text{and} \quad y^{<q>} = \sum_{p:(q,p)\in E} x^{<p>}$$

where p, q: web pages in base set

$x^{<p>}$: Authority value of web page p

$y^{<q>}$: Hub value of web page p

(q, p): there exists a directed hyperlink from web page q to web page p

E: the set of directed hyperlinks in the base set collection

In contrast to Page Rank, the normalization process of HITS is done after calculating the Authority and Hub values. In the original algorithm, no weights are involved in the mathematical model. Therefore, when we insert combined weights into the equation of Authority and Hub, we do not need to normalize the

weights in the equation, which is different from the modifications in Page Rank. We, however, have found that if we simply multiply the combined weights to the Hub of the object's parents or the Authority of the object's child, the latter objects will always have less Hub and Authority scores than their ascenders. Since our weights are between zero and one, in order to produce fair HITS scores, we propose to divide the combined weight by the minimum possible value of combined weight, that is, the product of all the minimum value of each kind of weights.

Considering that our base set is built up based on the assumptions of Kleinberg's HITS algorithm [3], there is no need to make important amendment for the iterative process of scores on top of the procedures presented in the traditional HITS. Therefore, our main concern should be on how we can improve the formulation.

The following is the Modified Hub equations:

$$\text{Hub y: } y^{<v>} = \sum_{u:(v,u)\in E} \frac{w_{hub\_v\_u}}{\min\_value\_of\_combined\_weight} x^{<u>}$$

$$\text{Authority x: } x^{<u>} = \sum_{v:(v,u)\in E} \frac{w_{authority\_v\_u}}{\min\_value\_of\_combined\_weight} y^{<v>}$$

where u, v: media object in the base set

$x^{<u>}$: Authority value of media object u

$y^{<v>}$: Hub value of media object v

(v, u): there exists a directed hyperlink from media object v to media object u

E: the set of directed object links in the base set collection

$w_{Hub\_v\_u}$ (or $w_{authority\_v\_u}$): combined weight of distance between objects, object types, embedded states, positions and sizes of the objects corresponding to media object u (or v), $w_{hub\_v\_u}$ (or $w_{authority\_v\_u}$)= $w_{object\_type} \times w_{embedded} \times w_{position} \times w_{size} / w_{layer}$

for which w_{object_type}: weight representing the object type corresponding to media object u (or v)

w_{layer}: weight representing the number of layers of object u relative to the root set objects when its parent is object v

$w_{embedded}$: weight showing whether the object v embeds the object u

$w_{position}$: weight representing the importance of position of object u in object v

w_{size}: weight reflecting the size of object u in object v

and the Authority values and the Hub values should be normalized so that their squares sum to 1, i.e., $\sum_{v \in baseSet} (x^{<v>})^2 = 1$ and $\sum_{v \in baseSet} (y^{<v>})^2 = 1$

The calculation method of combined weights for Modified Hub is similar to Modified Page Rank. The combined weights in Hub and Authority also depend on the distance of the object link, the size of the media object in its parent, the position where the media object is located in its parent and whether the object link is embedded or not in its parent. The only difference is the weight of object type.

Traditional Hub score is the summation of total Authority of its children objects. When involving the object type weight in the combined weight, we intuitively think Hub value of a media object may be affected by the type of objects it points to. Therefore, we include the type weight of its children objects in the combined weight. We do not include the object's own type in calculating the Hub value because the principle of Hub is rather focused on how well an object cite to some meaningful and authoritative objects. Changing the resultant value of the sum of Authority is less relevant to this principle.

For the original Authority score, it equals to the summation of total Hub of its parent objects. Similar to the modification in Hub, we involve type weights of the object's parents as one of the factors in calculating the combined weight, so as to reflect the influence on the type of parent objects point to the media object. The purpose of Authority value of an object is rather focused on the representation of the objects giving citation to it. Therefore, the object's own object type is relatively less useful in the scoring of Authority and is not included in our modification.

Same as the steps as Kleinberg's HITS, the Modified Authorities are also calculated first. This is the I Operation (in-link operation) of HITS modeling. After the set of Modified Authorities is obtained, we use them to generate the Hub values for each object from our Modified Hub equation. This is the O operation (out-link operation) in the HITS calculation. Then we normalize the Authority and Hub value of each object so that the squared sum of all Authorities and that of all

Hubs equal to one. The process iterates several times so that stable values of Hub and Authority can be obtained. In our experiment, we have set the number of iteration to be 20.

Sample Diagrams illustrating the basic operation of Modified HITS Algorithm is shown in Fig. 3.3 and Fig. 3.4. In Fig. 3.3, objects q_1, q_2, and q_3 are the parent objects of object p. The Authority score of p is the sum of weighted Hub scores of q_1, q_2 and q_3. The weight $w_{authority\_q_1\_p}$ is the combined weight of q_1 corresponding to p. It is obtained from object type of q_1, embedded state, position and size of p in q_1 and layer of p with parent q_1, and the calculation is the same for $w_{authority\_q_2\_p}$ and $w_{authority\_q_3\_p}$.

Similarly, in Fig. 3.4, objects p_1, p_2, and p_3 are the children objects of object q. The Hub score of q is the sum of weighted Authority scores of $p1$, $p2$ and $p3$. The weight $w_{hub\_q\_p_1}$ is the combined weight of p_1 corresponding to q. It is obtained from object type of p_1, embedded state, position and size of p_1 in q and layer of p_1 with parent q, and that is similar for $w_{hub\_q\_p_2}$ and $w_{hub\_q\_p_3}$.

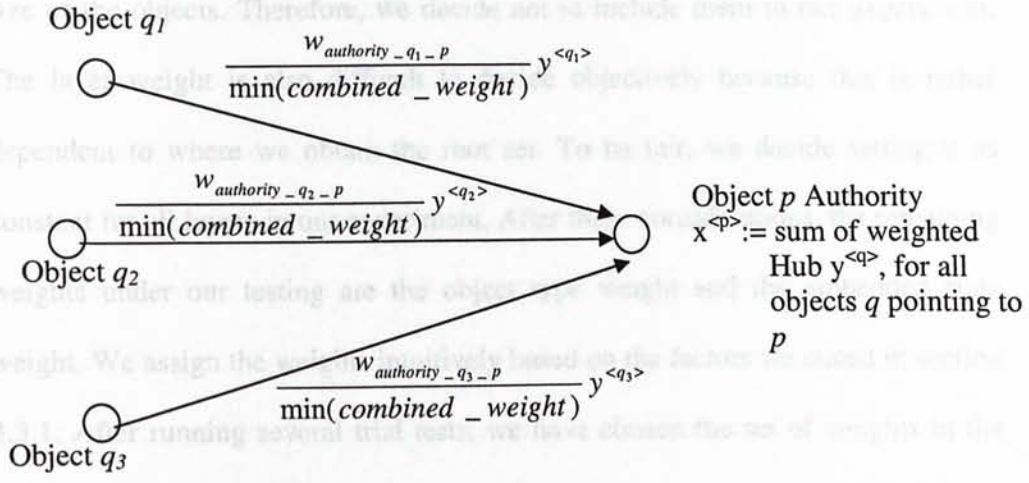

Fig. 3.3: Basic operation of Authority calculation

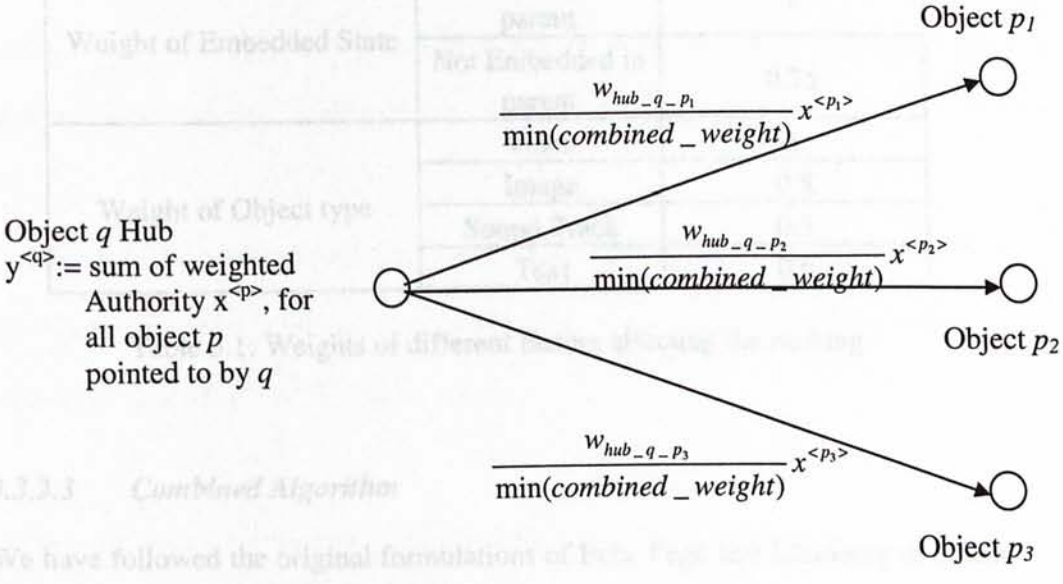

Fig. 3.4: Basic operation of Hub calculation

All the above modifications are the preliminary design. When we try to review different types of weights, we can find that some of the weights are related to the content-based analysis rather than pure link analysis, for example, the position and

size of the objects. Therefore, we decide not to include them in our experiment. The layer weight is also difficult to decide objectively because this is rather dependent to where we obtain the root set. To be fair, we decide setting it as constant for all layers in our experiment. After these considerations, the remaining weights under our testing are the object type weight and the embedded state weight. We assign the weights intuitively based on the factors we stated in section 3.3.1. After running several trial tests, we have chosen the set of weights in the Table 3.1.

Weight of Embedded State	Embedded in parent	1
	Not Embedded in parent	0.75
Weight of Object type	Video	1
	Image	0.8
	Sound Track	0.5
	Text	0.6

Table 3.1: Weights of different factors affecting the ranking

3.3.2.3 Combined Algorithm

We have followed the original formulations of Brin, Page and Kleinberg to obtain the traditional Page Rank, Authority and Hub values for our experiment. We have found that though the Page Rank and Authority are both ordering links, the resulting ranks of Page Rank and Authority to the same links are sometimes much different. Therefore, we also try to compute the combination of Page Rank scores and the Authority scores to obtain a new ranking indicator, the Combined Rank scores. There are many different kinds of methods which combine two ranking

methods together, such as the Similarity Merge method used by Yang [19] in combining the text-based and link-based retrieval methods, and the PageRank-HITS Algorithm proposed by Diligenti, Gori and Maggini [13]. However, there is no clear evidence from literatures for which fusion method performs better than the others. So for simplicity, we directly use the weighted sum of the Page Rank value and the normalized Authority value of each media object. Authority scores are normalized so that the sum of Authority value of all objects equals to one.

The formulation of the Combined Rank is shown below:

$$CombinedRank(u) = a \times R(u) + b \times normalized\_x(u)$$

where a, b are the weights of importance of Page Rank value and Authority value relative to the combined ranking score of media object u, and $a + b = 1$.

The formulation of Modified Combined Rank is the same as original Combined Rank, only the Page Rank and Authority scores are now changed to the Modified Page Rank and Modified Authority scores.

Obviously, when we rank the objects, more emphasis is put on the objects' in-links as the two components in the equation of Combined Rank, the Page Rank value and the Authority value, both depend mainly on the in-degree. This is reasonable because we usually measure the degree of relevance of objects to user's query in terms of number of references points to those objects. Having more references means that it is more recommended by others and so, it can be more relevant to the query.

Chapter 4. Experimental Results and Discussions

4.1 Experimental Setup

The main theme of this experiment is to compare the difference in two traditional web page ranking algorithms: Page Rank and HITS, and their fusion method: Combined Rank, against their corresponding modified object link analysis ranking algorithms: Modified Page Rank, Modified HITS and Modified Combined Rank. We have tried to crawl and rank the object links based on 10 different query topics which are shown in Table 4.1.

	Query Keywords
1	Nemo
2	PABF 2004
3	The sound of Music
4	The Miracle Box
5	Titanic
6	Lion King
7	Brother Bear
8	Mickey cartoon
9	Tom Cruise
10	Emma

Table 4.1: Query keywords

For each query, we have collected around four hundred object links in different media as the base set for different modified algorithms to work on. The method of

crawling and filtering the base set has already been stated in the pervious chapter. Object information such as the object type, parent links, children links, embedded state, the object type of parents and children are also stored for further analysis.

Then we use the base set to build a directed network with weights on each edge. Ten sets of Modified Page Rank scores, HITS scores and Combined Rank scores can be obtained. For each set of results, we have divided them into 4 groups according to their object type. The objects within each group have been ranked in the descending order of score.

Among these ten base sets, usually there are around 400 object links for each set. We use these 400 links to generate another set of ranking scores from the traditional algorithms. These scores have also been ranked in descending order.

For all setups, only the top thirty objects in each object type are returned to the user. Our evaluation is based on the degree of relevance in these highest-ranked results.

4.1.1 Assumptions for the Experiment

Several assumptions should be stated in advance before carrying out the experiments. This is because our major settings in the experiment and the design of different object ranking algorithms are based on these assumptions:

(a) A hyperlink from object A to object B is a recommendation of object B by the author of object A. Otherwise, the use of link analysis in multimedia information retrieval cannot be reliable

(b) If object A and object B are connected by a hyperlink, they may be on the same topic. We assume that most of the image, video and sound objects are meaningful and related to the content of its web page parents. This is a core element which gives much influence on the quality of the base set.

(c) The size of the base sets is representative enough to the queries

(d) Sufficient number of objects with different media type should be included in the base set of each query

(e) The source of our root set is of high quality in terms of the relevance to the user query

(f) Our method is able to crawl different types of object links in most types of HTML files.

(g) The difference in object type and embedded states of objects reflects the degree of relevance to the users.

4.2 Some Observations from Experiment

Before we move on to the evaluation of this experiment, we would like to point out here some observations in our experiment. Throughout the experimental process, we have come across several characteristics of the base sets and the returned result set.

4.2.1 Dangling links

In most of the cases, video clip links except for .swf files (Flash) and sound track links become dangling links and their Hub values are usually 0. This is mainly due to the nature of these objects. These objects are usually larger in file size and load slower in the web browsers. Thus, not much people use these objects for hyperlink reference. Therefore, they hardly have out-links.

Also, other than Flash movies, the majority of video objects inside our base set are Quicktime movies, Real Player movies and Mpeg movies. Most of them are film trailers of the query terms. They are usually not embedded in its parents. Even when there exists out-links inside the movies, we cannot pick them up when the movies are broadcast as our system do not support the techniques in extracting information from movies. This is another reason why most of the video object links in our base set are dangling.

4.2.2 Good Hub = bad Authority, Good Authority = bad Hub?

Besides, that an object is a good Hub does not imply that it is a poor authoritative object. The Hub and Authority values are in a mutually reinforcing relationship, but do not affect each other on the same object. An object's Hub value only affects its children's Authority values, while its Authority value only influences its parents' Hub values.

4.2.3 Setting of weights

Before understanding relevance and importance of different object types and embedded states, we set all the media objects' links as of the same weight. Therefore, we can see that the relative differences in the ranking scores of all the objects are very small and insignificant to distinguish the importance and relevance among them. Objects with different object types will have the same ranking scores under the same physical conditions, i.e., same in-degree and out-degree. In order to distinguish the difference between the objects with different object types and connectivity formats, we have tried different values of the object type weights and embedded state weights to the ranking algorithms so as to obtain an optimal weight setting for our system.

For example, we make the embedded links have higher weights than the hyperlinks directing to outside objects. This is simply because the objects appearing inside its parent object are usually with higher relevance than objects from external sources. Besides, by observation, video clips are often more relevant to the query topic, followed by images and texts. Sound tracks are the least relevant to the query topics relative to other media.

By and large, the values of these weights are only generated based on our observation and intuitive thinking. They may not be the most suitable ones in the calculation of different scores.

4.3 Discussion on Experimental Results

We have got ten sets of query results. Each includes the top thirty ranked in every group of object type. For the traditional Page Rank, HITS and Combined Rank algorithms, only results with one object type—text—is obtained. Our evaluation is, therefore, based on the comparison between the performance of traditional algorithms and that of their corresponding modified methodology. We also investigate into the relevance of the result set in different object types for different modified ranking methods. For the HITS algorithm, because we are now focused on finding out the highly recommended objects or the most authoritative objects, only the Authority values are taken into account. Therefore, only the top thirty Authorities are ranked and evaluated in the HITS and Modified HITS algorithm.

4.3.1 Relevance

The degree of relevance is a subjective measure to indicate the "aboutness" and "appropriateness" of an object. Different users may differ about the relevance or non-relevance of a particular object to a given query. Therefore, when we ask the user to measure the relevance of our returned results, we have provided some objective guidelines for the users to follow. These include the number of query terms or query term related items, for example, the trailers of the query film or the posters of the query actor, appear in the object, or whether the object is part of a large combined object and that combined object includes the query terms or query term related items, etc. After the users send the feedback of the relevant objects in

each search query and ranking methods, we use them to calculate the precision and recall of each query in different ranking algorithms.

4.3.2 Precision and recall

Both precision and recall are common ways in measuring the performance of information retrieval algorithms. Precision measures the portion of relevant objects in the returning result set and recall aims at finding out the portion of relevant object in the returning result set relative to the total number of relevant objects in the base set.

The formulation in calculating precision and recall are as follows:

$$\text{Precision} = \frac{number\_of\_relevant\_objects\_in\_return\_set}{total\_number\_of\_objects\_in\_result\_set}$$

$$\text{Recall} = \frac{number\_of\_relevant\_objects\_in\_return\_set}{total\_number\_of\_relevant\_objects\_in\_base\_set}$$

In order to counterbalance the biases of personal feedback, for each query, we have at least take an average of two precision values and recall values for carrying out the t-test later.

The average precision and recall for 10 queries of difference algorithms are shown in Table 4.2 and a graph illustrating the average precision and recall is provided in Fig. 4.1.

	Average Precision for 10 queries	Average Recall for 10 queries
Page Rank(text)	0.842	0.126
Authority(text)	0.644	0.096
Combined Rank(text)	0.887	0.159
Modified Page Rank(text)	0.868	0.139
Modified Authority(text)	0.634	0.093
Modified Combined Rank(text)	0.924	0.179
Modified Page Rank(image)	0.633	0.200
Modified Authority(image)	0.567	0.179
Modified Combined Rank(image)	0.667	0.210
Modified Page Rank(video)	0.980	0.980
Modified Authority(video)	0.980	0.980
Modified Combined Rank(video)	0.980	0.980

Table 4.2: Average Precision and Recall for different algorithms in ranking different media objects

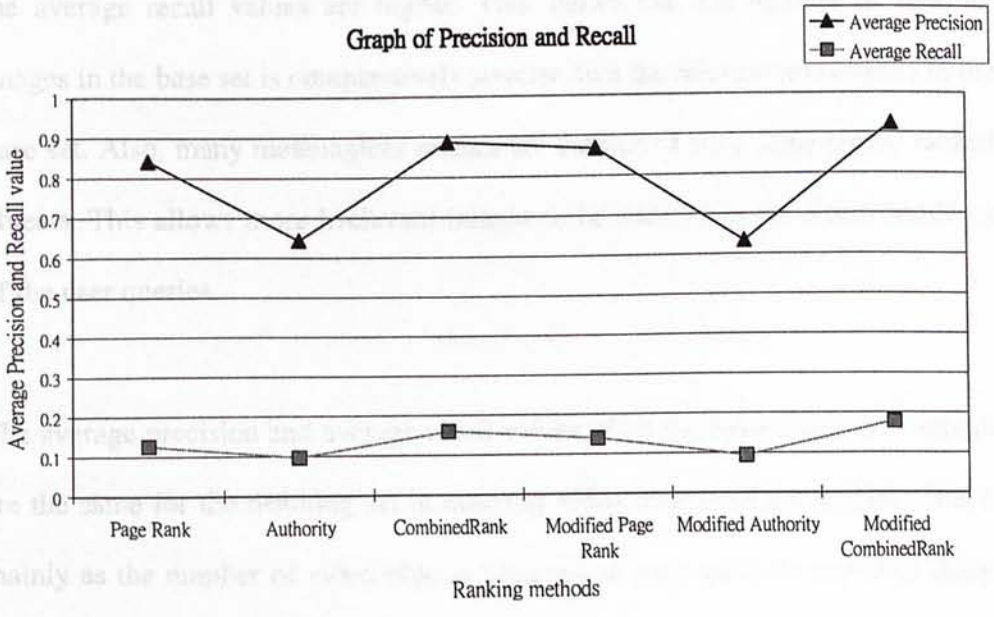

Fig. 4.1: Graph of Average Precision and Recall for all ranking algorithms

We can see that improvements are made on the Modified Page Rank Algorithm when compared with the traditional Page Rank Algorithm, and the Modified Combined Rank Algorithm with the original Combined Rank Algorithm, in terms of the average precision and the average recall for the text object retrieved for the ten queries. For the Modified Authority on text objects, we can see that there are improvements in four queries but it performs worse than the traditional one in the other six queries. The recall value of the three modified algorithms are quite low because of the large portion of relevant text objects in the base set, usually there are more than 200 relevant text objects in the base set.

Providing that we cannot find a benchmarked link-based algorithm to rank the images, videos and sound tracks, we can only comment on the average precision and recall values alone. The average precision values of the modified algorithms in ranking image objects are relatively lower than that in ranking text objects, but the average recall values are higher. This shows that the number of relevant images in the base set is comparatively smaller than the relevant text objects in the base set. Also, many meaningless images are connected with some highly ranked objects. This allows more irrelevant images to be added into the return result set of the user queries.

The average precision and average recall values of all the three modified methods are the same for the resulting set in ordering video objects of the queries. This is mainly as the number of video objects obtained in each query is less than thirty. As a result, our experiment shows no significant difference among the three

modified algorithms in ranking video. However, if the number of video objects in the base set is over thirty, we believe that there may be significant difference in the average precision and recall among all the modified algorithms.

No average precision and recall value is provided for the resulting set of sound track ranks, because only 2 queries contains sound track objects in their base set. Therefore, no meaningful comparison can be made for the algorithms.

Generally, the performance of Modified Combined Rank is much better than the others among our three modified algorithms, in both ranking text objects and image objects. In contrast, the Modified Authority is the worst in both ordering text objects and image objects. This may be due to the choice of number of iterations in calculating the HITS scores or the weight settings in the Modified HITS algorithm.

4.3.3 Significance testing

We have run the one-tailed paired t-test for the precision and recall values of each traditional algorithm with its modified version. A one-tailed paired t-test is used to find how the quality of the resultant set of data can be significantly improved with the modification of algorithms. The result of the one-tailed paired t-tests of the precision and recall values for the three algorithms and their modified versions is shown in Table 4.3.

	Precision	Recall
P(T<=t) for the significant level of Modified Page Rank improves the Traditional Page Rank	**0.083***	**0.092***
P(T<=t) for the significant level of Modified Authority improves the Traditional Authority	0.253	0.305
P(T<=t) for the significant level of Modified Combined Rank improves the Original Combined Rank	**0.022****	**0.073***

* p value < 0.1, improvement is significant

** p value < 0.05, improvement is highly significant

Table 4.3: Result of One-tailed Paired T-test

From Table 4.3, we can see that the p-value of the precision and recall of traditional Page Rank and Modified Page Rank are 0.083 and 0.092 respectively. This means that the Modified Page Rank performs significantly better than the traditional Page Rank.

For the Modified Authority and the traditional Authority, the p-value of precision and recall are 0.253 and 0.325 respectively, which shows that there is no significant proof for whether the Modified Authority or the traditional Authority out-perform the other.

The p-value of the precision and recall of original Combined Rank and Modified Combined Rank are 0.022 and 0.073 respectively. This means that the Modified Combined Rank is relatively higher in the significance level to perform better than the original Combined Rank.

Apart from t-test, we have also done an Anova test to demonstrate the similarity of the performance of the three modified ranking algorithms in ranking the text objects. The null hypothesis for the Anova is the performance of all the three scores is similar to each other. The result of Anova by using precision and recall of ranking text object is demonstrated in the Table 4.4.

Testing value	*P-value*
Average precision of 3 modified algorithms in ranking text objects	**9.07E-15***
Average recall of 3 modified algorithms in ranking text objects	2.37E-11*

* p<0.01 The 3 groups are extremely significant to show that their value are not similar.

Table 4.4: Result of Anova showing the similarity in the three modified ranking algorithms (ranking text objects)

In both Anova of precision and recall of the ranking in text objects by the three modified algorithms, the p-value is smaller than 0.01, which implies that the performance of these three algorithms are significantly dislike to each other. This is reasonable because the two traditional ranking algorithms, Page Rank and HITS, are using different mathematical modeling for the iteration process although both of them are dependent on the linkage between object nodes. The number of iteration and the way of normalization may also result in the difference in the ranking of them. CominedRank, for which ranking is based on the weighted sum of Page Rank and Authority values, therefore hardly follows either Page Rank or Authority's ranks. As a result, the Modified Page Rank, Modified Authority and Modified Combined Rank algorithms have significantly difference in the ranking.

4.3.4 Ranking

The ranking of objects among the traditional algorithms and the modified ones is another evaluation tool for the performance of these ranking algorithms. We have

compared the ranking difference of Page Rank and Authority with their modifications. From our experiment, we can find more significantly improved examples for the Modified Page Rank against the traditional Page Rank. Less regular pattern of improvements in Modified Authority values to the traditional Authority can be obtained. Therefore, in this section, we will only provide the example of improvement in object ranks by Modified Page Rank.

ID	Link	Rank
t0	http://www.pixar.com/featurefilms/nemo/	1
t1	http://www.imdb.com/title/tt0266543/	2
t2	http://www.apple.com/trailers/disney/finding_nemo/trailer/	3
t3	http://disney.go.com/disneyvideos/animatedfilms/findingnemo/index2.html	4
t4	http://www.amazon.com/exec/obidos/tg/detail/-/B00005JM02?v=glance	5
t5	http://quizilla.com/users/wgryph/quizzes/What%20Finding%20Nemo%20Character%20are%20You%3F/	6
t6	http://bima.astro.umd.edu/nemo/	7
t7	http://www.e-nemo.nl/	8
t8	http://www.disney.de/DisneyKinofilme/nemo/	9
t9	http://www.newmet.nl/	10

Table 4.5: Top ten pages of traditional Page Rank without using object model

ID	Link	Rank
m0	http://disney.go.com/disneyvideos/animatedfilms/findingnemo/index2.html	1
m1	http://www.pixar.com/featurefilms/nemo/	2
m2	http://www.imdb.com/title/tt0266543/	3
m3	http://www.apple.com/trailers/disney/finding_nemo/trailer/	4
m4	http://www.amazon.com/exec/obidos/tg/detail/-/B00005JM02?v=glance	5
m5	http://bima.astro.umd.edu/nemo/	6
m6	http://www.disney.de/DisneyKinofilme/nemo/	7
m7	http://www.newmet.nl/	8
m8	http://www.apple.com/trailers/disney/finding_nemo/trailer/3_fullscreen.html	9
m9	http://www.apple.com/trailers/disney/finding_nemo/trailer/1_large.html	10

Table 4.6: Top ten text objects of Modified Page Rank using object model

We take the query "nemo" as our first example. This example shows an improvement of object rank by using the Modified Page Rank algorithm. Table 4.5 and 4.6 are part of the result sets of query "nemo" ordering by the traditional Page Rank scores and our Modified Page Rank. Object m0 is the official website of the film "Finding Nemo". Therefore, it should be the most relevant to the user among all other objects in the base set. However, according to the ranking of traditional Page Rank, it is less relevant than the text objects m1, m2 and m3.

The use of object type and embedded state properties in our Modified Page Rank algorithm is the source of improvement in ranking m0. We can find that majority of the children and parent links of m0 are images and embedded videos, while those in m1, m2 and m3 are images and unembedded text. As a result, the ranking of m0 will be higher than that of m1, m2 and m3 in the Modified Page Rank, for which the difference in object type and embedded states for parent and children objects can have effects on the ordering process.

Another discussion on ranking is about the Combined Rank. Some people may doubt that if Combined Rank Algorithm is the weighted sum of Page Rank and HITS, why the combined rank of an object does not the same as the average of its Page Rank rank and Authority rank? In Table 4.7, we stated an example of the ranking results of Modified Page Rank, Modified Authority and Modified Combined Rank.

ID	Link	Page Rank	Authority	Combined Rank
x0	http://www.imdb.com/title/tt0266543/	3	3	1
x1	http://disney.go.com/disneyvideos/animatedfilms/findingnemo/index2.html	1	142	2
x2	http://www.pixar.com/featurefilms/nemo/	2	152	3
x3	http://www.apple.com/trailers/disney/finding_nemo/trailer/	4	143	4
x4	http://www.amazon.com/exec/obidos/tg/detail/-/B00005JM02?v=glance	5	297	5
x5	http://bima.astro.umd.edu/nemo/	6	298	6
x6	http://www.disney.de/DisneyKinofilme/nemo/	7	284	7
x7	http://www.newmet.nl/	8	299	8
x8	http://www.imdb.com/title/tt0266543/board/threads	14	1	9
x9	http://www.imdb.com/Sections/Years/2003	27	2	10

Table 4.7: Top ten text objects of Modified Combined Rank using object model and their corresponding Page Rank and Authority ranking

This example used "nemo" as the query keyword. From Table 4.7, we can see that the ranks of the top ten text objects of Modified Combined Rank are not the same as average ranking of Modified Page Rank and Authority. The highest-ranked object x0 was ranked in the third position in terms of relevance in both Modified Page Rank and Modified Authority. Some of the lower-ranked objects in Modified Authority, such as object x1 to x7, because of their high Modified Page Rank scores, they can attain a comparatively higher weighted sum than the x8 to x9, though x8 and x9 are the two top-ranked objects by Modified Authority.

Considering the difference in iterative process and algorithm design of Modified Page Rank and Modified Authority, the rank of an object calculated by Page Rank and HITS can be significantly independent to each other. For instance, the number of iteration in HITS is under controlled, in reverse, that in Page Rank is

uncontrolled and the algorithm iterates until the ranking scores reach stability. This may make the authority scores not optimize and makes the authority ranks and Page Rank ranks different. Moreover, the calculation of authority scores is based on hub values of objects' parents. The initial value of hub and authority scores of the root set links may also have affects on the overall authority values. This can also be another factor which leads to the difference in Authority and Page Rank scores Therefore, the Modified Combined rank does not necessarily equal to the average of the rankings in Modified Page Rank and Authority.

Most important of all, we would like to state here that the above observations and statistical analysis are only based on our assumptions, our implemented systems and the queries we used in the experiments. Therefore, there may be some other observations that we cannot conclude here due to our limited data set.

4.4 Limitations and Difficulties

No matter how well people can plan for the precautions to run an experiment, design the methodology used in experiment and organize the set up of the experiment, there will still be some limitations and difficulties in carrying out the experiment. The following are limitations and difficulties we have faced as the experiment proceed:

4.4.1 Small size of the base set

Due to limited computer resources, it is difficult for us to build a larger base set which is rich in different types of media object. As a result, some evaluations become insignificant. One of the examples happens in calculating the precision and recall for the video objects and sound tracks. The size of base set takes much influence in the quality of the base set and the linkage structure of the base set. More links means a higher chance to build up a highly-linked network, from which more representative and realistic rank of the objects in WWW can be obtained easily.

4.4.2 Parameter settings

The number of iterations that should be run by the HITS and the Modified HITS, the weights of object types and embedded states, etc. are decided by our subjective intuition and trials. There may be better parameter settings for our proposed methodology which have not been discovered by us.

4.4.3 Unable to remove all the meaningless links from base set

As we have seen from the precision values of our three modified ranking algorithms to order image objects, we have discovered our limitation in filtering out unwanted links from our base set. In addition, the complexity in gathering links from the script languages in the HTML files is also an important reason why we cannot gather more video objects in our base set. More web authors tend to use VB script or Javascript to embed the video links in the HTML tag. Since our

crawling system cannot grab the object links written inside the scripts, we have probably missed some of the useful objects inside the scripts.

4.4.4 Resources and time-consuming

The preprocessing step of this experiment is both resource and time-consuming. The preprocessing steps include choosing the suitable query, building up the computer system for crawling and generating different ranks. We nearly spend half of our research period to do the preprocessing. The time for crawling is the longest in comparison with other computer-involved task. And we always face the problem of running out of memory. That is the reason why we need to limit our base set to a small size in order to reduce the computational cost.

4.4.5 TKC Effect

Inevitably, our object graph forms the tightly-knit community (TKC) effect, due to our crawling techniques and also, we need to involve objects with different types, even they are in the same domain. Some researchers suggest removing the links with same domain in the base set so as to reduce the TKC effect. However, this method is not applicable to our system, as many images and video clips come from a few domains. If we delete these media objects, the number of image objects and video objects will dramatically decrease to 1-tenth of the original number.

4.4.6 Continuously updated format of HTML codes and file types

Development of Internet technology is rapid and unpredictable. More web pages are now formatted by scripts like JSP and XML. Also, more different types of software developed for object processing and viewing also lead to new file extensions for object link. Therefore, it is difficult for our crawling system to collect and identify all the object links in all text objects.

4.4.7 The object citation habit of authors

There is no definite instruction to restrict how authors of web pages design their pages' linkages to different objects. They may also point to some objects which are totally unrelated to the content of their pages. The organization of the web pages is also another reason for the difficulties in filtering useful link. Some authors usually mix objects with different contents together. This confused our crawling systems and the results is we may miss some useful object links because their neighbor paragraphs or table rows do not contain the query keywords, and we may collect some meaningless object links owing to the existence of query keyword in their nearby areas.

Chapter 5. Conclusion

5.1 Contribution of our Methodology

Our research has tried to import the idea of multimedia nodes to the directed network graph and migrate the traditional link analysis algorithms, Page Rank and HITS to the new directed object graph, with new definition of attributes in the web graph nodes and the linkages. We have proposed several factors which potentially affect the ranking of an object, as well as listed out our observations through the experiment. These can be good examples for later researchers who are interested in multimedia information retrieval.

Our Modified Page Rank and Modified Combined Rank are proved to have significant improvement in ranking text objects than the traditional ones. Some other previous modification of Page Rank, such as Haveliwala's Topic-sensitive Page Rank [6], may also be applied on our algorithms.

5.2 Possible Improvement

Previously proposed methodology, for instance, the BHITS [16] or WHITS [5] can be tried to apply on our Modified HITS algorithm to increase the quality of the objects in the returned result set in terms of relevance. We can introduce other weights to our modified algorithms depending on the number of repetition of objects with same domain. This may be helpful in relieving the TKC effect of our

query result set. Moreover, we can try to further improve the ranking of media objects by considering the content of the anchor texts, filename of hyperlinks of the object, as well as the position and size of the objects when they are inside their parents. We have proposed this in former chapters but have not implemented in our system as we think that it is content-related rather than link-based.

Apart from the above mentioned points, we find that running our current crawling and ranking algorithms are both time and cost-consuming. A good search engine should be both high in retrieving speed and good at finding relevant objects. Thus, reducing the time and computer resources used by our system for each query is also one of the possible improvements of our system. Some of the solutions for this problem are restricting the number of iterations in running the ranking methods, controlling the error significance level in the calculations, including only the most important weights in adjusting the ranking values, adjusting the size of base set, etc. But the effectiveness of them still needs to be proved.

5.3 Conclusion

To conclude, multimedia searching is a new topic in the Internet information retrieval. Using the idea of link analysis on multimedia object graph in information retrieval is still a germinating topic nowadays. There are still many untouched topics for scientists to explore, such as designing better network model for easy adaptation of the currently developed link analysis algorithms, or proposing new link analysis methodology which is suitable for use in multimedia object graph. Undoubtedly, a high-quality searching algorithm is necessary to

satisfy users' query on multimedia information. What we can do is to try our best to include as much users' expectations on information retrieval as possible and enhance the searching quality while designing our own searching algorithms.

Bibliography

1. Monika R. Henzinger, "Hyperlink Analysis for the Web", I*EEE Internet Computing*, January-February 2001, P.45 - P.50
2. L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", *Stanford Digital Library Technologies, Working Paper* 1999-0120, Standford University, Palo Alto, California, US, 1998.
3. Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", *Proceedings of the 9th Annual ACM-SIAM Symposium Discrete Algorithms*, ACM Press, New York, US, 1998, P.668 - P.677
4. Jun Yang, Qing Li, and Yueting Zhuang, "Octopus: Aggressive Search of Multi-Modality Data Using Multifaceted Knowledge Base", *Proceedings of the 11th International World Wide Web Conference*, Hawaii, 2002, P.54 – P.64
5. Longzhuang Li, Yi Shang, and Wei Zhang, "Improvement of HITS-based Algorithms on Web Documents", *Proceedings of the 11th International World Wide Web Conference*, Hawaii, 2002, P.527 - P.535
6. Taher H. Haveliwala, "Topic-Sensitive PageRank", *Proceedings of the 11th International World Wide Web Conference*, Hawaii, 2002, P.517 – P.526
7. Micheal Chau, and Hsinchun Chen, "Creating Vertical Search Engines Using Spreading Activation", accepted by *IEEE Compute*r, forthcoming
8. Julien Gervey, and Stefan M Ruger, "Link-based Approaches for Text Retrieval", the *10th Text Retrieval Conference (TREC"10)*, Gaithersburg, Maryland, 2001, P.279 – P.285
9. Jeffrey Dean, and Monika R. Heinzinger, "Finding related pages in the World Wide Web", *Elsevier Science B.V.*, 1999, P.1467 – P.1479
10. Brian Amento, Loren Terveen, and Will Hill, "Does Authority Mean Quality? Predicting Expert Quality Ratings ", *Proceedings of the 23rd International ACM SIGIR Conference Research and Development in Information Retrieval (SIGIR00)*, ACM Press, New York, 2000, P.296 – P.303

11. R. Lempel, and S. Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect", *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, The Netherlands, 2000, P.387 - P.401
12. Gopal Panduragan, Prabhakar Raghavan, and Eli Upfal, "Using PageRank to Characterize Web Structure", *the 8th Annual International Computing and Combinatorics Conference (COCOON)*, Singapore, 2002, P.330 – P.339
13. Michelangelo Diligenti, Marco Gori and Marco Maggini, "Web Page Scoring Systems for Horizontal and Vertical Search", *Proceedings of the 11th International. World Wide Web Conference*, Hawaii, 2002, P.508 – P.516
14. Andrew Y. Ng, Alice X. Zheng, and Micheal I. Jordan, "Stable Algorithms for Link Analysis", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New Orleans, Louisiana, US, 2001, P.258 – P.266
15. Sergey Brin, and Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine", *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 1998, P.107 – P. 117
16. K. Bharat, and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, P.104 – P.111
17. Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas, "Finding Authorities and Hubs from Link Structures on the World Wide Web", *Proceedings of the 10th International Conference on World Wide Web*, Hong Kong, Hong Kong, 2001, P.415-429
18. Jacques Savoy, and Yves Rasolofo, "Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections", *Proceedings of the TREC'9 Conference*, Gaithersberg, Maryland, 1997, P.489-502
19. Kiduk Yang, "Combining text- and link-based retrieval methods for Web IR" , *Proceedings of the TREC'10 Conference*, 2001, P609-618
20. Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst Simon, "PageRank, HITS and a unified framework for link analysis", *Proceedings of*

the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002, P. 353-354

21. Ronny Lempel, and Aya Soffer, "PicASHOW: Pictorial Authority Search by Hyperlinks on the Web", *ACM Transactions on Information Systems (TOIS)*, Volume 20, 2002, P. 1-24
22. Chris Ding, Hongyuan Zha, Xiaofeng He, Parry Husbands, and Horst Simon, "Link Analysis: Hub and Authorities on the World Wide Web", *Society for Industrial and Applied Mathematics (SIAM) Review*, Volume 46, P. 256-268
23. Zheng Chen, Li Tao, Jidong Wang, Wenyin Liu, and Weiying Ma, "A Unified Framework for Web Link Analysis", *the 3rd International Conference on Web Information Systems Engineering (WISE'00)*, Singapore, Singapore, 2002, P. 161-172
24. Taher H. Haveliwala, "Efficient Computation of PageRank", *Technical Report*, 1999
25. Soumen Chakrabarti, Bryon Dom, Prabhakar Raghavan Sridhar Rajagopalan, David Gibson, Jon Kleinberg, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text", *Proceedings of the 7th international conference on World Wide Web 7*, Brisbane, Australia, 1998, P. 65-74
26. Krishna Bharat and Geoge A. Mihaila, "When Experts Agree: Using Non-affiliated Experts to Rank Popular Topics", *ACM Transactions on Information Systems (TOIS)*, Volume 20, 2002, P. 47-58
27. Julien Gevrey, and Stefan M Ruger, "Link-based Approaches for Text Retrieval", *Proceedings of TREC'10*, Gaithersburg, Maryland, 2001, P. 279-285
28. Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan, "Searching the Web", *ACM Transactions on Internet Technology*, Volume 1, 2001, P. 2-43
29. Soumen Chakrabarti, Byron E. Dom, David Gibson, Jon Kleinberg Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "Mining the Link Structure of the World Wide Web", IEEE Computer, Volume 32, 1999, P. 60-67

30. Matthew Richardson, and Pedro Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in Page Rank", *MIT Press*, Volume 14, Cambridge, MA, 2002

31. Google Search Appliance, Search Tools Product Report, http://www.searchtools.com/tools/google-app.html

Appendix

A1. One-tailed paired t-test results

	Traditional Page Rank	*Modified Page Rank*
Mean	0.8421	0.8678
Variance	0.000649	0.003324
Observations	10	10
Pearson Correlation	0.360025	
Hypothesized Mean Difference	0	
df	9	
t Stat	-1.50506	
P(T<=t) one-tail	0.083284	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.166567	
t Critical two-tail	1.833114	

Table A1.1: Result of paired t-test of precision in traditional Page Rank and Modified Page Rank

	Traditional Authority	*Modified Authority*
Mean	0.6437	0.6336
Variance	0.001521	0.00081
Observations	10	10
Pearson Correlation	0.096093	
Hypothesized Mean Difference	0	
df	9	
t Stat	0.694091	
P(T<=t) one-tail	0.252572	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.505145	
t Critical two-tail	1.833114	

Table A1.2: Result of paired t-test of precision in traditional Authority and Modified Authority

	Original Combined Rank	*Modified Combined Rank*
Mean	0.8868	0.9243
Variance	0.001375	0.001162
Observations	10	10
Pearson Correlation	-0.00513	
Hypothesized Mean Difference	0	
df	9	
t Stat	-2.34842	
P(T<=t) one-tail	0.021709	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.043417	
t Critical two-tail	1.833114	

Table A1.3: Result of paired t-test of precision in original Combined Rank and Modified Combined Rank

	Traditional Page Rank	*Modified Page Rank*
Mean	0.1256	0.1388
Variance	0.000448	0.000305
Observations	10	10
Pearson Correlation	-0.12295	
Hypothesized Mean Difference	0	
df	9	
t Stat	-1.43724	
P(T<=t) one-tail	0.092246	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.184492	
t Critical two-tail	1.833114	

Table A1.4: Result of paired t-test of recall in traditional Page Rank and Modified Page Rank

	Traditional Authority	*Modified Authority*
Mean	0.0959	0.0932
Variance	0.000176	0.000355
Observations	10	10
Pearson Correlation	0.538089	
Hypothesized Mean Difference	0	
df	9	
t Stat	0.527477	
P(T<=t) one-tail	0.305308	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.610617	
t Critical two-tail	1.833114	

Table A1.5: Result of paired t-test of recall in traditional Authority and Modified Authority

	Original Combined Rank	*Modified Combined Rank*
Mean	0.1594	0.1788
Variance	0.000588	0.000841
Observations	10	10
Pearson Correlation	-0.04189	
Hypothesized Mean Difference	0	
df	9	
t Stat	-1.59012	
P(T<=t) one-tail	0.073136	
t Critical one-tail	1.383029	
P(T<=t) two-tail	0.146272	
t Critical two-tail	1.833114	

Table A1.6: Result of paired t-test of recall in original Combined Rank and Modified Combined Rank

A2. Anova results

Summary

Groups	*Count*	*Sum*	*Average*	*Variance*
Modified Page Rank	10	8.678	0.8678	0.003324
Modified Authority	10	6.336	0.6336	0.00081
Modified Combined Rank	10	9.243	0.9243	0.001162

Table A2.1: Summary of Anova showing the similarity in the three modified ranking algorithms (precision in ranking text objects)

Anova

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.475161	2	0.237581	134.5809	**9.07E-15***	2.51061
Within Groups	0.047664	27	0.001765			
Total	0.522825	29				

* p<0.01 The 3 groups are extremely significant to show that their value are not similar.

Table A2.2: Result of Anova showing the similarity in the three modified ranking algorithms (precision in ranking text objects)

Summary

Groups	*Count*	*Sum*	*Average*	*Variance*
Modified Page Rank	10	1.374	0.1374	3.78E-05
Modified Authority	10	0.932	0.0932	0.000355
Modified Combined Rank	10	1.738	0.1738	0.000314

Table A2.3. Summary of Anova showing the similarity in the three modified ranking algorithms (recall in ranking text objects)

Anova

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.032583	2	0.016292	69.1668	**2.37E-11***	2.51061
Within Groups	0.00636	27	0.000236			
Total	0.038943	29				

* p<0.01 The 3 groups are extremely significant to show that their value are not similar.

Table A2.4. Result of Anova showing the similarity in the three modified ranking algorithms (recall in ranking text objects)