Using Virtual Documents for NTCIR-4 Web Information Retrieval Task

Yinghui Xu Kyoji Umemura
Toyohashi University of Technology
xyh@ss.ics.tut.ac.jp umemura@tutics.tut.ac.jp

Abstract

The Web is a large collection of heterogeneous pages. Web documents are not always descriptive and accurate in content. In addition, a significant difference between the problems of Web search and traditional text search is the availability of hyperlinks between pages. A page on the Web might possibly be cited by or cite other pages. When evaluating a page, the neighborhood of the page might be a part of the input. In this paper, in addition to the explicit information unit (page content), a new information unit, a virtual document, is introduced in our systems, which is mainly organized by the associated anchor-text of in-bounds links to a page and also its title data. We analyzed the utility of virtual document for Web searching. Three searching function based on virtual document are proposed in our study:

- *We propose a way to weight query terms through term entropy in the virtual document collection space.*
- *Ranking algorithms are configured to index virtual documents as separate queryable data and our system considers it an important indicator of page relevance in addition to the relevance score of practical Web documents.*
- *Our system implements a modified version of link analysis which employs literal matching between information units of the NTCIR-4 Web data.*

The experiment results show: Our Web searching system that uses our proposed ranking function works well. The new information units, virtual documents, play an important role in improving information retrieval results. Query term weighting using term entropy on virtual document space is effective in improving searching results. The combination of evidence from actual documents and virtual documents can improve searching results beyond either information source alone. The Query-independent score, which is calculated by our proposed link analysis model, could also obtain modest improvements through our tentative re-ranking methods.

Keywords: *PageRank, Information Retrieval (IR), Virtual document (VD), Strong Connected Core (SCC)*

1 Introduction

There is growing frustration with traditional IR systems applied to Web data due to the particular characteristics of Web resources. IR systems work with finite document collections, and the worth of a document with regard to a query is critical for the document. Non-Web documents are self-contained units, and are generally descriptive and clear regarding their contents. In contrast, because the Web is a large collection of heterogeneous pages, Web documents are not always descriptive or clear regarding their contents. In addition, one significant difference between the problems of Web search and traditional text search is the availability of hyperlinks between Web pages. A page on the Web might possibly be cited by or cite other pages. When evaluating a Web page, the explicit textual content on the page itself might not be enough to reflect the on-topic information, and the neighborhood of the page might be a part of the input.

Hyperlinks are being actively used to improve Web search engine ranking. A hyperlink has two components: the destination page, and the associated anchor-text describing the link. The personal nature of the anchor text allows for connecting words to destination pages. Anchor-text has been utilized by commercial search engines, such as Google and Altavista, to improve Web search. In addition to the anchor description of the hyperlink, the title tag is also important from the search engine perspective, as it not only communicates the theme of the web page to the human visitors, but is also considered very important by search engine crawlers [3]. Therefore, to explore the value of this kind of information to the Web searching task of NTCIR-4, in addition to the explicit information unit (page content), a new information unit, a virtual document for a page, is introduced in our system, which is mainly organized by textual information from both anchor and title. This work attempts to utilize the virtual document to improve information retrieval results in three ways.

First, in the Web convention, a user tends to submit a very short query, and the terms in the query are seldom repeated. Query-term weighting scheme is important for Web IR system. How can we distinguish the importance of query terms having the same frequency? The unique information resources of the virtual document allow it to share the characteristics of both anchor and title. Recent study [7][12] examined several aspects of anchor-text (e.g., its relationship to title, the frequency of queries that can be satisfied by anchor-text alone, and translating queries using anchor-text mining). They showed evidence that anchor-text summaries, on a statistical basis at least, look very much like real user queries. Similarly, observations on the relationship between the title data and queries were reported by Jin, Haupmann, and Zhai [9]. They pointed out that document titles bear a close resemblance to queries as well. Therefore, a virtual document collection provides a kind of simulated information space for analyzing user query data. In this paper, we propose using the term entropy in VD space to distinguish the importance of query terms.

Second, the ranking algorithms used in our experiment are configured to index virtual documents as separate queryable data, and our system considers this an indicator of page relevance in addition to the relevance score of existing Web documents.

Finally, we are interested in investigating the effectiveness of link analysis to improve search results. The PageRank algorithm [1] used in the Google search engine plays an important role in enhancing the quality of its results by employing the explicit hyperlink structure. Note that the PageRank model is only based on the hyperlink structure without considering the textual information that is carried by the link between connected pages, the transition probability from a given page to its outgoing links is weighted equally. The transition probability might become more accurate if we consider some practical aspects of human search behavior. During a user's surfing procedure, the information obtained through reading a page and the literal information of outgoing links will help the user weight those options because the search might provide a literal match between the words in the user's mind and the words on the page. We might as well regard the virtual document as one kind of information resource for a user's idea of Web page due to the particular characteristic of anchor and title data in VD. Therefore, we propose a literal matching biased-link analysis model. A virtual document is used to model a user's idea of Web pages, and literal matching is used to measure the transition likelihood of those outgoing links in a page. Our intuition is that inbound links from pages with a similar theme to a page have a larger influence on its page rank value than links from unrelated pages. An experiment using a query-independent score calculated by our proposed link model for improving Web searching results was also performed.

The experimental results show that our Web searching system using our proposed ranking function works well. The new information units, virtual documents, play an important role in improving information retrieval results. Query-term weighting using term entropy in virtual document space is effective for improving search results. The combination of evidence from actual documents and virtual documents can improve search results better than either information source alone. The query-independent score, calculated by our proposed link analysis model, could also obtain modest improvements through our proposed reranking procedures.

This paper is organized as follows: in Section 2, we describe our system and architecture in detail. In Section 3, we provide statistical information about the Web data processed by our system, experiment results and analysis. Lastly, we offer our conclusions and point out our future plans in Section 4.

2 Architecture and system description

Three kinds of data resources in the NTCIR-4 Web corpus [6] are used in our system. One is html page collection with EUC encoding. Another is the ”linklist” file containing linked-pair lists from the Web corpus, which are designed to emulate a snapshot of real search engine circumstances. And the third is the ”Doclist” file that provides the mapping table from URL to doc-name. At the beginning of the preprocessing, a DOM tree-based parsing scheme was adopted for extracting the textual data from Web pages. Then the morphological analyzer, Chasen [14], was used for obtaining segmented terms from Japanese Web pages. To reduce the size of the dictionary of Web corpus and index file size, only segmented words that belong to noun groups were considered, and words in the Web corpus are case-insensitive. The architecture of our system is shown in Fig. 1. Below we introduce the important modules in our system.

2.1 Document generator

One of the distinguished features of our system is the virtual document generator. What is a virtual document? There are a number of ways to define VD. It is dependent on information processing tasks. The concept of the virtual document was introduced by Glover et al. [8], and they used VD for a Web pages classifying and describing task. In our approach, the virtual document of a given page, which is comprised of the expanded anchor-text from pages that point to the page and the title words on the page itself, acts as a separate information unit for Web searching. The definition is

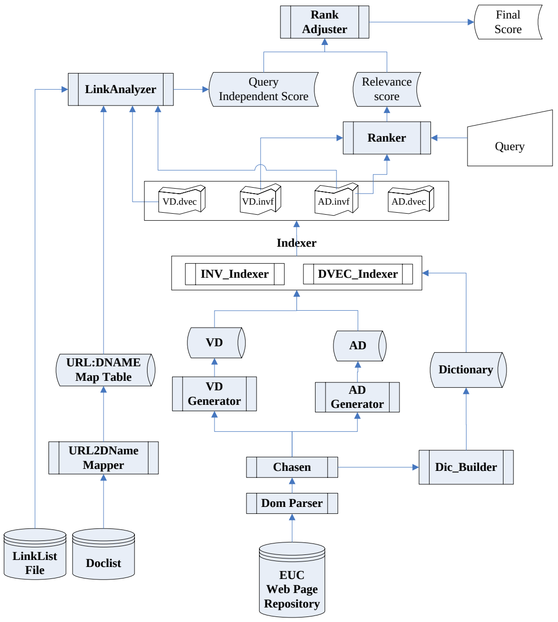

Figure 1. Implementation architecture

as follows.

$$AnchorText\,(i,j):\ \textit{set of terms that appear in}$$
$$\textit{and around the anchor of the link from page } i \textit{ to } j.$$
$$TitleText\,(j):\textit{set of terms that appear in}$$
$$\textit{the title tag of page } j.$$
$$VD\,(j):\ \textit{set of terms in the virtual document } j.$$
$$VD\,(j)=\bigcup_{i}\left(AnchorText\,(i,j)\right)\cup TitleText\,(j)$$

This solid building method relates three steps: First, we create the link text table that includes the three elements, $\langle URL_i, URL_j, DT\rangle$. This means that the page with URL_i to the page with the URL_j contains the description text (*DT*). The DT is extracted based on Document Object Model (DOM)[3] tree structure. The left and right sibling nodes with the text properties of the anchor tag ”a” node, and all text information contained in the closed anchor tag is extracted as the description data. Considering the case that a sibling node with text properties around the anchor tag node may be over several lines, and may deviate from the main author’s intent due to the poor page structure, only the text information around the anchor tag within one phrase is kept for description data of an anchor link in our experiment. Next, we also extract text information from the title tag. Lastly, a virtual document of the URL_j is organized by integrating textual information from all associated description texts of its inbound hyperlinks and its own title text.

2.1.1 Characteristics of Virtual Document

The specific information resources of VD for a page give it a number of significant advantages for Web searching:

- The size of a virtual document collection is much smaller than that of total documents collection data. Thus, processing it is much faster than processing the total document data.

- The virtual document provides a summarization of the target document within the context of the source document, and the process of creating a virtual document for a page might be a good approximation of the type of summarization sought by the user of the search system in most queries.

- The virtual document of a page that aggregates a large amount of anchor-text information pointing to it tends to have higher in-degree (number of links pointing to it) and higher ranks based on link analysis. Thus, this collective wisdom of what the page contains could help refine Web search results.

- It is well known that the general similarity metric may fail to capture the correct relationship between two connected pages due to the heterogeneous characteristics of unstructured Web pages. Yet, our intuition about VD space tells us that the VD information resources are somewhat homogeneous.

2.1.2 Functionality of Virtual Document

Base on the characteristics of the VD mentioned above, the functionalities of using the VD in our system are as follows:

- VD is a separate information unit in addition to the actual page content. It allows one to set up different weighting from the actual document text information and to investigate whether VD-based searching can improve information retrieval results.

- VD collection is a way of approximating collected queries information. It is used to determine the importance of terms in a query.

- VD is a type of a representative summarization of Web pages. In our proposed literal matching biased-link analysis, it is used to model user’s views of a page to decide the transition probability.

The actual functions using VD are described in ”Ranker” section 2.2.

As for the actual document (AD) is extracted based on the actual textual information in the DOM tree of Web pages, denoted by:

$$AD\,(j):\ \textit{set of terms in actual document of page } j.$$

2.2 Ranker

2.2.1 Baseline Ranking Function

According to some reports about Web Information Retrieval, the Okapi model has proven to be efficient for content-based Web searching. Accordingly, we used Okapi's BM25 [4] as our baseline for comparison. The equation used in our system may be stated:

$$RS\left(Q,d\right) = \sum_{w\in Q} \frac{tf}{\left(tf+0.5+\frac{1.5\times dl}{ave\_dl}\right)} \times \frac{\log_2\left(0.5+{}^{N}\!/_{df}\right)}{\log_2(1+\log_2(N))}$$

$$RS\left(Q,d\right): \ represent\ the\ relevance\ score\ between\ query\ "Q"\ and\ page\ "d"$$

2.2.2 Query Term Importance Based Ranking Function (QTIBRF)

In a general information retrieval system, especially for a long topic, query term frequency is used to indicate the terms importance for the relevance ranking function. However, in actual Web searching, the input information of user usually tends to be short and terms used are seldom repeated. Query-term frequency based ranking function may fail to capture the main purpose of user's request. For example, for the query "google, pagerank", "PageRank" should be the term that reflects the main purpose of a user request. How can reasonable term weighting for each query term be set up? Considering the characteristics of VD collection, entropy based term weighting [10] is used in our system. First, the entropy of a term in a query is calculated by the following equation:

$$VDTF\left(w,j\right) = \#\left\{w|w\in VD\left(j\right)\right\}$$
$$VDDF\left(w\right) = \#\left\{j|w\in VD\left(j\right)\right\}$$
$$P\left(w,j\right) = \frac{VDTF(w,j)}{\sum\limits_{k=1}^{N} VDTF(w,k)}$$
$$VDEtropy\left(w\right) = -\sum_{j=1}^{N} p\left(w,j\right)\log_N P\left(w,j\right)$$

$where:$

$N: \ number\ of\ virtual\ documents\ in\ collection$ using N as the base of log function for normalization.

Then, each term in a query is weighted by the quotient obtained by dividing the entropy value summation of all terms in the query by the entropy value of the term itself, denoted by the equation: $QTW\left(w\right) = \frac{\sum\limits_{w\in Q} VDEtropy(w)}{VDEtropy(w)}$. The lower the term entropy is, the greater its importance when compared with other terms in the query. The calculated term weighting, which is regarded as the query term importance factor, is integrated into the Okapi ranking function. The augmented Okapi Model is:

$$QTIBRF\left(Q,d\right) = \sum_{w\in Q} QTW\left(w\right)\times \left(\frac{tf}{tf+0.5+1.5\times {}^{dl}\!/_{ave\_dl}} \times \frac{\log_2\left(0.5+{}^{N}\!/_{df}\right)}{\log_2(1.0+\log_2 N)}\right)$$

2.2.3 Score Merge Ranking Function (SMRF)

In our system, for a given Web page P, there are two kinds of representation units, a virtual document VD_P and a actual document AD_P. It is natural to consider merging the ranking score of the two searching processes performed on both the virtual document collection and the actual document collection, respectively. A simple linear merging scheme was adopted as follows:

$$P = \left\{VD_P, AD_p\right\}$$
$$SMRF\left(Q,P\right) = QTIBRF\left(Q,VD_p\right) + \alpha\times QTIBRF\left(Q,AD_p\right)$$

2.2.4 Literal Matching Biased-Link Analysis

A link analysis that makes use of the hyperlink structure for ranking Web resources is an important Web searching function. The two best-known algorithms that perform link analysis are HITS [11] and PageRank [1]. The latter, used in Google, has proven its efficiency in practical World Wide Web searching. In our system, the literal matching biased link analysis (LMBLA) module has the same purpose of bringing order to the Web through query-independent ranking as does Google's PageRank but uses a different calculation mechanism. We concentrate on investigating a way to take the textual information of Web pages into account for link analysis. Hyperlinks in a page might serve different roles. We divide the hyperlinks in a page into two types, informative link, termed as InforLink, and referential link, termed as referLink. We are interested in the links with literal matching between pages, because the purpose of such kind of links is to point to similar, more detailed, or additional information. As for the referential links, they are the links in a page that have no literal matching with its target. We aim at assigning link weights through the literal information between a page's contents and the virtual document contents of its target pages. The calculation mechanism is defined as: given a page P and its outgoing sets $S = \left\{s_1, s_2, \cdots, s_m\right\}$, transition odds from p to s_k are determined by:

$$TranOdds\left(p\rightarrow s_k\right) = SIM\_IR\left(VD_{s_k}, p\right) + SIM\_JACCARD\left(VD_{s_k}, VD_p\right)$$

$where:$

$$SIM\_IR\left(VD_{s_k}, p\right) = OKAPI\_BM25$$
$$SIM\_JACCARD\left(VD_{s_k}, VD_p\right) = \frac{\left|VD_{s_k}\cap VD_p\right|}{\left|VD_{s_k}\cup VD_p\right|}$$

Based on the calculated values that indicate transition likelihood for all possible connections on a page, we can assign a transition probability to the page and regard it as the link weight in the Markov chain. We then use the same processing procedure as the original PageRank to calculate the principle eigen-vector of the transition matrix. The link allocation method is shown in the following Equations. Parameter γ is used for adjusting the probability that the surfer tends to follow those links with literal matching information. In this paper, the optional value of γ will be determined from the experiment results. The calculating equation is shown as follows:

$$
\begin{array}{l}
LMBLA\left(j\right) = \\
\left(1-\lambda\right)1/N + \lambda \sum\limits_{i \in B_j} LMBLA\left(i\right) prob\left(i \to j\right) \\
\quad B\left(i\right):\ set\ of\ pages\ which\ link\ to\ page\ i; \\
prob\left(i \to j\right) = \left\{ \begin{array}{l} \frac{\gamma \times TranOdds(i \to j)}{\sum\limits_{k \in InforLink(i)} TranOdds(i \to k)} \\ \frac{(1-\gamma)}{\#referLink(i)} \end{array} \right. \\
where: \\
\quad \gamma\ represent\ the\ probability\ that\ transition \\
\quad follows\ \text{informative}\ link. \\
\quad InforLink\left(i\right) = \left\{ j | LinkType\left(i \to j\right) = 1 \right\} \\
\quad referLink\left(i\right) = \left\{ j | LinkType\left(i \to j\right) = 0 \right\} \\
\quad LinkType\left(i \to j\right) = \left\{ \begin{array}{l} 1,\ if: A \wedge B \wedge C \\ 0,\ otherwise \end{array} \right. \\
\quad\quad \left\{ \begin{array}{l} A: VD\left(j\right) \neq \emptyset;\ \text{B: Page (i)} \neq \emptyset; \\ C: \left\{ w | w \in \left(VD\left(j\right) \cap Page\left(i\right)\right) \right\} \neq \emptyset \end{array} \right.
\end{array}
$$

2.2.5 Rank Adjuster

Rank adjuster is performed on the top 1000 return results which are obtained through the relevance-ranking function, and the adjustment is based on the query-independent score computed by our proposed link analysis model. Two kinds of rank-adjusting schemes are implemented in our model. The first is based on a simple linear score combination, denoted as RA1:

$$
\begin{array}{l}
For\ a\ given\ query\ Q\ and\ return\ doc.\ sets\ R: \\
RA1\left(Q, P_i\right) = SMRF\left(Q, P_i\right) \\
\qquad\qquad + \eta \times \frac{\log(LMBLA(P_i) \times N)}{\log(1.89)} \\
\quad where:\ P_i \in R
\end{array}
$$

In the RA1, the influence of the query-independent score on final page rank score is processed according to the in-degree distribution of NTCIR-4 Web pages, which follows the power law distribution [2] and the power value is 1.89.

In the rank adjuster model 2, termed as RA2, the relevance score obtained by SMRF is smoothed by the rank of both the query-dependent sequence and the query-independent sequence. There are two rank lists, one sorted by SMRF scores and the other sorted by LMBLA scores. We assume that the higher the summation of the two rank values of a page is, the lower the score assigned to that page. The greater the difference is between the two ranks of a page, the smaller the adjustment made on the page relevance score. The calculation equation of RA2 is as follows:

$$
\begin{array}{l}
For\ \text{a given query Q and return doc sets R:} \\
\quad \tau_1:\ document\ in\ R\ sort\ by\ SMRF\ score \\
\quad \tau_2:\ document\ in\ R\ sort\ by\ LMBLA\ score \\
\quad \tau_k\left(i\right):\ rank\ of\ i\ in\ \tau_k \\
RA2\left(P_i, Q\right) = SMRF\left(P_i, Q\right) \\
\qquad\qquad - \eta \times \frac{\tau_1(P_i) + \tau_2(P_i)}{|\tau_1(P_i) - \tau_2(P_i) + 1|} \\
\quad where: P_i \in R
\end{array}
$$

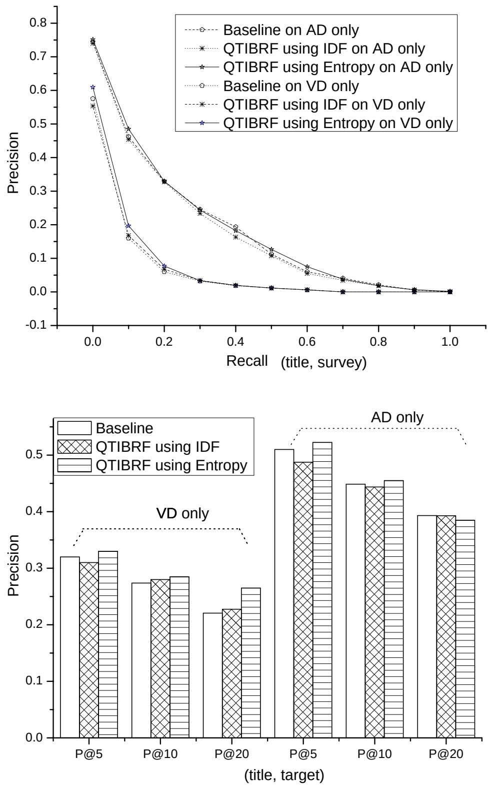

Figure 2. Comparison results of relevance ranking using term entropy, IDF and baseline for searching on AD only and VD only. The upper is for topic "title" using "Survey" type. The bottom is for topic "title" using "Target" type.

3 Experiment results and analysis

Our experiments are conducted for an Information Retrieval Task [5], which refers to ”Survey Retrieval Task” and ”Target Retrieval Task”. Two kinds of user models are employed for evaluations: user model U1, where a user attempts to comprehensively find documents relevant to his/her information needs, and user model U2, where the user requires just one or only a

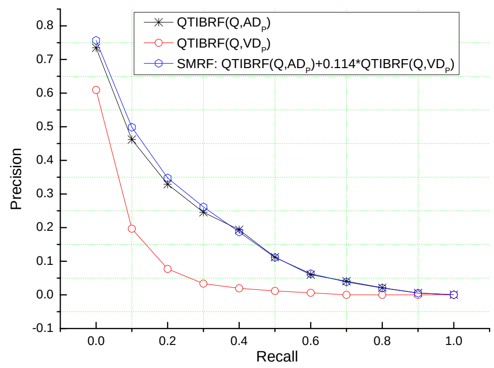

Figure 3. Comparison of recall precision between SMRF and QTIBRF

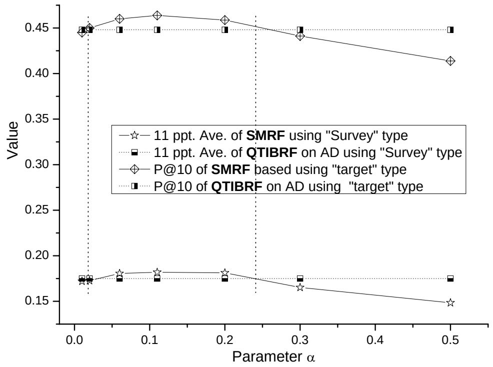

Figure 4. Comparison of precision obtained by SMRF under different α

few relevant documents. Corresponding to these two user models, two kinds of relevant judgment files are provided by task organizers. One is termed the ”Survey” type where 35 queries with 5,674 correct answers (relevant to a particular query) are included and the other is denoted as the ”Target” type, where 80 queries with 9,244 correct answers are contained. According to the organizer description on user models, in our experimental system, for the ”Survey” type, the 11 ppt. average precision is adopted to measure IR efficiency, while for the ”Target” type, precision of the top returned documents is used as the evaluation metric. As for the query data, the ”Title” in a topic is used in this paper because it is similar to the query terms used in real Web search engines.

3.1 Experiment Using QTIBRF

To investigate the effectiveness of the proposed ranking function using the term entropy in VD space, in addition to the baseline experiment without query-term importance consideration, the ranking function using query-term weighting by its inverse document frequency (IDF) [13] in VD space is performed as well. IDF is usually applied to represent the ability of a term to be discriminated in a set of documents. The weighting method is denoted as: $QTW(w) = VDIDF() = \log \frac{N}{VDDF(w)}$. The comparison results for topic ”title” using ”Survey” type and ”Target” type evaluation are both plotted in Fig. 2. It shows that the term entropy based ranking function achieves the best results when searching AD and VD collection respectively. Considering the characteristics of IDF-based weighting, the underlying assumption is that a term is uniformly distributed among pages, while entropy-based weighting will take this into account. For the VD of a given page, this contains the anchor text of its inbound links, and each of its inbound links carries the same description data as the others. Such web conventions give a particular term in a virtual document high term frequency. Therefore, benefiting from the consideration of term distribution in VD space, the weight of a term attained from its entropy value is more representative for the importance of a term than from IDF. What is more, we note that searching the AD space using QTIBRF did not add improvements in precision at the top 20 return documents based on the ”Target” type evaluation. This indicates that the QTIBRF model is more adaptable for improving the results of searching on VD space.

3.2 Experiment Using SMRF

We continue to investigate whether the combination of the two scores obtained through the QTIBRF model for both VD and AD collection can provide more improvement than any single method of searching. From another standpoint, we are trying to determine whether the new information unit, VD, can boost the precision of normal full-text searching. Comparison tests were done and the results for both the ”Survey” type and the ”Target” type are shown in Fig. 5. For the ”Survey” type evaluation, the comparisons of precision at each recall level are plotted in Fig. 3. This shows that the distinct improvements obtained through the combined scores came out at recall to be less than 0.3 and there are no clear differences at higher recall level. Such phenomena can show that VD-based searching could not find more relevant pages at a recall over 0.3 (precision near 0). Similarly, the contribution from improving the merging score is much smaller or nothing. Experiments to analyze the most suitable parameter used in our merging function were also performed. The results plotted in Fig. 4 show that the SMRF model, using a suitable parameter from 0.05 to 0.25 can obtain greater precision than QTIBRF performed on AD alone. Such wide range of suitable parameters indicates that our merging scheme is robust, and that searching using our introduced VD space could help boost searching precision in actual Web page collection. Through our reviews of experimental results, the

Coll. Space	Model	"Title", "Survey" type	"Title", "Target" type		
		11 ppt Ave precision	P@5	P@10	P@20
AD	Baseline	0.1739	0.5010	0.4487	0.3931
VD	Baseline	0.049	0.3200	0.2738	0.2206
AD	QTIBRF	0.1753	0.5100	0.4550	0.3850
VD	QTIBRF	0.0551	0.3300	0.2850	0.2431
AD+VD	SMRF, $\alpha = 0.114$	0.1826	0.5275	0.4625	0.3985

Figure 5. Comparison results of QTIBRF and SMRF

Model		Topic "title", "Target" type		
		P@5	P@10	P@20
SMRF ($\alpha = 0.114$)		0.5275	0.4625	0.3985
RA1	η =0.01	0.5125	0.4637	0.3950
	η =0.02	0.5075	0.4625	0.3956
	η =0.001	0.5150	0.4625	0.3950
RA2	η =0.01	0.5150	0.4637	0.3975
	η =0.02	0.5150	0.4650	0.3994
	η =0.03	**0.5150**	**0.4685**	**0.3996**
	η =0.04	0.5125	0.4662	0.3994

Figure 6. Results of comparison between SMRF and re-rank using LMBLA

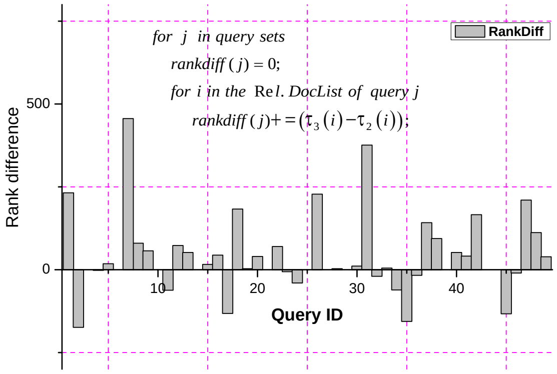

Figure 7. Rank difference of relevant files for each query

case of a parameter equal to 0.114 generates the highest efficiency. Therefore, we chose it for our comparison.

3.3 Experiment Using LMBLA

First of all, we would like to present our experiment for evaluating the effectiveness of our proposed link analysis model. The simple direct comparison was performed based on rank value of the right answer (relevant files) in the result sets between PageRank and our approach. Among the 294 relevant documents of the 35 random selected queries from the "Target" type relevant judgment file, we got 182 "WIN", 97 "FAIL" and 15 "EQUAL". The rough metric is defined as:

$$
\begin{aligned}
&R:\ return\ document\ sets\ for\ a\ given\ query\\
&\tau_2:\ document\ in\ R\ sort\ by\ LMBLA\ score\\
&\tau_3:\ document\ in\ R\ sort\ by\ original\ PageRank\ score\\
&\tau_k(i):\ rank\ of\ i\ in\ \tau_k\\
&\begin{cases} WIN: \tau_3(i) > \tau_2(i) \\ FAIL: \tau_3(i) < \tau_2(i) \\ EQUAL: \tau_3(i) = \tau_2(i) \end{cases} \quad , i \in R
\end{aligned}
$$

To observe the difference in degree clearly, the plot is based on the summation of rank difference, $\tau_3(i) - \tau_2(i)$, of right answers for each query as shown in Fig. 7. It shows clearly that the rank value based on LMBLA is less than that of the original PageRank in most cases. Therefore, our proposed model will possibly works better than the original PageRank for NTCIR-4 Web data.

PageRank is well used to adjust the top rank sequence as a kind of query-independent evidence [15]. For the same reason, our system uses the LMBLA score as a kind of query-independent evidence. Evaluating the usefulness of query-independent evidence in boosting search results is complicated by the need to combine the query-independent score with the query-dependent score. There is a risk that a spurious negative conclusion could result from a poor choice of combining functions. In this paper, two kinds of tentative adjustment experiments are performed. Our aim is to gauge the possibility of improvements in searching results using our proposed link analysis model for NTCIR Web data. Because rank adjustment is performed on the top 1,000 return sets to boost the precision of search results, the precision at the top 5, 10 and 20 documents using the ”Target” type is used for the evaluation. The experimental results using two re-ranking methods based on several parameters are shown in Fig. 6. Although there are no clear improvements from using RA1, we obtain modest improvements using the RA2. The slight improvements and small suitable parameter indicate that the query-dependent relevance score is much more important than query-independent evidence for final search results of the Information Retrieval Task. Such indications are consistent with the presentation of Information Retrieval task from the overview of Task organizer [5]. They presented that the Information Retrieval Task is similar to traditional ad-hoc retrieval at the point of focusing on the topical relevance. Because our proposed LMBLA model shows its advantage over original PageRank, it is reasonable for us to expect that the LMBLA might possibly do a good job for ”Navigational Retrieval Task”, another Sub-Task in NTCIR-4 Web. What is more, from our re-ranking experimental results, it indicates that the rank information of both lists (document lists sorted by relevance score and link analysis score respectively) might be a useful factor for rank adjusting scheme.

4 Conclusions

In this paper, a new information unit, the VD, provides an important information resource for our proposed ranking functions, and ranking functions using VD in addition to AD work well in our Web searching system for NTCIR-4 Web Information Retrieval Task. We conclude that: query-term weighting using term entropy on VD space is effective for improving search results. The combination of evidence from both actual and virtual documents can improve search results beyond the use of either information source alone. As for our proposed LMBLA model, although only modest improvements were obtained through re-ranking experiments, it shows advantages over the original PageRank and we are currently experimenting with it to perform retrieval, which will hopefully yield much greater effectiveness in the future.

5 Acknowledgments

We would like to thank the organizers of the NTCIR Web Task for their efforts in constructing this test data. This research was supported by the 21st Century COE program ”Intelligent Human Sensing”.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks*, pages 309 – 320, Amsterdam, 2000.

[3] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, San Francisco, 2003.

[4] W. B. Croft. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, BOSTON, 2003.

[5] K. Eguchi, K. Oyama, AkikoAizawa, and H. Ishikawa. Overview of the information retrieval task at nctir-4 web. In *Working Notes of the Fourth NTCIR Workshop Meeting*, Tokyo, 2004.

[6] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. System evaluation methods for web retrieval tasks considering hyperlink structure. In *Proceeding of 12th Internaltional World Wide Web Conference (WWW2003*, page 344, Budapest, 2003.

[7] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 459 – 460, Toronto, 2003.

[8] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, and G. Flake. Using web structure for classifying and describing web pages. In *International World Wide Web Conference*, pages 562–569, Hawaii, May 7–11 2002.

[9] R. Jin, A. G. Hauptmann, and C. Zhai. Title language model for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42 – 48, Finland, 2002.

[10] H.-Y. Kao and S.-H. Lin. Mining web information structures and contents based on entropy analysis. *IEEE Transaction on Knowledge and Data Engineering*, 16(1), 2004.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[12] R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceeding of the 13th international conference on World Wide Web*, New York, 2004.

[13] C. D. Manning and H. Schutze. *Foundation of statistical natural language processing*. The MIT Press, Cambridge Massachusetts, London, England, 1999.

[14] Y. Matsumoto and A. K. etc. Japanese morphological analysis system chasen version 2.2.1. 2000.

[15] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286 – 313, 2003.