Expansion of Web Search based on Hyperlink Linkage Analysis

Seung Kim∗, TaeSo Lim**, Wookey Lee***, Suk Ho Kang∗

∗Dept. of Industrial Engineering, Seoul National University
**Dept. of Computer Engineering, Sungkyul University
***Dept. of Industrial Engineering, Inha University

Abstract
A conventional Web search engine only can retrieve Web pages that contain whole given query keyword. So, when the popularity of the given query keyword is low or, many numbers of the query keywords are given, query results can be degraded abnormally. Also, even though partial query result pages exist in a short link structure, a conventional search engine may not retrieve these pages because these pages not contain query keywords in their own unit single pages.

In this paper, we propose a framework that web search result is presented as Web pages and their link structure (Web pages networks), not as a unit single web pages. For the purpose of this end, we propose an algorithm that explores the link structure among the partial search result pages. And we propose a new ranking scheme of these search result networks. We expect this proposed framework can also treat a multi-topic query effectively.

1. INTRODUCTION

During the web search, a user may suffer with some sort of difficulties. These difficulties can be cleared in the viewpoint of IR(Information Retrieval) performance measure and query type. Kleinberg [1] has classified web search query type as the specific, broad-topic and similar-page queries.

A specific query is a kind of that well enough qualified so query results tend to be specific, precise and narrow. On the other side, a broad topic query is a loosely qualified so query results tend to be broad and wide range. In the case of a specific query is given, search engine may retrieve a more precise result compared to the other case. But, the possibility of a low recall result may be also enlarged. So, this type of query may lead to a scarcity problem. On the other hand, a broad-topic query may lead to an abundance problem. In this case, too high recall result may be obtainable. So, plenty number of the query result pages may be given to a user and he/she may suffer with selection problem among the result pages. In company with a scarcity and a abundance problem, the conventional search engine based on VSM(Vector Space Model) can’t handle with a similar-page query also.

In most cases, hyperlinks encode a considerable amount of latent human (web master) judgment. In fact, the creators of Web pages create link to other pages, usually with an idea in mind that the linked pages are relevant to the linking pages. Therefore, a hyperlink, if it was reasonably created, reflects the human semantic judgment and this judgment is objective and independent of the synonymy and polysemy of the words in the pages. This latent semantics, once revealed, could be used to find deeper relationships among the pages, as well as to find the relevant pages for a given page [2].

Abovementioned problem such as a scarcity problem is mainly caused by the fact that the most of the search engines which currently used are try to find out only some Web pages related with search queries as a search result. These pages are subject to include a whole part of the query words. But, in case of that query words are spread out among the shortly linked pages, current search engine can’t retrieve such shortly linked pages as the results.

Also, this insufficient capability of the search engine is revealed in its result page ranking too. If a certain page includes some part of the query words and if it has a short hop linked pages with the other part of the query words, it must be higher relevance more than otherwise one. But current search engine can’t distinguish this difference. Because, searching unit of the search engine is always only a single page.

The motivation of our research is that the web searching unit must be enlarged from a single web page to a web page set which is linked together from web pages. For the sake of this ends, we propose a rich web search mechanism based on linkage analysis. Under this mechanism, a user can be provided a search result as page sets not single pages. Also, more precise search result ranking can be possible.

2. FORMAL NOTATION
In this paper, we borrowed the idea from the *Newton's law of universal gravitation*. The law of universal gravitation states that the force of gravity is proportional to the mass of an object and inversely proportional to the square of the distance between the objects. Based on this intuition, we can assume a Webpage as an object, the mass of an object as the querying relevance of a Webpage which endowed by a conventional search engine, and the distance between the objects as a link length between two Webpages.

We describe the *gravitational information retrieval* querying model and start with the definitions for the terms and notations used in the rest of the paper:

- The WWW (World Wide Web) is modeled as a directed graph $G(N, A)$, where $N$ is the set of Webpages, $A$ is the hyperlinks connecting these Webpages. (We also define the undirected Web as $G^u(N, A^u)$, where $A^u$ is the same as $A$ except that the edges in $A^u$ are undirected.)
- The dictionary, $D$, is a finite set of keywords that can be used for querying the Web. A query, $Q$, which is a list of keywords, i.e., $Q = \{K_1, K_2, \ldots, K_n\}$. And $Q$ is a subset of $D$, i.e., $Q \subseteq D$.
- There is a keyword-to-page mapping $k$: $K \rightarrow N$, which lists the set of Webpages that contain the keyword. A Search Engine like Google [3] or Altavista [4] is can be treated as a sort of this mapping function.
- There is a mapping, $s$: $N \rightarrow N'$ which selects a certain page from the set of Webpages, $N$.
- Finally, we also assume that there is a cost function, $w$: $G^u \rightarrow \text{real}$, which computes the weight of the Webpages.

We denote the set of Web pages containing a given keyword, $K_i \in D$, as $[R_i]$ called a Rootset, i.e., $[R_i]=k(K_i)$. Furthermore, for a given $[R_i]$, $[R_i]^h$ is defined as the set of Webpages that can reach through a link to (or from) at least one of the pages in $[R_i]$ with a link path of length $h$, for $h$ positive integer.

For example, the trivial case is when $h=1$ such that $[R_i]=[R_i]^1$.

Now, if we can find consecutive minimal links length $h$ among for all $s([R_i]^h)$, $\forall i$, we can construct a sub undirected Webgraph $G^u(N, A^u)$, where $N$ is the set of $s([R_i]^h)$, or $s(k(K_i))$, $\forall i$, and $A^u$ is hyperlinks connecting all $s([R_i]^h)$, $\forall i$.

A conventional search engine returns querying results as Webpages. But, in this paper, we treat this sub Webgraph $G^u$ as a querying result for a given query. Also, this result sub Webgraph $G^u$ should be ranked by using proper cost function $w$. We use gravitational information retrieval querying model, i.e., the weight of the result sub Webgraph $G^u$ is proportional to the weight of a each Webpage and inversely proportional to the square of the hop distance $h$ between each Webpage.

The rest of this paper is organized as follows. In Chapter 3, we present the algorithm that can construct the result sub Webgraph $G^u$ in detail. In Chapter 4, we propose a ranking function for the result sub Webgraphs. In Chapter 5, preliminary experimental results are discussed. Finally, we conclude the paper.

### 3. ALGORITHM

In this paper, we assume that a query is composed of the several keywords, and a document collection set is composed of Webpages that are hypertext linked together. We try to find out a query result as a form of the Webpages and their link structure. We can see a Webpage as a node and a link between Webpages as an arc. So, from now on, we call this Webpages and their link structure as a network. In chapter 3, we propose an algorithm that can be used to find out networks as the given query results. Also, the ranking scheme for the result networks is presented in the following chapter 4.

The algorithm works as follows. First, the algorithm disassembles the given query into the keywords and each keyword is used to retrieve corresponding relevant Webpages set by using a conventional search engine like Google or Altavista. Then, the algorithm searches the link structure among the Webpages in the retrieved Webpages sets. Finally, the relevant Webpages and their link structures for the query are obtained as the result of the query.

The document collection set is constructed by following procedures. First, the query is disassembled into $n$ query keyword $K_1, \ldots, K_n$ ($n$ is a size of query terms). And, each keyword, $K_1, \ldots, K_n$ is used to obtain a each relevant page set by using a conventional search engine. The number of the pages in this