KOMOS – A Keyword Occurrence Method for Ordering Search Results

Poomagal S
Lecturer
Department of Information Technology
PSG College of Technology
Coimbatore
Tamilnadu, India

Hamsapriya T
Professor
Department of Information Technology
PSG College of Technology
Coimbatore
Tamilnadu, India

ABSTRACT
Information retrieval is the most common objective of internet users. For a given query, existing search engines will return a long list of search results ranked by their relevance to the given query. In most of the existing search engines, Page Rank algorithm is used to find the relevance of the documents to the given query. The problem with this method is that it considers only the incoming and outgoing links and the popularity of the document. The proposed method eliminates this problem by considering occurrence of the given term in domain name, URL, TITLE tag, META tag and inside the document in addition to the popularity of the document. It makes the relevant documents to have higher ranking.

Keywords
Page Rank, Term frequency, Domain name, URL .

1. INTRODUCTION
Information from the web can be accessed by means of a search engine. When the user enters a query, search engines will return list of titles and snippets from the result pages according to the rank of each result. With the vast amount of information available online, searching the relevant document requires the user to go through many titles and snippets. These titles and snippets are available in the title tag and meta tag and returned from the corresponding pages. Ordering of these titles and snippets returned by existing search engine is computed by the Page Rank algorithm.

Page Rank algorithm has proved to be a very effective algorithm for ranking results of web search algorithms. The ordering produced by this algorithm creates a problem of finding the relevant documents in a returned list of documents. Since the algorithm calculates ranks according to the links, frequency of access. For some documents the frequency of access may be small since its presence is at the end of the list of returned result documents but it may be relevant to the given query. The problem of this less frequency of access arises since the user of the search engine usually has the patience to see only the first 30 or 40 pages from the returned results.

The solution to this problem is to rank documents according to the occurrence of a keyword in domain name, URL, title tag, meta tag and inside the document in addition to the popularity of the documents. Normally, document relevant to the given query will have the keyword in the root domain name, URL, TITLE tag and META tag and have high term frequency. This method orders the search results so that most relevant documents are placed at the beginning of the list. The user can select the needed document quickly so that the search time can be minimized.

The remainder of this paper is organized as follows: Section 2 gives an overview of the related work. Section 3 explains the steps involved in the proposed method. Section 4 presents the results obtained. Section 5 concludes this paper and future work is also mentioned.

2. RELATED WORK
In the past, Ordering of search results was done by using the number of links. It has been observed that web information tends to have a hierarchical structured reflected in URLs [1, 2, 3, 4]. This is due to the fact that many web servers simply export a file system, and it has become a common practice for humans to group related files together in a single directory, and for administrative delegation of authorship to be done at the directory level. This hierarchical structure is even present on servers whose content is not stored in a hierarchical file system, as the URL standard was originally designed to incorporate hierarchical structure when it is natural to do so [5, 6].

In [2] it was further observed that URLs can often be grouped into "compound documents" that represent a single unit of information, and that such compound documents tend to consist of URLs that agree up to the last '/' character. Thus it is perhaps natural to group together URLs that agree up to the last delimiter as a single information node, and construct a DirRank graph. The nodes in this case correspond to URL prefixes up to the last '/' character (or other delimiter), and there is an edge from one node to another if there is a link from a URL in the source virtual directory to a URL in the destination virtual directory. In effect this groups URLs at a finer level of granularity than entire hostnames, but still often conforms to a human-designed hierarchical organization of information. The "dynamic" URLs containing a '?' character tend not to follow this hierarchical organization of information, but such URLs are often an indicator for the existence of an underlying database capable of serving an enormous number of URLs, and these are often still closely related to each other and it is natural to group them. Thus we can extend the hierarchical grouping to include all URLs that agree up to the last '/' or the last '?', whichever occurs first. In our crawl of a billion URLs, over 79% contained five or fewer URLs, and over 87% contained fewer than ten URLs. Whenever a directory contains a large number of URLs, it tends to be either the archive of a mailing list or else a schema-driven set of dynamic URLs. In this also only the links and dangling links are considered for ranking. Entropy maximization can also be used for ranking documents [7, 8]. In [9], block rank algorithm is used to find the rank based on the block structure of a web.

Google Page Rank is based on the number of links google has found pointing to a certain web page. Links can be from other web sites as well as from other pages in the same site. The rank varies from 0 to 10. 0 means less (or zero) number of incoming links and 10 means the highest number of incoming links.

Google Page Rank is assigned to each web page of a site, not only to the main domain. Almost all new sites get 0 rank at the beginning and there are only few sites that have a PR of 10 or 9. Most of the established sites are around PR4.

Google calculates the PR of a web page (say P) [10] using Probabilistic distribution as,

$$PR(P)=(1-d) + d (PR(P1) / L(P1) + (PR(P2) / L(P2) + ...PR(Pn) / L(Pn))$$

where d is called Damping Factor and is usually 0.85**.** PR(P1), PR(P2) etc are the Google Page Ranks of web pages linking to Page-P and L(P1), L(P2) etc are the total number of outgoing links (links to internal web pages or to other web sites) on those pages.

There are several factors for calculating the rank of any page [11]. They are,

1. Keyword focused anchor text external links
2. External link popularity
3. Keyword use in the title tag
4. Keyword use in the root domain name
5. Keyword use in the URL of the page

Literature survey reveals that term frequency in the entire web page is given less importance. Contents in the meta tag is also useful to identify the relevancy of documents since it contains key terms present inside the document. This method considers the occurrence of a keyword in the root domain name, URL, TITLE tag, META tag and entire document.

3. PROPOSED METHOD

The proposed method assigns ranks to the pages based on the term frequency and also the keyword occurrence in root domain name, URL name, title tag and meta tag.

Proposed Algorithm
1. **Extract keywords from the user's query**
2. **Create a list L1 of documents that will contain the keyword in its root domain name**
3. **Create a list L2 of documents that will contain the keyword in its URL**
4. **Create a list L3 of documents that will contain the keyword in its title tag**
5. **Create a list L4 of documents that will contain the keyword in its meta tag**
6. **Create a list L5 of documents that are not included in the above lists L1 to L4.**
7. **Assign ranks to the documents in each list based on the frequency of a keyword inside the documents**

For creating each list, domain name, URL, title tag and meta tag contents are extracted and compared with the keyword. Depending on the occurrence of the keyword in each of the four categories, lists are framed.

The title and meta tag are also considered in addition to the domain name and URL since they have more significance on identifying the relevance of each document to the given query. In web pages, title and meta tags contains the key terms present inside the document.

The documents which does not fall under lists 1 to 4 are grouped under list 5. List 5 will contain only less number of documents as most of the documents will come in any of lists 1 to 4.

Once the lists are formed, ranks are assigned to documents in each list based on the frequency of keyword inside the documents. Keyword frequency is calculated by extracting the terms from each document and comparing those terms with the keyword. Since web documents enclose text contents inside HTML tags, extracting terms is done by removing these tags, Regular expression is used for this purpose. Stopwords are removed from the extracted terms and stemming is also performed.

3.1 Illustration of the Proposed Algorithm
Keywords are extracted from the query and the list creation is done based on the keyword occurrence as follows,

$L1 = d_i$ where $1 \leq i \leq n$ and t is in $DN(d_i)$
$L2 = d_i$ where $1 \leq i \leq n$ and t is in $URL(d_i)$
$L3 = d_i$ where $1 \leq i \leq n$ and t is in $Title(d_i)$
$L4 = d_i$ where $1 \leq i \leq n$ and t is in $Meta(d_i)$
$L5 = d_i$ where $1 \leq i \leq n$ and d_i not in other lists
where,
n = Number of search results
t = Keyword given in the query
$DN(d_i)$ = Domain name of the ith document
$URL(d_i)$ = URL of the ith document
$Title(d_i)$ = Title tag of ith document
$Meta(d_i)$ = Meta tag of ith document

After the list creation, ranks are assigned to documents in the following way. For instance the number of documents in each list is taken as,

List 1 = L
List 2 = M
List 3 = P
List 4 = Q
List 5 = R

Table 1 shows the sample number of documents in each list and the ranks assigned to it.

List Number	Number of documents	Rank assigned
1	L	1 to L
2	M	L+1 to L+M
3	P	L+M+1 to L+M+P
4	Q	L+M+P+1 to L+M+P+Q
5	R	L+M+P+Q+1 to L+M+P+Q+R

Table 1 : Ranks assigned to each list

Pseudo-code of the proposed method

```
1: D  ← input documents
2: for all d ε D do
3: if the keyword occurs in domain name
4:       append the document into list1
5: else if the keyword occurs in URL name
6:       append the document into list2
7: else if the keyword occurs in title tag
8:       append the document into list3
9: else if the keyword occurs in meta tag
10:      append the document into list4
11: else
12:      append the document into list5
13: end if
14: end for
```

```
15: x=0
16: for each list from 1 to 5 do
17: i=0
18: for each d ε D do
19: count=0
20:  for each term t ε d
21:  compare each term with the stop
         words list
22:    if any match found then
23:       remove the stop words;
24:    end if
25:   apply Stemming Algorithm
26: if keyword ==extracted term then
27:       add 1 to count
28: enf if
29: end for
30: freq[i]=count
31: i=i+1
32: end for
33: for i= x to |d| do
34:  Assign ranks to the documents
        according to freq[i] from x to |d|
35: end for
36: x=x+|d|
37: end for
38: end for
```

3.2 Extracting Terms

Text documents are different from web documents. Web documents are unstructured. In addition to text contents it contains tag contents. These contents must be removed to extract terms from documents.

Given a character sequence and a defined document unit, tokenization is the task of breaking into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization:

        Input: Friends, Romans, Country men, lend
                 me your ears;
        Output: Friends Romans Country men lend
                 me your ears

These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence. A term is a type that is included in the IR system's dictionary. The set of index terms could be entirely distinct from the tokens, for instance, they could be semantic identifiers in a taxonomy, but in practice in modern IR systems they are strongly related to the tokens in the document.

Contents from the tags are extracted by removing the tags with the use of regular expressions. After extracting the terms from the documents, the terms can be analyzed to remove stopwords and the remaining terms are given to stemmer.

3.3 Stop word removal and Stemming

Stopwords does not provide any useful information to identify the relevancy of documents. So they can be removed to avoid confusions. Some common stopwords are is, was, are, were, what etc., For stopword removal, first the database of stopwords will be created and the terms extracted from the web pages will be compared with the stopwords list. Stopwords found in the extracted terms are removed.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am,are,is	$\Rightarrow$ be
car, cars, car's, cars'	$\Rightarrow$ car

The result of this mapping of text will be like:

the boy's cars are of different colors $\Rightarrow$ the boy car be of differ color

However, the two words differ in their flavor. Stemming usually refers to a heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. If confronted with the token saw, stemming might return just *s*, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun.

For stemming, Porter's algorithm is used. Stemming Algorithm is a simple utility that reduces English words to their word stems - without the "ing", "ings";"s" etc.,

Table 2 shows the stemmed words for the query "cloud computing".

redplaid, cloud, computing, oracl, weblogic, comput, easy, google, chelsio, communications, host, free, trial, wikipedia, alpha, wmf, definit, whati, word, http, sdefinit, sid, gci, html, noodp, concept, articl, describ, impact, variety, compani, countri, industri, refer, amazon, amzn, blackbaud, blkb, blackboard, bbbb, concur, technolog, cnqr, geographi, page, googl, goog, hp, intel, intc, intern, busi, machin, ibm, new, result, salesforc, forc, platform, leader, servic, instant, time, valu, complet, independ, noydir, really, mean, infoworldth, big, trend, sound, nebul, fuzzy, view, proposit, perspect, profession, index, follow, web, app, engin, applic, architectur, infrastructur, soa, orient, softwar, sun, microsystem, open, sourc, discov, brought, world, popular, starlight, howstuffwork, quot, work, smart, essenti, futur, let, file, internet, learn, benefit, drawback, computingbarry, warwick, raleigh, ibmu, public, v, templat, gener, corpor, copyright, expo, sy, media, journalsy, computingcloud, n, june, stamford, ct, youtub, explain, computingexplain, video, produc, present, christoph, barnatt, author, explainingcomput, associ, professor, compu, utility, saa, haa, explainedcloud, confus, term, want, talk, technology, buzzword, boil, section, ec, gogrid, netflix, gmail, server, dedic, vmware, virtualizationturn, datacent, secur, privat, virtual, add, capacity, demand, deploy, extern, built, vm, ware, gklqyx, kxtfwei, optim, data, center, lab, consolid, lifecycl, provid, continuity, pc, solut, dell, increas, capabl, flexibility, choic, eiwatch, content, enterpris, aspx, ng, publish, elast, ecamazon, aw, deliv, set, form, reliabl, scalabl, inexpens, pay, includ, simpledb, sq, fp, simpl, storag, grid, mechan, turk, queue, flexibl, payment, devpay, alexa, develop, forum, user, group, tool, api, unveil, centr, india, nist, gov, security, divis, csrc, websit, involv, differ, project, webpag, base, visit, resourc, clearinghous, event, driver, mobileag, pki, encrypt, ae, mode, ipv, alert, virus, organ, link, mission, staff, itl, divisiion, network, policy, manag, oper, solutions, ubuntu, plain, english, common, craft, product, computerworld, knowledg, voic, inform, c, right, reserv, action, command, viewpag, amp, pagepath, polici, eng, document, faq, demystifi, worldnetwork, bring, great, promis, industry, key, question, answer, level,

agreement, supplement, jon, brodkin, schneier, onlin, rackspac, mosso, power, site, support, live, chat, today

Table 2 : Stemmed words

3.4 Term Frequency Calculation

Term frequency denotes the number of times the given keyword occurs in each document. This is calculated by using a loop. The terms extracted are compared with the keyword and if both are equal then the frequency count is incremented.

4. EXPERIMENTAL RESULTS

For experimentation, first 200 results returned by an existing search engine is taken and rearranged. For the first 20 documents, the ordering produced by the proposed method is somewhat similar to the ordering produced by an existing search engine since documents with high relevancy tend to appear in the 20 results. So the documents from 21 to 40 produced by both the methods are shown below.

Results (21 to 40) generated by an existing search engine for the query "apple"

TUAW -- The Unofficial **Apple** Weblog
16 Dec 2009 **...** Comments on Mac products and events, and plenty of tips and advice. www.tuaw.com/ - Cached - Similar
Apple -- TUAW
Apple today released a new digital RAW compatibility update for some Canon and Nikon cameras that shoot photos in RAW format. **...** www.tuaw.com/category/**apple**
MobileMe
Access and manage your email, contacts, calendar, photos, and files on the web at me.com. Not a member? You can try MobileMe free for 60-days.www.me.com/mail/ - Cached
Top **Apple** News, Videos and Images
Top **Apple** News, Videos and Images as voted on by the Digg Community. Discover the latest news, breaking news and popular news.digg.com/**apple** - Cached - Similar
Apple Inc. - Google Finance
More Upside in **Apple** Inc. (AAPL) - Braunie - The Market Guardian | Monitoring The **...** Stock Market Analysis: **Apple** Inc AAPL Stock Analysis - 12/16/09 **...**www.google.com/finance?q=NASDAQ:AAPL - Cached - Similar
Battery Exchange Program iBook G4 and PowerBook G
National and regional resellers, catalogers, and **Apple's** on-line and retail stores **...** Or, you can use **Apple** System Profiler to confirm your serial number. **...**www.support.apple.com › Support - Cached
Apple - Support - Downloads
Need to update your **Apple** software? Learn how to use Software Update on your **...** Learn about reporting security issues to **Apple**, checking security on your **...**support.**apple**.com/downloads/ - United States - Cached - Similar
Show more results from support.apple.com
Apple II History Home
Welcome to the **Apple** II History! The core of information presented here was originally a series of articles printed in the early 1990's in the Metro **Apple ...**apple2history.org/ - Cached - Similar
Apple II History Chap 3
16 May 2003 **... Apple** II standard No hacker is completely satisfied with a project he is working on as long as it is possible to make improvements that make **...apple**2history.org/history/ah03.html - Cached - Similar
Show more results from apple2history.org
apple-history.com

A brief history of the company that changed the computing world forever. Look in the gallery for a description and picture of every **Apple** computer ever **...**www.**apple**-history.com/ - Cached - Similar
Welcome to the **Apple** Store - **Apple** Store (U.K.)
Perfected. Make it uniquely theirs — only at the **Apple** Online Store. **....** You can also order from The **Apple** Store by calling 0800 048 0408. **...**store.**apple**.com/uk - United Kingdom - Cached - Similar
Find Your School - **Apple** Store (U.S.)
Apple offers special education pricing on **Apple** computers, software and select third party products to College Students, or Teachers, Administrators and **...**store.**apple**.com/us-hed/findyourschool?node=home/... - United States - Cached
Show more results from store.apple.com
AppleInsider | **Apple** Insider News and Analysis
14 Dec 2009 **...** Though previous studies found the iPhone was outselling Windows Mobile devices and had a greater presence online, **Apple's** handset finally **...**www.**apple**insider.com/ - Cached - Similar
FT.com / Markets data / AAPL:NSQ: Performance for **Apple** Inc
Get detailed information on **Apple** Inc (AAPL:NSQ) including quotes, charts, earnings details, company background, key stats, financial statements, **...**markets.ft.com/tearsheets/performance.asp?s=us:AAPL - Cached - Similar
apple - Slashdot: News for nerds, stuff that matters
FruitWorm writes in with word of a vulnerability in Java that has been patched by everyone but **Apple**. "Security researchers say that Mac OS X users are **...**slashdot.org/**apple**/ - Cached - Similar
The Power][be your Best!
The "**Apple** II Information and Pictorial Reference", featuring detailed information and pictures of all kinds of s**Apple** II hardware, including several rare **...**www.**apple**2.org/ - Cached - Similar
The **Apple** Collection, Desktop Pictures, Mac Links, Bookstore and **...**
6 Dec 2009 **...** Every thing about **Apple** Collection all designed to be fun and to add a bit of color to your browsing. Best Mac Desktop Pictures. www.the**apple**collection.com/ - Cached - Similar
Green my **Apple** campaign - Greenpeace
We love **Apple**. **Apple** knows more about "clean" design than anybody, right? **...** That's why we **Apple** fans need to demand a new, cool product: a greener **Apple**. **...**www.greenpeace.org/**apple**/ - Cached - Similar
APPLE INC (AAPL:US): Stock Quote & Company Profile - BusinessWeek
15 Dec 2009 **...** Get **APPLE** INC (AAPL:US) stock research & investing information. **...** SmarTone Telecommunications Holdings Ltd. Inks Deal with **Apple** to Sell **...**investing.businessweek.com/research/stocks/.../snapshot.asp?... - Cached - Similar
APPLE INC (AAPL): Stock Quote & Company Profile - BusinessWeek
The company, formerly known as **Apple** Computer, Inc., was founded in 1976 and **... Apple** Inc. Releases Two Updates, Java for Mac OS X 10.6 Update 1 and Java **..**investing.businessweek.com/businessweek/research/.../snapshot.asp?... - Cached

Results(21 to 40) generated by the Proposed method for the query "apple"

apple - Slashdot: News for nerds, stuff that matters
FruitWorm writes in with word of a vulnerability in Java that has been patched by everyone but **Apple**. "Security researchers say that Mac OS X users are **...** slashdot.org/**apple**/ - Cached - Similar

The Power][be your Best!

The "**Apple** II Information and Pictorial Reference", featuring detailed information and pictures of all kinds of **Apple** II hardware, including several rare **...**
www.**apple**2.org/ - Cached - Similar

The **Apple** Collection, Desktop Pictures, Mac Links, Bookstore and **...**

6 Dec 2009 **...** Every thing about **Apple** Collection all designed to be fun and to add a bit of color to your browsing. Best Mac Desktop Pictures.
www.the**apple**collection.com/ - Cached - Similar

Green my **Apple** campaign - Greenpeace

We love **Apple**. **Apple** knows more about "clean" design than anybody, right? **...** That's why we **Apple** fans need to demand a new, cool product: a greener **Apple**. **...**
www.greenpeace.org/**apple**/ - Cached - Similar

Find Your School - **Apple** Store (U.S.)

Apple offers special education pricing on **Apple** computers, software and select third party products to College Students, or Teachers, Administrators and **...**
store.**apple**.com/us-hed/findyourschool?node=home/... - United States - Cached

Apple (**Apple**) on Twitter

apple dot twitter at gmail dot com. **...** **Apple** is using Twitter. Twitter is a free service that lets you keep in touch with people through the exchange of **...**
twitter.com/**apple** - Cached - Similar

Apple info29 Oct 2009 **...** A profile of **Apple**, featuring data from the NW 200 survey and the latest breaking news on **Apple**.
www.networkworld.com › Company profiles - Cached - Similar

Apple Tablet To Redefine Newspapers, Textbooks and Magazines **...**

30 Sep 2009 **...** Steve Jobs said people don't read any more. But **Apple** is in talks with several media companies rooted in print, negotiating content for a **...**
gizmodo.com/.../**apple**-tablet-to-redefine-newspapers-textbooks-and-magazines

The **Apple** Museum | Home

Dedicated to the historical background of **Apple** Computer, Inc. Reviews and specifications of all **Apple** products.
www.the**apple**museum.com/ - Cached - Similar

Apple -- Engadget Mobile

Multifl0w project teaches **Apple** a thing or two about iPhone multitasking **...** **Apple** lets devs bite their nails in real time as iPhone apps navigate approval **...**
mobile.engadget.com/tag/**apple**/ - Cached

Apple Computer Inc. News - The New York Times

20 Oct 2009 **...** News about **Apple** Inc. Commentary and archival information about **Apple** Inc. from The NewYorkTimes.
topics.nytimes.com › Business › Companies - Cached

FIONA **APPLE**

Official website provided by Sony Entertainment. Site features tour information, latest videos and album releases.
www.fiona-**apple**.com/ - Cached - Similar

Community-powered support for **Apple**

7 Dec 2009 **...** View topics about products & services from **Apple**. **...** **Apple** have not resolved Time Capsule internet connection dropping issues **...**
getsatisfaction.com/**apple** - Cached - Similar

Apple

Apple Inc. is a company that develops and markets consumer electronics. It consistently features creative designs and marketing strategies for their **...**
www.mahalo.com/**apple** - Cached - Similar

Apple Pugetsound Program Library Exchange (**A.P.P.L.E**)

A monthly magazine published by the **Apple PugetSound Program Library Exchange** (**APPLE**) user group. Back issues, group news and a product catalog with **...**
www.call**apple**.org/ - Cached - Similar

Newton's **Apple**

Whether its a question about dinosaurs or space travel, robot camp or car engines, this PBS show finds the answer. Has info on episodes, multimedia science **...**
www.newtons**apple**.tv/ - Cached - Similar

Apple's MobileMe drops support for IE 6 - (37signals)

Apple announced a rebranding of their .Mac service earlier this morning. It's now called MobileMe and features online versions of Mail, iCal, **...**
37signals.com/svn/.../1072-**apple**s-mobileme-drops-support-for-ie-6 - Cached

TheAppleBlog

11 Dec 2009 **...** TheAppleBlog, published by and for the day-to-day **Apple** user, is a prominent source for news, reviews, walkthroughs, and real life **...**
the**apple**blog.com/ - Cached - Similar

Big **Apple**

Big **Apple** - A unit of Express retail services Pvt Ltd is Delhi's first ever company owned Fruit, Vegetable and Grocery retail chain. **...**
www.big**apple**fresh.com/ - Cached - Similar

Apple Bottoms is a clothing line that caters to females of all **...**

With stylish and trendy tops, jeans, dresses and jackets; **Apple** Bottoms stands behind it's creed of The clothes should fit a woman, not the woman fit the **...**
www.**apple**bottoms.com/ - Cached - Similar

Documents in the list (21-40) produced by the proposed method does not contain some documents which are available in the corresponding list produced by an existing search engine. The reason for this is that those documents are treated as highly relevance documents and they are thrown to the previous list (1-20). Similarly some documents in this list produced by the proposed method are not available in the list produced by an existing search engine. The reason is that the documents with less relevance are thrown to this list from the set of documents 1 to 20.

Existing method and the proposed method are compared in terms of search time. Time taken to search the relevant document is reduced by using the proposed method since it assigns ranks to the documents in a way such that the relevant documents tend to appear at the beginning of the list of returned results.

5. CONCLUSION

The proposed method presents the search results in a way such that all relevant documents are displayed before other less relevant documents. This method is simple since it considers only 4 criterias and keyword frequency for assigning rank and this is an efficient method because the difficulties with the links are removed. This work concentrates on queries with single keyword. This can also be implemented for the queries with a phrase. In future, Semantics of the keywords can also be considered.

6. REFERENCES

[1] Nadav Eiron and Kevin S. McCurley, 2003, "Locality, hierarchy, and bidirectionality in the web". In *Workshop on Algorithmsand Models for the Web Graph*, Budapest.

[2] Nadav Eiron and Kevin S. McCurley,2003, "Untangling Compound documents on the web". In *Proc. 14th ACM Conf. on Hypertext*, pages 85–94.

[3] Richardson M and Domingos P, 2002, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", volume 14. MIT Press, Cambridge, MA, 2002.

[4] Arasu A, Novak J, Tomkins A and Tomlin J, 2002, "PageRank Computation and the structure of the web: Experiments and algorithms". In Proceedings of the Eleventh International World Wide Web Conference, Poster Track.

[5] T. Berners-Lee, R. Fielding, and L. Masinter. Uniformresource identifiers (URI): Generic syntax.http://www.ietf.org/rfc/rfc2396.txt. RFC 2396.

[6] Mei Q, Shen X and Zhai C, 2007, "Automatic Labeling of multinomial topic models". In Proceeding of the Knowledge Discovery and Data Mining Conference (KDD'07).

[7] John A. Tomlin, 2003, "A new paradigm for ranking pages on the world wide web". In *Proc. 12th World Wide Web Conference*,pages 350–355, Budapest.

[8] Phan X H, Nguyen L M and Horiguchi S, 2008, "Learning to classify short and sparse text and web with hidden topics from large-scale data collections", In Proceedings of the International World Wide Web Conference (WWW'08).

[9] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub, 2003, "Exploiting the block structure of the web for computing pagerank", Technical report, Stanford University.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and TerryWinograd, 1998, "The PageRank citation ranking: Bringing order to the web". Technical report, Stanford Digital Library Technologies Project, Paper SIDL-WP-1999-0120 (version of 11/11/1999).

[11] Monica Bianchini, Marco Gori, and Franco Scarselli. 2005. "Inside PageRank", ACM Transactions on Internet Technology, 5(1):92–128.