Link Analysis to discover relevant documents using Information Retrieval

Hemangini S. Patel
Bhagwan Mahavir College of Computer Application (BCA)
Bharthana, Vesu
Surat, India

Apurva A. Desai
Department of Computer Science
Veer Narmad South Gujarat University
Surat, India

ABSTRACT
In recent year's growth of World Wide Web become faster to cross all expectations; World Wide Web is becoming most valuable resources to information retrieval and knowledge discovery from Web. It is a fertile area for web mining research; an emerging challenge for web mining is the problem of mining richly qualitative documents, where the objects are linked via multiple types of relations. These links provide additional context that can be helpful for web mining tasks. Traditional link analysis treats all hyperlinks equally and makes the assumption that links are endorsement, so there is need to only extract links which are valuable. Unfortunately, this assumption does not incorporate in present World-Wide Web. Hyperlinks are not identical; they may be created in different contexts and for different purposes. By using novel characteristics of web page and hyperlinks to help a search engine focus on relevant and high quality content. The important hyperlink features—topicality —is proposed.

General Terms
Web search, Link Analysis.

Keywords
Web Mining, Web Structure Mining, Topicality, Information Retrieval, Link Analysis, Anchor Text, WWW.

1. INTRODUCTION
Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). Web mining tasks can be classified into three categories: Web content mining, Web structure mining and Web usage mining. An emerging challenge for data mining is the problem of mining richly structured datasets, where the objects are linked in some way. The goal of Web Structure Mining is to generate structural summary about the web site and web page. Web Structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different web sites. Most of the web information retrieval tools only use the textual information, while ignore the link information that could be very valuable [1].

Link mining is a newly emerging research area that is at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. Link mining tasks are broadly categorized into following tasks. They are: Object-Related Tasks: (a) Link-Based Object Ranking,(b) Link-Based Object Classification, (c) Object Clustering (Group Detection) and (d) Object Identification (Entity Resolution); Link-Related Tasks: (a) Link Prediction; Graph-Related Tasks: (a) Sub graph Discovery, (b) Graph Classification, (c) Generative Models for Graph [2]. These tasks are exploring exciting and rapidly expanding of this field as described by Getoor and Diehl [3].

Link mining research holds promise for many different areas, including commercial and business enterprises, personal information management, web search and retrieval, medicine and bioinformatics, and law and security enforcement [3]. With the explosive growth of the WWW, most searches tend to retrieve a large number of Web documents. Information retrieval systems aim to find relevant resources in large collections of documents. Searchers on the Web often aim to find key resources about a topic. Finding such results is called topic distillation. Resources on the World Wide Web are interlinked with hypertext, and so contain a link structure, and each link is accompanied by some descriptive text, called anchor text. This search task was initially defined by authors as aiming to find quality documents on a broad query topic. The task was later adopted and tested within the framework of the TREC (Text REtrieval Conference) Web Track to find key resource pages [4]. TREC Web Track's Topic Distillation Task was introduced in 2002 and one of its objectives was to capture a typical web search, where users consider entry pages to relevant sites as "more valuable than isolated pieces of relevant text". To be more specific, the aim of the task was to identify key resources on a broad topic [5]. Link analysis has been shown to be useful in identifying high quality documents within the hyperlink graph structure of documents. Wu et al. [4] have shown a new approach to improve topic distillation by exploring the use of external sources of evidence: link structure, including query dependent in-degree and out-degree; and web page characteristics, such as the density of anchor links. It was found that anchor density as a source of evidence lead to only small improvements over the baselines. Anchor density provides comparatively weaker evidence as compared to the page characteristics, URL depth and URL length. This might be due to the fact that many organizations have their own web page design patterns like recurring navigation bar at the top or side of the screen, advertising banners. It was suggested that the finer analysis of page structure, which ignores these global navigational links and computes anchor density only on links inserted by the author of a page, may improve the performance of this evidence source. Hence, there is a need to carry out a more detailed analysis of page structure to determine whether more refined models of anchor density can lead to larger improvements in search performance.

2. BACKGROUND

As the web is growing rapidly, the users get easily lost in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfil their needs. Web mining technique is used to categorize users and pages by analysing users behaviour, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge [6].Web search engines uses various sources of evidence to rank web pages matching a user query, most notably textual content and the link structure of the web. The latter is particularly beneficial in helping to address the abundance of relevant documents on the Web by determining the authority of a page based on the links that point to it. Hyperlinks between the world wide documents can be used to determine the relative authority values of documents for various search queries. This process of finding quality pages is called topic distillation.

Topic Distillation first computes a query specific sub-graph of the Web, by including pages on the query topic in the graph and ignoring pages not on the topic. Then the algorithm computes a score for every page in the sub-graph based on hyperlink connectivity. That is, each page is given an authority score, computed by summing the weights of all incoming links to the page. For each such reference, its weight is computed by evaluating how good the referring page is as a source of links. Katz and Li in [7] use a three step approach – (i) Document Keyword Extraction, (ii) Keyword propagation across pages connected by links, and (iii) keyword propagation through category tree structure – to automatically distil topics from the set of documents belonging to a category or to extract documents related to certain topics.

The information from hyper linked Webpages is a rich resource for topic distillation. There is the triangular relationship between the hub web page, anchor text, and authoritative web page. Therefore, a good hub web page is the one that is directed to a good authoritative web page through the authoritative anchor text; an authoritative anchor text is the one that connects good hub web page and good authoritative web page; and a good authoritative web page is the one that a good hub web page directs to through a good authoritative anchor text. Such a mutual reinforcement relationship between hubs and authorities helps the mining of authoritative Web pages and automated discovery of high quality Web structures and resources [8]. The prosper of the World Wide Web presented many new research directions involving link structure analysis. Two independent efforts in the late 1990 that have profound influence on link analysis were Brin & Page's PageRank [9] and Jon Kleinberg's work on HITS [10].

The HITS (Hyperlink Induced Topic Search) algorithm based on mutual reinforcement relationship provides an innovative methodology for Web searching and topics distillation. The HITS algorithm discovers the hubs and authorities of a community on a specific topic or query. A research work on link analysis of hyperlinked documents, HITS was applied to the research area of topic distillation and several kinds of link weights were involved to indicate the significance of links in hyperlinked documents. In the work of Bharat and Henzinger [11], the metrics of similarity of whole contents in linked documents were applied on link weights and the use of text surrounding the links as keyword-based evidence to determine a weight for each link. They found the nepotism problem and the topic drift problem in HITS. The nepotism problem arises because mutually reinforcing relationships between two hosts give undue weight to the opinion of a single person. The topic drift problem is that the most highly ranked authorities and hubs tend not to be about the original topic. If the expanded set contains irrelevant documents to the query topic and those documents are well connected, the topic drift problem arises. Chakrabarti et al [12] combine the TFIDF-weighted model and micro-hub to represent the significance of anchors in regions with information needed. The model gave higher weights to hyperlinks in relevant DOM (Document Object Model) sub-trees than hyperlinks in irrelevant DOM sub-trees to the query topic. This approach reduces topic drift and helps in identifying parts of a Web page relevant to a query. Rafiei and Mendelzon [13] define a new measure called "reputation" of a page and compute the set of topics for which a page will be rated high. Haveliwala [14] proposed a "Topic-Sensitive PageRank", which pre-computes a set of PageRank vectors corresponding to different topics. Choi and Kim [15] analyzed the hyperlink graph structure using hierarchy concept tree to solve the mixed hubs problem that remained in the Bharat's algorithm. They tried to find the relationship in documents connected by hyperlinks using content analysis and assign weights to hyperlinks based on the relationship. They evaluated the algorithm using 50 topics on WT10g corpus and obtained 25 to 46% improvements. The use of anchor text in topic distillation overcomes the drawback of one language for the same web page. It is being a general Multilanguage description method. Experiments on investigating several aspects of anchor text contains following features: 1. An anchor text can be used as a reserved topic to link web page. 2. The number of feature items in anchor text is far less than those in HTML text. 3. There are usually more than one evaluation anchor texts for the same URL. 4. Anchor text, overcomes the drawback that there is only one language version for the same web page [16]. Zhong et al. [17] studied the hypertext-based topic analysis models and community-based topic distillation. They described the THTA (topic hub and topic authority) model and ETHTA (extension THTA) model based on hypertext topic distillation. A comparative analysis was carried out between THTA, ETHTA and TOPHITS. It was concluded that the hypertext-based topic distillation contributed to improving the quality of topic distillation. THTA, ETHTA models are more advanced than HITS model and TOPHITS model.

Lempel and Moran [18] observed that HITS approach is weak to link spam and the TKC (tightly-knit community) effect due to the inter-dependency between hub and authorities, which will push pages within a tightly-knit community to high rankings even though the pages are not relevant. As a solution, they propose a Stochastic Approach for Link-Structure Analysis model (SALSA) to survive from the TKC effect by making the coupling between hubs and authorities much less strict. In SALSA model, a node's authority (hubness) will be distributed among the targets equally during the propagation; while in original HITS, every target always gets the entire authority (hubness) from the node [18].

3. RESEARCH PROBLEM AND HYPOTHESIS

The Web is the major collection of information in the world, which consists of tens of billions of widely visible web documents distributed across millions of web sites world-wide. It is growing drastically with a speed of millions of new pages per day. In order to find useful information from the Web, required a powerful search tool. A modern Web search engine satisfies this need by making the enormous quantities

of data searchable and making the best data findable. Today, searching the WWW has become an everyday task for millions of people around the world, with hundreds of millions of queries issued per day to the search engine.

Due to wide area of web searching for specific query will give large amount of pages or millions of relevant results. Rather than all set of relevant pages; user is being interested in most authoritative pages. Through keyword based matching analysis search doesn't distinguish between authoritative and non-authoritative pages that appear to be relevant. However a simple keyword based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can lead to a huge number of document entries returned by a search engine, many of which are only marginally relevant to the topic or may contain materials of poor quality. Second, many documents that are highly relevant to a topic may not contain keywords defining them. This is referred to as the polysemy problem [8]. So a keyword-based web search engine is not sufficient for the web discovery, then web mining should be implemented in it. Compared with key-word based web search, web mining is more challenging task that searches for web structures, ranks the importance of web contents, and mines web access patterns.

Topic distillation is the part of web structures which analysis the hyperlink graph structure to identify mutually reinforcing authorities (popular pages) and hubs (comprehensive lists of links to authorities). The best-known algorithms model in topic distillation is the Web graph at a coarse grain which considers whole pages as single nodes. Such models generates topic drift or contamination problem. The problem gets especially severe in the face of increasingly complex pages with navigation panels, advertisement links and decoration. This problem can be avoided by analysing topic distillation with mark-up tag trees constituted with HTML pages and hyperlinks between pages, which identifies sub-trees based on hyperlink coherence to the query. These sub-trees get preferential treatment in the mutual reinforcement process [12]. Here focus is on finding relevant pages through finding most relevant links to the related topic. This links are surrounded by some text, an anchor text. Page contains many links in which some are related to topic and some are basic links like navigation links, advertisement links. Here idea is that to remove unnecessary links and focus on related topical link only with the help of texts surrounding links.

Many researchers have shown that HITS provides good search results for a wide range of queries in topic distillation. But, this method may encounter some difficulties by ignoring textual contexts. The problems faced in the HITS are:- (i) This algorithm does not have an effect of automatically generated hyperlinks, (ii) The hyperlinks pointing to the irrelevant or less relevant documents are not excluded and cause complications for updating hub and authority weights, (iii) A hub may contain various documents covering multiple topics. The HITS algorithm faces problem to concentrate on the specific topic mentioned by the query. This problem is called drifting, and (iv) Many web pages across various web sites sometimes points to the same document. This problem is referred to as topic hijacking. Such problems can be overcome by replacing the sums of authority weights and hub weights with weighted sums, scaling down the weights of multiple links from within the same site, using anchor text (the text surrounding hyperlink definitions in Web pages) to adjust the weight of the links along which authority is propagated, and breaking large hub pages into smaller units. By using the VIPS (Vision-based page segmentation) algorithm, extraction of page-to block and block-to-page relationships and then construct a page graph and a block graph. Based on this graph model, the new link analysis algorithms are capable of discovering the intrinsic semantic structure of the Web. Thus, the new algorithms can improve the performance of search in Web context. The block-to-page relationship is obtained from link analysis. Web page generally contains several semantic blocks, different blocks are related to different topics. Therefore, it might be more reasonable to consider the hyperlinks from block to page, rather than from page to page [8].

Previous research on topic distillation has investigated various sources of external evidence for improving search effectiveness. These approaches were evidence based on link structure, and evidence based on web page characteristics. When study was carried out for topic distillation taking link evidence for full text and anchor text collection within the TREC Web track evaluation framework, anchor density led to only small improvement over the baseline as compared to page characteristic, URL depth and URL length. The reason for this was pointed out to the adoption of certain web page design patterns by organizations which may have a recurring navigation bar at the top or side of the screen. To overcome this problem, one can ignores these global navigational links and computes anchor density only on links inserted by the author of a page. Hence, it is required to study more detailed analysis of page structure to determine whether more refined models of anchor density can lead to larger improvements in search performance. Moreover, this study used limited .GOV collection by a partial crawl of Web pages from the .gov domain. Therefore, study is required to analyze whether the .GOV collection is representative of the whole Web, and whether one can generalize the findings from this study to full-scale Web search. Hence, future study can be feasible with the TREC 2009 ClueWeb dataset [4]. Fujii [19] have proposed a method for improving anchor-based retrieval system, which computes the probability that a document is retrieved in response to the given query identifies synonyms of query terms in the anchor texts on the Web and uses these synonyms for smoothing purposes in the probability estimation. They found that anchor-based retrieval method improved the accuracy of existing methods.

An approach to automatically extracting the web query terms through mining the Web anchor texts to finding effectiveness of a link-based ranking method. A link-based ranking method use link sources, targets and possibly anchors to generate the ranked list. Link methods are those which make some use of the hypertext structure of the Web. Using link methods, a document's ranking is based (at least in part) on its incoming and outgoing links. A hyperlink is a relationship between two documents or two parts of the same document. The link's anchor appears to the user in the source page. If the user selects the anchor, their browser will display the target document. A ranking method, given a query and a set of documents, generates a ranked list of documents. Link methods can be divided into three classes, depending on which of these alternate assumptions they rely: recommendation, topic locality and anchor description [20]. The topic locality assumption is that pages connected by links are more likely to be about the same topic than those which are not, this assumption is often true. Using such methods, a page which is link-adjacent to likely relevant pages may be ranked more highly. The anchor description assumption is that the anchor text of a link describes its target.

4. AIM AND SCOPE OF THE STUDY:

- The main aim of the research work to obtain topicality through text surrounding the links in hyperlink structure of web page to get more refined quality pages.
- To study most existing traditional algorithms in topicality and to get more refined models for quality results.
- To study the hyperlink structure by differentiating it based on topicality and importance into topic-specific sub structures.
- To study more detailed analysis of page structure to determine whether more refined models of anchor density can lead to larger improvements in topicality analysis of link evidence as anchor text collection at partial as well as full-scale Web search.
- WHITs algorithm is implemented based on HITs algorithm by considering anchor text as external source of evidence.

5. RESULTS AND EXPERIMENTS

Initially considered short term queries such as one term and two term queries. Dataset shown in below table is collected from google using AJEX web search API and Bing web search API for the short term query Q. Root set contains top 100 links or Nodes (t) after removing duplicate links. By using nodes of Root set collection of out-links, anchors of out-links and in-links and titles of in-links are performed to build base set (S).

Table 1: Experimental data for various queries

Query (Q)	Nodes (t)	Out link	In link	Links	After normalization Base Set (S)
Java	102	11546	1912	13458	10806
Jaguar	102	16527	744	17373	12711
Harvard	95	27243	4271	31514	13192
Search engine	100	8264	2273	10637	9152
Kyoto Uni.	94	6393	700	7093	6070

Table 2: Average authority weights by HITs and WHITs algorithms

Query(Q)	HITs	WHITs	WHITs-HITs
Java	0.1609	0.2350	0.0740
Jaguar	0.1329	0.2128	0.0799
Harvard	0.0894	0.2510	0.1616
Search Engine	0.1035	0.1042	0.0007
Kyoto University	0.1035	0.1042	0.0007
Average	0.1683	0.1989	0.0306

Table 1 shows the dataset collection detailed results. Table 2 shows results obtained from HITs and WHITs algorithms. Experimental results proves that WHITs outperforms for short term queries.

6. CONCLUSION

Link information is very helpful in finding relevance of web pages. So the link analysis is important for discovering relative information using hypertext and some external evidences and effectively improves the performance. The external source of evidence used is anchor text in WHITs algorithm and it outperforms for short term queries. The major part of performance measures are discovered for related documents. Web search tasks are also evaluating like named page finding, entry page finding, ad-hoc search task etc. Test collections are also important for retrieval process.

7. REFERENCES

[1] Pandia M., Pani S.K., Padhi S.K., Panigrahy L. and Ramakrishna R. 2011. A Review Of Trends In Research On Web Mining, International Journal of Instrumentation, Control & Automation (IJICA), 1(1), 37-41.

[2] Srinivas K., Reddy L.K.K. and Govardhan A. 2010. A Theoretical Approach to Link Mining for personalization, International Journal of Computer Science Issues, 7(3), 41-42.

[3] Getoor L. and Diehl C. P. 2005. Link Mining: A Survey, SIGKDD Explorations, 7(2), 3-12.

[4] Wu M., Scholer F. and Turpin A. 2011. Topic Distillation with Query-Dependent Link Connections and Page Characteristics, ACM Transactions on the Web, 5(2), 3-25.

[5] Tsikrika T. and Lalmas M. 2005. Best Entry Pages for the Topic Distillation Task, qeen marry, university of London.

[6] Jain R. and Purohit G. N. 2011. Page Ranking Algorithms for Web Mining, International Journal of Computer Applications (0975 – 8887), 13(5),22-25.

[7] Katz V. and Li W.S. 1999. Topic Distillation on hierarchically categorized Web Documents. In Proc. of the 1999 Workshop on Knowledge and Data Engineeering Exchange, IEEE.

[8] Gupta M., Tomar V., Verma J. And Roy S. 2011. Mining databases on world wide web, IJCSI, 560-564.

[9] Page L., Brin S., Motwani R. and Winograd T.1998. The PageRank citation ranking: Bringing order to the Web. Unpublished draft.

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

[11] Bharat K. and Henzinger M. R. 1998. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. Proc. of 21th ACM SIGIR Conf. on Research and Development in Information Retrieval, 104-111.

[12] Chakrabarti S., Joshi M. and Tawde V. 2001. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. Proc. of 24th ACM SIGIR Conf. on Research and Development in Information Retrieval, 208-216.

[13] Rafiei D. and Mendelzon A.O. 2000, What is this Page Known for? Computing Web Page Reputations, In Proceedings of Ninth International WWW Conference, Amsterdam.

[14] Haveliwala T. 2002. Topic-Sensitive PageRank. In Proc. of the 11th International World Wide Web Conference, Honululu, Hawaii.

[15] Choi I. and Kim M. 2003. Topic Distillation using Hierarchy Concept Tree, SIGIR'03, Toronto, Canada, ACM 1-58113-646-3/03/0007,371-372.

[16] Eiron N. and McCurley K. S. 2003. Analysis of Anchor Text for Web Search. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. Toronto, Canada, 459-460.

[17] Zhong J. K., Zhao L., Qiong W. Y. and Zhong G. J. 2008. An Algorithm of Topic Distillation Based on Anchor Text, International Symposium on Electronic Commerce and Security, IEEE computer society, DOI 10.1109/ISECS:11-15.

[18] Lempel R. and Moran S. 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In Proceedings of the 9th International World Wide Web Conference, Athens, Greece.

[19] Fujii A. 2008. Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval, WWW 2008 / Refereed Track: Search - Query Analysis, Beijing, China, 21-25

[20] Craswell N., Hawking D. and Robertson S. 2001. Effective Site Finding using Link Anchor Information SIGIR'01, New Orleans, Louisiana, USA. ACM 1-58113-331-6/01/0009, 250-257.