**Literature review: Search engine’s result**

**correlation improvement using intelligent retrieval**

(blue text by Eric, green text by Yufeng Shi, black text by Gang Wu)

**Abstract**

Nowadays, network has been the necessary part for our daily life and searching on the web search engine has also became the main way to acquire knowledge for us. In statistic, about 380 new websites are created every per minute [1]. Another amazing figure is the website spams take up 30% of these websites. Thus, when people use search engine to find the results of their queries, it become much harder for them to find the most relevant answer, which wastes much time for users. Currently, the most popular searching engine technologies are PageRank and Hyperlink-Induced Topic Search (HITS), the efficient technologies on web mining [2] to help people to find the most relevant result they want. However, because of the drawbacks of itself, the disadvantage of them is more and more obvious. In a recent survey, there are 40% of interviewees reflected that they often or always looked one more page of results when they used search engine [3]. Based on this, a novel and innovative way for information retrieval are developed-intelligent retrieval. Intelligent retrieval, a new technology which combines the link-based, content-based and individual relevance for ranking the all websites to satisfy the user’s demand on query to retrieval the most relevant results. In this literature, we mainly compare three representative methods to analyze their own advantages and indicate something they should improve for researchers to study for.

**Introduction**

Due to the development of computer network and rapid accumulation of the information on the web, users are suffering from finding accurate and relevant information. From large amounts of information on the web, only a small part is valuable for users. Therefore, the Web Mining technology have been developed to overcome this problem. This technology made users to extract information from large database by using sort algorithm. It calculates the relevance between document information and user retrieval information. Also, with increasing number of users on the web, the number of queries submitted to the search engines are also increasing exponentially. So the First Generation sort algorithm, which consider word frequency on the appearing documents moves to Second Generation sort algorithm, which is based on link analysis such as citation. To present information more efficiently and accurately, Page Ranking methods are applied to arrange information by relevance and importance [10]

The current search engine is difficult to meet the needs of users

With the development of computer network, people mainly rely on information retrieval systems which are search engines to obtain resources and information. However, Users encounter a lot of obstacles in the searching process, and it is difficult to get a lot of useful information [3]. If users use keywords in searching process, they will get a large set of retrieval results, and only a little information can be valuable for them [1]. More important, it is waste of time for users to filter out irrelevant information from retrieval results [2].

Challenges in search engine

1. Massive data storage

Some large websites own thousands of pages. When the search engine crawls the page, it must store the data effectively. Also, the data structure must be reasonable, highly scalable, and access speed requirements are very high [7].

2. search engine update

The Internet is a dynamic content update. Every day, many people post new content on the Internet or update old content. However, the search engine program takes a lot of time to update the program every time [8].

Thus, the purpose of this literature review is to combine the personal relevance and content feature with link-based algorithm to optimize the search engine. We introduce a new term called intelligent retrieval to achieve this goal. Intelligent retrieval, a latest technology, is a hot topic in research area because it can solve the semantic understanding and user feedback to return the more precise retrieval result [1]. Also, by applying this kind of technology, it will be a huge improvement on increasing efficiency on search result and compressing calculating time as much as possible [2]. The primary question to drive our research is What are the different ways to achieve intelligent retrieval and how they work to information retrieval to help researchers optimize it better.

**Web Mining**

Duhan et al[1] refer Data Mining is the extraction of interesting(non-trivial, implicit, previously unknown and potentially useful) information or patterns from large databases. Also they refers Web Mining is the application of data mining techniques to discover and retrieve useful information from the WWW documents and services. According to Rawat and Pathak[2], Web Mining has following three types :

* 1. Web Structure Mining

Web Structure Mining uses the hyperlinked structure of the web documents and generates the structure summary about the web documents in the form of web graph, where web pages are represented as nodes and hyperlinks between web documents which are represented as edges. In this technique, back links and forward links play vital role in the retrieval of the web page. Web Structure Mining can be divided into two part as Hyperlink and Document Structure. A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. Document Structure focus on automatically extracting content within a Web page which is organized in a tree-structured format, based on the various HTML and XML tags within the page.

* 1. Web Content Mining

Web Content Mining extract information from the contents of web documents or result pages produced from search engine. The web document may contain text, multimedia or structured records like lists and tables.

* 1. Web Usage Mining

Web Usage Mining extract information such as user navigation pattern and the useful from the secondary data stored in server like access log, agent log, cookies, user profile data, metadata. The interactions of the user while surfing on the web derives to all kinds of data.

2.1 PageRank Algorithm

The PageRank algorithm was proposed by Brin and Page at Stanford University. This idea came from the citation analysis. The concept of PageRank is the calculating importance of research publications according to the number of citations they have. Normally, Web domain has a link which is connected another Web page as citation. From the point of view of the cited page this link is called a back-link. PageRank does more than counting back-links in that it assigns different ranks to back-links. As a result, a Web page can get a high rank if it has many back-links. PageRank considers both ranks scores and back-links, it can be treated as both quantitative and qualitative method. Now the PageRank Algorithm considers more than 25 billion web pages for ranking web pages.[3]

PageRank is a measure based on the number of back links to a page. In the Google, PageRank is displayed on the toolbar of the browser if the Google toolbar is installed. But the toolbar will display a Page Rank of 0 to 10 for the page, 0 being an unnoticeable page and 10 a highly visible page. Therefore, a page has a high rank if the sum of the ranks of its backlinks is high. If the addition of all the ranks of the back links is large then the page it is provided has large rank. [6]

The advantages of PageRank algorithm are following[4] :

1. Low Query time cost: PageRank’s query-time cost of computing and synthesizing importance score for a web page is low.

2. Low vulnerability at localized links: PageRank is activated using the entire Web graph, rather than a small subset, it is strong on localized links for spam information.

3. High Efficiency: PageRank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. This method is more efficient than other web page search algorithms.

4. Feasibility: PageRank algorithm is feasible in today’s scenario since it performs computations at crawl time rather than query time.

2.2 The Drawbacks of PageRank

It is obvious PageRank is efficient Web search method, but according to Sharma and Bhagat[5], there are some drawbacks remain. PageRank uses links between webpages to specify scores, but this structure can cause problems. Sometimes Web pages form a group without forming outside links. This is called spider traps. If the hypertext points to a page with no outgoing links, this is called Dangling link. In such case, web pages just don’t form links themselves. This is called Dead Ends. These problems are mainly caused by the structural features of PageRank that focus on links between web pages.

Unlike other common studies about PageRank, from Agarwal and Khan’s research[6], they refer “Rank Sink” as the representative problem of PageRank. Noticed the representative problem of PageRank, which is called Rank Sink. They refer that Rank Sink can be occurred when two or more web pages are connected to each other and form a cycle. The problem is, if this cycle did not refers to other web pages but are just referred to by other web pages outside of the cycle, they would accumulate rank but never distribute any rank.

Even though PageRank is used widely, W. Xing[7] insists there is a gap between the results of the calculations and the results of the actual web, which means some links in a real web page may be more important than are the others. He proposed extended PageRank algorithm, a Weighted PageRank(WPR), which can assign larger rank values to more important or popular pages instead of dividing the rank value of a page evenly among its outlink page. Each outlink page gets a value proportional to its popularity.

3.1 Hypertext induced Topic Selection (HITS)

Hyperlink Induced Topic search also known as HITS algorithm was introduced by Klienberg. HITS is a search query dependent algorithm that ranks the web page by processing its entire in-links and out-links. It classifies web pages by two types, Hubs and Authorities. Ranking web pages by HITS is decided by analyzing topic relevant texts. When the user issues a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings of the expanded set of pages. This set of web page is named as authority if the web page is pointed by many hyperlinks and a web page is named as HUB if the page point to various hyperlinks.[8]

Y. He et. al[9] re-define HITS as two metrics for each page: authority weight and hub weight. These two weights are computed iteratively to determine the importance of a certain page. HITS algorithm use a traditional text search engine to get a root set of pages related to the query topic. Then extend the root set in order to acquire a larger base set, namely that add the pages that point to the pages from root set as well as the pages that pointed by the pages from the root set. Construct a Web adjacency graph, and acquire the authority weight and hub weight through iterative computation according to the mutually reinforcing relationship between authority pages and hub pages. Then the algorithm sort pages and acquire the authority source of information on search topic based on the authority weights.

The advantages of HITS algorithm are following [4]:

1. HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.

2. The ranking can be combined with other search engine's results.

3. HITS is sensitive to user query.

4. Important pages are obtained on basis of calculated authority and hubs value.

5. HITS induces Web graph by finding set of pages with a search on a given query string.

3.2 The Drawbacks of HITS

Punit & Kanu[8] and Grover & Wason[4] both refer several drawbacks of HITS. HITS is a query dependent algorithm, but the query time evaluation is normally costs very high. Also, It has risks of giving high rank of hub and authorities which contain irrelevant information, due to flaws done by web designer. In addition, HITS brings up a traditional search engine to obtain a bunch of related pages for getting hubs and authorities pages, it takes query times, so it is vulnerable to treat most recent or current processing of information.

They mention “Topic Drift” also can interrupt HITS’s function. There are irrelevant pages in the root set and they are strongly connected. Since the root set itself contains non-relevant pages, this will reflect on to the pages in the base set. Also, the web graph constructed from the pages in the base set, will not have the most relevant nodes and as a result the algorithm will not be able to find the highest ranked authorities and hubs for a given query.

**improve quality of result accuracy in search engine**

Research shows that current search engine cannot satisfied the requirement of users, so we need to find new methods to improve quality of result accuracy in search engine.

Key concept: Web structure mining is also called link mining. It is mainly used in the field of information retrieval on the WWW and based on the topology of the hyperlink which can be used to improve the quality of search engine queries [2]. In general, the search engine query results are usually relatively large, the content is a lot of irrelevant information query, it can analyze and classify the similarities and relationships of hyperlinks on different websites to rank search results. Web search for structural analysis can improve the efficiency of spiders crawling the web, the search strategy is to crawl along the hyperlink with the highest priority should be Web Page Rank, making it the shortest path, minimum time that the document up to date information [1]. It also generates structural summary about website and webpage by analyzing the link. For example, Hits and PageRank are the popular web structure mining algorithms.

 L.Yan [4] want to find new method to detect a large number of similar image Web page which is analyzed by IP address, URL, and link structure analysis model of combination. Then the author [4] proposed Web structure mining which presents several integral subtopics, such as graph structures and searching, as well as content categorization and classification techniques to set a sound foundation in order to explore, extract, and analyze Web information data[5]. Also, several researchers proposed “Web Structure Mining” which can Collect and analyze web search engine statistics on each company to predict links which are users need [4]. The benefits of the method are that the proposed algorithm increases the degree of relevance than the original one, and decreases the query time efforts of topic-sensitive PageRank [4]. However, the method is still a link for the analysis of Web pages and is still not based on the content of Web pages [4].

Another method is “link-based clustering” which is similar to the first method based on link analysis. However, Web page clustering aims at separating irrelevant pages and joining relevant pages into semantically meaningful groups, which is different from clustering in other fields [6]. Wang [6] proposed a link-based clustering algorithm by analyzing the number of in-links or out-links in common. Firstly, we need to divide cluster web search results into n cluster based on the number of in-links or out-links [6]. Even two related pages cannot have many same in-links and out-links, we found that the short texts attached to the in-links of pages are very similar which reference as in snippet. Also, we found that an in-snippet contains a short text describing the main topic of the page of the in-link, divide pages to group more precisely.

Then, researchers made several relevant experiments on “link-based clustering”. In many circumstances, it's hardly to predict how many clusters in the end. Similarity threshold is pre-defined to determine whether one page is a member of one cluster, and it could be easily defined and adjusted [9]. when the similarity between the page and the correspondent cluster is above the similarity threshold, each page is assigned to existing clusters [6]. If none of current existing cluster meet the demand, the page under consideration become a new cluster itself [6]. When new members are introduced to the cluster, the centroid vector which is used to calculate the similarity should be recalculated in incremental way. Due to one page could belong to more than one cluster, it is limited to top N(N=10) clusters based on similarity value [6]. As a result, all pages joining clustering procedure are processed sequentially and whole process is iteratively executed until it converges (centroid of all clusters are no longer changed) [7]. At the same time, if they share majority members, the final clusters are generated by recursively merging two base clusters [8].

Finally, the researchers use two measures to evaluate the quality of clusters. One is entropy which provides a measure of "goodness" or "purity" for clusters. It compares the clustering result to known classes. Low entropy suggests high-quality of the clustering result which has high intra-cohesiveness [6]. On the contrary, high entropy suggests that the low quality of clustering result which may covers more than one sub-topic. Another is f-measure which calculates the accuracy of test in statistic [6]. Based on the experiment results, I found that the proposed algorithms improve the relevance of search results and save computation time [1]. However, this method just based on links and in-snippet, and it do not analyze content and snippet, so the retrieval result is not enough accurate [6], and I think we could focus on how to combine links, snippet, content and in-snippet together to get more precise clustering results.

The Third link-based method is Aggregate PageRank (APR) by aggregation of PageRank and HITS to calculate the importance score of web page [9]. The reserchers proposed this methodology consist of four steps: (1) we extracted all the hyperlinks associated with the target URL and then refined it to eliminate all irrelevant links before Assigning a unique numeric value to every web page. (2) we need to apply PageRank algorithm which is based on the number of incoming links of each page to compute the importance score of score of web pages [10]. On the other word, A page with higher number of incoming links has higher importance score than the page with lower number of back links. (3) we need to apply the HITS algorithm which is based on Authority score and Hub score to compute the importance score of web pages [10]. (4) Finally, we need to combine PageRank and HITS to get final rank of each web page by using a specific formula. The result of the experiment shows that the proposed method gives useful and a quite good and accurate results. However, the experiments just analyze part websites, and I think we need to carry out comprehensive performance analysis of Apr by using different websites.

**Intelligent retrieval**

Intelligent retrieval, a new generation of information retrieval system [3]. The most obvious difference of it with the current information system (link-based) is it can provide individual services to simplify the retrieval results [4]. At the same time, with entering user’s query demand, semantic understanding and user feedback are calculated by this system so that it can adjust and update itself to meet more precise user’s individual requirement [5]. Thus, this is why it is named intelligent retrieval.

In Sathya’s research [7], K-Means clustering algorithm can make the results more relevant and decrease the response time when search engine retrieves the results. K-Means clustering designs a brief algorithm to achieve the division on n observations into k clusters and then to find which cluster is closest to mean value. Not like previous methods which just retrieve the out-link and in-link pages, K-Mean calculates the weight of these co-occurrence terms. The below formula is the way to calculate to the mean of weight of co-occurrence terms

  [12]

As a result, the threshold values of which are 80, 30 and less than 30 will be defined as most relevant, relevant and irrelevant clusters. By discarding these irrelevant clusters, this kind of intelligent retrieval technology could decrease the response time on retrieving and get more relevant results.

Xue and Yan [8] used the novel ways to get the intelligent retrieval. multi-agent information retrieval system provides the users more accurate results than traditional search methods. This system includes user agent(with function of sending query to communication agent), communication agent(for interpreting channels between users and webs), mining agent(the duty for fining the users’ interest and rewrite the query), personal agent(for storing the users’ individual hobbies and access records), intelligent evolution agent(calculating the feature vector of users), information retrieval agent(querying request data from resource webs), group agent (extracting users who have the similar queries for clustering)and clustering agent (analyzing the above data and returning query results). by comparing the correct rate of acquired documents in different intelligent retrieval system, Xue and Yan showed multi-agent system had better performance on information retrieval, as the average correct rate of this intelligent retrieval system is 81.6%, much higher than any other existing intelligent retrieval system. Not only this, with the higher frequency the users use this technology, the nine agents could adjust and coordinate with each other so that it can make the retrieval process become more intelligent.

There is much research which focus on the link-based and individual relevance to achieve intelligent retrieval. However, not many articles are related to content-based feature to rank page because of its complexity. Firstly, we should make the definition of content-based technology. Contend-based ranking technology, a new retrieval way which consider the context of the websites’ text frequencies and similarity for ranking the page. Although there are not many articles which discusses the content-based technology, Chouhan [10] presents some basic principles about how it works. Chouhan presented the entire rank model to simulate the whole process about how to achieve content-based retrieval. From query interface, then though search technique, to Page ranking, this is a simple process to achieve retrieval. For Page ranking, there are many specific ways to finish this step, and one of most important way is similarity and frequency estimation. There are main two formulas to calculate similarity. One is 

Another one is 

By calculating the frequency of words and sentences, different websites will be grouped separately. About the effect of this kind of algorithm, it uses the memory consumption, time consumption and precision as indicators to compare the effects with traditional PageRank algorithm. The results showed that although it need a little more time to get retrieval back, the precision on content-based technology has some improvements compared with traditional PageRank.

**Conclusion and discussion**

By applying link-based and individual relevance technologies, it will be some improvements on increasing efficiency on search result and compressing calculating time as much as possible. However, the true intelligent retrieval is combing the link-based, intelligent retrieval and content-based features, which gives the future researchers the directions that they should do a lot efforts on achieving content-based technology. In this way, the information retrieval will never bother users to spend much time to find the most relevant answers about their queries.

For the future, the researchers should consider more factors about intelligent retrieval like content-based algorithm and self-organizing feature map neural network technology. Also, the future work will add new parameter in searching algorithm using web usage mining to improve the search result this include data as server access logs, referrer logs, client-side cookies, user profile and meta data which retrieved from user’s surfing pattern.

Reference:

[1] N. Duhan et. al, “Page Ranking Algorithms : A Survey”, 2009 IEEE International Advance Computing Conference(IACC 2009), vol

[2] J. Srivastava et. al, “Chapter 3 Web Mining-Concepts , Applications & Research Directions.”

[3] M. Sangeetha and K. S. Joseph, "Page ranking algorithms used in Web Mining," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, 2014, pp. 1-7.

[4] N. Grover and R. Wason, “Comparative Analysis of Pagerank and HITS Algorithms”, International Journal of Engineering Research &amp; Technology (IJERT), vol.1(8), 2012

[5] Sharma, S., &amp; Bhagat, A. “Research on Ranking Algorithms in Web Structure Mining”, International Journal of Knowledge Based Computer Systems, 3(2), 13–20, 2015.

[6] B. B. Agarwal and M. H. Khan, "Analysis of Rank Sink Problem in PageRank Algorithm", International Journal of Scientific & Engineering Research, vol. 4(11), 2013, pp. 251-256

[7] W. Xing and A. Ghorbani, “Weighted PageRank algorithm”, Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004, Fredericton, NB, Canada, pp. 305-314, 2004

[8] P. Patel and K. Patel, “A Review of PageRank and HITS Algorithms”, International Journal of Advance Research in Engineering, Science & Technology(IJAREST), vol. 2(1), January 2015

[9] Y. He et al. “Improvement on HITS Algorithm”. Applied Mathematics Information Sciences, vol. 6-3S(3), pp. 1075-1086, 2012[10] L. Z. Xiang, “Research and Improvement of PageRank Sort Algorithm based on Retrieval Results

[1] M. Sathya, J. Jayanthi, and N. Basker, "Link-based K-Means clustering algorithm for information retrieval," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 1111-1115.

[2]L. Z. Xiang, "Research and Improvement of PageRank Sort Algorithm Based on Retrieval Results," 2014 7th International Conference on Intelligent Computation Technology and Automation, Changsha, 2014, pp. 468-471.

[3] S. G. Pawar and P. Natani, "Effective utilization of page ranking and HITS in significant information retrieval," International Conference for Convergence for Technology-2014, Pune, 2014, pp. 1-6.

[4] L. Yan, Y. Wei, Z. Gui and Y. Chen, "Research on PageRank and Hyperlink-Induced Topic Search in Web Structure Mining," 2011 International Conference on Internet Technology and Applications, Wuhan, 2011, pp. 1-4.

[5] Freire, M. M., & Periera, M. (2008). Encyclopedia of Internet technologies and applications. Hershey, PA: IGI Global.

[6] N. Yang, Y. Liu and G. Yang, "Clustering of Web Search Results Based on Combination of Links and In-Snippets," 2011 Eighth Web Information Systems and Applications Conference, Chongqing, 2011, pp. 108-113.

[7] De, S. S., Dehuri, S., & Wang, G. (2012). Machine learning for social network analysis: A systematic literature review. IUP Journal of Information Technology, 8(4), 30-51.

[8] Gilmore, C., & Haydaman, J. (2016). Anomaly detection and machine learning methods for network intrusion detection: An industrially focused literature review. Paper presented at the 292-298.

[9] Y. Wang and M. Kitsuregawa, Link based clustering of web search results, Proceedings of the Second International Conference on Advances in Web-Age Information Management, 2001.

[10] P Pooja Devi, Ashlesha Gupta, and Ashutosh Dixit, “Comparative Study of HITS and PageRank Link based Ranking Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering

[1] https://www.millforbusiness.com/how-many-websites-are-there/

[2] L. Zhang, "An intelligent information retrieval algorithm based on knowledge discovery and self-organizing feature map neural network," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-4.

[3] L. Z. Xiang, "Research and Improvement of PageRank Sort Algorithm Based on Retrieval Results," 2014 7th International Conference on Intelligent Computation Technology and Automation, Changsha, 2014, pp. 468-471

[4] B. Xue and G. Yan, "Research on multi-agents information retrieval system based on intelligent evolution," Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, Changchun, 2012, pp. 1042-1045.

[5] B. Xue and G. Yan, "Research on multi-agents information retrieval system based on intelligent evolution," Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, Changchun, 2012, pp. 1042-1045.

[6] M. Sathya, J. Jayanthi and N. Basker, "Link based K-Means clustering algorithm for information retrieval," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 1111-1115.

[7] R. M. Esteves, T. Hacker and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, Bristol, 2013, pp. 17-24.

[8] B. Xue and G. Yan, "Research on multi-agents information retrieval system based on intelligent evolution," Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, Changchun, 2012, pp. 1042-1045.

[9] X. Xu, J. Xiong and C. Cheng, "The model and the security mechanism of the information retrieval system based on mobile multi-agent," *2010 IEEE 12th International Conference on Communication Technology*, Nanjing, 2010, pp. 25-28.

*[10] J. Singh Chouhan and A. Gadwal, "Improving web search user query relevance using content based page rank,"*2015 International Conference on Computer, Communication and Control (IC4)*, Indore, 2015, pp. 1-5.*