Vol.4.No.1, July, 2015
22 - 35,ISSN:0975-5446] [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

An Anchor Based Information Retrieval For Link Analysis: A Survey

PATEL Hemangini S.
Bhagwan Mahavir college of Comp. App.
BCA, Bharthana, vesu.
hemanginipatels@gmail.com

DESAI Apurva A.
Department of Computer Science
Veer Narmad South Gujarat University,
Surat.
aadesai@vnsgu.ac.in

Abstract

The growth of World Wide Web (WWW) is being increasing continuously and becomes a valuable resource to information retrieval and knowledge discovery from Web in the field of web mining. A challenge become emerging for extracting richly qualitative documents, particularly where the documents have linked via multiple types of relations. These links provide additional environment that can be helpful for web mining tasks. Hyperlinks may be created for different purposes and in different contexts. Conventional link analysis does all links likewise and assumes that links are endorsement. Hence, there is a need to extract links that are valuable. In recent days, the use of anchor text in link analysis has been found to increase the performance of web search significantly to extract a valuable link. In the present study, novel characteristic of the web page such as anchor text based information retrieval has been studied to help the search engine to extract a pertinent and valuable link.

Keywords: Information Retrieval, Link Analysis, Topic Distillation, Anchor Text

1. Introduction

The goal of Information Retrieval (IR) is to retrieve all relevant documents from a collection of documents to a user query. Most of the web IR tools only use the textual information while ignores the link information that could be very valuable [1]. Most of research in IR has become successful in developing and refining techniques that are word-based [2]. But the word-related search engines bear usual problems in IR, such as synonymy, polysemy (word with more than one meaning), authorship, persuasion, keyword spamming and context sensitivity become especially severe on the web[3].
Resources on the WWW are interlinked via hypertext, so, hold a link formation, and each link contains some explanatory text, called anchor text. Page et al. [4] have shown the effectiveness of anchor text for a web search. Initially McBryan shown the utilization of anchor text has been valuable for a web search. The search task was initially defined as planning to discover superior documents on a wide query topic and

later it was accepted and experienced inside the agenda of the TREC (Text Retrieval Conference) Web Track to discover key resource pages [5]. Finding key resources around the topic is known as a topic distillation. The topic distillation task of TREC Web Track was introduced in 2002 to capture web search, where users considered entry pages to relevant sites as more valuable than isolated pieces of relevant text. Hence, the aim of the task was to identify key resources on a broad topic [6]. Wu et al. [5] have shown a new approach by exploring the use of external sources of evidence such as; link structure, query dependent in-degree and out-degree to improve topic distillation; and web page characteristics, such as the density of anchor links. Another external evidence, an outgoing link pointing to retrieved documents that are a query-dependent has also been made known to be valuable for topic distillation [7]. Anchor text is tested by TREC Web track contestant for named page finding, homepage, ad-hoc search tasks, and topic distillation. It was observed from experimentation with the TREC web track data via anchor text on its own can direct to significant performance increases in topic distillation [5,7]. Hence, the aim and contributions of this study is to review of how anchor text can be used to get better search superiority to help the search engine to extract a pertinent and valuable link.

2. Information retrieval systems

Information retrieval (IR) finds the stuff of an amorphous nature that satisfies an information require from huge collections. Finding the essential web page on the web with successfully and expertly has become a challenge. In the traditional IR activity, users can research topics of interest [8], where a primary unit of analysis is a document. A search engine uses various sources for ranking web pages corresponding a user query particularly textual content and the link formation of the web. Now days Hyperlink and a citation have a significant effect on an influenced hypertext search and ranking on the Web [9]. Hyperlinks between the www documents are utilized to specify the relative authority values of documents with reference to different search queries. This process that finds quality pages is called topic distillation.

2.1 Topic Distillation

Traditional topic distillation with hyperlink has been applied to a macroscopic Web model where the hyperlinks are edges and documents are nodes in a directed graph [10]. At first, Topic Distillation task calculates a query specific sub-graph of the Web by comprising pages on the query topic within the graph and ignoring irrelevant pages and then calculates a grade for every page in the sub-graph related on link connectivity. For computing a score, each page is specified an authority score, which is obtained by computing the summing of the weights of all inward links to the page. To each of this reference, its weight is calculated by evaluating how superior the referring page is as a source of links. Katz and Li [11] have used a three-step approach to extract topics from the set of documents belonging to a group automatically; *(i) Document Keyword mining, (ii) keyword broadcast through category tree structure, and (iii) Keyword broadcast across pages related by links.*

The information from hyper linked WebPages is a rich resource for topic distillation. There is a triangular association between the hub, authority and anchor text. Hence, a page pointing to a fine authority page through the authoritative anchor text is a fine hub page; an authoritative anchor text connects a fine hub page and a fine authority page, and a page which is pointing to a fine hub page through the authoritative anchor text is a fine authority page. The mining of authority pages and automated discovery of high quality Web structures and resources assisted by such kind of mutual correlation between hubs and authorities [12]. The development in the new research direction in the WWW involves the link structure analysis. Efforts had been made in late 1990 that have

a thoughtful power on link analysis were Brin & Page's PageRank [4] and Jon Kleinberg's work on HITS [13].

3. Linked Based Analysis

Link mining is an emerging research area used in various fields like; commercial and business enterprises, web search and retrieval, personal information management, law and security enforcement, and medicine and bioinformatics [14]. The significant achievements of linked based analysis during the last few years for the WWW are; (i) Independent of specific applications, researchers made comprehensive measurements on the WWW. Formulated models for formulation, creation and destruction of nodes, as well as links, observed statistics of WWW, (ii) Link-based analysis is a prime tool to select the top few authoritative pages from a large response set received using text-based methods, and (iii) Links-based analysis is useful for clustering and classifying pages [15]. There are the numbers of proposed algorithms based on link analysis, but the significant algorithms are HITS, Page Rank, and SALSA. The HITS (Hyperlink Induced Topic Search) algorithm that related on mutual reinforcement association gives a modern tactic for a Web search and topics distillation. The HITS algorithm searches the hubs and authorities of the area on a particular query or topic. In the research area of link analysis of linked pages, an algorithm HITS is useful to examine the area of topic distillation, and numerous types of link weights are drawn in to specify the importance of links in linked pages. In the work of Bharat and Henzinger [16], the metrics of correspondence of whole contents in the linked pages were applied to link weights and the text nearby the links as keyword-related facts used to find out a weight for each link. They found two types of problems in HITs as a nepotism problem and the topic drift problem. The first problem occurs as mutually reinforcing associations among two hosts provides unnecessary weight to the estimation of a single one. The *topic drift* problem occurs if the majority well-graded authorities do not connect, and an extended set contains unrelated documents to the query topic. To overcome topic drift, Chakrabarti et al.[17] combine the TF-IDF(Term Frequency- Inverted Document Frequency) weighted model and micro-hub to correspond to the importance of anchors in regions with information required. The model considers hyperlinks in relevant DOM (Document Object Model) sub-trees have higher weights than links in unrelated DOM sub-trees to query topic. This approach helps in discovering parts of a Web page related to query and reduces *topic drift* problem. Rafiei and Mendelzon [18] have defined a latest determine called "reputation" of a page and calculate the set of topics for which the page would be ranked high. Haveliwala[19] has proposed a "Topic-Sensitive PageRank," which pre-calculates a set of PageRank vectors equivalent to dissimilar topics. Choi and Kim [20] have analyzed the use of *hierarchy concept tree* by hyperlink graph structure to overcome the problem of *mixed hubs* in the Bharat's algorithm. They tried to discover the association in documents linked by hyperlinks with the use of content analysis and allocate weights to hyperlinks related to the association. They evaluated the algorithm via 50 topics on WT10g corpus and got 25 to 46% improvements. Borodin et al. [21] differentiated between algorithms according to the scope of their imitative target graph. According to their opinion, PageRank operates on the whole web so it is query-independent, and HITS and SALSA had a base set consisting of web pages relevant to a given topic so it is query-dependent. The findings of Modi and Desai has been used to classify information need that can optimally retrieved by the individual algorithm and to suggest the alternate for improving search results such that PageRank calculated and refreshed off-line and not relevant to query term so not suitable for concept searching and finding web communities instead HITs and SALSA are outperforms[22]. Modi [23] combined link structure information and text content of the documents to make a judgment about document relevancy for web IR and to motivate research on exploring hidden semantics, categorization of results, upgrading sub-dominant community. Link mining approach has been used for ranking

and categorization of documents and usage analysis for re-organization of outcomes to increase precision rates for top listed results. Existing algorithms for web page categorization based on post-processing categorization technique that applied in broad topic query to provide clustering algorithm to organize documents into semantic group based on URL, Title, snippets and link structure of documents on web graph was analyzed by Modi and Desai[24] for machine learning to provide assistant to web search agent. In other study, they also identified concealed information on web which may be hidden semantics uncovered by general-purpose search engine so that limited resources available for search assistant to increase satisfaction of user to provide intelligent search assistant[25]. Another model targeted towards broad topic query using eigenvector measures to find groups of web pages having common interest via link structure of base result set to identify set of highly linked group of web pages and identify the dominant and sub-dominant groups using the principal component analysis [26]. According to the association between authority and cocitation in HITS, Xie and Huang [27] proposed a new hyperlink weighting scheme to depict the strength of the relevancy between any two web pages to combine hyperlink weight normalization and random surfing schemes as used in PageRank to justify the new model. In the new model based on co-citation, the pages with stronger relevancy are assigned higher values, not just depending on the out links; it combines features of HITS and PageRank. An improved HITS algorithm based on the full text, the anchor text and the close textual context of the hyperlink firstly computes the relevance of arbitrary two pages based on page topic similarity and meta-information similarity. Then, by using the relevance, a new adjacency matrix is constructed to calculate authorities and hubs iteratively. This new algorithm improves the efficiency and quality of query, reduce the theme drifts[28].

3.1 HITS (Hyper-link Induced Topic Search)

Jon Kleinberg's HITS algorithm, based on link analysis, which ranks Web pages [13]. Clever a model search engine utilized it [29, 78] for an IBM research project. It was antecedent to Page Rank. The HITS algorithm uses WWW as the directed graph as G(V,E),where G represents directed graph with set V as vertices indicates pages, and set E as edges indicates to link. It gives weight on the terms of queries to link and end point of the link. Combining the anchor to make the weight to the link and a big hub is break into parts for one topic[30]. Figure1 shows triangular association between hub-authority and anchor.

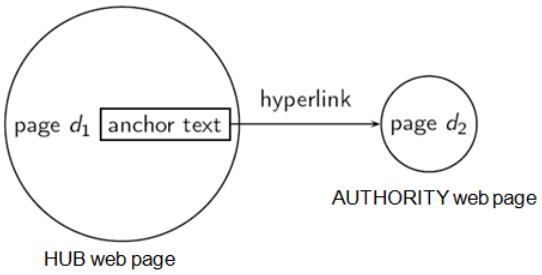

Figure 1: An illustration of modified HITS

To find a web page as Authority or Hub, HITS follows Authority-Hub update rules [31]. For a particular set of web pages retrieved to a search query topic, the HITS algorithm first structures n by n adjacency matrix A, whose element m(i,j) is 1 if page i associates to j and 0 otherwise, and then iterates the following equations [32].

$$a_i^{(t+1)} = \sum_{\{j:j\to i\}} h_j^{(t)} \quad ; \qquad (1)$$

$$h_i^{(t+1)} = \sum_{\{j:i \to j\}} a_j^{(t+1)} \quad ; \qquad (2)$$

Determination of ultimate hub-authority scores of nodes are done after infinite repetitions of the algorithm. By hub-authority update rule directly and iteratively, values are obtained. Hence, it is essential to normalize the matrix after each iteration[33]. A depiction of hubs-authorities calculation for a specific topic of a user query is shown in figure 2.

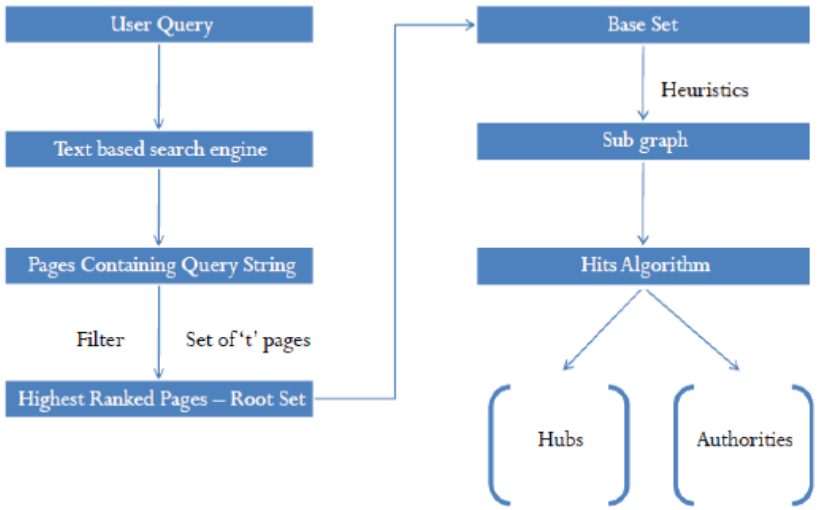

Figure 2: Computation of hubs and authorities for a specific topic by a normal query p

3.2 PageRank

Page et al.[4] have proposed an algorithm known as page rank algorithm. This algorithm ranks the web pages in search engine Google. The Page Rank algorithm utilizes link structure of the web pages. The algorithm assigns a numerical weight or rank to each page of a hyperlinked set of documents with the purpose of finding its relative importance within the set. This algorithm is query-independent and it operates on the whole Web, and assigns a PageRank to every web page [20]. It is related on the thought that if a page holds essential links to it then the links of this page to the other page are also to be regarded as essential pages. Back link is utilized by Page Rank to decide the rank score. It provides a large rank, if the summation of all the ranks of the back links is large then the page. A basic PageRank is defined as follows [34]:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (3)$$

Where, *u* is a page, B(*u*) represent the set of pages that directed to u, PR(*v*) and PR(*u*) are grade counts of page *v* and *u*, correspondingly. N*v* is the amount of outer links of page *v*, and *c* is a factor use for normalization. In a PageRank, the grade count of a page p is uniformly divided along with its outer links and the values given to the outer links of a page p are sequentially used to compute the grades of the pages directed to by p. An illustration for back links is shown in figure 3. Here, Q is the back link of P & R and P & R is the back links of S.

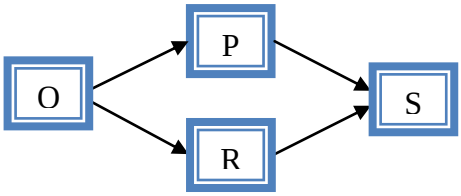

Figure 3: An illustration of backlinks

The rank of a page p can thus be written as:

$$PR(p) = d/n + (1 - d) \sum_{(q,p)\in G} \frac{PR(q)}{Outdegree(q)} \quad (4)$$

Where, n is a magnitude of nodes in the directed graph and OutDegree(q) is the magnitude of links on a page q. This comes to be seen as a stochastic analysis of a random walk on the Web graph. In equation 4, the first term in the right hand side corresponds to the likelihood that a random Web surfer arrives at a page p by a bookmark, or type the URL, or may have a particular page as his/her homepage. Here, d is the probability that a random surfer chooses the URL directly, rather than traversing a link, and 1 − d is the probability that a person arrives at a page by traversing a link. The second term match to the possibility of arriving at a page by navigating a link. This thought is illustrated in figure 4 by how the PageRank of a page p to be calculated [35]. Arbitrary, suppose any page P that has pages Q1 to Qn directed to it (inward link). Then, PageRank can be computed by below equation.

$$PR(P) = (1- d) + d(PR(Q1) /C(Q1) + ...+ PR(Qn /C(Qn)) \quad (5)$$

Where, the factor d is a damping factor and sets it to 0.85 (to prevent the other pages having a lot weight, this total vote is "damped down" by multiplying it by 0.85). C(Qi) is the amount of links departing out of page Qi, and PR(Qi) denotes the PageRank of the Pages Qi that links to page P. The PageRanks structures a probability distribution over the Web pages, therefore, the sum of all Web pages' PageRank will be one [36].

3.3. SALSA

SALSA algorithm is proposed by Lempel and Moran [3]. It is a combination of PageRank and HITS. It is a probabilistic extension of the HITS algorithm. The random walk starts from some authority's graph, alternating between the hubs and authority sides then proceeds by alternating between backward and forward steps. When at node on the hub side, the algorithm selects one of the outgoing links uniformly at random and moves on the authority. Random walk is done subsequently on hyperlinks in both direction two part random walks Hub walk and authority walk.

4. Anchor text and models in topic distillation

The anchor text in topic distillation overcomes the drawback of individual language for the identical web page. It can be used as a general Multilanguage description method. Other features of anchor text are; (i) The amount of quality items in the anchor text is extremely fewer in comparison with HTML, (ii) It is a reserve topic to link page, and (iii) The same URL contains more than one consideration anchor texts[37]. Zhong et al.[38] studied the hypertext-based topic analysis models and area-based topic distillation. They described the THTA (topic hub and topic authority) model and ETHTA (extension THTA) model related to hypertext topic distillation. A comparative

analysis was carried out between THTA, ETHTA and TOPHITS. It was concluded that the use of hypertext-related topic distillation improves the superiority of topic distillation. THTA, ETHTA models are more superior to HITS and TOPHITS model. By the use of annotations preface on IBM shows that it obtains better search quality. Dmitriev et al.[39] used session queries as implicit annotations, rather than explicit annotations and anchor to end pages and found that explicit annotations were difficult to obtain but very useful, implicit annotation could directed to boost in early precision.

Kraft and Zien [40] verified that anchor text is better than document collection to refine a query. Also, by exploiting the similarity between anchor text and search queries, high-quality refinements for one and two-term queries, which represent most of typical Web search engine's traffic. Lu et al.[41] proposed an approach to successfully exploits the anchor-text resources and partially solves the existing difficulties in term translation for applications of anchor text in the area of cross-language information retrieval where mining of anchor texts and link structures is used for automatically extracting translations of Web query terms[41]. An anchor text is also used as relevant queries to train a retrieval model [42]. Lempel and Moran [3] observed that HITS approach was weak to link spam and the TKC (tightly-knit community) effect; thereby it will drive pages within a tightly-knit community to high rankings even if the pages are not relevant, due to the inter-dependence between hubs-authorities. So, to get a solution, SALSA model was proposed to bear from TKC effect by weak combination between hubs-authorities. Here, node's authority (hubness) is distributed among the target equality during propagation. Whereas in original HITS, node provides the entire authority (hubness) to every target [3]. Anchor text and its applications have been usually experienced and tested by TREC Web track members for named page finding, homepage, ad hoc search tasks, and topic distillation[7].It was concluded from the use of anchor text lead to substantial performance gain from experiments with the TREC web track data [43,44, 45,46,47]. Eiron and McCurley [37] have showed an arithmetic study of anchor texts. It was originated that anchor texts resemble real-world user queries in language in their term allocation and length. The use of anchor texts was observed more valuable than headings or titles that were adopted in conventional text-based search. Westerveld et al presented a related method [48] but by unigram language model as an alternative to the BM25 model. Another model investigated by Fujii [49] with combination of text-related and the anchor-related retrieval method. In this model, anchor texts were cutting down to specific terms and the weight of each specific term was obtained by taking into account the weight of each anchor text towards the document and the weight of the term in the anchor text. In this retrieval method the probability is computed that a document is retrieved in response to a given query, synonyms of query terms in the anchor texts on the Web is identified, and synonyms are used for smoothing purposes in the probability assessment. Dou et al. [50] considered the relationship between target sites and anchor texts, and the anchor from related sites should be lower weighted than the ones from unrelated sites should be higher weighted. This work was determined on how to find ranking the anchor linked to the same page. For ranking the anchors of the web page new method was proposed by Dai and Davison [51], they consider historic trends of anchor texts into relation and added the power of time in turn into the weights of anchor texts. Metzler et al. [52] proposed a way to overcome anchor text sparsity by making enriched document with anchor texts that combined across the link graph so the effectiveness of retrieval can be significantly enhanced by auxiliary anchor text-enriched document representations. The resolution proposed by Metzler et al. [52] for aggregated anchor text prepared of anchor texts was; (i) derived outer of the Web site, and then (ii) linkage to Web pages inner side of a Web site, and then (iii) linkage to the Web page in query. This can be utilized to assist moderate the trouble of navigational, intra-server linkage with anchor texts similar to 'next' or 'click here.' An improvement to Web page depictions using aggregated anchor texts an establishment link paths, Weninger et al.[53] showed that the in turn from link paths can be utilized to get a better item search in site-specific search, and map Web

pages to database records. According to Omara et al.[54], the anchor text would be analyzed and used to enhance the clustering and labeling process. Lee et al. [55] proposed to recognize an intention of a query is navigational or informational, pretentious that transactional is integrated with navigational by using an allotment of clicks and that of anchor text. Their method mainly takes benefit of the subsequent heuristic in click-through and Web data to carry out the task. If a query is navigational, then the intent of the search query is to discover an exact web page. Thus, the widely held of clicks with the query ought to be on a particular URL in the click-through data. For example, the anchor texts "IBM" should direct to the URL ibm.com. In other words, the click ACM Transactions on intellectual Systems and allocation of a navigational query is skewed, and so is the anchor allocation of a navigational query. Yi and Allan[56] proposed utilizing content likeness between web pages and revealed a web page's probably misplaced in-link anchor text by utilizing its most similar web pages' in-link anchor text. Zhang et. al.[57] have evaluated page importance by using anchor texts and added in turn to anchor-related retrieval algorithm and showed that the models have high scores as compared to baseline BM25 model. Also, it was considered that the links as of the related sites have extremely dependent relative, and anchor texts from dissimilar sites had dissimilar weights and showed that ASRM that consider this variation among dissimilar sites i.e. *Homepage Retrieval Task* proved enhanced than the other two models LRM and ALRM in most of the cases. Liu et al.[58] proposed the enhanced HITS (I-HITS) algorithm based on likeness and esteem, which distinguishes the significance of links with the likeness of pages and the query topic and the esteem of pages. Theoretical analysis and investigational results both demonstrates that the I-HITS algorithm acts better in search precision and escapes the topic drift successfully. Furthermore, they found anchor text was utilized to depict the target document, not to depict the existing document [1,10] and it was summarized that the focus of the end document with an elevated degree of precision [11]. Hence, directed to decrease the computational complication, calculating the likeness of the end page and the query is cut down by computing the likeness of the anchor text and the query. Using the given individual name and aliases proposed method will calculate anchor texts-related co-occurrences among them, and will construct a word co-occurrence graph by making links between nodes indicating name and aliases in the graph based on their earliest order links with each other. The graph mining algorithm to discover the step distances between nodes will be used to recognize the link orders between name and aliases. To rank the anchor texts according to the co-occurrence statistics in order to identify the anchor texts in the first order links by Ranking SVM. The web search engine can enlarge the query on a personal name by tagging aliases directed to their links with name to retrieve all related results thereby getting better recall and significant MRR evaluated to that of earlier planned methods [59]. Berardi et al. [60] use anchor text in Blog distillation and sentiment analysis for decisive the force of a hyperlink on the weighting task and the sentence in which the anchor text was embedded were analyzed. In order to find out a response score for the hyperlink, calculate a weighted sum of the response scores of all the portions that emerge in the sentence including the anchor text, then use weights that are a declining purpose of the distance of the portion of the anchor text, according to the assumption that the closer a portion is to the hyperlink, the supplementary it is linked to it. This distance is itself calculated as a weighted sum, where each token between the anchor text and the portion has its weight depends on its nature; for instance, mood-altering elements such as "instead" are allotted a high weight. Song et al.[61] showed the approach of effectively improve tail query search e by expanding URLs with a textual context (by the click-through graph, anchor text web page titles, etc.) to map URLs to concepts. Once both queries (head-tail alike) and URLs are in same concept space, they computed their similarity as a relevance measure. Dang and Croft [62] introduce to utilize anchor text to reproduce parts of a log due to a likeness of anchor text to queries. This approach leads due to anchor are liberally available can be an effective replacement for a query log and study the use of a range of

query improve techniques (replacement and growth) using standard TREC. And show that using anchor text as a simulated query log is as slightest as effective as a real log. Al-akashi and Inkpen [63] presented custom indexing and ranking model, meta-terms or anchor texts extracted from the titles and URLs for indexing the contents of web documents, developed for the TREC 2012 web track. This approach is more sophisticated and robust for processing all types of queries. Samar et al.[64] proposed a study at a page level, and do not stop at just revealing the misplaced pages, but also propose to recover a depiction of these by using anchor text. Anchor text has previously enriched the web page content, mostly to get better retrieval. Combinations may consist of anchor text, but also the imitative structure of source and target sites, assigned categories or other extractable features. Only a small fraction of the revealed URLs can be improved by their anchor text. The integrated move towards to jointly search and hyperlinking of video snippets [65] is related to methods motivated by content-based schemes for multimedia recordings, which gives the majority related audio-visual parts to a certain text query or to another part, related to words. The similar approach is utilized for hyperlinking, but in adding up they use the visual concepts noticed in the anchor part and the indexed ones directed to re-rank responses based on visual likeness. In other words, while textual descriptions are suitable for queries for known-thing search (rather than asking users for visual query examples), anchors with better multi-modal content can valuably be used for hyperlinking.

An approach is given by Guisado-Gamez et al.[66] to expand the query anchor text of most important twenty links could be utilized in the work to rate the significance of the links, and then, comprise the strength of associations in their community detection algorithm which results in a major enhancement in terms of precision. It is believed that such knowledge could be established the community detection procedures to get better the quality of the system. They obtained major enhancements compared to the baseline using the path analysis of the Wikipedia and the anticipated community search algorithm. Jeong et al.[67] developed a method for automatically extracting the title of a Web page using the anchor text and the link information of the Web pages for searching or categorizing. They verified that using anchor text, achieves a high degree of accuracy showing 79.33% better results as the number of Web pages exponentially grows. A system of Dean et al.[68] may determine an extent to which a document is selected when it is included in a set of search results, search engine may generate (or alter) a score associated with a document based, at least in part, on information relating to a manner in which anchor text changes over time. Gasparetti et al.[69] analyzed the text snippets associated to links during a browsing session due to perception depends on the links anchor, that text can be considered strongly correlated to the user needs may be valuable for profiling users in personalized systems. It may introduce noise and degrade potential representations of user interests. Large retrieval systems, such as Google, can collect anchors from incoming links by sifting through a corpus of billions of pages and, thus, filtering out less useful information. Geng et al. [70] presented a variety of link-hiding techniques and organized them into taxonomy. They analyzed the prevalence of link-hiding techniques on the web. If an anchor text is not present in the text Vector, the hyperlink is recognized as hidden link. And, of course, the relative position of anchor text should also be taken into account. Weninger et al.[71] aimed to find the complete set of entity-pages in a Web site by each sorted bag of anchors the database is queried with the peak ranked anchor as the search term. Searching throughout the link path in downward order is particularly essential because the most explanatory anchors will emerge at the top of the list, and once a match is originated no need to continue searching. In fact, they discover that it is very likely that the peak ranked anchor text will hold a match to the database. Craswell et al.[72] introduced numerous robust algorithms for query rewriting in that their broad approach is to acquire a huge linguistic dataset, the ClueWeb09 anchor data, and build transformation and goal models. They applied these to produce revise of robust04, GOV2, and ClueWeb09 Web

Track queries, and complete retrieval supported by a fusion of novel and revise queries. They note that move towards based on exterior linguistic resources can potentially be enhanced merely by accumulating extra-linguistic data. Kong et al.[73] proposed to extract heterogeneous features by explicitly considering the users heterogeneous data within the networks, i.e., social, spatial, temporal and text information. They invent the suggestion difficulty for anchor links as a stable matching difficulty between the two sets of user accounts in two dissimilar networks. A useful answer, MNA (Multi-Network Anchoring), is derived to assume anchor links with respect to the one-to-one constraint. Wide experimentation on two real-world heterogeneous social networks shows that their MNA model constantly better than commonly-used baselines on anchor link prediction. Lee and Croft [74] used social anchors and test collection based on ClueWeb09 together as a corresponding source of evidence to usual anchors for ad-hoc search and perform studies on the comparative impact of the community anchor features, and establish that query-dependent features can be further significant for retrieval performance than the autonomous ones. Dai [75] proposed to incorporate anchor text for improving search relevance by three aspects: (i) the relationship between anchor with a same page, (ii) the relationship between anchor and target pages; and (iii) the relationship between similar anchors pointing to similar pages. They incorporate aspects into a unified optimization framework, to enhance basic anchor text importance. Experimental results show it significantly improved baseline and useful for answering informational and bi-term queries.

In Kamps et al.[76] experiments, full-text suffers from spam while anchor text is less targeted by spammers. If search results rerank by combining the retrieval score with the spam score, it can improve the effectiveness of anchor text that, does not suffer from spam for locating key resources. According to the experiments of [77] for the initial text-based run, anchor text is very effective as it has more relevant documents in the top 20 ranks than full-text runs, which cover more diverse aspects of the search topic.

5. Conclusions

Most of the web information retrieval tools only use the textual information while ignores the link information that could be very valuable. The aim and contributions of this study were done to explore the link structure for ranking and utilization of anchor text in IR. Experimentation with the TREC web track data via anchor text has shown significantly increased its performance in topic distillation. Hence, how anchor text can be utilized to get better search superiority in topic distillation was reviewed. By considering anchor texts, one can further improve Web search rankings, especially for the navigational queries, named page finding, homepage, and ad hoc search tasks. The further research on this area by using links to assist the search engine to extract pertinent and valuable link.

References

[1] M. Pandia, S. K. Pani, S.K. Padhi, L. Panigrahy and R. Ramakrishna, A review of trends in research on web mining, Int. Journal of Instrumentation, Control & Automation. 1(2011) 37-41.

[2] M. Henzinger, Link analysis in web information retrieval, IEEE Data Engineering Bulleitin. (2000)1-6.

[3] R. Lempel, S. Moran, The stochastic approach for link-structure analysis (SALSA) and the TKC effect, Computer Networks. 33(2000) 387–401.

[4] L. Page, S. Brin, R. Motwani, T. Winograd, The page rank citation ranking: Bringing order to the web. (1998).

[5] M. Wu, F. Scholer, A. Turpin, Topic distillation with query-dependent link connections and page characteristics, ACM Trans. on the Web. 5(2) (2011) 3-25.
[6] T. Tsikrika, M. Lalmas, Best entry pages for the topic distillation task, qeen marry, university of London. (2005).
[7] M. Wu, D. Hawking , A. Turpin, F. Scholer, Using anchor text for homepage and topic distillation search tasks, Journal of the American Society for Information Science and technology. 63(6) (2012) 1235–1255.
[8] N. Craswell and D. Hawking, Web information retrieval, Information Retrieval: Searching in the 21st Century, Wiley. (2009) 85-101.
[9] R. Jain and G. N. Purohit, Page ranking algorithms for web mining, International Journal of Computer Applications. 13 (2011) 22-25.
[10] S. Chakrabarti, Integrating the document object model with hperlinks for enhanced topic distillation and information extraction, In proceedings of the 10th international conference on World Wide Web, Newyork, USA. (2001) *211-220.*
[11] V. Katz and W.S. Li, Topic distillation on hierarchically categorized web documents, In Proceeding of the 1999 Workshop on Knowledge and Data Engineeering Exchange, Chicago. (1999) 34-41 .
[12] M. Gupta, V. Tomar, J. Verma And S. Roy, Mining databases on world wide web, International Journal of Computer Science. 8 (2011) 560-564.
[13] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM. 46(5) (1999) 604–632.
[14] L. Getoor and C. P. Diehl, Link mining: A survey, SIGKDD Explorations. 7(2) (2005) 3-12.
[15] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan-Kauffman Publishers, 2002.
[16] K. Bharat and M. R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, In proceeding of 21th ACM SIGIR Conference on Research and Development in Information Retrieval. (1998) *104-111*.
[17] S. Chakrabarti, M. Joshi and V. Tawde, Enhanced topic distillation using text, markup tags, and hyperlinks, In proceeding of 24th ACM SIGIR Conference on Research and Development in Information Retrieval. (2001) 208-216.
[18] D. Rafiei, A. O. Mendelzon, What is this page known for? Computing web page reputations, In Proceedings of 9th Int. WWW Conf., Amsterdam. (2000) 1-17.
[19] T. Haveliwala, Topic-sensitive pageRank, In Proceeding of the 11th International World Wide Web Conference, ACM. (2002).
[20] I. Choi and M. Kim, Topic distillation using hierarchy concept tree, *SIGIR'03,* ACM. (2003) 371-372.
[21] A. Borodin, G. Roberts, J. Rosenthal, P. Tsaparas, Finding authorities and hubs from link structures on the world wide web, Proceedings of the 10th International World Wide Web Conference. (2001) 415–429.
[22] N. Modi and A. A. Desai, Analysis of link-structure algorithms for web mining, Journal Of Veer Narmad South Gujarat University. IV B(2006)32-42.
[23] N. Modi, Link-Structure and semantic analysis of hyperlink graph towards categorization of web communities, PhD Thesis, VNSGU (2010).
[24] N. Modi and A. A. Desai, Web page categorization: Machine learning for web IR, Innovative Dimension for Business and Info. Tech. iDBiT: (2008) 29-31.
[25] N. Modi and A. A. Desai, Analyzing spatial locality on web graph: concept searching with search agent, Proceedings of ICETAETS. (2008) 828-831.
[26] N. Modi, Finding communities on social network of web pages using eigenvector method,Vnsgu Journal Of Science And Technology. 3(2) 2012, 89-97.
[27] Y. Xie, and T. Z. Huang, A model based on cocitation for web information retrieval, Mathematical Problems in Engineering. (2014)1-6.
[28] X. Tian, Y. Du, W. Song, W. Liu, and Q. Meng, Improvements of HITS algorithm based on content analysis, Journal of Comp. Info. Sys. 10(10) (2014) 4049-4058.

[29] S.B. Boddu, V.P Krishna Anne, R. R. Kurra, D. K. Mishra, Knowledge discovery and retrieval on world wide web using web structure mining, In Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference, IEEE. (2010) 532-537.
[30] M. K. Hussein and M. H. Mousa, An effective web mining algorithm using link analysis, (IJCSIT) Int. Journal of Computer Science and Information Technologies. Vol. 1 (3) (2010) 190-197.
[31] M. P. Selvan, A .C. Sekar, A. P. Dharshin, Survey on web page ranking algorithms, International Journal of Computer Applications. 41(19) (2012) 1-7.
[32] A. Y. Ng, A. X. Zheng and M. I. Jordan, Stable algorithms for link analysis, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. (2001) 258-266.
[33] N. Radhika and K. Vidya, Association rule mining based on ontological relational weights, Int. Journal of Scientific and Research Publications. 2(1) (2012) 1-5.
[34] D. K. Sharma and A. K. Sharma, A comparative analysis of web page ranking algorithms, (IJCSE) Int. Journal on Comp. Sci. and Eng. 2(8) (2010) 2670-2676.
[35] J. Srivastava, P. Desikan, V. Kumar, Web mining – concepts, applications and research directions, In Foundations and Advances in Data Mining, Springer-Verlag Berlin Heidelberg StudFuzz 180 (2005) 275–307.
[36] S. Brin and L. Page, The anatomy of a large-scale hyper textual Web search engine, Computer Networks and ISDN Systems. 30(1–7) (1998)107–117.
[37] N. Eiron and K. S. McCurley, Analysis of anchor text for web search, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. (2003) 459-460.
[38] J. K. Zhong, L. Zhao, W. Y. Qiong, G. J. Zhong, An algorithm of topic distillation based on anchor text, International Symposium on Electronic Commerce and Security, IEEE computer society. (2008)11-15.
[39] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita, Using annotations in enterprise search, In Proc. of the 15th int. Conf. on WWW, ACM. (2006) 811-817.
[40] R. Kraft and J. Zien, Mining anchor text for query refinement, In Proceedings of WWW. (2004) 666-674.
[41] W.-H. Lu, L.-F. Chien, and H.-J. Lee, Translation of web queries using anchor text mining, ACM Transactions on Asian Language Information Processing (TALIP),1(2) (2002)159–172.
[42] R. Nallapati, W.B. Croft and J. Allan, Relevant query feedback in statistical language modelling, In Proceedings of CIKM. (2003) 560-563.
[43] N. Craswell and D. Hawking, Overview of the TREC-2004 web track. In Proceedings of the 13th Text Retrieval Conference,TREC 2004. (2004)1-9.
[44] N. Craswell, D. Hawking, and S. Robertson, Effective site finding using link anchor information. In Proceedings of the 24th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval ACM. (2001) 250–257.
[45] B.D. Davison, Topical locality in the web, In Proc. of the 23rd Annual Int. ACM SIGIR Conf. on Research and Development in IR. (2000) 272–279.
[46] M. Koolen and J. Kamps, The importance of anchor text for ad hoc search revisited, In Proceedings of 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2010)122–129.
[47] V. N. Anh and A. Moffat, The role of anchor text in ClueWeb09 retrieval, In TREC. (2010).
[48] T. Westerveld, W. Kraaij and D. Hiemstra, Retrieving web pages using content, links, urls and anchors, In Proc. of 10th Text REtrieval Conf. (2001) 663–672.
[49] A. Fujii, Modeling anchor text and classifying queries to enhance web document retrieval, In Proceeding of WWW ’08, ACM. (2008) 337–346.

[50] Z. Dou, R. Song, J.-Y. Nie and J.-R. Wen, Using anchor texts with their hyperlink structure for web search, In In Proc. 32nd Annual Int'l ACM SIGIR Conf. on Research and Dev. In Information Retrieval, (2009) 227-234.
[51] N. Dai and B. D. Davison, Mining anchor text trends for retrieval, In Proceedings of ECIR. (2010) 127–139.
[52] D. Metzler, J. Novak, H. Cui, and S. Reddy, Building enriched document representations using aggregated anchor text, In Proc. of the 32nd int. ACM SIGIR conference on Research and development in information retrieval. (2009) 219-226.
[53] T. Weninger, C. Zhai, J. Han, Building enriched web page representations using link paths, ACM.(2012) 53-62.
[54] F. A. Omara, N. A. El-Fishawy, M. Amoon and S. El-kazaz, Analysing anchor links to enhance the web snippet clustering technique, The 8th Int. Conf. on INFOS Advances in Software Engineering Track, IEEE. (2012) SE-7 – SE - 11.
[55] U. Lee, Z. Liu, and J. Cho, Automatic identification of user goals in web search, In Proceedings of the 14th Int. Conf. on WWW'05, ACM. (2005) 391–400.
[56] X. Yi and J. Allan, A content based approach for discovering missing anchor text for web search, In Proc. of the 33rd int. ACM SIGIR conf. on Research and development in information retrieval, ACM. (2010) 427-434.
[57] Y. Zhang, K. Lei and Li. Huang, Using anchor text refined by page importance to improve web retrieval, The 7th Int. Conf. on CS & Edu., IEEE. (2012) 1200-1203.
[58] X. Liu, H. Lin and C. Zhang, An improved hits algorithm based on page query similarity an page popularity, journal of computers, 7 (1) (2012) 130–134.
[59] B. Rama Subbu Lakshmi, R. Jayabhaduri, Automatic discovery of association orders between name and aliases from the web using anchor texts-based co-occurrences. International Journal of Computer Applications. 41(19) (2012) 30-35.
[60] G. Berardi, A. Esuli, F. Sebastiani, and F. Silvestri, Blog distillation via sentiment-sensitive link analysis, In Natural Language Processing and Information Systems ,Springer-Verlag Berlin Heidelber. (2012) 228–233.
[61] Y. Song, H. Wang, W. Chen, S. Wang, Transfer understanding from head queries to tail queries, In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management , ACM. (2014) 1299–1308.
[62] V. Dang and W. B. Croft, Query reformulation using anchor text, In Proc. of the 3rd ACM Int. Conf. on Web Search and Data Mining, WSDM'10. (2010) 41-50.
[63] F. H. Al-akashi And D. Inkpen, Term impact-based web page ranking, In Proc. of the 4th Int. Conf. on Web Intelligence, Mining and Semantics ACM. (2014).
[64] T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. de. Vries, Uncovering the unarchived web, SIGIR'14,ACM. (2014) 1199-1202.
[65] C. Bhatt, N. Pappas, M. Habibi, A. Popescu-Belis, Multimodal re-ranking of content-based recommendations for hyperlinking video snippets, In Proceedings of International Conference on Multimedia Retrieval, ICMR '14, ACM. (2014).
[66] J. Guisado-Gámez, D. Dominguez-Sal, and J. L. Larriba-Pey, Massive query expansion by exploiting graph knowledge bases for image retrieval, Proceedings of International Conference on Multimedia Retrieval, ACM. (2014)
[67] O. R. Jeong, J. Oh, D. J. Kim, H. Lyu, and W. Kim, Determining the titles of Web pages using anchor text and link analysis, Expert Systems with Applications. 41(9) (2014) 4322-4329.
[68] J. Dean, P. Haahr, M. Henzinger, S. Lawrence, K. Pfleger, O. Sercinoglu, and S. Tong, Document scoring based on query analysis, *U.S. Patent No. 8,639,690*. Washington, DC: U.S. Patent and Trademark Office. (2014)
[69] F. Gasparetti, A. Micarelli, G. Sansonetti, Exploiting web browsing activities for user needs identification, International Conference on Computational Science and Computational Intelligence, IEEE. (2014) 86-89.

[70] G.G. Geng, X.T. Yang, W. Wang, and C.-J. Meng, A taxonomy of hyperlink hiding techniques. (2014) 1-12.
[71] T. Weninger, T. J. Johnston, and J. Han, The parallel path framework for entity discovery on the web, ACM Transactions on the Web. 7(3) (2013) 16(16.1-16.29).
[72] N. Craswell, B. Billerbeck, D. Fetterly, and M. Najork, Robust query rewriting using anchor data, WSDM'13. (2013) 335-344.
[73] X. Kong, J. Zhang, and P. S. Yu, Inferring anchor links across multiple heterogeneous social networks, In Proc. of 22nd ACM international conf. on information & knowledge management CIKM'13, ACM. (2013)179–188.
[74] C.-J. Lee and W. B. Croft, Incorporating social anchors for ad hoc retrieval. In Proc. of 10th Conf. on Open Research Areas in IR.OAIR'13. (2013)181–188.
[75] N. Dai, Building contextual anchor text representation using graph regularization. In AAAI. (2012) 24-30.
[76] J. Kamps, R. Kaptein and M. Koolen, Using anchor text, spam filtering and wikipedia for web search and entity ranking. (2010).
[77] R. Kaptein, M. Koolen and J. Kamps, Result diversity and entity ranking experiments: Anchors, links, text and wikipedia, Amsterdam Univ (Netherlands) Intelligent Systems Lab Amsterdam. (2009).
[78] The CLEVER project, IBM, http://www.almaden.ibm.com/cs/k53/clever.html