10.31341/jios.48.1.1 UDC 004.775
Open Access Preliminary Communication

Relevancy between Anchor Text and Wikipedia: A Web Search Framework

Falah Al-akashi *falahh.alakaishi@uokufa.edu.iq*
Faculty of Engineering
University of Kufa, Najaf, Iraq

Diana Inkpen *Diana.Inkpen@uottawa.ca*
Faculty of Engineering
University of Ottawa, Ottawa, Canada

Abstract

The overall volume of data available on the Internet is growing rapidly while finding relevant documents is becoming increasingly difficult. Moreover, queries entered by users are unique, unstructured and often ambiguous while the process has changed dramatically from standard query languages that governed by strict syntax rules to unstructured strings. In Web information retrieval, search paradigms used term occurrences to weight document content prior to any boosting stage. PageRank algorithm, for instance, was used integrated techniques to enhance post retrieval document relevancy to adequately compromise the overall process in two stages. Nevertheless, hypertexts in Web have been used for improving the quality of search results for the most common type of queries. Our main premise is that hypertexts play an important role for ranking documents in IR such as margining between user queries and consensus hypertext. We propose a new algorithm that uses term impact technology for compromising hypertext weighting in Web along with Wikipedia for efficiently find most relevant documents among large set of results. Our experimental results showed that Wikipedia could efficiently improve document relevancy rank when combined with hypertexts for exhibit robust and very good short-term process capability.
Keywords: World Wide Web, Entity Weighting Schema, Data Fusion, Anchor Text

1. Introduction

Over the past two decades, the benefits of technology advances on the Internet have grown exponentially. Although casual Internet surfing is typically enough to determine the relevancy of documents, there are huge volume of information and innumerable links available online can complicate the privileges of users for processing access to information requests. In addition, search engines faced high volume of difficulty to determine the relevance rank of documents on the Internet, whilst information has become useless unless involved to a validation technique of weighing and sorting. From the first indexing approach in 1990 (Archie) to the modern search algorithms used today, the problem for deciding the relevance of information

remains a critical issue because Web is fuelled by vast troves of data and becomes a surveillance platform and gatekeepers of innovation. An effective way to compute the relevancy of documents is to determine the empirical evidences in the content of the documents using an automated algorithm for predicting web topics. Researchers had developed an efficient method for computing the relevance of hypertext along with a link structure technique to weight a hypertext graph (Yadah et al. 2009). Often, weighting documents is assigned for every on-bound and outbound links on a directed graph, and the inter-links between webpages have been improved to govern the information that target the links for observing the results. We believe such algorithms might be very useful when incorporated with other existing algorithms as web usage mining has been increased exponentially in the last decade. This tremendous collection of web pages along with the degree of topics between websites could pose challenges to the information retrieval systems.

Often, indexing starts by terms have supposedly been appeared more significant. Because search engines are software programs, they work according to the rules established by their designers to determine which terms are usually important in a broad range of documents. First, the titles of pages usually convey specific and helpful information to a specific user/audience. Second, terms positioned at the beginning of documents give more weight as they repeat several times on the document. Third, hypertexts could significantly enhance Web page retrieval by improving the semantic retrieval of Web data (Halpin and Lavrenko, 2011). Fourth, optimizing hypertexts could improve Web ranking and reduce search engine impact traffic. Sometimes hypertext used typically to indicate the subject matter of pages to which they were topically linked (Yadav et al, 2009); for instance, the keywords in hypertexts could improve the relevancy of target pages. This pattern of Web data has been exploited by search engine algorithms to enhance the relevancy of target pages due to some relevancy of keywords appeared on some pages, this was boosted our previous framework (Al-akashi and Inkpen, 2014) using the linked keywords on the target URLs. Traditional search engines, e.g. Google, Yahoo, etc., pretended to exploit the content of hypertexts for a valid indexed page, which configured to index hypertexts in a separate table to make it evident. Recent studies showed that weights given to hypertext have been optimized among traditional search algorithms. The basic technique to materialize a portion of a web page is to navigate from known hyperlinks using regular expressions while hypertexts played a vital boosting to improve the retrieval performance (Craswell et al., 20sf88u11). The experiments have been conducted that ranking based on hypertexts was twice as effective ranking based on document contents in finding the main entry point of a specific topic. Hypertexts are considered relevant and consequently the Page Rank algorithm was applied in hypertext contents to improve document relevancy because hypertexts typically match or share similar properties of titles and that's why documents targeted by different hypertexts. This is a significant indicator that hypertexts acquire high relevant text for targeting the relevancy of documents. Retrieval approaches that leverage extra evidences include a well-known PageRank algorithm (page et al., 1999; Kleinberg, 1999) and HITS (Boytsov and Belova, 2012) to estimate the credibility of pages based on the number of ongoing and outgoing links. However, TREC Web track provided a

subset collection that involved more hyper textual content than the whole collection (Margin and Jaap, 2011). Other studies showed that hypertext annotators somewhat useful for the homepage finding task (McBryan, 1994). Other external evidences such as outgoing links pointed to the retrieval documents were also useful for topic distillation tasks (Chakrabrti et al., 1998). As a result, links and hypertext are important for increasing the relevancy of ranking algorithms but queries entered by users showed more challenges to determine user intent since they focus on multiple viewpoints at different times. Moreover, keywords entered by users might involve problem for distinguishing between some terms spelled similarly when involved different meanings, synonyms, or sometimes not entered by a query string. For instance, a query about 'heart' in the class of diseases might return results related to a term 'cardiac'.

The rest of this paper is organized as follows: Related work will be outlined in Sections 2. Hypertexts and Hyperlinks evidences are elaborated in Section 3. Section 4 will discuss our overall approach. Section 5 will discuss our experimental results. Hypertexts diversification will be demonstrated in Section 6, and finally, conclusion and future insights will be outlined in Section 7.

2. Related Work

Hypertexts are used for various Web information retrieval tasks. When the Internet became established, McBryan (1994) proposed that hypertexts were important to a Web searching. Most current Web search engines use hypertexts as evidences to improve search relevancy ranks. Contextual text in a specific vicinity of the hypertexts would automatically compile lists of authoritative Web resources for a range of topics (Atsushi, 2008). Nadav and Kevin (2003) conducted experiments to investigate several aspects of hypertexts, including their relationship to titles, the frequency of the queries that could be satisfied by hypertexts alone, and the homogeneity of the results acquired by using hypertexts only. They found that hypertexts were typically less ambiguous than other types of texts and enabled more coherent and focused results for queries solved by hypertexts more than for those based on other features in the corpus. In addition to providing a better match than titles or content, hypertexts also have a potential delivery on authoritative search results. Atsushi (2008) proposed modeling hypertexts and classifying queries to identify synonyms of query terms in the hypertexts. He also assumed using synonyms for smoothing purposes could enhance the document relevancy ranks. As hypertext window considered important for other tasks (Davison, 2000; Attardi et al, 1999; Shuming et al, 2009; Ganeshiya and Sharma, 2014); Shuming et al., (2009) proposed that hypertext were not necessarily needed in window to improve relevant/non-relevant documents. Although they approached windows implicitly, they adopted Pseudo URLs and machine learning approaches to exploit the citation relationship and extract pseudo hypertexts for academic articles. They also proposed that the extracted pseudo anchors were useful for improving search performance. Clarke and Cormack (1995) explored the important of hyperlinks and various resources or parameters could be used to effect on-page ranking factors, e.g. in-links and out-links, anchor text, anchor-related text,

etc. The hypertext in a cloud computing paradigm is important since they induced topic selection for addressing the abundance problem (Albi and Silvester, 2017).

3. Hypertext vs. Hyperlink

Hyperlink and hypertext are a fundamental core of the internet and also a foundation of SEO. Using hypertext evidences to rank web documents rather than a document full-text search seems significant and constant for increasing effectiveness of homepage finding task and topic distillation task (Hiemstra and Hauff, 2010). Removing aggregated hypertexts length normalization altogether or normalizing according to a full-text document length was also found to improve the retrieval effectiveness. The most effective usage of hyperlink scores was to reduce the corpus size without decreasing homepage finding tasks. Document evidences should include full-text evidences and other useful document level evidences while Web based evidences might use incoming hypertext and other helpful external document features. For a home page finding task, URL depth features are measured literally for re-ranking documents within URLs' lengths under n% characters, or by adding a normalized URL length score to a query dependent score. The results of such experiments showed that (i) the importance of both hypertext and URL length for home page finding tasks, (ii) both PageRank and in-degree depth performed similarly and are highly correlated, (iii) using hypertext evidences to rank documents rather than document full-text provide significant effectiveness improvements in homepage finding and topic distillation tasks, and (iv) hyperlink recommendation evidences are far less effective than URL based measures.

However, search engines have been used multiple evidences for facilitating the retrieval performance while popular models focused on full-text document content for evidences. Regardless of document content is relevant, other content, such as videos, audios, graphics, etc., were also indexed (Anh and Moffat, 2010). While some approaches focused on hypertext, the visible keywords in hyperlinks and text appeared within bounds of tags when linked to other documents. They could provide search engines and users with relevant contextual information regarding the content of target websites. It was observed that hypertext in web documents were very useful to improve web search quality for most types of queries. Search engines use external hypertext for reflecting the issue how people view pages and what topics of pages they focused on? While websites typically could not control links, hypertext used in websites is useful, descriptive, and relevant. If many websites assumed that a particular page is relevant for a given set of terms, the page might be ranked highly even if terms were not appeared in the text itself. Search engines algorithms have had been developed extremely and they can now identify more metrics to determine relevant web pages. One important metric is a link relevancy, means how a topic in page 'A' is related to page 'B' if two pages are linked together and took users and search engine robots/crawlers from page 'A' to page 'B'. A highly relevant link could improve the rank of both pages 'A' and 'B' for queries related topics. Links pointed to a content related to the topic of source pages often send stronger relevance signals than links pointed to unrelated content. For example, a page about the best lattes in

Seattle might pass a better relevance signal to Google when it links to the coffee shop's websites rather than web sites with pictures of baby animals. Traditional search engines considered different hypertext variations that linked back to the original articles, and used them for additional weights for what topics of articles are and which search queries are relevant. If too many inbound links contain similar hypertexts, it could appear suspicious signs for links were not acquired truly. In general, it is still best practice to obtain and use keywords and text and topics of specific anchors. However, search engine optimization might achieve better results when search for a variety in hypertext phrases rather than similar keywords each time. Current studies proposed that indexing documents based on their term frequency is not adequate to determine their relevancy. Some search engine algorithms use different strategies to increase the impact of documents on some websites rather than others, for example, Google PageRank algorithm used inbound and outbound hyperlink in their experiments (Kamps et al., 2010) whereas impact of terms on document's header was used to weight the associated document's content (Serdyukov and Vries, 2009). However, developing algorithms to improve retrieval effectiveness and answering the prompt queries from large collections is critical especially when algorithms preserved to improve precision and recall of algorithms. Several systems are already used different relevance ranking models than conventional information retrieval models.

4. Our Overall Approach

The TREC Web track coordinators at the University of Twente provided researchers with a dataset from the ClueWeb09 collection named subset (A) which was 500 million documents, approximately involved 2.5 billion hyperlinks and hypertexts in a file of 10 Gigabytes (Cormack, 2010). This was a set of links used experimentally in the articles of 440,678,986 documents, 87% of English documents had hypertexts while the collection of subset (B) was 45,077,244 documents (approximately 90% of collection B). To process such corpus, hyperlinks referred to the same destinations were normalized by discarding the duplicated links from the whole collection. Then, hypertexts and hyperlinks had been compromised to focus on the user's intention. However, researchers showed that indexing only hypertexts outperformed indexing the whole collection.

4.1. Weighting Hypertexts in Web

Hyperlinks in the shared hypertexts could be used to represent the unique URLs for the target documents, but it is important to compute the frequency of each shared link to calculate the number of outbound hyperlinks. To do this, all hyperlinks have been aggregated and combined in a single hash-table using a unique identifier while markup characters (e.g., https://, http://, www) stripped out initially. According to the Page Rank algorithm, the procedure was repeated recursively to compute the inbound and outbound value of each page or URL. From our perspective, computing the impact of each hyperlink is not limited by inbound and outbound values, which means, the impact is not specified by the Web designer solely but also by the user behavior. In

some cases, documents with only few links were visited more often according to the Alexa.com global rank. We approached this fact by weighting the total ranking value of each URL by creating a three featured hash table: hypertext, hyperlink, and its frequency. Recursively, the hyperlinks would aggregate each key shared similar URLs. Table 1 shows an example of frequency and Alexa networking weights of websites in a subset of aggregated outbound links.

Hypertext	Hyperlink	Local Weight	Global Weight
Bbcarabic	bbcarabic.com	0.19	0.61
BBC Arabic	bbcarabic.com	0.13	0.61
BBC News From bbcarabic.com	bbcarabic.com	0.04	0.61
prestige news mission	bbcarabic.com	0.08	0.61
Daily News BBC	bbcarabic.com	0.06	0.61
CNN Arabic	cnnarabic.com	0.12	0.48
CNN News	cnnarabic.com	0.05	0.48
CNN	cnnarabic.com	0.01	0.48
Hawaw	hawaw.com	0.01	0.00
Kids hats	flickr.com	0.06	0.67
handy pouches	flickr.com	0.02	0.67
Shopping bag	flickr.com	0.03	0.67
Arabic news	aljazeera.net	0.70	0.57
Arabic news	alarabiya.net	0.56	0.52

Table 1. Hypertext, local weights and global weights

Anchors that shared similar hyperlinks could combine the corresponding text to improve their weight efficiently. First, we stripped out all inappropriate phrases (e.g., click here) or implicitly contain that phrases since they decrease the impact of target documents. Second, the phrases that shared ongoing links are considered important to determine the keywords and the descriptions of the target documents. They referred to good resources for weighting document keywords rather than weighting document content itself. However, all phrases in hypertexts shared similar ongoing links that combined together to annotate the overall description of hyperlinks in which texts were changed to lowercase. For example, the hypertext "T" in hyperlink "L" was combined together to treat it as a bag of words. However, weight assigned to the similarity between the hypertext "A" and a query string "B" might be different than those between the content of the document "A" and the query string "B". In other words, the importance of hypertexts in the same collection was treated and weighted differently. The relevancy of hypertext "T" that pointed to hyperlink "L" was computed as follows:

$$Score(D) = \beta \sum_{i=1}^{n} W(T_i, L) \quad (1)$$

where β denotes the weighted factor used to normalize the weight of phrase 'T' in anchor 'L', and n is the number of phrases pointed to the target document 'D'.

The following algorithm shows how to combine and seize the similar phrases.

SELECT *d.URL, e.URL, a.LABEL*
FROM Document *d* SUCH THAT
'www.mysite.com' →* *d*,
Document *e* SUCH THAT *d* → *e*,
Anchor *a* SUCH THAT *a.base* = *d.URL*
WHERE *a.href* = *e.URL* **AND** *a.label* = 'label';

Vector Space Model (SVM) was used to map each document on a unique vector to be counted once a time, as shown in Table 2.

Hypertext	Hyperlink	Local Weight	Global Weight
Bbcarabic, bbc, arabic, news, from, bbcarabic.com, prestige, mission, daily, bbcarabic.com	bbcarabic.com	0.50	0.61
Cnn, arabic, news	cnnarabic.com	0.18	0.48
kids, hats	flickr.com	0.06	0.67
Handy, pouches	flickr.com	0.02	0.67
Shopping, bag	flickr.com	0.03	0.67
arabic, news	aljazeera.net	0.70	0.57
arabic, news	alarabiya.net	0.56	0.52

Table 2. Hypertexts combination, local weights and global weights

Figure 1 shows the execution time for indexing 2.5 billion of anchors in 7 days.

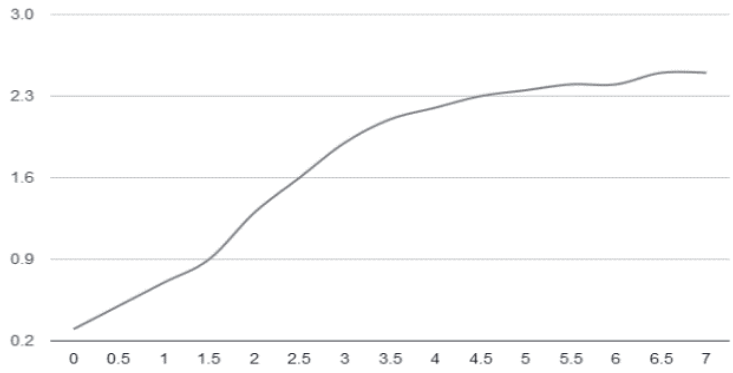

Figure 1. Hypertext indexing period within 7 days

4.2. Weighting Links in Wikipedia

Hypertext in a conventional or classical Web is different from that available in Wikipedia. Wikipedia repository is a significant source for information retrieval since

contains information very worthy. Researchers have had used external links in Wikipedia to successfully retrieve high relevant pages potentially works better than searching hyper information for the same task (Kamps et al., 2010)), and also, its content used for entity and home page finding tasks (Serdyukov and Vries, 2009). Most registered websites are referenced or cited by Wikipedia writers on the External Links and References sections to use URLs to link directly to any web pages. No page should be linked from a Wikipedia article unless its inclusion is justifiable according to the Wikipedia's guideline and copyright role. Longevity of links is a one of the most important feature that considers the link is likely to remain relevant and acceptable to the article in the foreseeable future. However, as Wikipedia is a major part of ClueWeb09 corpus, we approached our index to involve all external web pages. Wikipedia uses a standard appendices and footers template of terms, such as 'Official', 'Website', or 'Home Page', to represent an external website of relevant resource pages (example below). A regular expression was used to parse the literals in the standard terms and extract the corresponding text. Often, the hypertexts refer to the titles of the remote pages which extracted to create a class tag of index nodes. Hence, the indexed node would enclose all target pages referenced by articles alongside with home pages.

```
== External links ==
* [https://example.com Official website]
* [https://example.net/link Interview with Subject] in ‘’Example Magazine’’
* [https://example.org/repository/Subject Subject’s papers] at [[Example Museum]]
```

4.3. Improving Quality and Relevancy

Based on our research, the algorithm in the previous section is not applicable for all purposes since hypertext involves other features particularly when similar hypertext targets different hyperlinks, unlike different hypertexts target similar hyperlinks. As shown in Table 1, 'Arabic news' points to different hypertexts. Alexa Network Global Rank[1] is a rich information repository that weighted resources based on website's visitors. As a result, our approach exploits such worth information for improving the relevancy ranks of all collected hypertexts. We proposed a weighted threshold for each hypertext, if it exceeded a frequency "45", it was discarded. Equation 2 used to strip out the irrelevant hypertexts and to compute the score of relevant documents:

$$P(D)=\frac{|A|^{\pi}}{\sum_{A\in D}|A|^{\pi}} \quad (2)$$

where π denotes ALEXA normalization impact factor on hypertext A.

The final retrieval weight of a document D pointed by a hypertext A is computed as follows:

$$Weight(D, A)=Score\ (D)\ .\ P(D) \quad (3)$$

[1] www.alexa.net

Some irrelevant hypertexts were weighted too high; to address this issue all duplicated terms on the hypertexts alongside with all the following references were stripped out in the earlier stage:

1. **Harmful hyperlinks:** When there an excessive short or long hypertext repeated with some promoting text that targeted some particular sources.
2. **One way hypertext backlink:** When some hypertexts have backlinks to external web pages.
3. **Keyword stuffing**: When hypertexts on a page associate with many resources and target a particular page but with different hypertexts. This causes an algorithm to pay more attention to avoid spam links or black hats (e.g., '*pay day loans*', '*buy now', 'viagra*').
4. **Poor frequency:** Some hyperlinks encountered poor frequency (e.g., 2 or 3), such links typically contain generic text, such as: '*click here*', '*read more*', '*check this out*', '*more info here*', '*login here*', '*email me*', '*admin*', etc.

Hence, the index was minimized to 77,048 tables, in which, each contained an approximately 1,000 vectors (80 Kb in size). Experimentally, we noticed that all hyperlinks pointed to a similar website should be combined in a similar index table and might be shared with other tables. This influenced our approach to use a hash table which is a block index orienting schema for improving the indexing performance and accelerating the query processing time. Each table was represented in a bag-of-words model (BoW) and defined by a particular identifier. The Apache Lucene[2] indexing framework, as a high performance and full-featured indexing library, then used to index all tables. As a result, 77,048 index tables were stored to be used later for a query processing. Figure 2 shows the tradeoff between hypertexts frequencies and types of relevancy. The fallacies of relevancy, spams for example, clearly fail to provide adequate result for true information. Although they often used to attempt to persuade users by irrelevant means, the predisposed users are apt to be fooled.

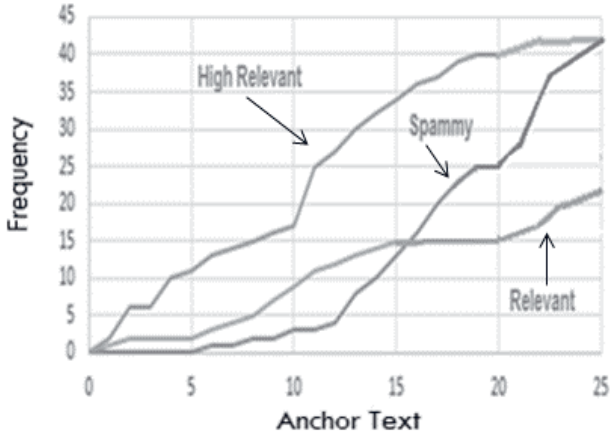

Figure 2. Hypertext relevancy

[2] www.luecne.com

4.4. Query Processing

Our experiments indicated that Apache Lucene is not an ideal tool for hypertext indexing because it deals with text like a bag of words and ignores other related features, such as bigrams and synonyms. Furthermore, the weight given to the similarity between hypertext and query might be different than the weight between documents and query because the relevancy of hypertext and query on document *D* might be different. Apatche Lucene typically uses vector space model for dealing with data as set of vectors and then invokes cosine similarity function for ranking, which is a suitable tool for indexing documents rather than hypertext. We deal with each index as a block content that composited each vector for each document. However, Lucene retrieved block of documents that matched every user's query despite the table implied several vectors (e.g., hypertext, hyperlink, etc.). Thus, we needed to examine vectors sequentially and track it more precisely to determine the relevant ones whose match a query string because the query string might be showed sparsely and distributed over the content of index table. A common way to measure the similarity of two vectors is to compute the similarity distribution of terms; by which, we used a cosine similarity function for vector similarity rather than for block or table similarity. Given two probability distributions, document D and query Q with a set of terms *t*, the cosine distance of two non-zero vectors was derived for two vectors of attributes as follows:

$$\text{Similarity (Q, D)} = \text{Q.D} = ||Q||\ ||D|| = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} Q_i^2}\ \sqrt{\sum_{i=1}^{n} D_i^2}} \quad (4)$$

where Q_i and D_i are components of vector Q and D, respectively.

Thus, the cosine similarity between a query Q and documents was ranged between 0 and 1.

4.5. Query Refinement

We emphasized how Wikipedia articles are connected to form an online resource in which Wikipedia writers join similar/ related articles using hyperlink. This means Wikipedia articles can expand current topic by transferring the current articles to other articles using interconnected hyperlink (inbound links); similarly, the target articles often point back to reliable articles using outbound links. For instance, if article *A* point to article *B* while article *B* point back to article *A*, we assume articles *A* and *B* are related topically. However, all ingoing and outgoing links in each article are indexed using a custom hash table index platform. The article titles would utilize the keys in the table whilst the outbound links would store the content of the corresponding keys and hence the index would map the query for processing. Let's consider an article *Global Warming* points to the articles *Carbon Dioxide*, *Air Pollution*, *Greenhouse Gas,* and *Alternative Fuel*; these articles would point back to the source article *Global Warming*. Similarly, the article *Automobile* points to three articles while the destination articles points back to the source article. This strategy was used to expand the original query and refine other results from the related topics.

Figure 3 shows an example of expanding a query *Global Warming* by *Carbon Dioxide*, *Air Pollution*, *Greenhouse Gas* and *Alternative Fuel*; whereas, the query *Automobile* was expanded by *Carbon Dioxide*, *Air Pollution* and *Greenhouse Gas*. However, query expansion was not used for all queries but only when the resulting list was short.

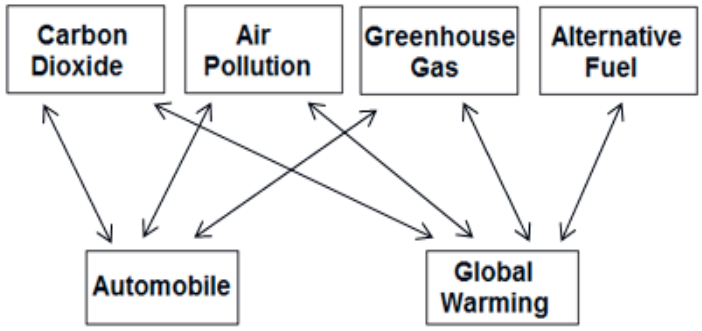

Figure 3. Query Expansion

5. Experimental Results

Often, traditional search engines are judged by some people and candidates. Human relevance judgment is popular for evaluating search algorithms in the market. The TREC Web Track queries were used to compare the results of our algorithm with the results of other algorithms in a task that annotated the best algorithm for each test set using relevant judgement. Our algorithm was filtered out scams and spams using an impact factors identified by ALEXA networking, in which, a document with a ranking score greater than the threshold 0.4 was considered a spam, therefore, discarded from the final results (Cormack, 2010). Table 3 and Table 4 show our best results compared with the results from the related algorithms. We used a precision score at different recall levels P@k to denote a proportion of relevant items amongst other items located on the first k retrieved documents, while Main Average Precision (MAP) denotes the effectiveness performance of adhoc retrieval tasks. The effectiveness performance of diversity tasks[3] for the same queries is shown in Table 5 and Table 6. Our run achieved a significant and substantial improvement in the effectiveness and relevance of retrieval performance in regard to 50 test queries (Table 7).

Run	Description	MAP	P@5	P@10	P@20
Muadanchor	Anchor only	0.0256	0.296	0.250	0.179
Our system	Anchor only	0.0293	0.321	0.284	0.307

Table 3. Effectiveness performance of Web track Adhoc tasks (query list 2011)

[3] The diversity task asked to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list.

.

Run	Description	MAP	P@5	P@10	P@20
Uma10BASF	Anchor + content	0.088	0.383	0.356	0.321
Uma10IASF	Anchor + content	0.087	0.394	0.358	0.319
Our system	Anchor only	0.113	0.471	0.477	0.421

Table 4. Effectiveness performance of Web track Adhoc tasks (query list 2012)

Run	nDCG@5	nDCG@10	nDCG@20	P-IA@5	P-IA@10	P-IA@20
mudvimp	0.220	0.241	0.268	0.091	0.073	0.061
Our system	0.227	0.411	0.392	0.131	0.112	0.195

Table 5. Effectiveness performance of Web track diversity tasks (query list 2011)

Run	nDCG@5	nDCG@10	nDCG@20	P-IA@5	P-IA@10	P-IA@20
Uma10BASF	0.275	0.336	0.379	0.162	0.152	0.131
Uma10IASF	0.281	0.335	0.380	0.165	0.144	0.130
Our system	0.311	0.370	0.472	0.201	0.179	0.218

Table 6. Effectiveness performance of Web track diversity tasks (query list 2012)

Test Queries	P@1	P@5	P@10	P@20
403b	1.00	0.80	0.60	0.60
angular cheilitis	1.00	0.60	0.70	0.50
Pocono	0.00	0.40	0.20	0.40
Figs	0.00	0.60	0.30	0.20
last supper painting	1.00	0.40	0.30	0.60
university of phoenix	1.00	1.00	1.00	1.00
the Beatles rock band	1.00	0.60	0.30	0.10
Grilling	1.00	0.60	0.50	0.60
furniture for small spaces	0.50	0.20	0.60	0.70
Dnr	0.00	0.60	0.50	0.80
Arkansas	1.00	1.00	1.00	1.00
hobby stores	1.00	0.80	0.70	0.40
blue throated hummingbird	0.00	0.60	0.30	0.15
computer programming	1.00	0.40	0.60	0.60
Barbados	1.00	1.00	1.00	1.00
Lipoma	1.00	0.40	0.80	0.80
battles in the civil war	1.00	0.40	0.70	0.80
Scooters	0.00	0.30	0.30	0.30
ron howard	1.00	1.00	1.00	1.00
becoming a paralegal	1.00	0.40	0.70	0.70
hip fractures	1.00	0.90	0.90	0.90
septic system design	1.00	1.00	0.60	0.50
rock art	1.00	0.60	0.50	0.40
sings of a heart attack	1.00	0.80	0.80	0.50
weather strip	1.00	0.60	0.40	0.30
best long term care insurance	1.00	0.90	0.80	0.70
pork tenderloin	1.00	1.00	0.70	0.70

black history	0.00	0.70	0.80	0.60
new york hotels	1.00	1.00	1.00	1.00
old coins	1.00	1.00	0.80	0.70
quit smoking	1.00	1.00	0.80	0.80
kansas city mo	1.00	1.00	0.80	0.70
civil rights movement	1.00	0.20	0.60	0.60
credit report	1.00	0.80	0.60	0.40
Unc	1.00	0.80	0.90	0.80
Vanuatu	1.00	1.00	1.00	0.80
internet phone service	1.00	0.70	0.60	0.40
gs pay rate	1.00	0.50	0.60	0.50
brooks brothers clearance	0.00	0.20	0.50	0.40
churchill downs	1.00	0.60	0.60	0.60
condos in florida	0.00	0.40	0.50	0.60
Porterville	0.00	0.70	0.50	0.20
dog clean up bags	0.00	0.30	0.50	0.70
designer dog breeds	1.00	1.00	0.60	0.50
pressure washers	1.00	1.00	0.70	0.80
sore throat	1.00	1.00	0.50	0.60
idaho state flower	1.00	0.70	0.50	0.50
indiana state fairgrounds	1.00	1.00	0.80	0.80
Fibromyalgia	1.00	0.30	0.50	0.50
ontario california airport	1.00	0.60	0.80	0.40

Table 7. The precision of 50 test queries

6. Hypertext Diversification

Some applications have become unpopular due to their focus on hypertext lack of diversity. Google, for example, devalued the backlinks of websites that excessively used promotional keywords as hyperlinks and placed links on spammer websites. This is why diversification is essential (i.e. the mixing of hypertext and backlinks quality). Research has been conducted to determine whether websites that experienced good rankings had optimized hypertext over instead of using natural mixed hyperlinks.

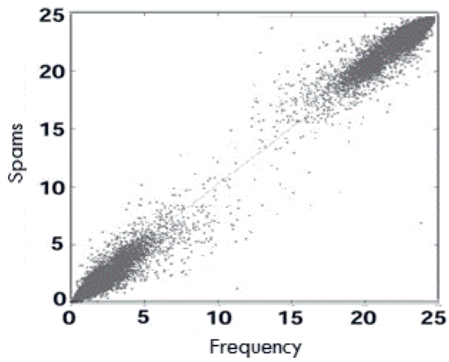

Figure 4. Anchor frequency vs. spam distribution

Figure 4 shows the trade-off between the frequency of hyperlinks in websites and spam validation. The percentage of some keywords (e.g., money) should not exceed the exact threshold. This does not mean a keyword could not be used more than once but it should not rotate that keyword one to five in 100% of the links acquired. To determine what types of hypertexts work from long term perspective and in what proportion, the competitive keywords could be selected, e.g. 'quick weight loss', 'real estate for sale', 'cheap hotels', or 'web design company'.

Figure 5 shows six major types of hypertexts involved different distributions in the selected dataset classified as follows:

- **Branded:** Brand hyperlinks use the actual name of a brand or business. For example, "Outreachmama.com" would use the brand hyperlink "OutreachMama". As the Internet grew, many companies adopted names that would capitalise on industry keywords, such as www.Londonseocompany.com, the branded hypertext might legitimately read as "Find out more about London SEO Company".
- **Exact match:** In some cases, it was difficult even impossible to determine if the hyperlink features exactly match text or simply a brand name. Other examples are domain names, such as "smoking-cessation.org" that ranked high for "smoking cessation".
- **Partial match:** Most websites successfully implemented the diversification strategy by mixing some targeted keywords with synonyms. The number of partially matched hyperlinks, and anchors with synonyms, and long tail keywords were available up to 50% across all websites.
- **Naked links:** As the name implies a naked URL without an http://, it is a URL that only uses "www" and excludes "http://" (e.g. www.youthnoise.com). The quantity of naked URLs in hyperlinks ranged from 2% to 15%. Many URLs (15-35%) as hyperlinks were used by websites with famous brands; such as, namecheap.com (40% of URLs in hyperlinks), hotels.com (53%) and smoking-cessation.com (90%).
- **Non-descriptive hyperlinks are rare:** The quantity of non-descriptive hyperlinks was low across all websites ranged between 2% to 5%.
- **Other languages:** The quantity of keywords in other languages was few ranged from 0.2% to 0.8% per website.

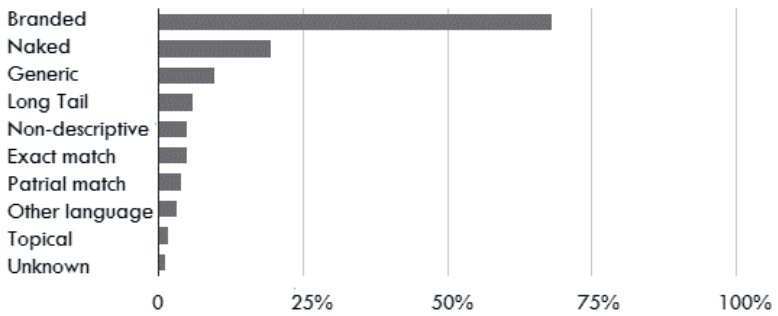

Figure 5. Types of hypertexts

7. Conclusion

We demonstrated how important of hypertext in Web documents for improving the quality of Web search for many types of queries and emphasized the importance of indexing in hypertext systems. We highlighted this by examining the properties of hypertext in a large subset of ClueWeb09 collection. Our main premise is that hypertext has an impact very similar to real user queries and consensus titles, and also, unlike linear text as closely resembles the networked and associational organization of information on the Web. Thus, understanding the relationships of hypertext in the documents leads to better understanding how to find high quality search result to a query string. We tested our approach experimentally in several types of queries from TREC Web track including the significant stream of queries versus the Web content. We conducted our experiments to investigate several aspects of hypertext, including the relationship to its domains, the frequency of queries that could be satisfied by hypertext alone, and the homogeneity of results acquired by hypertext index. We think that a reason makes hypertext more efficient for Web search is that most users used short queries and tend to choose minimal number of terms that annotated by meta-tags or concisely summarized pages as the same way that Web designers choose hypertext.

References

[1] Yadav, D., Sharma, A., and Gupta, J. (2009). "Topical web crawling using weighted anchor text and web page change detection techniques". *WSEAS Transactions on Information Science and Applications*, 6(2).

[2] Halpin, H. and Lavrenko, V. (2011). "Relevance feedback between hypertext and Semantic Web search: Frameworks and evaluation". *Journal of Web Semantics*, 9(4):474-489. DOI:10.1016/j.websem.2011.10.001.

[3] Al-akashi, F. and Inkpen, D. (2014). "Term Impact-Based Web Page Ranking". *In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS)*. DOI; 10.1145/2611040.2611060.

[4] Craswell, N., Hawking, D., and Robertson, S. (2001). "Effective Site Finding using Link Anchor Information". In *Proceedings of the SIGIR'01*.

[5] Page, L., Brin, S, Motwani, R., Winograd, T. (1999). "The PageRank Citation Ranking: Bringing Order to the Web". *Stanford InfoLab*, DIO; 1.1.31.1768.

[6] Kleinberg, J. (1999). "Authoritative sources in a hyperlinked environment". *Journal of the ACM*, Vol. 46, No. 5, Pp 604-632.

[7] Boytsov, L. and Belova, A. (2012). "Does Category (A) Anchor Text Improve Category (B) Results". In *Proceedings of the Twenty- first Text Retrieval Conference (TREC)*.

[8] Marijn, K. and Jaap, K. (2011). "Reducing Redundancy with Anchor Text and Spam Priors". *In Proceedings of the Twentieth Text REtrieval Conference (TREC).*

[9] Upstill, T., Craswell, N., and David, H. (2003). "Predicting Fame and Fortune: PageRank or In-degree". *ACM Transactions on Information Systems (TOIS).* Vol. 21, No. 3, Pp. 286-313).

[10] McBryan, O. (1994). "Genvl and wwww: Tools for taming the web*".* In *Proceedings of the First International World Wide Web Conference*, Pp. 79-90.

[11] Chakrabarti, S., Dom, B., Gibsonm D., J., Kleinberg, P., and Rajagopalan, S., Gibson, D. Kleinberg, J. (1998). "Automatic resource list compilation by analyzing hyperlink structure and associated text". *Computer Networks and ISDN Systems*. Vol 30, No. 1-7, Pp. 65-74.

[12] Nadav, E. and Kevin, S. (2003). "Analysis of Anchor Text for Web Search". In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (Pp. 459-460).

[13] Atsushi, F. (2008). "Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval". In *Proceedings of the 17th International World Wide Web Conference.* DOI: 10.1145/1367497.1367544.

[14] Amitay, E. (1998). "Using common hypertext links to identify the best phrasal description of target web Documents". In *Proceedings of the SIGIR'98 Post Conference Workshop on Hypertext Information Retrieval for the Web*. Melbourne, Australia.

[15] Haveliwala, T., Gionis, A., Klein, D., and Indyk, P. (2002). "Evaluating strategies for similarity search on the web". In *Proceedings of the 11th International Conference on World Wide Web.* Pp. 432–442.

[16] Davison, B. (2000). "Topical locality in the web". *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 272- 279), New York, ACM.

[17] Attardi, G., Gulli, A., and Sebastiani, F. (1999). "Theseus: categorization by context". In *Proceedings of the 8th International World Wide Web Conference*.

[18] Shuming, S., Fei, X., Mingjie, Z., Zaiqing, N., and Ji-Rong, W. (2009). "Anchor Text Extraction for Academic Search". In *Proceedings of the Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, ACL-IJCNLP (Pp. 10–18).

[19] Ganeshiya, D. and Sharma, D. (2014). "A survey: hyperlink analysis in webpage ranking algorithms," *International Conference of Soft Computing*

Techniques for Engineering and Technology (ICSCTET), Pp. 1-8, DOI: 10.1109/ICSCTET.2015.7371192.

[20] Costa, J., Rodrigues, J., Simões, T. and Lloret, J. (2016). "Exploring Social Networks and Improving Hypertext Results for Cloud Solutions," *Mobile Network Application,* 21, 215–221. DOI: 10.1007/s11036-014-0513-z

[21] Clarke, C., Cormack, G., (1995). "Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System". *TechRep MT-95-01*, University of Waterloo.

[22] Albi, D., and Silvester, H. (2017). "PageRank Algorithm". *IOSR Journal of Computer Engineering (IOSR-JCE)*, Vol. 19, No. 1, Ver. III (Pp. 01-07).

[23] Hiemstra, D. and Hauff, C. (2010). "MIREX: MapReduce Information Retrieval Experiments". Technical Report TR-CTIT-10-15. ISSN 1381-3625. http://eprints.eemcs.utwente.nl/17797

[24] Anh, V. and Moffat, A. (2010). "The Role of Anchor Text in ClueWeb09 Retrieval". In *Proceedings of the Nineteenth Text Retrieval Conference Proceedings (TREC).*

[25] Kamps, J., Kaptein, R., and Koolen, M. (2010). "Using Anchor Text, Spam Filtering and Wikipedia for Web Search and Entity Ranking," in *Proceedings of the 19th Text Retrieval Conference (TREC)*, USA.

[26] Serdyukov, P. and Vries, A. (2009). "Delft University at the TREC 2009 Entity Track:
Ranking Wikipedia Entities," in Proceedings of the 18th Text Retrieval Conference (TREC), USA.

[27] Cormack, V., Smucker, D., and Clarke, L. (2010). "Efficient and effective spam filtering and re-ranking for large web datasets".
http://arxiv.org/abs/1004.5168, 2010.