Journal of Web Engineering, Vol. 14, No.5&6 (2015) 503-524

MODIFIED PAGERANK FOR CONCEPT BASED SEARCH

PAVAI G

Anna University, Chennai, Tamil Nadu, India
pavai_gops@yahoo.co.in

UMAMAHESWARI E

Anna University, Chennai, Tamil Nadu, India
umavasanth28@gmail.com

GEETHA T V

Anna University, Chennai, Tamil Nadu, India

tv_g@hotmail.com

Received November 14, 2014
Revised April 10, 2015

Traditional PageRank algorithm computes the weight for each hyper-linked document, which indicates the importance of a page, based on the in-links and out-links. This is an off-line and query independent process which suits a keyword based search strategy. However, owing to the problems like polynymy, synonymy etc.., existing in keyword based search, new methodologies for search like concept based search, semantic web based search etc., have been developed. Concept based search engines generally go in for content based ranking by imparting semantics to the web pages. While this approach is better than the keyword based ranking strategies, they do not consider the physical link structure between documents which is the basis of the successful PageRank algorithm. Hence, we made an attempt to combine the power of link structures with content information to suit the concept based search engines. Our main contribution includes, two modifications to the traditional PageRank Algorithm, both specifically to cater to the concept based search engines. Inspired by the topic sensitive PageRank algorithm, we have multiple PageRanks for a document, rather than just one for each document, as given in the traditional implementation of the PageRank algorithm. We have compared our methodologies with an existing concept based search engine's ranking methodology, and found that our modifications considerably improve the ranking of the conceptual search results. Furthermore, we performed statistical significance test and found out that our Version-2 modification to the PageRank algorithm is statistically significant in its P@5 performance compared to the baseline.

Key words: PageRank, Semantic web based search, Concept based search, Physical link, Concept link, UNL
Communicated by: G.-J. Houben & Y. Deshpande

1 Introduction

The amount of information on the web is rapidly growing. Information retrieval systems are responsible for providing information of interest to users of search engines, who expect results relevant

to their queries to be available in the first few pages. This is the driving force behind search engine optimization. Some of the popular web search engines like Google, use keyword based search [5].

Users type a query consisting of a few keywords describing their information need. In the offline processing phase, an index is created with the list of keywords for each of the documents in the web document collection. Keyword based Information Retrieval systems perform a term based match, between query words and the index of the documents that have already been pre-processed and indexed to return matching documents. Along with the presence of keywords in the index, other factors like co-occurring terms, frequency of the keyword in a document, position weight of the keyword [9] etc., are also used in ranking a web page, to increase the relevancy of the results for a query.

Web Retrieval is complicated due to the large and dynamic content of the web. Web search engines usually serve millions of users and process millions of queries every day. It is very unlikely that all the users have similar interests and search for similar information. Even though the web has a huge number of resources, users are unable to get the relevant results quickly. Therefore, different methods have been proposed to improve the relevance of the search results for the particular user's query. One such method is the ranking of Web pages based on some criteria, that would help increase the relevance of search results. Apart from improved search ranking algorithms, researchers have gone into ideas like concept based search as in Haav and Lubi [9], which is based on the fact that the meaning of a word depends on its conceptual relationships to objects in the world rather than considering other types of relations like the linguistic or the context based relations that are found in the texts or dictionaries, Semantic web as per Madhu et al. [15], which is a meaning based search rather than the exact keyword match which is ambiguous at certain instances, Personalized web search as Liu et al. [14] say is providing different results to different users even for the same queries by keeping track of each user's interest based on their search behaviour, Collaboration which is a generalized form of personalization as in Freyne et al. [8] provides search results by exploiting the repetition and regularity within the query space of a group of people who have similarity in taste, spatial data based search as given by Noack [18] which deals with making the search results relevant to the search engine user's geographic locations in the case of geographically dependant queries like "petrol bunk" which would certainly mean a petrol bunk that is closer to the user's current location, and temporal search as per Efendioglu et al. [7] where the search results are based on temporal classification of the documents in order to answer queries like " important sports events in India in 2006 ", etc.., for the improvement of search results.

The simplest technique to evaluate a page in the context of search engines is determining the content similarity between the query words and the indexed documents. Based on concepts borrowed from the social network and academic citation analyses [21], researchers have attempted to use hyper-link and user behaviour information available in the web pages for evaluating a page. Hyper-links have been manually created by authors of web pages to link to other web pages whose author considers these pages important and relevant. One such attempt to rank a web page using hyper-links is the very successful Google's PageRank [5], which ranks pages based on the number of out-links from a page and the number of in-links to that page. Primarily, PageRank calculation is based on the heuristic on which an author creates hyper-links to pages, that are considered important, and therefore a page that is pointed to by many pages is considered important by many authors, and hence is an important page.

PageRank has been used to improve the performance of the traditional keyword based search engines. But a serious problem with this kind of a physical link based ranking approach is that, the links are manually created and so can be used by any author to boost the rank of one of his web page, by simply giving an out-link to that page from various other web pages that he authors (whether the link is appropriate or not). In addition, the author may create hyper-links from web pages to pages that may not be semantically related to the page in question, but may be just an indication of the flow of information. We try to resolve this problem, by considering not only the physical links between web pages which may be misleading, but also by considering how well the two pages are conceptually connected, so that an author cannot simply boost an unimportant web page's rank. Despite these problems with the physical links, its huge success in Google [5] motivated its use in our ranking algorithm along with concept-based ranking elements.

The Concept based Search Engine that we consider for our work is COREE, a concept based Tamil search engine. COREE uses the Universal Networking Language (UNL) [2] to represent concepts and relations present in the documents by appropriate enconversion [2]. The success of PageRank in Google Search engine motivated us to develop a ranking strategy using PageRank along with the benefits of the semantics provided by the UNL representation. However, in our work, we modify the individual page based concept importance of web pages used by CoRee search engine discussed in section 2.2. We extend the Page Rank algorithm where we consider the concept oriented links between pages and the concept relatedness between pages connected by physical links to find the PageRank of a page. This inter page based PageRank can be determined only as an offline process and not directly during the online search and rank process.

In this work we have used a combination of physical hyper-links as well as document-document conceptual links and modified the PageRank to determine the importance of pages which will help in ranking relevant results for the user. Concept relevance can well be used as a standalone feature in ranking. However, this case mainly speaks about a document's relevance to a concept. It falls under the category of content based search. Our proposed methods are not only using content information of pages to compute the relevance of that particular page to a concept, we also find conceptual links between pages. This is like using the physical link information for ranking. However, the type of link we use is not physical and we use content information to form the invisible links. Hence, our modified algorithms try to get the best of both worlds [content based information and link information].

The next section gives a brief account of COREE and the basic Concept Representation Language it uses – UNL (Universal Networking Language) which is followed by the related work section that describes about works that describes existing modifications to PageRank which is followed by the description of our methodologies followed by the experimental results and discussion.

2 Current Practice and Research

This part gives a brief description about the Universal Networking Language (UNL) and CoRee – a concept based search engine for the Tamil language. We would like to describe the working of CoRee as we consider the ranking methodology of CoRee as the baseline against which we compare and evaluate our work.

2.1 Universal Networking Language (UNL)

UNL basically consists of concepts (C) called "Universal Words" (UWs) which are inter-linked with other UWs to form sentences [24]. Such links represent "relations" (R). Relations describe the role of each word in a sentence. There are 46 generic relations in UNL including mod, plc, plf, pld, iof, ben etc., In addition to the above semantic constraints "UNL attributes" associated with UWs help express the speaker intention. "UNL Expression" is the UNL notation for a sentence. Here UNLKB [24], a Knowledge Base is used for defining the semantics of UWs which gives hierarchical relations between concepts as well as an inference mechanism based on the inclusion of relations between concepts to help resolve ambiguity. In document processing each sentence in all documents are converted into a set of UNL graphs representing concepts and their associations.

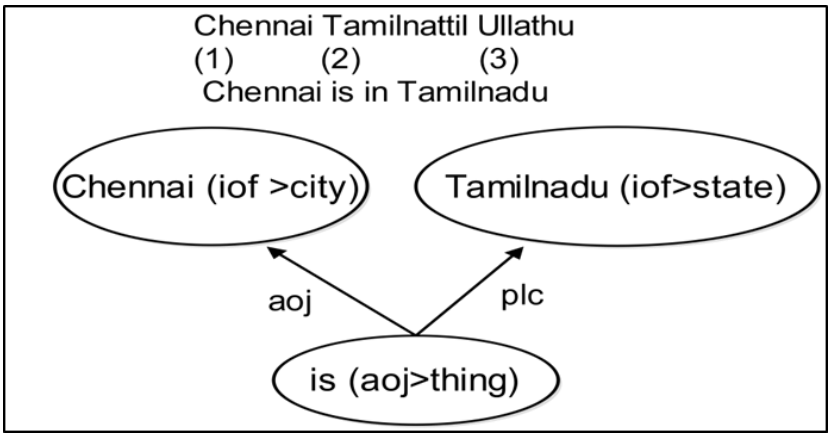

Figure 1 Example 1: UNL graph of a Tamil Sentence (Transliteration and translation given).

A binary relation between the concepts of the UNL graph is expressed as follows:

<relation> (<uw1>, <uw2>)

The above relation is interpreted as follows: <uw1> holds the relation in <relation> with <uw2>. Similarly <uw2> plays the role indicated by the <relation> and held by <uw1>. The UNL representation is in the form of semantic networks with hyper-nodes, where nodes represent concepts and arcs represent the relations between concepts. Concepts are annotated with information such as POS, frequency of occurrence within the document, document ID, sentence ID, concept position ID etc., The UNL representation for the sentence (example 1) "Chennai TamilNattil Ullathu (Tamil) - Chennai is located in TamilNadu (English)" is shown in figure 1. This example has three UWs Chennai(iof>city), locate(aoj>thing), TamilNadu(iof>state) and these UWs are connected with the UNL relations such as aoj, plc to represent the conceptual relationship between the UWs. Here, the UNL constraints (iof>state), (aoj>thing), (iof>city) expresses the word level semantics.

2.2 UNL Based Concepts Extraction and Indexing

Balaji et al. [2] discuss the process of UNL Enconversion of Tamil sentences in web pages. They use morpho-semantic features of the words and their preceding and succeeding context along with morphological suffixes which indicate case relations of nouns, POS tag and word level semantics to

enconvert web pages to UNL representation. The UNL representation, represents each sentence as a semantic graph with concepts as nodes and UNL relations as links. It is this UNL semantic graph that we take as input for our work. It is from this graph that we obtain the concepts and relations that determine our PageRank.

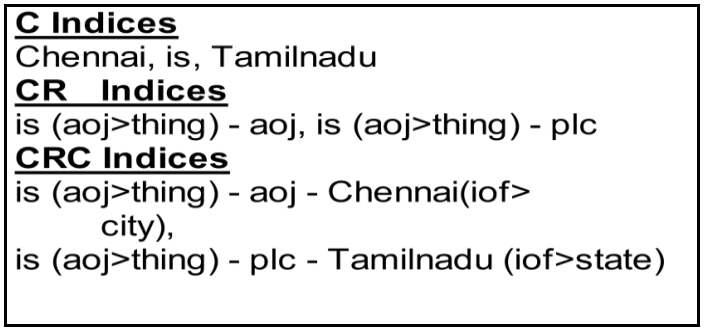

Figure 2 Example for C, CR, CRC in a sentence.

Umamaheswari et al. [23] have used this graph to build a conceptual index consisting of three components – CRC index (representing all Concept-Relation-Concept components in the documents), CR index (representing all Concept-Relation components in the documents) and C representing all concepts in the documents index for their rank and search process. These components for example 1 is also shown in figure 2. Here the concept indices are Chennai(iof>city), locate(aoj>thing), TamilNadu(iof>state), concept-relation indices are Chennai(iof>city)-aoj, locate(aoj>thing)-aoj, locate(aoj>thing)-plc, TamilNadu(iof>state)-pla and concept-relation-concept indices are locate(aoj>thing)-aoj- Chennai(iof>city), locate(aoj>thing)-plc- TamilNadu (iof>state) which are stored in the C , CR and CRC Indices respectively. The UNL expression of the above given example is given in figure 3.

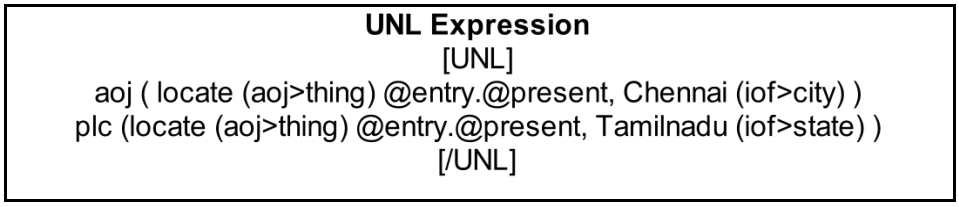

Figure 3 Example 1: UNL Representation of a Tamil Sentence.

2.3 Existing Concept Based Search Engine CoRee and its Ranking

We have designed a modified PageRank algorithm that can be used with COREE to improve its search and ranking modules. COREE [23] first builds a document representation by enconverting important sentence constituents contained in the documents into a UNL representation [24]. The search and ranking procedure used by COREE [23] considers only content based matches between the query and documents although it uses concepts and concept relations as well as index based conceptual query expansion to determine this match. A three level ranking methodology is followed here.

Level 1: Degree of match categorization prioritizing documents based on complete match (CRC match), partial CR match or Concept only(C) match.

Level 2: Concept association categorization that is based on whether the match is a term match, concept match or an expanded concept match.

Level 3: Ranking is based on index based features like the frequency of occurrence of the term and concept in the document, position weight, Named Entity (NE) weight and Multiword (MW) weight.

2.4 Existing Modifications to PageRank

While the previous sections gave an account of our baseline concept based search engine – CoRee, this section gives the various existing modifications to the PageRank method. PageRank is basically a query independent measure of the importance of a web page that is a method for computing a popularity (or importance) ranking for every web page based on the graph of the web. The concept behind PageRank is that a page is considered important if it receives many links from pages which are in turn important. A link from a page A to page B is considered as a vote or recommendation of the author of A for B and therefore the number of incoming links to a page gives a measure of the importance and authority of the page. However in addition PageRank also considers a page to be more important if the sources (web pages) of its incoming links are important [5]. Word co-occurrences, font size, etc. are also considered in the ranking process. In addition, the paper proposes data structures for quicker access to the index and the inverted index.

PageRank is based on the random walk (surfer) model of user behaviour. Brin and Page [5], defined the PageRank of a web page as given in equation 1.

$$PR(A) = (1 - D) + (D * \sum_{n=1}^{N} \frac{PR(T_n)}{c(T_n)}) \qquad (1)$$

where,

PR(A) – Represents the PageRank of a Web Page A.

D – Indicates the damping factor (which is normally set to 0.85)

T1….TN – are the pages that link to page A.

C –represents the total number of out-links from the page Tn, Where n =1 to N.

The damping factor D is the probability that at each page, the random surfer will choose a random page, instead of following the links.

Many modifications to the original PageRank have been proposed. While some of the modifications strive to reduce the time complexity, others attempt to use other relations between web pages to redefine the PageRank. Though our work modifies the PageRank to suit concept based search, we have covered other works done using PageRank in this section. First, we discuss about the attempts to optimize PageRank calculation followed by usage of PageRank for applications other than search which in turn is followed by the actual modifications to PageRank. One of the main problems with the PageRank [5], is the re-computation of the already converged PageRanks repeatedly in every iteration,

thereby increasing the time complexity of the PageRank computation process. According to Kamvar et al. [11], fewer pages converge quickly, and those that take time to converge are actually pages with a higher PageRank. In addition, the web consists of millions of pages, and so the iterative PageRank calculation is a time-consuming process. Kamvar et al. [11] proposed two methods to overcome these defects of PageRank. The first is the Adaptive PageRank process that identifies the converged pages periodically, and constructs an adjacency matrix only for pages that have not converged, which is lesser in size than the original adjacency matrix, thereby decreasing the iteration cost. The second method is the modified adaptive PageRank, where pages with out-links from the already converged pages are not recomputed again. However, the problem with this approach is that pages that have converged in an earlier iteration sometimes have the chance of exhibiting variations. The authors also claim that this problem could be dealt with by phase-by-phase adaptive PageRanking.

PageRank calculations of sub-graphs were explored by Wu and Raschid [25] and Bar-Yossef, Z. and Mashiach [3]. Wu and Raschid [25] proposed two different algorithms for calculating the PageRanks of the sub-graphs. The first one is the Ideal PageRank where the PageRanks of the external pages is assumed to be known. However, in case this assumption is wrong, the authors have proposed the Approx PageRank, where the PageRank of the external pages is not available. Here, the authors consider the authority flow from the external pages giving equal importance to all of them.

Other variation of the PageRank algorithm uses the same mathematical calculation of the traditional PageRank algorithm, but is used to resolve Word Sense Disambiguation [17]. The modification is in the graph construction, where each of the words of the WordNet is a vertex and the edges between the vertices denote whether the two nodes have any semantic Relation between them [1]. In addition, Thelwall and Vaughan [22] have used the PageRank calculation as given by Brin and Page, [5] but the difference is in what is considered as a page. Thelwall and Vaughan [22] suggest that either i) every HTML file can be considered as a page or ii) all HTML files within a directory can be considered a page, or iii) all HTML files in the same domain can be considered a page, or iv) all web pages belonging to a particular University can be treated as a page.

Another problem with PageRank [5] is that the content of the page is not considered, and only the physical links between pages are considered for calculating the importance of a page. Though Brin and Page [5] say they consider word co-occurrences, font size, bold or not etc., it does not seem to be greatly useful for searches other than Keyword based searches (Jeffrey Xu Yu.et.al,2002). A few researchers have attempted to modify the PageRank to suit concept- based or semantic network-based searches. Lin [13] claims that using related document networks rather than the manually created hyperlinks will be a better approach since these are automatically computed by content similarity algorithms. Therefore, changes in the documents will be changing the page scoring automatically, while this is not the case with the traditional PageRank. Diesy et al [6] proposed a solution, that uses Ontology relations between the query words and the ontology relations in the documents, and takes into account not only the relevancy factor that is determined as the prime factor for the ranking, but also the hit count of the words given in the query.

The major problem associated with PageRank is that heavily linked pages, whether relevant to the topic of the query or not, have a high PageRank. In order to deal with this problem, Haveliwala [10] proposed the topic sensitive PageRank, that has multiple PageRank vectors, each associated with a topic and the author has considered sixteen predefined topics. The problem with this approach is that

multiple PageRank vectors occupy a much larger space than required by one vector. In addition, the computational cost is many times that of a single vector. Therefore, this method is a trade-off between relevance on the one hand, and space and time requirements, on the other hand.

Another modification of PageRank by Qiu et al., [20], is the use of the link vector to record the contact times between nodes and using them in the PageRank calculation. Personalized PageRank [26] is a popular modification to the PageRank that involves online calculation of personalized PageRank score for a query. Borgs et al. [4] use a multi-scale sampling scheme that uses fast personalized PageRank and a new local randomized algorithm for quicker convergence. A notable modification to the personalized PageRank called the incremental and accuracy aware personalized PageRank by Zhu et al., [26], reduces the online PageRank computation time. Maehara et al. [16] exploit the graph structures for quicker computation of personalized PageRanks. Our work in a way is similar to the Personalized PageRank that it computes PageRank value for all concepts it has 'offline' against that of computing a personalized PageRank value for each query online. The added advantage with our method is the calculations being done offline.

Our work is closely related to the work by Kurland and Lee [12] by the way that they have also used links other than the hyper-links for ranking. In fact, they have used language model (keyword) based similarities between documents whereas; we have considered the concept similarities between documents. Our work uses the concept (and/or relation) similarity measure between documents and computes PageRank for every such concept (and/or relation) in a way much similar to the Topic Sensitive PageRank where instead of topics, we consider individual concepts (and/or relations).

3 Research Approach

This section describes about the conceptual links and the modifications done to CoRee to suit our work. In this work we have used physical hyper-links along with the semantic information for ranking documents for concept based search. One of the major disadvantages of the PageRank [5] is its heavy dependence on the physical link as the major factor in the ranking process. The problem with this approach is that, physical links are biased. It depicts that the particular author of the web page is interested in this particular link to a certain web page. But there may not be any relevance between this web page and the web page associated with the out-link. This may mislead the PageRank algorithm, since an author may try to boost the rank for one of his web pages, by giving a huge number of in-links to that particular page.

3.1 Concept based links

However, the physical links are not the only links between two pages. Other virtual links like i) keyword based links where the common keywords between two documents determine the strength of the link between them, and ii) concept based links where the common concepts between two documents determine the strength of the link between them, do exist. These types of links are likely to provide better results in some cases, than using only physical links, since they depend purely on the similarity of the content of the pages, and let no other external factor like the user's intention to modify the rank of a page. Figure 4 clearly distinguishes between the three types of links – physical links, keyword links and conceptual links.

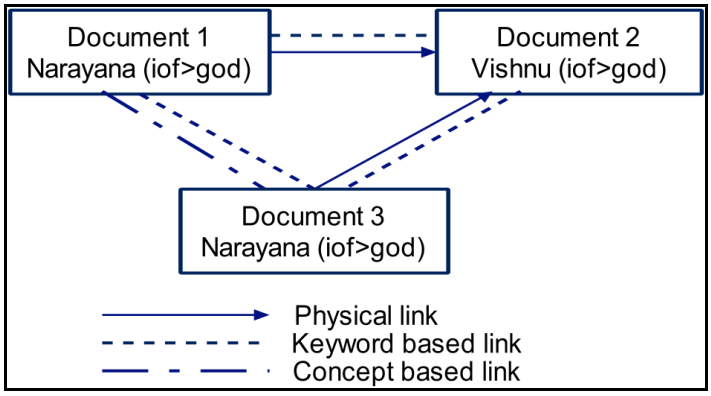

Figure 4 Physical, keyword based and concept based links.

In figure 4, there are two physical out-links from document1 to document2 and another to document3. This is a unidirectional link where the link starts from the page with the anchor text and ends at the page it is pointing to. There exists a keyword based link between two documents if they have exactly the same keywords. The keyword based link exists between document1 and document2 alone, since the same word occurs in these two documents only. There exists a concept based link between two pages, if they have the same or different keywords pointing to the same Universal word. For example, Narayana is an Indian God who is also called Vishnu. The concept based links exist between all the three documents since both Narayana and Vishnu point to the same concept. Figure 4 also depicts the keyword-based and the concept-based links as bidirectional links.

We use concept links in our work, in order to adapt to the concept based search. Our definition of how conceptual similarity between two documents is determined is based on UNL representation. In our approach, we boost a page's rank if it is physically and/or conceptually strongly connected to other documents in the corpus. The purpose of using physical links in our work, in spite of its ability to mislead the ranking process, is because of its reasonably good performance. Concept Link based PageRank algorithm version-1 (with physical links) boosts a pages rank, if it has physical links and the page is conceptually strongly connected to other physically linked pages, whereas, the Concept Link based PageRank algorithm version-2 (with/without physical links) boosts a page's rank if it has strong physical and/or conceptual connections with other pages. This PageRank is used for query dependant ranking as described in section 5.3.

3.2 Modifications to CoRee

We have used the concept-relation extraction of CoRee as such. We have a similar index structure to CoRee (C, CR and CRC indices). However, we have an additional index called the CXC index (representing all Concept-Don't care-Concept components in the documents). An example of this is "Chennaiyin KadaRkaRai" (Tamil) [Chennai's beach – English], "Chennaiyil KadaRkaRai" (Tamil) [Beach in Chennai – English], "ChennaiyiluLLa KadaRkaRai" (Tamil) [Beach located in Chennai – English], "Chennaiyudaiya KadaRkaRai" (Tamil) [Beach of Chennai – English] etc., are CRC's that almost mean the same but for the relations they have. Hence, we decided that the CXC index (concepts with any relations) would bring out closer relationships between the document-document and the

query-document pairs. This can be understood as 'Concept pairs occurring together in documents are considered in our work no matter what the relation is between them'. We will be using our new ranking methodology in place of the existing ranking methodology of CoRee.

Section 4 outlines the details of concept link determination, and Section 5 gives a detailed account of our modifications to the existing PageRank approach.

4 Concept Link Determination

This section discusses the method that is used to determine the conceptual links between two pages. As given earlier, we use the concept links that exist between the web pages. At first we calculate the concept link value between two physically connected pages. Then, we calculate the concept link values for pages which do not have physical links. The concept link values with the physical links and without them, are used to calculate the modified PageRank value. The conceptual link is a bidirectional link between two web pages. Statistical techniques are used here to determine the degree by which two pages are conceptually linked.

CLW(A,B), (Concept Link Weight) is the weight of the conceptual link between two pages A and B, which is determined by finding the commonly occurring concepts, and conceptual relations between the documents.

$$\mathrm{CLW}(A,B) = \sum_{i=A,B} \sum_{j} \frac{N^{j}}{TN_{i}^{j}} \qquad (2)$$

where A is a web page, B is also a web page that is pointed to by an out-link of page A, and j refers to the type of concept-relation index and it takes four values namely, C, CR, CRC and CXC – which are the common numbers of Cs [concepts], CRs [concept – relations], CRCs [concept – relation – concept] and CXCs [concept – don't care – concept] respectively. N is the concept-relation index value for each j common to both the web pages A and B. TN is the total number of concept-relation of each j for each web page i.

We have considered C, CR, CRC and CXC for the ranking purpose which itself is a contribution in this work since other conceptual ranking strategies consider topics as described by Haveliwala [10], or at the most keywords occurring together as given by Kurland and Lee [12]. However, we have used the C, CR and CRC indices used by Umamaheswari et al [23] for our ranking with an addition of CXC index which is capable of bringing out the maximum relationship between documents thereby helping our ranking algorithm perform better. The common number of concepts (and/or relations) in the pages A and B, are divided by the total number of concepts (and/or relations) in page A, and also by the total number of concepts (and/or relations) in page B. This is done to calculate the importance of the common concepts (and/or relations) with respect to the total number of concepts (and/or relations) in page A, and also with respect to the total number of concepts (and/or relations) in page B, i.e., how important the common concepts (and/or relations) are between the pages A and B in page A as well as in page B. This is done so that the conceptual link between two web pages is higher, if the common concepts (and/or relations) between the two pages are also the highly important concepts (and/or relations) of both the pages. Both the above said factors are added as given in equation 2, so that the

concept link value between the two pages is boosted, even if one of the pages A or pages B has a greater importance for the common concepts(and/or relations) in its content.

5 Modified Versions of the PageRank Algorithm

This section gives a detailed description of the modifications that we have done to the PageRank algorithm. Section 5.1 describes the Concept Link Based PageRank Algorithm-Version 1 (With Physical Links) and section 5.2 describes the Concept Link Based PageRank Algorithm- Version 2 (With/without Physical Links)

5.1 Concept Link Based PageRank Algorithm- Version 1 (With Physical Links)

The traditional PageRank algorithm is modified to accommodate the concept links in the calculation of the PageRank. In addition to the physical link weight, the concept link weights are also added, while calculating the PageRank. By doing this, the PageRank calculation that was originally biased towards the web page creator's thoughts, that certain web pages are closely relevant to his web page, is changed to accommodate the concept (and/or relations) wise relation between contents of the physically connected pages as well.

The modified PageRank that includes the concept link weight is given in equation 3.

$$PR_{withCLW}(A) = PR(A) + \sum_{i=1}^{n} \frac{CLW(A, T_i)}{C(T_i)} \qquad (3)$$

where,

$PR_{withCLW}(A)$ – PageRank value with Concept Link Weight,

Ti – is a web page that is an in-link to A,

C – Is the total number of concepts (and/or relations) in page Ti,

n – represents the number of in-links to the page A,

PR(A) and $CLW(A,T_i)$ are obtained from equations 1 and 2 respectively. We add the CLW component, so that the content information is used along with the physical link information for determining the PageRank of a document. We use equation 4 to calculate the PageRank for every page. This method suffers from the same defect as that of the Topic Sensitive PageRank Algorithm [10] in occupying an excessive amount of space to hold the PageRank values. This is because, the number of PageRanks a document has is equal to the number of Cs, CRs, CRCs and CXCs a page has. Though this raises the issue of space constraint, we would say the effect would have been much higher for a [nd x nC] matrix, or a [nd x nCR] matrix or a [nd x nCRC] matrix, or a [nd x nCXC] matrix, where nC is the number of Cs in all the documents, nCR is the number of CRs in all the documents, nCRC is the number of CRCs in all the documents, where nCXC is the number of CXCs in all and 'nd' is the total number of documents in the corpus.

This method has the advantage of ranking the same page differently for different concepts as against the single PageRank for a web page in the traditional version [5]. This gives our work the edge of query dependence rather than the query independence in traditional PageRank [5]. By query independence, we mean the reflection of the actual concept (and/or relations) of the query word in the document and we are not considering the query as such in any of our equations. This is the essence of the Topic Sensitive PageRank [10]. But in our work, we have considered ranking a page for all its concepts separately rather than the sixteen pre-defined topics in their work. Apart from the importance of concepts (and/or relations) with respect to a page, we determine the importance of every C, CR, CRC and CXC with respect to the corpus using equation 4.

$$\text{Concept} - \text{Weight}(c, A) = \frac{FR_c + PW_c}{TN_c} * T_c \qquad (4)$$

Concept-Weight(c, A) represents the importance of the Cs, CRs, CRCs and CXCs with respect to the document specific properties, such as the position, frequency of occurrence of the concept within a page and its importance across the entire corpus,

c is the concept-relation index and it can take four values namely C, CR, CRC and CXC,

FR represents the frequency count of a concept-relation 'c' in page A,

PW represents the position weight of the concept-relation 'c' in page A,

TN is the total number of concept-relation 'c' in page A,

T is the total number of concept-relation 'c' in the entire corpus.

The purpose of calculating the concept weight for every single C, CR, CRC and CXC is to determine how strong the C, CR, CRC or CXC is, in the current document as well as in the entire corpus. After calculating the Concept-weight, we add it with the PageRank with concept link weight obtained from equation 3, to obtain equation 5.

$$\text{PageRank}(A, c) = PR_{\text{withCLW}}(A) + \text{Concept} - \text{Weight}(c, A) \quad (5)$$

The purpose of adding the Concept-Weight(concept, page) to the PageRank component is to obtain different PageRanks for each of the C, CR, CRC and CXC for the same page, i.e., instead of a single PageRank for each document as proposed by most versions of the PageRank Algorithms, we have multiple PageRanks for a document, one for each of its C, CR, CRC and CXC. This is similar to the Topic sensitive PageRank, where each page has sixteen different PageRank values for sixteen predefined concepts. The difference we have shown in our work is that, we have as many PageRanks for a page as the number of concepts in that page. This helps to retrieve pages that are conceptually more relevant to the user queries.

The PageRank thus calculated using the frequency of the concept (and/or relations) in the document, the weight of the concept (and/or relations) in the document, and the co-occurring words of

the query in the documents are also considered for ranking a page, apart from the traditional physical links and our proposed concept links. The above said scores are calculated and the C, CR, CRC and CXC indices are built along with the calculated PageRank values for each of them.

Our Concept Link based PageRank algorithm version-1 (with physical links) improves the rank of the pages that are physically as well as conceptually linked. But, this has the disadvantage of a lesser PageRank value for the pages that are conceptually connected, but are not connected by any physical links. To tackle this problem, we have proposed another modification to the PageRank algorithm, called the Concept Link based PageRank algorithm version-2 (with or without physical links), that considers pages with strong conceptual connections, whether or not they are physically connected. We have added a boosting value based on the conceptual similarity between the documents. If the conceptual similarity between the documents exceeds certain threshold the boosting value will be 2. The threshold is set experimentally by iteratively checking the PageRank values.

5.2 Concept Link Based PageRank Algorithm- Version 2(With/Without Physical Links)

The version 2 of our algorithm differs from the version 1 in one way that version 1 using equation 2, calculates the conceptual link between two pages A and B only if A and B are physically connected whereas, version 2 calculates the concept link values between all N X N documents. As, we have mentioned earlier, this method further increases the space complexity. But as our experimental evaluation shows, this method outperforms both the existing and the version 1 of our proposed methodology. We have an improvement in the performance but at the cost of space and a considerable amount of offline processing time.

The concept links so calculated can be used as such for the HITS algorithm along with the physical links in the ranking process or using the concept links alone for the calculation of ranks using the HITS algorithm. We can further improve our system by filtering unwanted links of the web pages by determining how strongly an anchor text is associated with the content of the web page it is pointing to so that unnecessary deviation in the search process is avoided.

5.3 Query Dependant Ranking

Sections 5.1 and 5.2 concentrated on the offline score calculation and index creation for the ranking process. Following this step, the online ranking or the query dependant ranking is carried out which is discussed in this section. Once a query is given to the search engine, it is enconverted into a UNL graph as described in section 2.2. This UNL query graph is checked for matches in all the four indices that we have created. There are three different types of matches for a query concept to a document and they are given as follows in the order of their importance as given by Umamaheswari et al., [23]. 1) Presence of exact query terms in the document, 2) Presence of concept (and/or relations) terms in the document and 3) Presence of expanded query terms in the document. All the documents (d) falling under the above categories are chosen and the set 'D' is formed. The score of all these documents ($d \epsilon D$) is calculated using equation 6.

$$\text{Score}_d = \sum_{i=1}^{n} \text{match}_i * \text{Con} - \text{ind} - \text{wt}_i * \text{PageRank}(d,i) \quad (6)$$

Where,

match_i - is the weight corresponding to the type of the query concept (and/or relations) in the document and is weighted as 1, .5 and .25 for the types 1, 2 and 3 (discussed above) respectively. The weights are given based on the importance of the types in deciding the relevance i.e., the query concept (and/or relations) and query expansion are considered important only in the absence of the exact query term and the query expansion is considered important only in the absence of the query concept (and/or relations).

n – is the number of C, CR, CRC and CXC in the query.

Con-ind-wt_i – is based the query concept's presence in the C, CR, CRC and CXC indices and takes the values of 0.25, 0.5, 1 and 0.75 respectively. (CRC has a value 1 since it is the exact match, whereas, the rest correspond to partial match).

Our methods differ from the baseline by two components namely the Con-ind-wt_i (which improves ranking by our newly introduced concept index – CXC) and PageRank(d,i) (which improves ranking based on physical and/ or concept links present between the documents). Once the scoring is done, the 'd' documents in set 'D' are ranked according to the descending order of their scores. Thus we get a ranked list of documents for a query.

6 Discussion

We implemented the above variations of the PageRank algorithm and tested the performance of our methodologies with a set of one million Tamil tourism documents obtained from the project "Cross Lingual Information Access" which was funded by DIT, New Delhi [27]. We have used seed URLs and query collections of the CLIA project and compared our methods against the existing UNL based ranking methodology [23]. The vital statistics of the UNL-based indices are as follows: Depth of crawl: 4, about 13,00,000 URLs obtained of which 11,86,631 documents were totally enconverted, Number of C indices: 64,255, Number of CR indices: 97,00,109, CRC indices: 1,94,62,856 and the new component of index we have added to CoRee - Number of CXC indices: 1,04,58,971. Statistics regarding the queries are as follows: Number of queries: 64, Sub-topics of queries and their number i) temple(23), ii) special about places(10), iii) natural world(13), iv) festival(6), v) Facilities(4) and vi) Entertainment(8).

The evaluation was done by a set of 10 post graduate students working in the area of Information Retrieval in the University's Computer Science and Engineering Department. We consider CoRee as our baseline for evaluation and not the traditional PageRank algorithm since PageRank was designed to specifically suit keyword based search whereas CoRee is a concept based search engine and the multilevel UNL Concept based ranking algorithm used by CoRee is designed for concept based search. On initial evaluations and through the literature [23], we found that concept based search performs better than the keyword based search thereby ruling down our option of considering the traditional PageRank as the baseline for our work.

Now, we discuss the similarities and differences of the baseline with our proposed ranking methods.

Similarities of our ranking methods with CoRee's ranking are as follows: 1) As described in section 2.3, CoRee uses a three level ranking, where in the level 1, partial (CR) and the exact match (CRC) of CoRee is also used by our methods. However, our methods employ yet another type of partial match (CXC) as described in Section 3.2, 2) CoRee's level 2 ranking based on the match type is employed by our methods (equation 6), 3) Content information, which forms CoRee's level 3 ranking is also used by our methods (equation 4).

Differences of our ranking methods with CoRee's ranking are as follows: 1) Use of a new component of index – CXC which improves the ranking of pages that have two concepts connected by different UNL relations, 2) Use of Physical link structure existing between the web pages in our ranking methods helps to rank higher the pages with many inlinks and conceptual links - which is not used by CoRee's existing ranking method and 3)Use of concept link structure existing between the web pages in our ranking methods that helps in ranking higher the pages with many concept links - which is not used by CoRee's existing ranking method.

We have used three different evaluation schemes to compare the performances of our algorithm versions with the baseline namely, precision, recall and statistical significance test using one-way Anova [19] which are discussed in the sections 6.1, 6.2 and 6.3 respectively.

6.1 Precision

$$\text{Precision} = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{retrieved documents}|} \qquad (7)$$

We calculate precision using equation 7. The experimental results show only a small improvement over the existing UNL based ranking methodology for our Concept-link based PageRank Algorithm version 1. The reason for this is while calculating the concept links, we consider concept-link values between two documents only if they are physically connected i.e. the pages concerned are connected by an in-link or an out-link only. This helps in boosting the rank of the pages that are connected by in-links and out-links with many pages with many common concepts (and/or relations) (equation 3 value is boosted).

Table 1 shows the distribution of queries that performed (P@5 values) best under the following cases 1) CoRee is best, 2) CoRee and CLBPR-Version 1 are best, 3) CLBPR-Version 1 is best, 4) CLBPR-Version 1 and CLBPR-Version 2 are best, 5) CLBPR-Version 2 and CoRee are best, 6) CLBPR-Version 2 is best and 7) all three are good. These cases are explained in detail as follows:

Certain queries like (2 queries) "மாமல்லபுரம்" (Mamallapuram) performed well with the base line whereas did not perform well on our proposed methods. On examining documents ranked for these queries, by all three methods, we found that documents rich in content information about the query concepts fared well in the case of baseline and such documents lacked physical link connections

and concept link connection between other pages and hence did not fare well with both the proposed methods.

There are certain queries (3 queries) like "**முதுமலை வனவிலங்கு காப்பகம்**" (Mudumalai National Park) which performed better on the baseline and CLBPR-Version 1 and not with CLBPR-Version 2. On closer examination of the documents listed for these queries by all the three methods, we conclude that good content information (given by equation 4) and many physical links are the reason for the good performance of the baseline and the CLBPR-Version 1 respectively. The reason for this fall in precision by CLBPR-Version 2 is because the first few ranked documents have high physical link weight (part 1 of equation 4) than conceptual weight (part 2 of equation 4) or a high concept link value (given by equation 2) but the query terms play little or no part in the concept link weight of the document. Though this problem exists with the version 1 algorithm as well, we have tried to improve a document's concept link value by considering not only the conceptual links between physically linked pages but also conceptually well connected pages even if they are not physically connected. This helped in improving the results as shown above.

For some queries (4 queries) like "**உலக்கை அருவி**" (Ullakkai Falls), CLBPR-Version 1 performed better than the baseline and the CLBPR-Version 2. We found that the highly ranked pages in such cases are physically strongly connected (many in-links) which accounts to a major contribution to the PageRank value calculated by equation 4. In such cases, the Baseline falls due to the presence of less content information and CLBPR-Version 2 falls due to the lack of more number of common concepts (and/or relations) between the physically connected web pages.

Some queries (18 queries) like "**பிரகாச மாதா ஆலயம்**" (Prakasha Matha Church) showed better P@5 and P@10 by both the proposed methods than the baseline. On close examination of the listed documents for such queries where both the proposed methods performed better, we found that those documents had many physical in-links and the connected pages share high CLW value with such pages (this boosts PageRank value in equation 3) and the also had high CLW (from equation 2) for the query concept.

There are certain queries (4 queries) like "**காரைக்கால்**" (Karaikkal), which performed better for the baseline and CLBPR version 2 but performed bad for the other. On examining the result pages of all the three methods, we found that the fall of CLBPR-Version 1 is due to the influence of a high number of physical links between pages which neither have strong concept links with other pages i.e. the presence of common concepts (Cs, CRs, CRCs or CXCs) between them is very less (equation 2 has less value which in turn affects Part-2 of equation 3) nor have good content information (equation 4 value is very less). In certain cases, we also found out that the conceptual strength for such pages comes from concepts (and/or relations) other than those that are present in the query i.e., the page is conceptually highly connected to many other pages but the query term has no or little role in the conceptual connection inspite of the conceptual connection between certain pages being high.

As expected, our version 2 of the algorithm worked better overall as the result of exploiting the concept links along with the physical links. For example, many queries (31 queries) like "**தாஜ்மகால்**" (TajMahal) and "**புதுக்கோட்டை - குடுமியான்மலை**" (Pudhukottai-Kudumiyanmalai), showed better P@5 and P@10 values when compared to the other two methods. On

a close examination of the documents listed for these queries, we found that those pages had high concept link value (given by equation 2) for the query concept (and/or relations) apart from the content features (from equation 4) considered for ranking.

Table 1 Number of queries showing best P@5 for CoRee, CLBPR-Version 1 and CLBPR-Version 2

S.No	Methods	Number of Queries showing best performance (P@5)
1	CoRee	2
2	CoRee and CLBPR-Version 1	3
3	CLBPR-Version 1	4
4	CLBPR-Version 1 and CLBPR-Version 2	18
5	CLBPR-Version 2 and CoRee	4
6	CLBPR-Version 2	31
7	CoRee, CLBPR-Version 1 and CLBPR-Version 2	2

Certain queries (2 queries) like "நான்காவதுபடை வீடு : சுவாமிமலை" (The Fourth House: Swamimalai), perform well on all three methods. The ranked documents prove that the reason for this is mainly the rich content information regarding the query terms with an addition of good number of physical links and strong concept link values naturally exist in these pages.

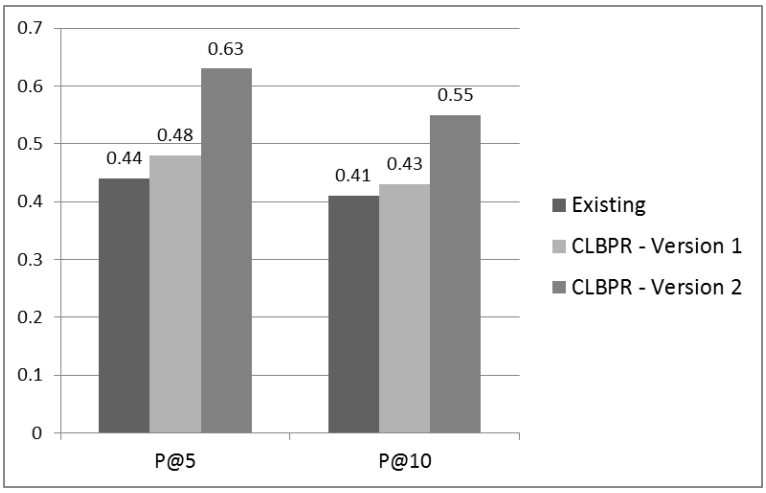

Figure 5 P@5 and P@10 for the existing, Concept-link based PageRank algorithm version 1(CLBPR-Version 1), Concept-link based PageRank algorithm version 2(CLBPR-Version 2).

We also found that the physically unconnected pages had rich conceptual connections among them. This valuable information was the motivating factor for our version 2 algorithm which calculated the concept link weight of each document with every other document. Though this process required extra computation time and space, there was a remarkable improvement in the precision values when compared to the existing method and the version 1 of our algorithm.

Figure 5 shows that the p@5 and P@10 for the three methods i)The existing UNL based ranking method, ii) Concept link based PageRank Algorithm version 1, iii)Concept link based PageRank Algorithm version 2. Figure 5 also shows the mean of all the P@5 values of all the three methods. Figure 5 clearly shows that our version 1 algorithm is only slightly better than the existing methodology and our version 2 algorithm shows a remarkable improvement of the P@5 values over the other two methods. Figure 5 also shows that our version 1 algorithm is only slightly better than the existing methodology and our version 2 algorithm shows a remarkable improvement of the P@10 values over the other two methods. On comparing the P@5 and the P@10 average values, there is a drop in the P@10 values than the P@5 values.

The clear picture of the statistics of the 64 queries considered is given in figure 6. As shown in the figure, CLBPR-Version 1 improves P@5 values of lesser number of queries (38) when compared to the CLBPR-Version 2 (53) against the baseline. A remarkable 82.8% improvement is achieved by the CLBPR-Version 2 algorithm on comparison with the baseline. Moreover, the number of queries for which the performance has degraded is very less in the case of CLBPR-Version 2 (5 queries - only 7.8%) when compared to CLBPR-Version 1 (15 queries - 23.44%) against the baseline. Both CLBPR-Version 2 and version 1 algorithms have certain queries for which the P@5 values remain the same as that of the baseline (11 queries - 9.38% and 6 queries - 17.18% respectively). An analysis of these query statistics also shows that the CLBPR-Version 2 algorithm performs better than the baseline as well as the CLBPR-Version 1 algorithm.

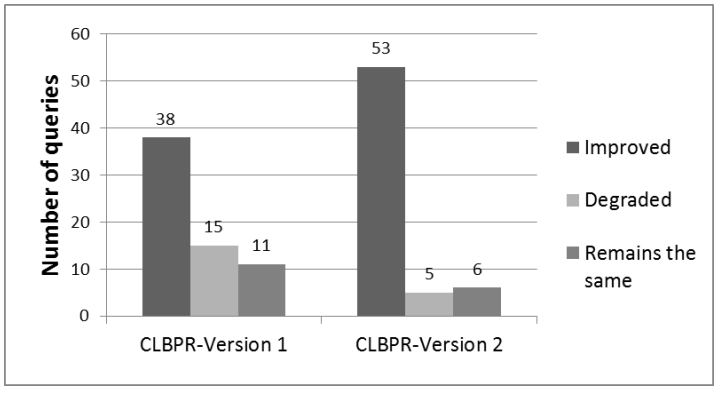

Figure 6 Queries improved and degraded by CLBPR-Version 1and Version 2 against the baseline.

6.2 Recall

$$\text{Recall} = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{relevant documents}|} \qquad (8)$$

We calculate Recall using equation 8. Our Recall evaluations showed minor variations for the three methods. Interestingly, both our proposed methods have better recall for 10% of the given queries so much so that a few highly relevant documents are retrieved by our proposed methodologies that were not retrieved at all by the existing methodology. We attribute this increase in the recall of our methods to the use of concept links as a major factor in ranking. The remaining queries have the same recall as that of the existing methodology i.e. our methodologies always have a recall that is greater than or equal to the recall of the existing method.

6.3 Statistical significance test using one-way Anova

Our next evaluation is the statistical significance test using one-way Anova [19]. We compared the performance of the three methods considered above. So our Null hypothesis is "There is a significant difference in the P@5 values of all the three methods".

We first conducted the experiment with a set of documents that were physically and conceptually not very tightly linked. As expected, this left us with no significance between the results based on the rejection of the null hypothesis considered above. The reason for this is the set of documents we have considered have poor physical and conceptual links thereby reducing the performance of both the versions of our algorithm to a mere content information based ranking similar to the baseline.

Table 2 Analysis of variance

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	0.265	0.132	3.733	0.038
Error	26	0.922	0.035		
Corrected Total	28	1.187			

Next, we conducted the experiment with a set of documents that were physically heavily linked but conceptually poorly linked. We expected no significance between the results using the light of the Precision evaluation done in section 6.1 since the our version-1 algorithm showed very less improvement over the baseline and the fact that the version – 2 of our algorithm performs almost the same as the version – 1 algorithm since the conceptual links between the pages are minimal and the ranking is centred completely on the physical links. The fall of version-1 algorithm is due to the highly physically linked pages being less relevant to the queries owing to the poor conceptual links.

Finally, we conducted the experiment with a set of documents that were conceptually heavily linked to other pages. We expected the version – 2 of our algorithm to be significant from the two other methods considered and the Anova tests have clearly shown that with the acceptance of the null hypothesis. The Analysis of variance is clearly shown in the table 2. The probability corresponding to the F value (Pr>F) is 0.038 and with this amount of risk only the null hypothesis is rejected. Therefore,

there is significance in the results. [We have not presented the Anova test results for the other two methods since they showed no significance].

Table 3 PRCcon / Fisher (LSD) / Analysis of the differences between the categories with a confidence interval of 95%:

Contrast	Difference	Standardized difference	Critical value	Pr > Diff	Significant
CLBPR Version-2 vs CoRee	0.236	2.726	2.056	0.011	Yes
CLBPR Version-2 vs CLBPR Version-1	0.138	1.593	2.056	0.123	No
CLBPR Version-1 vs CoRee	0.098	1.164	2.056	0.255	No
LSD-value:			0.176		

However, Anova only brings out the presence or absence of significance and other tests such as the Fisher's LSD or Tukey's HSD should be carried out to find the actual significance. We have used the Fisher's LSD method to compute the actual significance. This is clearly depicted in table 3. The last column of the table 3 says that the comparison of CLBPR Version – 2 against CoRee is statistically significant. Whereas, the other two comparisons show an absence of the statistical significance. This is also clearly evident from table 4 that shows group formation between the three considered methods. CLBPR Version-2 and CoRee are distinct apart from sharing space with CLBPR Version-1 in common in their respective groups.

Table 4 Group Formation

Category	LS means	Groups	
CLBPR Version-2	0.599	A	
CLBPR Version-1	0.461	A	B
CoRee	0.363		B

The document sets considered for the three Anova tests are subsets of the original set considered for the precision and recall calculation. Documents with high physical links and conceptual links were identified for this purpose. The query sets for these three Anova tests were also subsets of the original query set considered for the precision and recall calculation, based on the document subset considered for the respective methods.

Following an extensive Anova evaluation we conclude that CLBPR Version-2 is better in cases where the underlying set of web pages considered are conceptually very highly connected. In the other cases, as it is evident from the analyses, we conclude that the performance of the three methods do not show any statistical significance. However, the overall Precision evaluation shows that the CLBPR Version-2 outperforms both the other methods.

Our methodologies were designed specifically for concept based search engines. Though Coree is a search Engine for the Tamil language, our work is not language-limited. This is because of the underlying UNL principle that is language independent thereby making our methodologies work for any natural language.

7 Conclusions and Future Work

Our Concept-link based PageRank algorithm version 1 finds the strength of the concept links between physically linked pages and Concept-link based algorithm version 2 finds the strength of the concept links between all the N*N pages whether they are physically linked with each other or not. Our version 1 algorithm was better than the existing methodology but due to the very small improvement margin with the existing system, we developed the version 2 which has proved to be much better than the other two methods especially when there exists a dense network of concept links between the web pages.

As our future work, we would like to work on using concept links to learn a ranking function that automatically ranks documents for the user queries using machine learning techniques. We also want to make use of the concept links between pages for other processes like detecting unwanted links in a web page using the concept link values so as to minimize their effect on the ranking mechanism. We would also like to scale the methodology for a larger document set and a larger query set for evaluation by taming the time and space limitations that we currently face.

Acknowledgements

We thank Anna University, Chennai, Tamil Nadu, India for supporting the project with Anna Centenary Research Fellowship. We also thank CLIA (Cross Lingual Information Access) project funded by DIT (Department of Information Technology), New Delhi, India, for providing us with seed URLs for the document collection, queries and other resources used in this work.

References

1. Agirre, E. and Soroa, A., Personalizing PageRank for Word Sense Disambiguation. in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 33-41, Athens (2009).
2. Balaji, J., Geetha, T. V., Parthasarathy, R. and Karky, M., Morpho-Semantic Features for Rule-based Tamil Enconversion. in International Journal of Computer Applications IJCA, 2011, 26, 11-18.
3. Bar-Yossef, Z. and Mashiach, L-T., Local Approximation of PageRank and Reverse PageRank. in SIGIR, ACM, 865–866, Singapore (2008).
4. Borgs, C., Brautbar, M., Chayes, J. and Teng, S.-H., A Sublinear Time Algorithm for PageRank Computations. in Bonato, A. and Janssen, J. eds. WAW 2012: LNCS, vol. 7323, 41–53. Springer, Heidelberg (2012).
5. Brin, S. and Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine. in Computer Networks and ISDN Systems, Elsevier, 1998, 30, 107-117.
6. Deisy, C., Rajeswari, A. M., Indra, R. M., Jayalakshmi, N. and Mehalaa Devi, P. K., A Novel Relation Based Probability Algorithm for Page Ranking in Semantic Web Search Engine. in 5th

International Conference on Information Systems, Technology and Management, 2011, 138–148, Gurgoan (2011).
7. Efendioglu, D., Faschetti, C. and Parr, T., Chronica: a temporal web search engine. in Proceedings of the 6th International Conference on Web Engineering, California (2006).
8. Freyne, J., Smyth, B., Coyle, M., and Balfe, E. and Briggs, P., Further Experiments on Collaborative Ranking in Community-Based Web Search. in Artificial Intelligence Review, 2004, 21, 229–252.
9. Haav, H-M. and Lubi, T-L., A Survey of Concept-based Information Retrieval Tools on the Web. 5th East- European Conference, ADBIS 2001, 29–41, Vilnius (September 2001).
10. Haveliwala, T. H., Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. in IEEE Transactions On Knowledge And Data Engineering, 2003, 15(4), 784-796.
11. Kamvar, S., Haveliwala,T. H. and Golub, G., Adaptive Methods for the Computation of PageRank. in Proceedings of International Conference on the Numerical Solution of Markov Chains, 2003, 31–44.
12. Kurland,O. and Lee, L., PageRank without hyperlinks: Structural re-ranking using links induced by language models. in SIGIR'05, Salvador (2005).
13. Lin, J., PageRank without Hyperlinks: Reranking with Related Document Networks. Technical Report LAMP-TR-146/HCIL-2008-01.
14. Liu, F., Yu, C. and Meng, W., Personalized Web search for improving retrieval effectiveness. in IEEE transactions on Knowledge and Data Engineering, Jan 2004, 16(1), 28-40.
15. Madhu, G., Govardhan, A. and Rajinikanth, T. V., Intelligent Semantic Web Search Engines: A Brief Survey. in International Journal of Web and Semantic Technology, 2011, 2(1), 34-42.
16. Maehara, T., Akiba, T., Iwata, Y. and Kawarabayashi, K., Computing Personalized PageRank Quickly by Exploiting Graph Structures. in Very Large Data Bases, Hangzhou (2014), Proceedings of the VLDB Endowment, 7(12), 1023-1034.
17. Mihalcea, R., Tarau, P. and Figa, E., PageRank on Semantic Networks, with Application to Word Sense Disambiguation. in Proceedings of the 20st International Conference on Computational Linguistics (COLING 2004), Geneva (2004).
18. Noack, D., Spatial Variation in Search Engine Results. in Proceedings of the 43rd Hawaii International Conference on System Sciences, Hawaii (2010).
19. One Way ANOVA - University of Wisconsin - Stevens Point. [Online]. Available: http://www.uwsp.edu/psych/stat/12/anova-1w.ht
20. Qiu, L., Liang, Y. and Chen, J., Finding Important Nodes in Social Networks Based on Modified PageRank. in Computer Science and Information Technology, 2014, 6(1), 39-44.
21. Rasolofo, Y. and Savoy, J., Term Proximity Scoring for Keyword-Based Retrieval Systems. in European Conference on IR Research, 207 – 218, Pisa (2003).
22. Thelwall, M. and Vaughan, L., New versions of PageRank employing alternative Web document models. in ASLIB Proceedings, 2004, 56(1), 24-33.
23. Umamaheswari, E., Geetha, T. V., Parthasarathi, R. and Karky, M., A Multilevel UNL Concept based Searching and Ranking. in Proceedings of WEBIST 2011, 282-289, Noordwijkerhout (2011).
24. UNL- www.undl.org.
25. Wu, Y. and Raschid, L., ApproxRank: Estimating Rank for a Subgraph. in Proceedings of IEEE International Conference on Data Engineering, Shanghai (2009).
26. Zhu, F., Fang, Y., Chang, K.C.-C. and Ying, J., Incremental and Accuracy-aware Personalized PageRank through Scheduled Approximation. in PVLDB, 2013, 6(6), 481–492.
27. (http://tdil.mit.gov.in/RFP.aspx)