An Approach to Page Ranking Based on Discourse Structures

Subalalitha Chinnaudayar Navaneethakrishnan, and Anita Ramalingam

Original scientific paper

Abstract— **World Wide Web (WWW) which is predominant source for Information Retrieval today (IR) is essentially a set of hyperlinked documents. A web page containing more number of related hyperlinks satisfy the user needs in a single page. The IR systems should give high priority to such web pages. While assigning a rank for a web page, existing web mining techniques such as Hypertext Induced Topic Selection (HITS) and Page Ranking algorithms focus on the number of in links and out links present in the web page. Instead of just relying on the number of links present in the web page, the discovery of semantic relations between the web page and the hyperlinks present in the web page can improve the quality of the IR systems. The Rhetorical Structure Theory (RST) is widely used to find the semantic relations between text fragments by analysing the discourse structure of a text. In this paper, we propose a novel approach to find the semantic relation between a web page and the links present in the web page using RST. The proposed approach uses RST based discourse relations to find the relation between a web page and the hyperlinks present in the web page. We have implemented and evaluated our approach on an IR system using 500 Tamil language and 50 English tourism domain specific web pages. A comparison between the proposed approach and an existing page ranking algorithm has also been done.**

Keywords— Discourse structure, Link Analysis and Rhetorical Structure Theory.

I. INTRODUCTION

With the increasing web pages on the internet, it is the onus on the search engines to retrieve accurate web pages relevant to the user query. As there are many search engines available, it becomes essential to retrieve relevant web pages in few clicks. Web mining techniques analyse the documents on the web and relate them with the user's behaviour in selecting the web documents. Web mining techniques are categorized as Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) [1]. WCM identifies valuable information from the content on the web mainly within a document. WUM tracks the user profiles and their behavior in searching the content on the web. WSM identifies the link structure between the documents (inter document level) based on the number of in links and out links present in the web documents. This paper focusses on the WSM. WSM is implemented either by using Hypertext Induced Topic (HITS) algorithm or Page Ranking algorithm. Both the algorithms are used to rank the web pages retrieved by an Information Retrieval system (IR*).* To rank a webpage, both HITS and page ranking algorithm focus on the number of in links and out links present in each web page. Instead of just counting the links, identifying the semantic relations between the web pages and the links will reveal the strength of the bond between them in terms of the related contents. This paper proposes such a technique to find the semantic relations between the web pages through the links and to rank the web pages according to the number of reasonable semantic relations identified.

An end user will look for a web page where his needs get fulfilled in a single page that contains more number of useful links. The proposed approach explores this factor also by finding the number of out links that have semantic relations with the web page and ranks a web page accordingly. To find the semantic relationship between the web pages and the out links present in the web page, the proposed approach makes use of Rhetorical Structure Theory (RST) [2]. The proposed approach is an initial attempt in page ranking, where a discourse theory is used in page ranking.

The rest of the paper is organized as follows. Section 2 describes the works that are related to the proposed work. Section 3 gives the back ground information of the proposed approach. Section 4 illustrates the proposed approach. Section 5 discusses the evaluation of the proposed approach and Section 6 lists the conclusions of the paper.

II RELATED WORK

This section discusses the works about the existing page ranking algorithms.

HITS which was originally developed by Kleinberg in 1998 ranks the webpages by analysing the in links and out links. The webpages pointed to by many hyperlinks are called *authorities*, whereas webpages that point to many hyperlinks are called *hubs* [3]. Authorities and hubs exhibit a mutually reinforcing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs. Hubs point to lots of authorities and authorities are pointed to by lots of hubs. Both Authorities and Hubs form a bipartite graph. Each Hub and authority are assigned scores. An authority web page pointed by many highly scored hubs is

Manuscript received October 7, 2016; revised November 3, 2016 and November 24, 2016.

Authors are with the Department of Computer Science and Engineering SRM University, Kattankulathur Chennai, India (E-mails: subalalitha@gmail.com, anita_kalai@yahoo.co.in).

a strong authority while a hub that points to many highly scored authorities is a popular hub . Let a_p and h_p represent the authority and hub scores of page *p*, respectively. Let *B* (*p*) denote the set of pages that points to page p and *I* (*p*) be the set of pages pointed by page *p*. The scores of the hubs and authorities are calculated using (1) and (2)[4].

$$ap = \sum_{i \in B(p)} hi \quad (1)$$

$$hp = \sum_{i \in I(p)} ai \quad (2)$$

Once an IR system retrieves a set of pages for a specific query, the HITS ranks each web page depending on its hub score and authority score. HITS calculates these scores from the retrieved documents set. HITS ignores the textual content of the web page and hence a popular web page that is not highly relevant to the given query gets a high score.

On the other hand, the PageRank algorithm which was originally developed at Stanford University by Larry Page and Sergey Brin in 1996 assigns a rank to each page belonging to a hyperlinked set of documents by measuring its relative importance within the set. Page Rank algorithm is query independent and it analyses the whole web to assign a rank to every web page [4]. If a page contains important in links then the out links of this page are also considered important. If an authoritative web page *A* links to page *B*, then *B* is also authoritative. Page Rank uses the back link in deciding the rank score. If the sum of all the ranks of the back links is high, then the page is given a high rank [4].

Page rank of a page p is calculated using (3). Initially, all the web pages are assigned the rank 1.

$$\sum PR(p) = (1 - d) + d \sum_{q \in B(p)} PR(q)/Nq \quad (3)$$

Where, PR(p) and PR(q) *are* the Page Ranks of pages p and q, B(p) is the set of pages that pages that points to p. Nq is the number of out links of page q. *d* is a damping factor which can be set between 0 and 1, but it is usually set to 0.85

The page rank is repeatedly calculated until the values of two consecutive iterations match. This forms the basic page rank algorithm. Various works have been reported based on the page rank algorithm where each of them adds a variation to the basic algorithm. For instance, Xing et al have proposed a weighted page rank algorithm which adds a weight to each out link and in link based on their popularity whereas, the basic page rank algorithm gives equal weightage to all links of a web page[5]. The weight of a link is calculated using (4)

$$WL(p,q) = I(u)/\sum_{p \in R(q)} I(p) \quad (4)$$

Where R(v) denotes the number of out links of q. The page rank using the weight is calculated using (5).

$$PR(p) = (1 - d) + d \sum_{q \in B(p)} PR(q)/Nq(WL(v,u) \quad (5)$$

Nidhi Shalya et al have enhanced the weighted page rank algorithm proposed by Xing et al by adding a content weight Wc[6] . Each web page q that is linked to the page p is assigned a content weight depending on the terms that matches the query. The weight calculates the ratio of the sum of the frequency of the query words and sum of the frequency of whole query with respect to page q and the page rank is calculated using (6)

$$PR(p) = (1 - d) + d \sum_{q \in B(p)} PR(q)/Nq (WL(v,u)Wc \quad (6)$$

It can be observed that from the works discussed on the page ranking algorithms assign a rank to a web page depending on the number of in and out links present in it. In addition to this the strength of each link is assessed based on the popularity of the links. Again the popularity is calculated based on the quantity of the links. Though Nidhi Shalya et al have involved the query terms matching in finding the strength of the links the actual relationship between each web page and the links are not analysed in any of the work discussed in this section. The semantic relationship between the web pages and the links present in them is still a mystery. The proposed approach attempts to find the semantic relationship between the web pages and their links and gives weightage to each web page depending on the number of semantically relevant links it contains. As an initial attempt, the proposed approach focusses only on the out links rather than the in links using RST

The next section discusses the basics of RST based discourse relations which facilitates the understanding of the proposed work.

III Overview of RST

RST was originally proposed by, Bill Mann, Sandy Thompson and Christian Matthiessen at the University of Southern California [2]. RST was designed as a descriptive theory for organizing an NL text using discourse relations. Discourse relations connect the coherent text fragments of a text document. The smallest text fragment is known as Elementary Discourse Unit (EDU). An EDU is essentially a clause or a sentence. The text fragments that are bigger than clauses, and sentences, which may comprise many EDUs, are called Complex Discourse Units (CDUs)[7] . The text fragments are categorized as the nucleus and the satellite. The nucleus represents the salient part of the text, while the satellite represents the additional information about the nucleus. The discourse relations fall into three categories, namely, subject matter relations, presentational relations and multi nuclear relations. In subject matter relations, a satellite is a question, a request, or a problem raised by the reader, which is satisfied or solved in the nucleus. In presentational relations, a satellite increases the reader's inclination in accepting the facts stated in the nucleus. A multi nuclear relation connects two nuclei instead of connecting a nucleus and a satellite. *Elaboration, Evaluation, Interpretation, Means, Cause, Result, Otherwise, Purpose, Solutionhood, Condition, Unconditional* and *Unless* are subject matter relations.

Antithesis, Background, Concession, Enablement, Evidence, Justify, Motivation, Preparation, Restatement and *Summary* are presentational relations. *Conjunction, Disjunction, Contrast, Joint, List, Multi Nuclear Restatement* and *Sequence* are multi nuclear relations. Fig 1 shows the RST representation for an English sentence given in Example 1. The Nucleus-DiscourseRelation-Satellite structure is denoted as NRS sequence in this paper.

Example 1: While I was watching TV, I happened to see him on screen

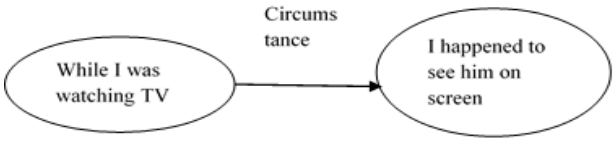

Fig 1.*NRS sequence* for Example 1

Recently an RST based discourse parser has been proposed using RST relations and constructs a language independent discourse structure for a given NL text using Universal Networking Language(UNL)[8]. The UNL provides a semantic representation for the NL text by relating the concepts of the text using UNL based semantic relations thereby forming a hyper graph that consist of concepts as nodes and UNL relations as edges[9]. The similarities that exist between the UNL relations and RST relations are exploited by the parser to build the discourse structure. Given an NL text of any language the discourse parser constructs a language independent discourse structure and hence the proposed approach is extendable to any language. The existing RST based discourse parser discourse parser finds the semantic relations between the consecutive sentences. In this paper, we extend the working of the existing RST based discourse parser to find the semantic relations between the web page and the hyperlinks present in the web page. Furthermore, the semantic indices are identified from the discourse structure. Each semantic index takes has the information regarding the number of useful hyperlinks present in a web page which is used for ranking the page in the IR system. Similar to the proposed approach, RST based indexing for IR system has been proposed by Farhi Marir and Kame1 Haouam[10]

IV PROPOSED WORK

Fig 2 shows the architecture of the proposed work of Each out link present in a web page is checked if it is semantically related to the web page. The web page is assigned a weight based on the number of the semantically related out links present in it. While indexing the web pages, the weights are stored along with the indices and hence the weight assignment process is done offline. The web page that contains the out link is denoted as the main document and the web page that corresponds to the out link is denoted as the linked document in this paper. The RST based discourse structures are constructed for the main document and the linked documents. From the discourse structure, the weight of a web page based on the number of semantically related out links is assessed as follows.

The input web page is given to the proposed model. The out links present in a web page are identified by the *Hyperlink Identifier*. The NL text content are extracted from these web pages and the *RST* based discourse parser constructs the discourse structure for the extracted text content. The *Topic Extractor* captures the theme of the web pages which are essentially a set of important concepts that convey the theme of the web page from the *NRS sequences* of both the main and linked documents. Since nouns convey the theme of a text, the theme concepts are chosen by finding the frequent noun concepts that occur in the nuclei of the *NRS sequences*. A theme of a text may be conveyed by more than one noun and hence the top five frequently occurring nouns are chosen as the theme concepts for a given text. Let the theme concepts of the main document be $\text{ThemeConcepts}_{\text{main}}$ and the theme concepts of the linked document be $\text{ThemeConcepts}_{\text{linked}}$. The *NRS sequences* that contain $\text{ThemeConcepts}_{\text{main}}$ and ThemeConceptslinked are extracted from both the main and the linked documents. These NRS sequences are ranked by the *NRS Sequence Analyser*. The ranking is done based on the factors as listed in the Table I. These factors are decided and ranked accordingly by analysing the discourse structures of about 100 text documents manually. The NRS sequences that gets the highest rank are chosen and a set of discourse relations R_{rel} is obtained from the discourse relation "*R"* present in the *NRS sequences, where* $R \epsilon R_{rel}$. The set R_{rel} expresses the semantic relations that exist between the main document and the linked document.

It can be observed from the table that the factor, *Presence of the* $\text{ThemeConcepts}_{\text{main}}$ *and* $\text{ThemeConcepts}_{\text{linked}}$ *in the nucleus of the NRS sequence* gets the top priority as the nucleus convey the salient part of the text and the consequently respective NRS sequences contain the semantic relations that links the salient part describing both documents. Both the factors, *Presence of* $\text{ThemeConcepts}_{\text{main}}$ *in the nucleus and the presence* of $\text{ThemeConcepts}_{\text{linked}}$ *in the satellite of the NRS sequence and Presence of* $\text{ThemeConcepts}_{\text{main}}$ *in the satellite and the presence of* the $\text{ThemeConcepts}_{\text{linked}}$ *in the nucleus of the NRS sequence* indicate that the both the main and the linked documents are related through a semantic relation. Since the

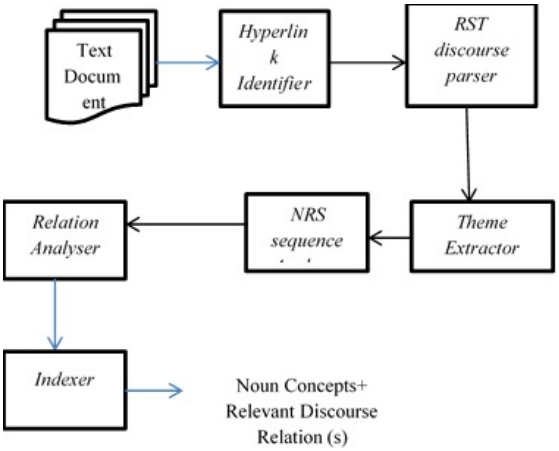

Fig 2. Architecture of the Proposed Work

TABLE I
FACTORS FOR RANKING NRS SEQUENCES

Factor	Rank
Presence of the ThemeConcepts$_{main}$ *and* ThemeConcepts$_{linked}$ *in the nucleus of the NRS sequence.*	1
Presence of ThemeConcepts$_{main}$ *in the nucleus and the presence* of ThemeConcepts$_{linked}$ *in the satellite of the NRS sequence.*	2
Presence of ThemeConcepts$_{main}$ *in the satellite and the presence of* the ThemeConcepts$_{linked}$ *in the nucleus of the NRS sequence.*	2
Presence of the ThemeConcepts$_{main}$ *and* ThemeConcepts$_{linked}$ *in the satellites of the NRS sequence*	3

importance of both the main and the linked documents are equally weighed, both the factors, *Presence of* ThemeConcepts$_{main}$ *in the nucleus and the presence* of ThemeConcepts$_{linked}$ *in the satellite of the NRS sequence and Presence of* ThemeConcepts$_{main}$ *in the satellite and the presence of* the ThemeConcepts$_{linked}$ *in the nucleus of the NRS sequence are equally ranked irrespective of the occurrence of* ThemeConcepts$_{main}$ in the nucleus *and* ThemeConcepts$_{linked}$ in the satellite *and vice versa*. The factor, *Presence of the* ThemeConcepts$_{main}$ *and* ThemeConcepts$_{linked}$ *in the satellites of the NRS sequence* denotes that the NRS sequence carry a text that contains an additional information pertinent to both the main and the linked documents and hence given the last priority.

The relations identified by the *NRS Sequence Analyser* are further analysed by the *Relation Analyser*. The RST based discourse relations that express a tightly coupled relationship between the web page and its hyperlink as far as tourism domain is concerned are, *Background, Enablement, Evidence, Justify, Preparation, Elaboration and Motivation.* This has been decided again through manual analysis of the discourse structures of the 100 text documents which were used to decide on the factors to extract the NRS sequences as discussed previously in this section. Hence, the *NRS Sequence Analyser* helps to find the semantic relations that link the web page and the hyperlink through the theme concepts, whereas, the *Relation Analyser* identifies the tightly coupled hyperlinks that will help the readability of the web page.

Finally, the indexer identifies a set of conceptual indices representing the main document. The indices convey the quality of the web page in terms of the number of semantically closer out links it contains. An index for a web page comprises a set of noun concepts (N), set of discourse relations (R) and number of tightly coupled out links. The discourse relations R are the ones that are analysed by the *Relation Analyser*, whereas the noun concepts refer to the frequent noun concepts that occur in the NRS sequences that contain the relations *R*. Since nouns convey the essence of a text, the noun concepts are chosen to be part of the index. Since an index needs to be concise, the number of noun concepts chosen to form the *index* is limited by choosing the frequent abstract noun concepts leaving behind its instances. The abstract noun concepts can be found by using any semantic knowledge base such as WordNet or ontology. The number of tightly coupled out links is the weight factor that is used to rank the web page when used by an IR system.

V EVALUATION

As mentioned previously, to construct the discourse structure, we have used the existing RST discourse parser [8]. We have evaluated the proposed approach in three different perspectives. First, we have evaluated our approach on an IR system to validate the correctness of our work. We have analyzed the accuracy of the semantic relations identified between the main documents and the linked documents. Furthermore, in order to examine the accuracy of the semantic indices identified, the precision is calculated for the top ten results. Secondly, we have compared the efficiency of the page ranking using the proposed approach with the existing page ranking technique [6]. Finally, we have analyzed that the in links and out links of the web pages in order to show that by just counting the number of in links and out links of a web page is not a sufficient factor to increase its rank in an IR system. This has been done by analyzing the semantic relations between the pages using the proposed approach to find if the web pages that are linked through hyperlinks are linked content wise.

The proposed approach has been tested on an IR system using 500 Tamil language web pages. Mean Average Precision (MAP) is used as the evaluation parameter for evaluating the discourse relations identified between the web pages and the hyperlinks [11]. In the IR jargon, MAP for a set of queries is the mean of the average precision scores for each query. This quality of the MAP score is used to evaluate the average number of correct discourse relations found between the web pages. MAP score is calculated at two levels namely, for each web page, and for all the 500 web documents. MAP score calculated for each web page gives the average of the precision value obtained in finding the discourse relation for each hyper link present in it, whereas, the MAP score calculated for the whole corpus gives the mean of the MAP score calculated for a single web page. MAP scores are calculated using human judgement using 5 experts in the discourse structure analysis. Table II shows the factors involved in the evaluation of the proposed approach.

TABLE II
FACTORS OBSERVED IN EVALUATION OF THE PROPOSED APPROACH

Factors	Values
Average number of semantic relations identified between a single web page and a hyperlink	5
Average number of hyperlinks present in a web page	30
Precision for the whole corpus	90%

RST based discourse relation such as, "*Elaboration", "Justification" and "List*" were observed frequently between the main and the linked documents.

The IR system was tested using ten queries, and the precision for the first ten documents (P@10) was used as the parameter for evaluation. The P@10 values calculated for ten queries are shown in Fig 3. The search is done by using various combinations of the noun concepts and discourse relations which help in retrieving the semantically closer documents to that of query and rank them accordingly. When the corpus size increases, the number of noun concepts and discourse relations becomes uncontrollable, and it becomes tedious in terms of storage. To overcome this problem, the size of the noun concepts in an index is set to be four by choosing the noun concepts according to their sentence position.

Since the document processing and indexing details of the Nidhi Shalya et al's work are not known, the documents are processed and indexed as done by the proposed approach using the same 500 Tamil language web pages. This is done in order to show that the proposed ranking approach can outperform the existing approach despite using the content weight for ranking. Table III lists the Discounted Cumulative Gain (DCG) values of the both the proposed and the existing approaches [6] computed for five queries for the first five top ranked web pages. The DCG gives a particular rank position *p* is defined as given in for the web page(7)[12]

Where, Reli is the graded relevance of the result at the position *i. The Reli* is 0 if the web page is completely irrelevant; 1 if the web page is partially relevant and 2 if the web page is completely relevant.

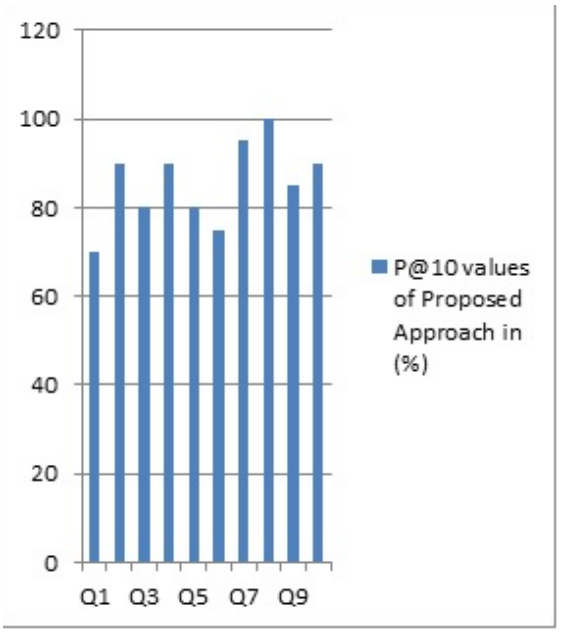

Fig 3. Graph Showing P@10 values for the Queries

$$DCGp = \text{Rel1} + \sum_{i=2}^{i=p} \left(\frac{Reli}{log2(i)} \right) \quad (7)$$

TABLE III
DCG VALUES FOR FEW QUERIES

Query$_1$		**Query$_2$**		**Query$_3$**	
IR$_{proposed}$	**IR$_{existing}$**	**IR$_{proposed}$**	**IR$_{existing}$**	**IR$_{proposed}$**	**IR$_{existing}$**
2	2	2	2	2	2
4	3	4	4	4	4
5.26	3.63	5.26	4.63	5.26	4
6.26	3.63	6.26	5.63	6.26	4
7.12	3.63	6.69	5.63	4.6	4

It can be observed that the performance of the proposed approach is better than the existing approach in terms of ranking. This is because a web page that satisfies the user in a single page gets a high graded relevance score and hence it is reflected in the DCG scores. A web page may contain irrelevant hyperlinks in order to increase its page rank. This has been eliminated by the proposed technique which analyses the semantic relation between the web page and the hyperlinks present in the web page through the RST discourse relations. This is further analyzed by examining the semantic relation between the web pages and the hyperlinks present in the web pages that has high in links and out links using 500 Tamil language web pages. This analysis has been done in order to show that it is insufficient to rank a web page by merely counting it's in links and out links as done by the existing page rank algorithms.

The number of out links is first identified for all the web pages. Each web page is checked if it contains the link (in link) of the rest of web pages. Each web page is tagged with the number of in links and out links count. The web pages having beyond three in links and out links (totally six) were examined further if they have semantically relevant out links using the proposed approach as discussed in section 4. The statistics of the analysis is shown in Table IV.

TABLE IV
STATISTICS OF THE FACTORS

Factors	**Values**
Number of web pages that had more than three in links and out links	378
Number of web pages that did not have a single relevant out link	88

Average number of out link count observed	30
Average number of in link count	6

It can be observed that out of 378 web pages that had more than six in link and out link count in sum, 88 web pages did not have even a single relevant in link in them. Furthermore, only two semantically relations were found on an average in single web page when the average number of out link is thirty. This shows that the semantic relevance between a web page and the hyperlinks present in it needs to be examined to increase its rank. The proposed approach finds the semantic relation between the web pages and their out links only. Also, the evaluation also focuses on checking the semantic relevance between the web pages and the out links. Analysing the semantic relevance between the web pages and the in links will improve the efficiency of the IR system even better.

Though the proposed approach finds the semantic relation between a web page and the hyperlinks present in the web page but can very well be used to find the semantic relations between the text documents. Furthermore, proposed approach can be implemented using any other discourse parser that can generate NRS sequences. The proposed approach is currently tested using less number of text documents which needs to be increased. Though the proposed approach is domain independent, in this paper the evaluation has been done using tourism domain specific documents. In order to prove the generic nature of the proposed approach, it should be tested with more number of expository texts.

VI CONCLUSION

A methodology to rank a web page based on the number of semantically related out links present in it has been proposed in this paper. The existing page ranking techniques ranks a web page based on the number of in links and out links present in a page and page ranking is done after the first round of retrieval is done using the user query to the IR system. The proposed approach assigns a weight to each web page based on the number of closely related out links while indexing phase itself and hence the user gets the ordered pages immediately after the query is given. The semantic relations between the web pages are identified by constructing a discourse structure using two types of semantic relations namely, RST based discourse relations. These semantic relations are identified using the language independent discourse parser named RST discourse parser [8]. RST based discourse relations capture the coherence of the text. The proposed approach explores many factors that involve the discourse structure in order to bring out the semantic coordination that exists between the web page and the hyperlinks present in the web page.

The proposed approach uses the discourse structures of the documents instead of doing a discourse analysis using the NL sentences. The proposed approach is a language independent by the using the language independent discourse structure that uses UNL [8]. Furthermore, the proposed approach can make use of any discourse structure that has an NRS sequence based analysis such as, RST and CST. The proposed approach is an initial attempt to page ranking using discourse structures. Other than page ranking in IR, the proposed approach suits well for NLP tasks such as, multi-document summarization. Since. the proposed approach is extendable to any language, the NLP tasks can also be made language independent.

REFERENCES

[1].Jaideep Srivastava , Robert Cooley , Mukund Deshpande , Pang-Ning Tan.”Web usage mining: discovery and applications of usage patterns from Web data”, ACM SIGKDD Explorations Newsletter, v.1 n.2, January 2000 .

[2] Mann, W., and Thompson, S. “Rhetorical structure theory: Toward a functional theory of organization”. Text 8(3):243–281,1988.

[3] Li, L.; Shang, Y.; Zhang, W. "Improvement of HITS-based Algorithms on Web Documents". Proceedings of the 11th International World Wide Web Conference (WWW 2002)

[4] Pooja Devi, Ashlesha Gupta and Ashutosh Dixit, "Comparative Study of HITS and PageRank Link Based Ranking Algorithms", *Proc of International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 2

[5] Xing, W., Ghorbani, A.A.”Weighted pagerank algorithm”, In: Proc. of the 2nd Annual Conference on Communication Networks and Services Research, pp. 305–314 (2004)

[6]. Nidhi Shalya, Shashwat Shukla and Deepak Arora,” An Effective Content Based Web Page Ranking Approach”, International Journal of Engineering Science and Technology (IJEST) , Vol. 4 No.08 August 2012.

[7]. Schauer, H. “From Elementary Discourse Units to Complex Ones”, in L. Dybkjær, K. Hasida and D. Traum (eds) Proceedings of 1st SIGdial Workshop on Discourse and Dialogue, pp. 46–55, Hong Kong,2000.

[8]].Navaneethakrishnan, S. C., and Parthasarathi, R. “Building a Language-Independent Discourse Parser using Universal Networking Language”. Computational Intelligence, 31: 593–618,2015

[9]. Uchida, H., Zhu, M., Della Senta, T. A gift for a millennium. IAS/UNU, Tokyo, 1999.

[10]Farhi Marir and Kame1 Haouam: “Rhetorical Structure Theory for content-based indexing and retrieval of Web documents” in Proc. of 2nd International Conference on Information Technology: Research and Education, 2004, ITRE 2004. Pages:160 – 164

[11] S. Robertson. A new interpretation of average precision. In SIGIR, pp. 689–690, 2008

[12]. Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, Tie-Yan Liu. 2013. A Theoretical Analysis of Normalized Discounted Cumulative Gain (NDCG) Ranking Measures. In Proceedings

Subalalitha C. Navaneethakrishnan has finished her PhD in the Natural Language Processing domain in the year 2014 in College of Engineering Guindy (CEG), Anna University, India. She was working as a Junior Research Fellow for the Indian Government funded project titled, “Cross Lingual Information Access” for about 3.5 years in CEG. She has published her research papers in various refereed International Journals, National Journals and International Conferences. Her research interests inclines towards the domains namely, Natural Language Processing and Computational Linguistics. She is currently working as Assistant Professor in SRM University, Tamil Nadu, India.

Anita Ramalingam is currently working as Assistant Professor in SRM University, Tamil Nadu, India. She has 10 years of teaching experience. She has published her research work in International Journals and national conferences. Her research interests include, Natural Language Processing, Theory of Computation and Cloud Computing.