



## Dynamic Navigation of Query Results Based on Concept Hierarchies using Topic-Sensitive Page Rank Algorithm

L. Lakshmi\*, Dr. P. Bhaskara Reddy (Guide), Dr. C. Shoba Bindhu (Co-Guide)

Department of CSE & JNTUA, Kukatpally,  
Hyderabad, Telangana, India

---

**Abstract**— *The web as we all know is an infinite source of information which includes massive collection of web pages and countless hyperlinks. The aim of the Web Structure Mining is to generate the structural abstract about the Web site and Web page. It tries to determine the link structure of the hyperlinks at the inter document level. This type of mining can be carried out at the document level (intra-page) or at the hyperlink level (inter-page). Search queries on educational databases, such as EBSCO (Elton B. Stephens Company Database on Education) often return a large number of results, only a small subset of which is relevant to the user. A natural way to organize educational citations is according to their The Library of Congress Subject Headings (LCSH) annotations. LCSH is a comprehensive concept hierarchy used by EBSCO. In this paper, we present a topic-sensitive Page Rank algorithm for dynamic navigation of query results, we pre compute the importance scores offline, as with ordinary Page Rank algorithm. However, we compute multiple importance scores for each page; we compute a set of scores of the importance of a page with respect to various topics.*

**Keywords**— *EBSCO, LCSH, Dynamic navigation, Page Rank algorithm*

---

### I. INTRODUCTION

During the past few years the World Wide Web has become the foremost and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs.

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information.

On the WWW, the use of structure mining enables the determination of similar structure of Web pages by clustering through the identification of underlying structure. This information can be used to project the similarities of web content. The known similarities then provide ability to maintain or improve the information of a site to enable access of web spiders in a higher ratio. The larger the amount of Web crawlers, the more beneficial to the site because of related content to searches.

In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites. The determined connection brings forth a useful tool for mapping competing companies through third party links such as resellers and customers. This cluster map allows for the content of the business pages placing upon the search engine results through connection of keywords and co-links throughout the relationship of the Web pages. This determined information will provide the proper path through structure mining to improve navigation of these pages through their relationships and link hierarchy of the Web sites.

With improved navigation of Web pages on business Web sites, connecting the requested information to a search engine becomes more effective. This stronger connection allows generating traffic to a business site to provide results that are more productive. The more links provided within the relationship of the web pages enable the navigation to yield the link hierarchy allowing navigation ease. This improved navigation attracts the spiders to the correct locations providing the requested information, proving more beneficial in clicks to a particular sit

### II. RELATED WORK

Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information. The first of these problems is irrelevant search results. Relevance of search information become misconstrued due to the problem that search engines often only allow for low precision criteria. The second of these problems is the inability to index the vast amount if information provided on the Web. This causes a low amount of recall with content mining. This minimization comes in part with the function of discovering the model underlying the Web hyperlink structure provided by Web structure mining.



An example concept hierarchy with low precision criteria

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links.

### III. EXISTING ALGORITHMS

The existing algorithms focused on effective navigation of query results while the proposed approach is to reduce redundancies and maximize structural coverage of sub topics in hierarchies that represent knowledge in the form of concept hierarchies.

#### Navigation and Cost Model

The navigation model of EduNav is formally defined in this section. Then the navigation cost model is presented, which is used to devise and evaluate our algorithms. After the user issues a keyword query, EduNav initiates a navigation by constructing the initial active tree (which has a single component tree rooted at the LeSH root) and displaying its root to the user. Subsequently, the user navigates the tree by performing one of the following actions on a given component sub tree.

---

```

EXPLORE( $I(n)$ )
  if  $n$  is the root
     $S \leftarrow \text{EXPAND } I(n)$  // that is  $S \leftarrow \text{EdgeCut}(I(n))$ 
    For each  $n_i$  in  $S$ 
      EXPLORE( $I(n_i)$ )
  else, if  $n$  is not a leaf-node, choose one of the following:
  1. SHOWRESULTS  $I(n)$ 
  2. IGNORE  $I(n)$ 
  3.  $S \leftarrow \text{EXPAND } I(n)$ 
    For each  $n_i$  in  $S$ 
      EXPLORE( $I(n_i)$ )
  else, choose one of the following: //  $n$  is a leaf node
  1. SHOWRESULTS  $I(n)$ 
  2. IGNORE  $I(n)$ 
    
```

---

topdown navigation model

The “optimal” valid EdgeCut is the EdgeCut that will lead to the minimum expected navigation cost, that is, the minimum average cost. In order to minimize the expected cost of TOPDOWN-EXHAUSTIVE navigation, we need to minimize the cost of EXPAND and of SHOWRESULTS. The cost of EXPAND is simply the number of component sub trees produced by the EdgeCut. The average cost of SHOWRESULTS over all component subtrees equals the sum of unique elements (citations) in every sub tree. Optimal cost can be computed by recursively listing all possible sets of EdgeCuts. This starts from the root and traverses every concept in the tree. This algorithm is expensive. To overcome this Opt EdgeCut algorithm is proposed which provides minimum expected navigation cost.

**Algorithm Opt-EdgeCut**

**Input:** The navigation tree  $T$

**Output:** The best EdgeCut

```

1 Traversing  $T$  in post-order, let  $n$  be the current node
2 while  $n \neq \text{root}$  do
3   if  $n$  is a leaf node then
4      $\text{mincost}(n, \emptyset) \leftarrow P_E(n) * L(n)$ 
5      $\text{optcut}(n, \emptyset) \leftarrow \{\emptyset\}$ 
6   else
7      $\mathbb{C}(n) \leftarrow$  enumerate all possible EdgeCuts
           for the tree rooted at  $n$ 
8      $\mathbb{I}(n) \leftarrow$  enumerate all possible subtrees
           for the tree rooted at  $n$ 
9     foreach  $I(n) \in \mathbb{I}(n)$  do
10      compute  $P_E(I(n))$  and  $P_C(I(n))$ 
11      foreach  $C \in \mathbb{C}(n)$  do
12        if  $C$  is a valid EdgeCut for  $I(n)$  then
13           $\text{cost}(I(n), C) \leftarrow$ 

$$P_E(I(n)) \cdot \left( \begin{array}{l} (1 - P_C(I(n))) \cdot L(I(n)) \\ + P_C(I(n)) \cdot (B + |S| + \sum_{s \in S} \text{mincost}(I_C(s))) \end{array} \right)$$

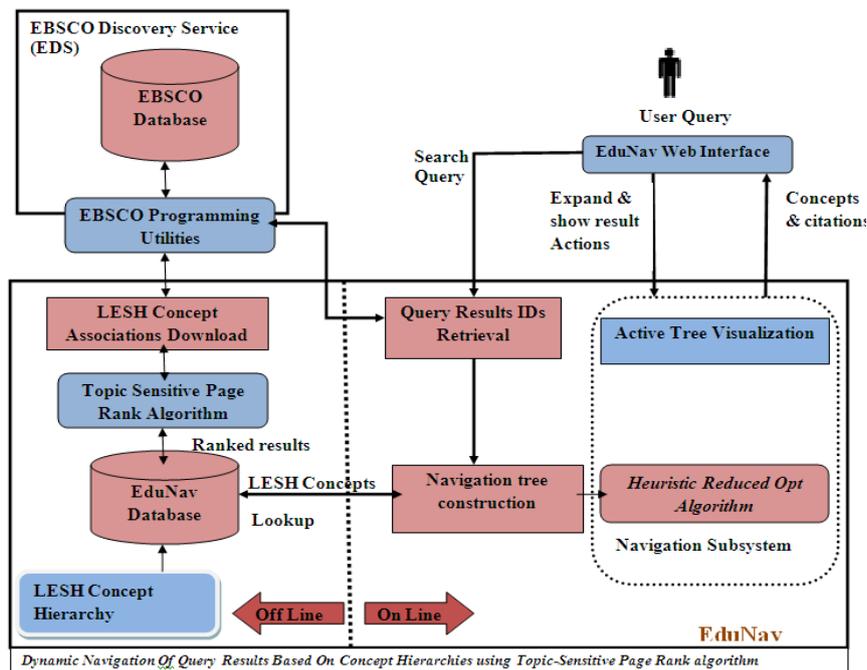
14        else
15           $\text{cost}(I(n), C) = \infty$ 
16         $\text{mincost}(n, I(n)) \leftarrow \min_{C_i \in \mathbb{C}(n)} \text{cost}(I(n), C_i)$ 
17         $\text{optcut}(n, I(n)) \leftarrow C_i$ 
18 return  $\text{optcut}(\text{root}, E)$  //  $E$  is the set of all tree edges

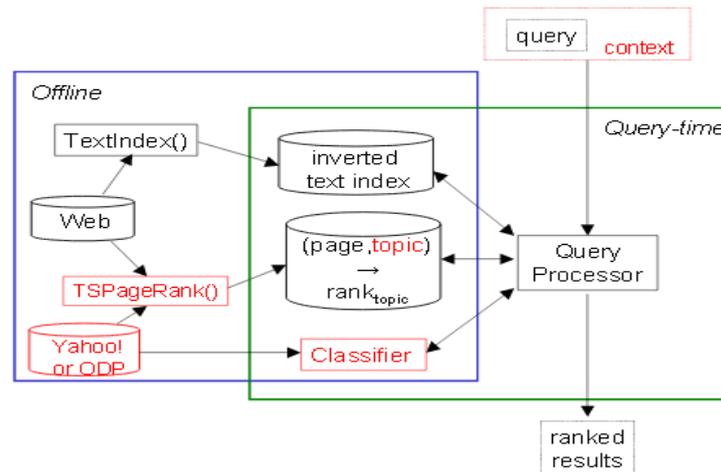
```

Optimal edge cut algorithm

**IV. PROPOSED SYSTEM**

In this system, we present a topic-sensitive Page Rank algorithm, we pre compute the importance scores offline, as with ordinary Page Rank algorithm. However, we compute multiple importance scores for each page; we compute a set of scores of the importance of a page with respect to various topics. At query time, these importance scores are combined based on the topics of the query to form a composite Page Rank score for those pages matching the query. This score can be used in conjunction with other IR-based scoring schemes to produce a final rank for the result pages with respect to the query.





Query processing of Topic- Sensitive page Rank algorithm

In addition we use EduNav novel search interface that enables the user to navigate large number of query results by organizing them using the LCSH concept hierarchy. First, the query results are organized into a navigation tree. At each node expansion step, EduNav reveals only a small subset of the concept nodes, selected such that the expected user navigation cost is minimized. In contrast, previous works expand the hierarchy in a predefined static manner, without navigation cost modelling. We show that the problem of selecting the best concepts to reveal at each node expansion is NP-complete and propose an efficient heuristic as well as a feasible optimal algorithm for relatively small trees. We show experimentally that EduNav outperforms state-of-the-art categorization systems with respect to the user navigation cost.

## V. OBJECTIVES

- A more accurate topic-sensitive Page Rank algorithm which computes the score of the pages according to the importance of the content available on the particular web page.
- Topic-Sensitive PageRank is based on the [PageRank](#) algorithm, and provides a scalable approach for personalizing search rankings using [Link analysis](#).
- PageRank is a link analysis algorithm that assigns a numerical weight to each object in the information network, with the purpose of “measuring” its relative importance within the object set.
- A biased PageRank score vector is computed for each predefined topic offline; and the probabilities that a query belongs to each topic are determined online, and the final query-dependent ranking is a weighted combination of the rankings for each topic.
- A comprehensive framework for navigating large query results from **EBSCO** using LCSH, an extensive concept hierarchy used for indexing citations to provide an easy to use and understand interface for user to search.
- A formal cost model for measuring the navigation cost incurred by the user.
- An efficient heuristic and a feasible optimal algorithm for minimizing the navigation cost. Experimental results validating the effectiveness of the EduNav system when compared to state-of-the-art categorization systems.

## VI. CONCLUSION

In this paper we explore topic sensitive page rank algorithm for effective navigation of query results. This will maximize structural coverage of sub topics in result hierarchy. At the same time it also focuses on reducing redundancy of the sub topics. Thus the proposed system achieves highly effective navigation of results. The results of this paper are compared with the results of an existing system meant for effective navigation of query results based on concept hierarchies. The results are obtained from EDS database which is one of the educational databases available. The existing system focused on the navigational cost. It is aimed at reducing navigation cost while the proposed system is aimed at maximizing structural coverage and minimizing redundancy in query results.

## REFERENCES

- [1] Abhijith Kashyap, Vagelis Hristidis, Michalis Petropoulos, and Sotiria Tavoulari “Effective Navigation of Query Results Based on Concept Hierarchies” Proc IEEE transactions, 2011
- [2] U.Sirisha\*, L.Lakshmi, P.Deepthi “Query Results Optimization Using Ontology and Result Diversification” Proc IJARCSSE, 2013.
- [3] EBSCO information service, <http://www.ebsco.com/>
- [4] TaherH.Haveliwala “Topic-Sensitive PageRank” Proc ACM 2012
- [5] K. Chakrabarti, S. Chaudhuri and S.W. Hwang: Automatic Categorization of Query Results. SIGMOD conference 2004: 755-766.
- [6] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage -based search result diversification. Journal of Information Retrieval, 2011
- [7] Web structure mining- data mining <http://www.web-datamining.net/structure/>