Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Query Results Optimization Using Ontology and Result Diversification

U.Sirisha*, L.Lakshmi, P.Deepthi

Department of CSE & JNTUH

India

Abstract--Web databases when queried result in huge number of records when users of query need a portion of those results which are real interest to them. This problem can be solved using concept hierarchies. Knowledge representation in the form of concepts and the relationships among them (Ontology) allows effective navigation. Recently Kashyap et al. presented provisions for categorization and ranking in order to reduce the number of results of query and also ensure that the navigation is effective. User should not spend much time to view the actual subset of records he is interested in from the avalanche of records that have been retrieved. For experiments, PubMed database which is in the public domain is used. The PubMed data is medical in nature and organized as per the annotations provided that is instrumental in making concept hierarchies to represent the whole dataset of PubMed. In this paper we apply structural diversification to concept hierarchy which represents knowledge in more comprehensive way. The goal of this paper is to maximize query sub topics coverage and also reduce redundancy in the results presented. We build a prototype application that demonstrates the proof of concept. The empirical results reveal that the structural relationships between query subtopics are reflect performance of diversification.

Index Terms–Concept hierarchy, effective navigation, annotated data, and result diversification

I. INTRODUCTION

The amount of data provided over World Wide Web (WWW) is increasing rapidly every year. In the past decade in started growing drastically. Especially biomedical data and the literature pertaining to it that reviews the aspects of biomedical data across the globe have seen tremendous growth in terms of quantity. Biological data sources such as [1], [2], and [3] are growing in terms of lakhs of new citations every year. The queries made by people associated with healthcare domain have to search such databases by providing a search keyword. The results are very huge in number and the users are not able to view all the records when they actually need a subset of them. This has led to users to refine query with other keywords and get the desired results after many trials. Here it has to be observed that user time is wasted in refining search criteria and also the navigation of query results which are abundant and bulky. The navigation cost is more as user has to spend lot of time in finding the required subset of rows from the bulk of search results. This problem has been researched in [1], [2], [3] and the problem is identified as information overload. Figure 1 shows an example for concept hierarchy.

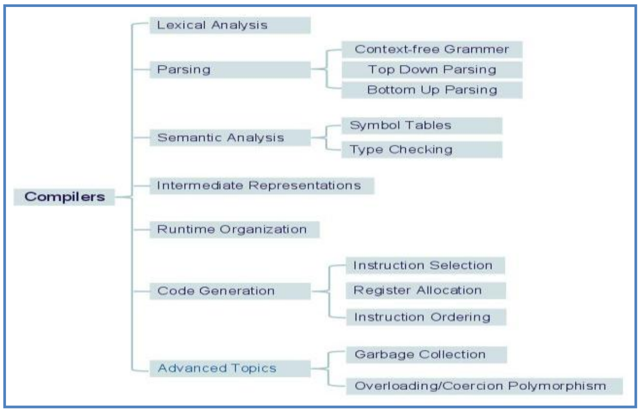


Fig. 1: Anexampleof concept hierarchy

The solutions are of two types namely categorization and ranking. However, these two can be combined to have more desired results. The proposed system is specially meant for presenting results in such a way that the navigation cost is reduced. For this purpose categorization techniques is used and concept hierarchies are built. The categorization techniques are supported by simple ranking techniques. The proposed solution uses citations as described in [4] and effectively constructs a navigation tree that can reduce cost of navigation and user's experience is much better when compared with existing systems that do not use these techniques. These techniques are being used by e-Commerce systems to let their users have smooth navigation to the results returned by such systems.

In this paper we proposed novel framework known as “Structural Diversification Framework” for redundancy of results and maximizing structural coverage of query sub topics based on concept hierarchies. The sample concept hierarchy is as shown in figure 1. First of all we determine the selection of nodes to be treated as sub topics for a query. Then based on the structural relationships among the sub topics, they are effectively organized. The structural similarity process is included as part of structural diversification for best results. The remainder of the paper is structured as follows. Section II reviewed literature. Section III described existing system. Section IV focused on the proposed system. Section V presented experimental results while section VI concludes the paper.

II. RELATED WORK

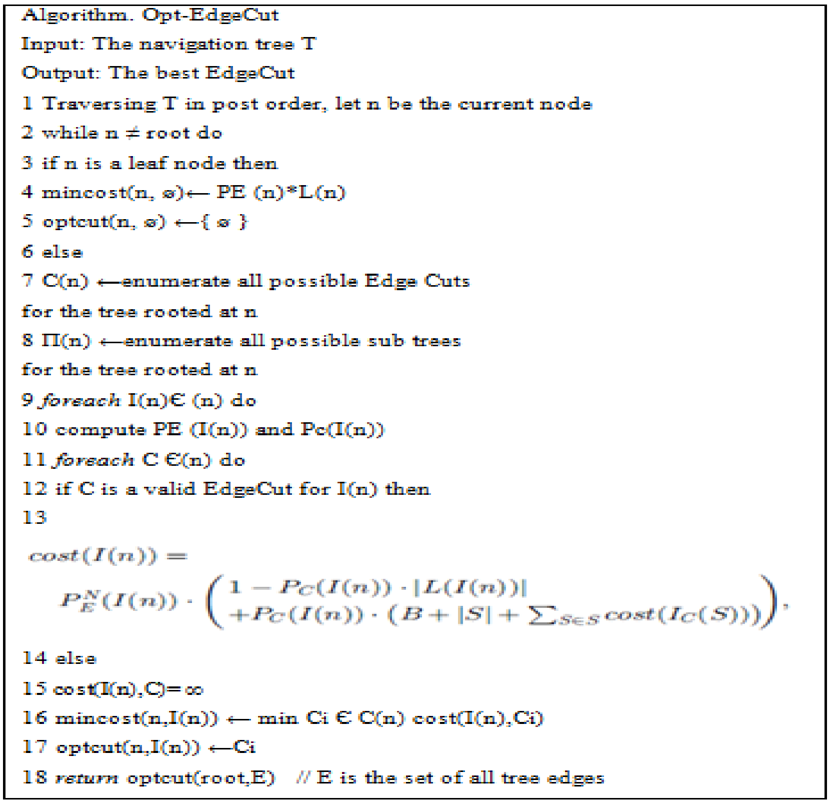
The aim of diversification is to given ranking to documents by using diversity and relevance. It has been proved in the literature that diversification based on subtopics has higher utility in real world applications. Its performance is better than other methods explored in [5], and [6]. These approach first identity sub topics prior to using various strategies to diversify the results based on the sub topics and the structural relationships among them [7], [5], and [8]. Many existing diversification methods make an assumption with respect to query sub topics. The assumption is that query subtopics are independent. However, it proved to be incorrect as there are relationships between them. The human assessors using TREC diversity collections [9], [10], [11] and query suggestion methods have evaluated this fact. This paper throws light further into re-examine the relationships between the subtopics of queries as part of the structural diversification process. The proposed work in this paper is similar to the structural relationships concepts proposed in [12], [13], and [14]. Query expansion applications also can use this kind of approach for finding semantically similar terms. More in-depth analysis is required in order to exploit the relationships among sub topics.

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. The following sub sections cover existing algorithms.

Best Edgecut Algorithms

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.



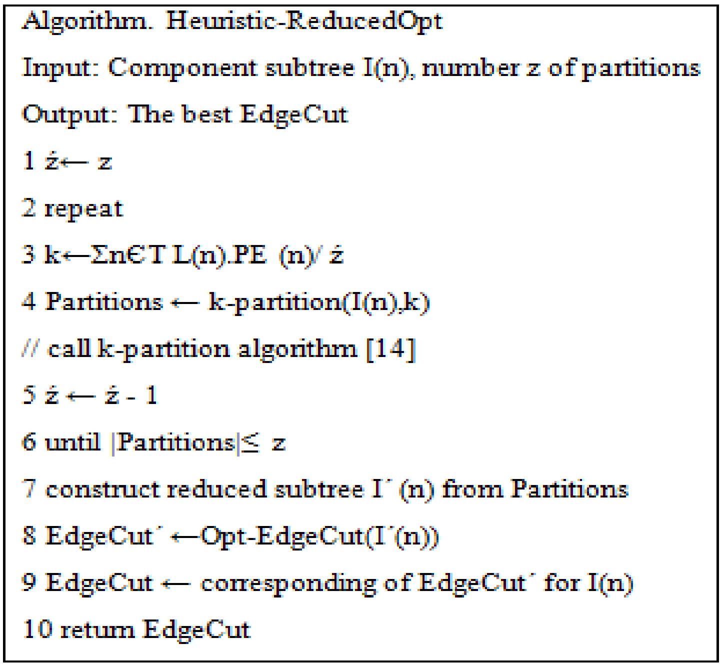
```
Algorithm. Opt-EdgeCut
Input: The navigation tree T
Output: The best EdgeCut
Traversing T in post order, let n be the current node
while n ≠ root do
    if n is a leaf node then
        mincost(n, ø)← PE (n)*L(n)
        optcut(n, ø) ←{ ø }
    else
        C(n) ←enumerate all possible Edge Cuts
        for the tree rooted at n
        Π(n) ←enumerate all possible sub trees
        for the tree rooted at n
        foreach I(n)Є (n) do
            compute PE (I(n)) and Pc(I(n))
            foreach C Є(n) do
                if C is a valid EdgeCut for I(n) then
                    $cost(I(n)) = P_E^N(I(n)) \cdot \begin{pmatrix} 1 - P_C(I(n)) \cdot |L(I(n))| \\ +P_C(I(n)) \cdot (B + |S| + \sum_{S \in S} cost(I_C(S))) \end{pmatrix},$
                else
                    cost(I(n),C)=∞
            mincost(n,I(n)) ← min Ci Є C(n) cost(I(n),Ci)
            optcut(n,I(n)) ←Ci
return optcut(root,E)   // E is the set of all tree edges
```

Listing 2 – Opt-EdgeCut Algorithm

The algorithm Opt-EdgeCut is supposed to compute the minimum expected navigation cost required for navigating the tree from bottom up in post order fashion.

Heuristic-ReducedOpt Algorithm

The Opt-EdgeCut algorithm is computational more expensive and can’t be practically used for most of the queries. Therefore, we proposed a new algorithm known as Heuristic-ReducedOpt as shown in listing 3.



```
Algorithm. Heuristic-ReducedOpt
Input: Component subtree I(n), number z of partitions
Output: The best EdgeCut
$\dot{z} \leftarrow z$
repeat
    $k \leftarrow \Sigma n \in T\ L(n).PE\ (n) / \dot{z}$
    Partitions ← k-partition(I(n),k)
    // call k-partition algorithm [14]
    $\dot{z} \leftarrow \dot{z} - 1$
until |Partitions| ≤ z
construct reduced subtree I´ (n) from Partitions
EdgeCut´ ←Opt-EdgeCut(I´(n))
EdgeCut ← corresponding of EdgeCut´ for I(n)
return EdgeCut
```

Listing 3 – Heuristic ReducedOpt Algorithm

This algorithm is based on k-partition algorithm [15]. This is adapted for our use here. The algorithm works in bottom-up fashion. For every node (n), the algorithm prunes heaviest children one by one till the weight of ***n***falls below k.

IV. PROPOSED STRUCTURAL DIVERSIFICATION

Diversification is meant for maximizing the sub topic coverage and minimizing redundancy in search results. Prior studies have shown that the structural diversification techniques outperformed other techniques [5], [6].

Concept Hierarchy based Subtopic Identification

Top ranked documents of a query are used to find sub topics of a concept hierarchy. Every top ranked document is assigned to similar node. The similarity between the documents is computed as follows.

$$sim(d,n) = \beta \cdot R(d,n) + (1-\beta) \cdot \frac{\sum_{n_j \in desc(n)} R(d,n_j)}{|desc(n)|}$$

Where number of descendents is represented as |desc(n)|. Relevance score between d and n is represented as R (d, NJ). With identified similarity among documents it is possible to find structural similarity between them.

Concept Hierarchy based Diversification

Having identified query sub topics not is it possible to find concept hierarchy based diversification. Given previously obtained documents D and query q, the ranking score is computed as follows.

$$\begin{aligned} Score(q,d,D) \quad &= \quad (1-\lambda) \cdot \sum_{s \in S(q)} [P(s|q) \cdot P(d|s) \\ &\cdot \prod_{d' \in D} (1 - SubCov(d',s))] + \lambda \cdot P(d|q) \end{aligned}$$

Where a set of sub topics for a query is represented by S(q). The relevance score of d with respect to given query is represented as P(d|q). The structural similarity for given two sub topics si andsj can be computed as follows.

$$\varphi(s_j|s_i) = \alpha \cdot f(|UP(s_i \rightarrow s_j)|) + (1-\alpha) \cdot f(|DOWN(s_i \rightarrow s_j)|)$$

Where number of elements in set X is represented by |X|. The number of down segments is represented by |DOWN(si→sj)|. The evaluation of this diversification is subjected to three constraints. The first constraint is that similarity of sub topic to itself is considered as smaller than other topics. The second constraint is that a subtopic ancestor's structural similarity should not be smaller than the ancestor's ancestor sub topic. The third constraint is that a subtopic's structural similarity should not be smaller than that of subtopic's ancestor subtopic. More details can be found in [16].

V. EXPERIMENTAL EVALUATION

For evaluating the proposed application, expansion time performance and average navigation cost are considered. The empirical studies are made in a PC with XP as operating system. Oracle 10 g is used as backend and Java is used to implement all algorithms.

Results of Existing System

The existing system focused on reducing navigational cost. The results are presented in figure 2.

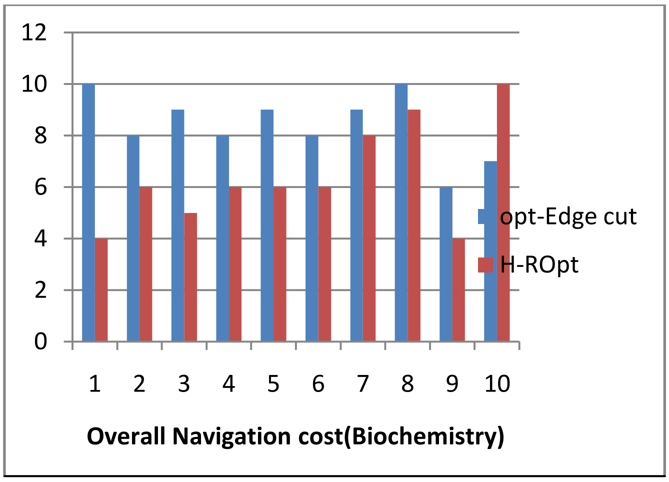


Fig. 2 – Comparison of number of concepts revealed

Fig. 2 shows the number of concepts shown when an EXAPND action takes place. The results revealed that our approach is superior to many other approaches.

Results of Proposed System

The proposed system exploits concept hierarchy for result diversification. It does mean that the results are improved in order to maximize structural coverage of sub topics while reducing the redundancy. The results of the proposed system are compared with the existing system in terms of structural coverage and redundancy of sub topics. Figure 8 visualizes the difference between the existing and proposed systems.

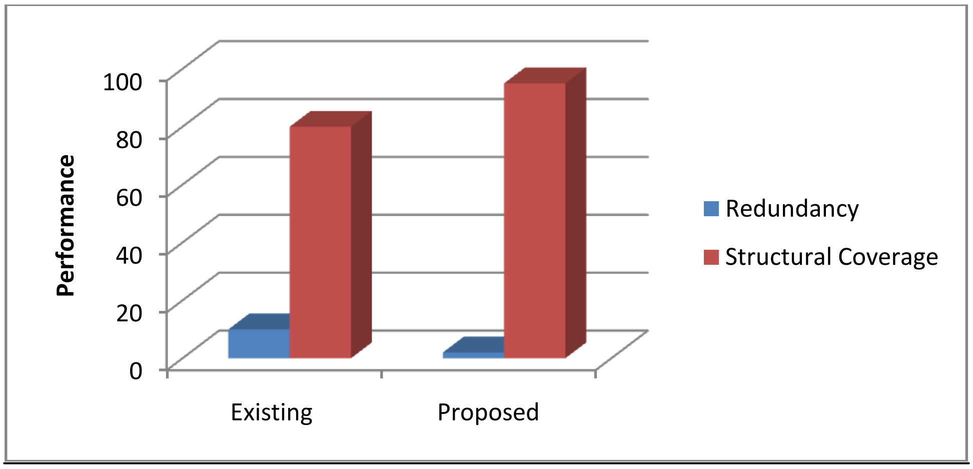


Fig. 3 –Performance Comparison

As seen in figure 3, it is evident that the existing system has redundancyin results and its structural coverage is also less when compared with the proposed system. The proposed system outperforms the existing system in terms of reducing redundancy and maximizing structural coverage which is the main aim of this paper.

VI. CONCLUSION

In this paper we explore concept hierarchy for result diversification. 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 PubMed database which is one of the biomedical 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. We built a prototype application which demonstrates the proof of concept. The empirical results revealed that the proposed approach is effective.

REFERENCES

[1] J S. Agrawal, S. Chaudhuri, G. Das and A. Gionis: *Automated Ranking of Database Query Results*. In Proceedings of First BiennialConference on Innovative Data Systems Research (CIDR),2003.

[2] K. Chakrabarti, S. Chaudhuri and S.W. Hwang: *Automatic Categorization of Query Results*. SIGMOD Conference 2004: 755-766.

[3] Z. Chen and T. Li: *Addressing Diverse User Preferences in SQLQuery- Result Navigation*. SIGMOD Conference 2007: 641-652.

[4] Medical Subject Headings (MeSH®). http://www.nlm.nih.gov/mesh/

[5] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of WWW'10*, pages 881–890, New York, NY, 2010. ACM.

[6] W. Zheng and H. Fang. A comparative study of search result diversification methods. In *Proceedings of DDR'11*, 2011.

[7] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proceedings of WSDM'09*, pages 5–14, New York, NY, 2009. ACM.

[8] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage-based search result diversification. *Journal ofInformation Retrieval*, 2011.

[9] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *Proceedings of TREC'09*, 2009.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. In *Proceedings of TREC'10*, 2009.

[11] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 Web Track. In *Proceedings of TREC'11*, 2011.

[12] M. Lalmas. *XML Retrieval (Synthesis Lectures on Information Concepts, Retrieval, and Services)*. Morganand Claypool, San Rafael, CA, 2009.

[13] H. Fang. A Re-examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL'08*, New York,NY, 2008. ACM.

[14] R. Thiagarajan, G. Manjunath, and M. Stumptner. Computing semantic similarity using ontologies. In *HPLabsTech Report*, 2008.

[15] (2008) Vivísimo, Inc. –Clusty. [Online].Available: http://clusty.com/

[16] Wei Zheng, Hui Fang, and Conglei Yao, “Exploiting Concept Hierarchy for Result Diversification”, CIKM 2012.