ISSN (Online) : 2319 - 8753
ISSN (Print)   : 2347 - 6710

IJIRSET

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Diversified Ranking Algorithm Based On Fuzzy Concept For Large Graphs

R.Kirubahari, T.R.Vedhavathy, S.Prabavathy

Dept. of CSE, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Dept. of CSE, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Dept. of IT, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Abstract— The main goal of ranking web pages is to find the interrelated pages. This interrelation of pages is represented as weighted graph. Ranking nodes on graphs is a fundamental task in information retrieval, data mining, and social network analysis. Many existing diversified ranking algorithms either cannot be scalable to large number of nodes due to time and memory requirements or reasonable diversified ranking measure. Graph-based algorithms are stationary distribution of the random walk on the graph. Diversified ranking measure was proposed for large graphs to identify the relevance and diversity among nodes. This measure is a nondecreasing submodular set maximization problem. To solve this problem an efficient greedy algorithm was developed with linear time and memory requirements, and the size of the graph to achieve near-optimal solution. This paper presents a system that provides users with personalized results derived from a search engine that uses link structures. The proposed system uses fuzzy document retrieval (constructed from a fuzzy concept network based on the user's profile) that personalizes the results returned from link-based search engines with the preferences of the specific users. Experimental results show that the developed method is scalable for large graph with linear time and space complexity and it has considerable efficiency in determining the rank of web pages.

Keywords— Diversified ranking, graph algoritms, Submoduler function, Fuzzy document retrivel..

I. INTRODUCTION

Ranking nodes on a graph is a primary task in information retrieval, mining task. Webpage [1] ranking in social networks provides personalized services for web search [3].Many existing graph-based ranking algorithms such as PageRank algorithm [3], [4] is a random walk on graphs. In this approach the node of a graph is ranked higher by considering the more links between the nodes. The random walk method is used to identify the top-K ranking list contain many similar nodes by considering only the relevance of the nodes. It reduces the ranking effectiveness and diversity is not achieved. To overcome THIS problem while designing the ranking algorithms, allowing the diversity in the top-K ranking list. We propose a novel diversified ranking approach, using query node the personalized PageRank is calculated and diversity is identified. The diversity ranking measure is used to identify the nodes that are not only relevant to the query but also dissimilar nodes. The diversified ranking measure is optimized by graph expansion, where the nodes will have a large expansion then the nodes will be dissimilar among them and the diversity is obtained. Thus the proposed novel diversified ranking measure is a nondecreasing submodular set function. By using the submodularity of this method an efficient greedy algorithm is developed with linear time and space complexity w.r.t the size of the graph. To optimize the linear time and space complexity a randomized greedy algorithm is developed by using the generalized diversified ranking measure by taking the K-step expansion.

II. RELATED WORK

A. *Incremental diversification for very large sets: a streaming -based approach*

A given user query often has a variety of intended meanings or associated information needs. Most existing approaches use approximation techniques and heuristics to increase the efficiency of diversification but cannot be applied for diversification of result streams. Diversification of results is impossible for systems that potentially process thousands of queries in parallel when they use algorithms with super-linear runtime and linear memory requirements. To solve problem using a streaming-based approach which processes items incrementally, maintaining a near-optimal diverse set at any point in time. This approach has linear computation and constant memory complexity with respect to input set size, and naturally fits for streaming applications. New incremental approach is linear in computational complexity and constant in memory complexity w.r.t the number of items, and therefore also very well suited for very large sets. In addition, our streaming-based computation enables very efficient diversification of queries over continuous data sources such as news streams or tweets, opening up further applications of diversification.

B. *Diversity in ranking via resistive graph center*

In this approach the formulation of diversity as finding centers in resistive graphs. Unlike in PageRank, we do not specify the edge resistances and ask for node visit rates. Instead of that identify the sparse set of center nodes so that the effective conductance from the center to the rest of the graph has maximum entropy. In marked deviation from prior work, our edge resistances are learnt from training data. Inference and learning are NP-hard, but we give practical solutions. The Markov walk approach can potentially get around this problem by directly comparing items without reference to the units of coverage. However, existing work hardwires edge weights rather than learn them. Markov walks express associativity a node i with large score linking to node j pulls up the score of j . However, diversity demands dissociative decisions: if i and j are both relevant and very similar, we should pick only one of them.

C. *Diversified ranking on large graph :An optimation view point*

A good top-k ranking list should identify both the relevance and the diversity. But this method does not exist, such a goodness measure for the graph data in the literature. Most of the existing works for diversified ranking on graphs are based on some heuristics. To find an optimal, or near-optimal, top-k ranking list that maximizes the goodness measure. Bringing diversity into the design objective implies that we need to optimize on the set level. It is usually very hard to perform such set-level optimization. we address these challenges from an optimization point of view. We propose a goodness measure which intuitively captures both (a) the relevance between each individual node in the ranking list and the query node; and (b) the diversity among different nodes in the ranking list. We further propose a scalable algorithm that generates a provably near-optimal top-k ranking list. To the best of our knowledge, this is the first work for diversified ranking on large graphs that has a clear optimization formulation and finds a provably near optimal solution, and enjoys the linearly scalability.

D. *DivRank: the interplay of prestige and diversity in information networks*

Many retrieval and mining tasks are concerned with finding the most important and/or relevant items from a large collection of data. These measures are known as centrality (or prestige) measures in general, with various instantiations like degree, closeness, between nodes, and more complicated measures such as the PageRank score and the authority score. These measures can be also combined with other features such as the relevance to a query. The diversity in top ranked results has been recognized as another crucial criterion in ranking. The need of diversity in ranking is even more urgent when the space of output is limited. The proposed system has a novel ranking algorithm, DivRank, based on a reinforced random walk in an information network. This model automatically balances the prestige and the diversity of the top ranked vertices in a principled way. DivRank not only has a clear optimization explanation, but also well connects to classical models in mathematics and network science.

III. DIVERSIFIED RANKING METHODOLOGY

A. *Personalized PageRank Algorithm*

Personalized PageRank [1] is a query dependent ranking algorithm. Given a query vector *r*, a graph G the personalized PageRank vector ω is calculated for each iteration using an equation:

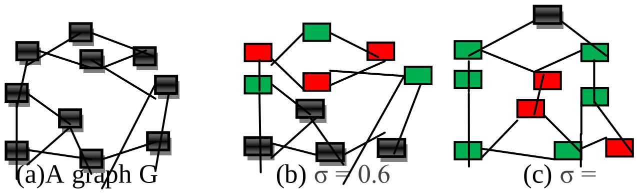
$$\omega = (1-\alpha)r + \alpha A \qquad (1)$$

Where α is a damping factor and A is the adjacency matrix of graph G. The PageRank vector ω is used to rank the nodes of the graph.

Stationary distribution of random walks is used by personalized PageRank for ranking the nodes in a graph instead of diversity. Markov chain [11] is formed by the random walk on graph, If a node gets hit frequently by random walks then that node is considered as high personalized PageRank score, thus the neighbors of that node also gets hit frequently and taken as high personalized PageRank score where the top-K ranking list contains many similar nodes and the personalized PageRank algorithm results in reducing the effectiveness of ranking among the nodes, so diversity must be provided for the application.

B. *Diversified Ranking Problem*

To improve diversity there are many ranking algorithms on graphs [5], [6], [7] but the algorithms cannot scale to large scale graphs or provide an effective ranking measure.



(a)A graph G (b) σ = 0.6 (c) σ = 0.9

Fig 1. Illustration of our idea: expansion ratio versus diversity. Red square nodes denote the selected nodes and green nodes are expanded nodes.

We propose a scalable diversified ranking algorithm to optimize the ranking process accurately. The diversified ranking problem is formulated as:

Notations and definition. Consider a graph G = (V , E), with a set of nodes V and a set of edges E, where the size of nodes is n = |V|.

Definition: Let S be a set of nodes. The expanded set of S is denoted by

$N(S) = S \cup \{v \in (V - S) | \exists\, u \in S, (u, v) \in$

The expansion is of a set of nodes, S, is the size of expanded set, N(S), denoted as |N(S)|. And the expansion ratio is defined as $\sigma = |N(S)$.

Expansion of a graph is based upon the directed and undirected topological structure. Expansion of expander graph is different from the expansion which is equal to the minimum expansion ratio among all expanded sets. The nodes are dissimilar to each other and they do not have any common neighbors in a graph since they have a large expansion ratio, so it is well known that better diversity is achieved in the set of nodes with large expansion ratio. Consider we take three nodes from Fig 1(b) and 1(c). In fig 1(b) the nodes are well connected and they have similar nodes while in Fig1(c) the nodes have no edge connected between them so they have dissimilar nodes, from the above figure it is clear that Fig1(c) is more diverse than Fig1(b) since Fig 1(c) has large expansion ratio.

In a graph a node must have high personalized PageRank and be dissimilar to the selected nodes. Diversified ranking measure is achieved by combining relevance and diversity which is called as maximum marginal relevance (MMR). A document which is used by MMR is relevant to the query and dissimilar in other words the document has high marginal relevance. To find a subset S of K nodes such as the nodes S with high personalized PageRank scores and expansion ratio of |N/S|/n is maximum. Our aim is to maximize the diversified ranking measure which is formulated below:

$$F(S) = (1-\lambda)\sum_{u \in S} w_{u} + \lambda \frac{|N(S)|}{n} \qquad (2)$$

w denotes the personalized PageRank scores of node u, and λ ∈ [0,1] is a parameter used to trade off the relevance and diversity. In (2) the first term is the personalized PageRank scores which is based on relevance and second term is the expansion ratio of ranking results. It is known that the system will focus on top-K ranking list but F(S) does not produce the accurate ordering of top-K ranks so we propose an algorithm for ordering the ranking list based upon combining the relevance and diversity scores.

C. Diversified Ranking Algorithm

Many algorithms are developed to overcome the diversified ranking problem such as NP hard. From the definition given below it is clear that the diversified ranking measure (F(S)) is a nondecreasing submodular set function. Using (F(S)) a near-optimal greedy algorithm is developed to overcome the ranking problem efficiently.

Definition: Let V be a finite set, a real valued function f(S) on the set of subsets of V, S, is called a nondecreasing submodular set function, if the following conditions hold.

- Nondecreasing. For any subsets S and T of V such that S⊆T⊆V, we have f(S) ≤ F(T).
- Submodularity. Let $\rho_j(S) = (S \cup \{j\}) - f(S)$ be marginal gain. Then, for any subsets s and t of v such that S⊆T⊆V and j ∈ V\T, we have $\rho_j(S) \geq \rho_j(T)$.

Greedy Algorithm:

```
Input: Graph G = (V, E), K, damping factor α,
        Adjacency matrix A, teleport vector r
        and parameter λ.
Output: A set S with K nodes.
Compute the personalized PageRank vector ω;
Initialize the answer set S ← Ø;
for each node $V_i$, initialize an indicator array Expan [i]←0;
for iter = 1 to K do
    max ← -1;
    maxIdx ← 0;
    for each node $V_i \in (V - S)$ do
        counter ← 0;
        for each neighbor node ($V_j$) of $V_i$ do
            if Expan[j] = 0 then
                counter – counter +1;
        if $(1-\lambda)\omega_i + \lambda \times counter/|v|) > max$ then
            max ← $(1-\lambda)\omega_i + \lambda \times counter/|v|$;
            maxIdx← I;
    S ← S ∪ {$v_{maxIdx}$};
    for each neighbor node ($V_j$) of $v_{maxIdx}$ do
        Expan[j] ← 1;
Return S;
```

In this greedy algorithm, First the personalized PageRank vector is identified by measuring the relevance of nodes. Then for each iteration a node with maximum mariginal gain is chosen and added to the answer set S. The iteration is repeated for K times and as a result ordering ranking list is produced based on $\rho_u(S)$ which satisfies the nondecreasing properties and node with top-K ranking score appers in top-K ranking list.

Complexity analysis of the greedy algorithm. The time complexity of algorithm is O(K|E|). Since the algorithm must visit all the nodes and its neighbors in the system which takes 2|E| time in worse case, so CELF framework is used to speed up the process and for space complexity is O(|V|+|E|) all the parameters such as input, output, personalized PageRank, answer set and array are put together and it is scalable for large scale graphs.

D. Generalized diversified ranking

F(S) considers only immediate nodes in set of nodes S. To generalize this diversified ranking measure F(S) by taking

k-step nearest nodes into rank, its denoted by $F_k(S)$ using k-step expanded set.

$$F_k(S) = (1-\lambda)\sum_{u\in S} w_{u} + \lambda \frac{|N_k(S)|}{n} \qquad (3)$$

Ranking algorithm using fuzzy concept:

We propose a system that searches web documents based on link information and fuzzy concept network. This method calculates the relevance of nodes using fuzzy logic with user profile. The search engine is able to select fitting websites for the user's query by processing fuzzy document retrieval system using the fuzzy concept network to represent the user's knowledge. The motivation of this research is to find a user profile to link based search engines. The fuzzy concept network provides the inference mechanism to calculate undefined relationships between concepts. Undefined information can be calculated using a transitive closure of the fuzzy concept matrix. This property can minimize the user's cognitive load to insert the relevance of concepts. This approach suffers from the inconvenience of obtaining information from the user, it can be improved by the use of a fuzzy concept network and various ranking algorithm were used such as,

- Personalized PageRank (PPR) algorithm which is used to evaluate the relevance of the nodes.
- Grasshopper (Gra) algorithm is used to achieve the diversity.
- Manifold ranking with stop points (MRSP) is similar to Grasshopper algorithm used in graphs.
- DivRank (Div R), this algorithm is used to achieve similar ranking performance using two various implementation such as pointwise DivR and cumulative DivR.
- Dragon (Dra) algorithm is used to optimize the diversified ranking measure but lacks topological explanation.
- Diversified ranking via resistive graph centers (RGC) algorithm tries to achieve diversity in ranking but cannot be scaled in large graphs.

Ranking algorithm using fuzzy concept provides high quality results and retrieval of document is faster while searching process of related documents are efficient. Users will be provided with accurate information by using fuzzy concept network.

IV. EXPERIMENTAL RESULTS

The personalized search engine selected the five most authoritative results as a source of personalization and produced a document descriptor of these documents. The ranking of these five documents was reordered with respect to the user's interest recorded on a user profile. A fuzzy concept network for a user was generated based on 20 degrees of relevance in the user profile, and unrecorded information was inferred from the transitive closure of the fuzzy concept network. The expanded document descriptor resulted from the multiplication of the document descriptor and the user's fuzzy concept network. The sum of the degree of relevance with respect to the concepts determined the new ranking of the documents as the final result.

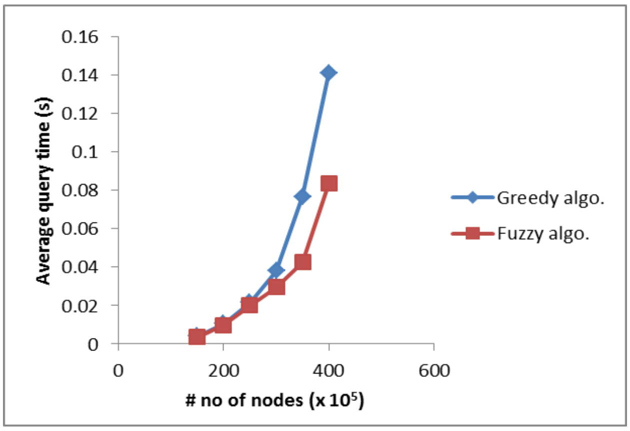


Fig.1 comparison of greedy algorithm with proposed algorithm

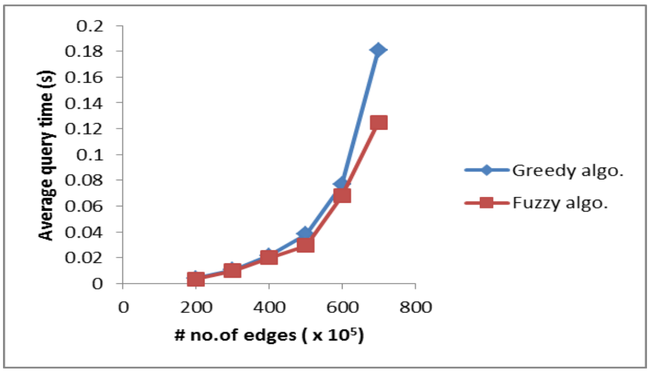


Fig.2 comparison of greedy algorithm with proposed algorithm

TABLE. 1 : SEARCH RESULTS FOR "JAVA"

Rank	*Authoritative result*	*Hub result*
1	www.java.sun.com	www.industry.java.sun.com/products
2	www.javalobby.org	www.java.sun.com/industry
3	www.javaboutique.internet.com	www.java.sun.com/casestudies
4	www.java.about.com/compute/java/mbody.htm	www.industry.java.sun.com/javanews/developer
5	www.javaworld.com	www.industry.java.sun.com/jug

V.CONCLUSION

To finding top-K diversified ranking on graphs using a novel diversified ranking measure, which captures both relevance and diversity. We prove the sub modularity of this measure and design an efficient greedy algorithm to achieve near-optimal diversified ranking. The proposed method has linear time and space complexity w.r.t. the size of the graph, thus it can be scalable to large graphs. We present a generalized diversified ranking measure and the proposed system uses fuzzy document retrieval (constructed from a fuzzy concept network based on the user's profile) that personalizes the results returned from link-based search engines with the preferences of the specific users. The developed method is scalable for large graph with linear time and space complexity and it has considerable efficiency in determining the rank of web pages.

VI.FUTURE ENHANCEMENT

Ranking nodes on graphs is a fundamental task in information retrieval and data mining task. . Diversified ranking measure was proposed for large graphs to identify the relevance and diversity among nodes with linear time and space complexity. To find relevant web documents for a given user, the proposed search engine uses link structures and a fuzzy concept network. For efficient searching, these link structures are stored in advance. The fuzzy document retrieval system personalizes the link-based search results with respect to the user's interests and reorders the ranking results. Future work will proceed as follows. Using the user's feedback about the search results, it is possible to change the value of the fuzzy concept network. This adaptation procedure helps to obtain better results. These preliminary results indicate that a soft computing method such as fuzzy logic can play a crucial role in information retrieval from the web, which provides an important platform for personalization of search engines.

REFERENCES

[1] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon, "Novelty and Diversity in Information Retrieval Evaluation", Proceedings of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08), pp. 659 – 666, 2008.

[2] A.Dubey, S.Chakrabarti, and C.Bhattacharyya, "Diversity in Ranking via Resistive Graph Centers" Proceedings of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 78-86, 2011.

[3] M.Drosou and E.Pitoura, "Search Result Diversification", ACM SIGMOD Record, vol. 39, pp. 41-47, 2010.

[4] S.Gollapudi and A.Sharma, "An Axiomatic Approach for Result Diversification" Proceedings of 18th International Conference on World Wide Web (WWW '09), pp. 381 – 390, 2009.

[5] R.H.Li and J.X. Yu, "Scalable Diversified Ranking on Large Graphs", Proceedings of IEEE 11th International Conference on Data Mining (ICDM), pp. 1152-1157, 2011.

[6] Q.Mei, J.Guo, D.R.Radev, "DivRank: The Interplay of Prestige and Diversity in Information Networks", Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10), pp. 1009 – 1018, 2010.

[7] E. Minack, W. Siberski, and W. Nejdl, "Incremental Diversification for Very Large Sets: A Streaming-Based Approach", Proceedings 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 585-594, 2011.

[8] H.Tong, J.He, Z.Wen, R.Konuru, C.Y.Lin, "Diversified Ranking on Large Graphs: An Optimization Viewpoint", Proceedings of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1028 – 1036, 2011.

[9] X.Zhu, A.B.Goldberg, J.V.Gael, D.Andrzejewski, "Improving Diversity in Ranking Using Absorbing Random Walks" Proceedings of Human Language Technologies: The Annual Conference, North America,
pp. 97 – 104, 2007

[10] X.Zhu, J.Guo, X.Cheng, P.Du, and H.Shen, "A Unified Framework for Recommending Diverse and Relevant Queries" Proceedings of 20th International Conference on World Wide Web (WWW '11), pp. 37 – 46, 2011

[11] S.Brain and L.Page,"PageRank: Bringing Order to the Web",technical report, stanford digital Library Project,1997.

[12] M.E.J. Newman, Networks: An introduction. Oxford Univ. Press, 2010.

[13] O.Haggstrom, Finite Markov chains and Algorithmic Applications. Cambridge Univ. Press,2002..

[14] H.Ma,M.R. Lyu, and I.King, "Diversifying Query Suggestion Results', Proceedings International conference in artifical intelligence,2010.

[15] P.Flajolet and G.N Martin,"Probilistic Counting Algorithms for Data Base Applications",J.Computer Sysyem Science, volume 31,no.2,PP.182-209,1985.