Study of Mrsp – A Comparison Study of Diversity, Relevance, Importance in Ranking

Neerukonda Venkatesh
PG scholar
Department of IT
Geeth anjali college of Engg. & Technology
Hyderabad, India

B.Srinivas
Associate Professor
Department of IT
Geethanjali college of Engg. & Technology
Hyderabad, India

Abstract: **- Information retrieval is an everyday task for different domain area users for specific task related to their area of interest for knowledge discovery (KDD), Decision makings etc. Data mining techniques plays a vital role for extracting information related to user search. Search Engines involves in retrieving of Users' queries by searching in multiple databases. As the users of the technology growing rapidly, the enormous collection of data in sort of data links & maintaining of that data by storing in databases became a big problem for companies to retrieve the Information Relevance to the given query, users are totally depending on the retrieval data. Here comes the actual problem in estimating the Relevance, Importance and divergence of the result obtained. Ranking is the primary concept in displaying any result on Web based Systems (WS), Automated Systems (AS), Information Retrieval Systems (IRS) by fulfilling the common properties. In this paper we discussed in ranking properties, divergence problems and studied an approach -Manifold ranking with sink points (MRSP) to address diversity as well as relevance and importance in ranking.**

Keywords: KDD, IRS, Ranking, MRSP.

I. INTRODUCTION

Ranking is typically used for displaying the Web-pages at client-side. Ranking has many applications in real world scenario in various fields like Data mining, Information Retrieval and Natural Language Processing. In order to retrieve a web page related to user query, which page has to be retrieved and displayed for user is the process at search engine done based on Relevance , Importance of given query. Redundancy, Diversity in results noticed as the basic problems of Ranking. Researchers from various domains have proposed many approaches to address this problem, such as Maximum Marginal Relevance (MMR) [1], subtopic diversity [2], [4], cluster-based centroids selecting [3], categorization- based approach [1], and many other redundancy penalty approaches [6], [5], [7]. However, these methods often treat relevance and diversity separately in the ranking algorithm, sometimes with additional heuristic procedures. In this paper, we studied a novel approach, named Manifold Ranking with Sink Points (MRSP), to address diversity as well as relevance and importance in a unified way. Specifically, our approach uses a manifold ranking process over data manifold, which can help find the most relevant and important data objects. Meanwhile, we introduce into the manifold sink points, which are objects whose ranking scores are fixed at the minimum score during the manifold ranking process. This way, the ranking scores of other objects close to the sink points (i.e., objects sharing similar information with the sink points) will be naturally penalized during the ranking process based on the intrinsic manifold. By turning ranked objects into sink points in the data manifold, we can effectively prevent redundant objects from receiving a high rank

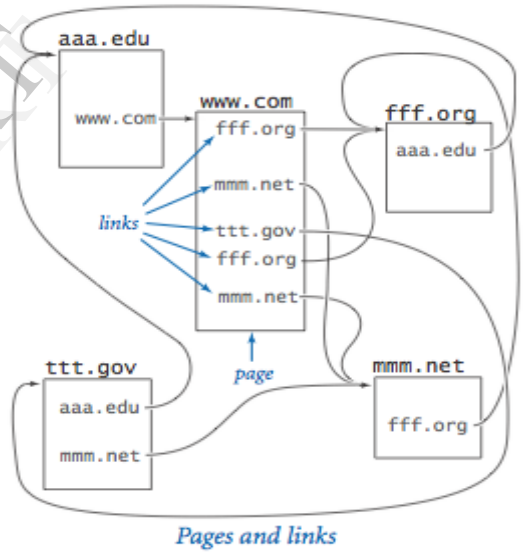

Fig 1: Indication of page Rankings flow.

II RANKING ON DATA MANIFOLDS

Manifold Ranking algorithm works based on two conditions:
1) Nearby data are likely to have close ranking scores
2) Data on the same structure are likely to have close ranking scores.

I) Description of Ranking Algorithm
A weighted network is constructed first, where nodes represent all the data and query points, and an edge is put between two nodes if they are "close." Query nodes are then initiated with a positive ranking score, while the nodes to be ranked are assigned with a zero initial score. All the nodes

then propagate their ranking scores to their neighbors via the weighted network. The propagation process is repeated until a global stable state is achieved, and all the nodes except the queries are ranked according to their final scores.

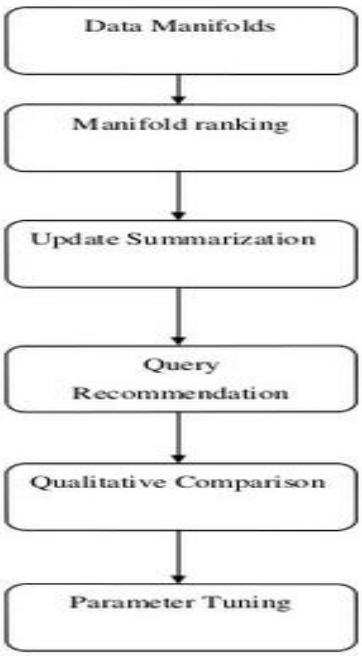

Fig: 2 Flow chart for Data manifolds

III.GOALS OF UPDATE SUMMARIZATION

Update summarization is a temporal extension of topic focused multi document summarization by focusing on summarizing up-to-date information contained in the new document set given a past document set.

Goals of update summarization.

Relevance: - The summary must stick to the topic users are interested in.

Importance: - The summary has to neglect trivial content and keep as much important information as possible.

Diversity: - The summary should contain less redundant information and cover as many aspects as possible about the topic.

Novelty: - The summary needs to focus on the new information conveyed by the later data set as compared with the earlier one under that topic.

1) Compute the similarity values $sim(x_i, x_j)$ of each pair of data objects x_i and x_j.
2) Connect any two objects with an edge if their similarity value exceeds 0. We define the affinity matrix W by $W_{ij} = sim(x_i, x_j)$ if there is an edge linking x_i and x_j. Let $W_{ii} = 0$ to avoid self-loops in the graph.
3) Symmetrically normalize W by $S = D^{-1/2} W D^{-1/2}$ in which D is the diagonal matrix with (i, i)-element equal to the sum of the i_{th} row of W.
4) Compute $\Omega = (I - \alpha S)^{-1}$, where $0 \leq \alpha < 1$.
5) Obtain the sub-matrices $\Omega_{11}, \Omega_{12}, \Omega_{21}, \Omega_{22}$ from Ω based on the free points and query points, and the corresponding trimmed vectors y_2.
6) Compute $f^* = \Omega_{22} y_2 - \Omega_{21} \Omega_{11}^{-1} (\Omega_{12} y_2)$.
7) Mark the object x_m with maximum score f_m^* as a new sink point.
8) If the pre-defined number of sink points K is not reached, go to step 5.
9) Return the sink points in the order that they get marked as sink points.

Fig:3 Steps Involved in MRPS Algorithm

IV. CONCLUSION

In this paper we studied different approaches for data manifold problems and approaches that helped to overcome the existing problems and focused on the working of MRSP algorithm which is one of the main approaches for fixing diversity problems. We studied how MRSP plays a major role in ranking pages.

REFERENCES

[1] H.T. Dang and K. Owczarzak, "Overview of the TAC 2009Summarization Track (Draft)," Proc. Second Text Analysis Conf. (TAC '09), 2009.

[2] J. Allan, R. Gupta, and V. Khandelwal, "Temporal Summaries of News Topics," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '01), pp. 10-18, 2001.

[3] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416, 2000.

[4] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna, "Query Suggestions Using Query-Flow Graphs," Proc. Workshop Web Search Click Data (WSCD '09), pp. 56-63, 2009.

[5] F. Boudin, M. El-Be`ze, and J.-M. Torres- Moreno, "A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization," Proc. Companion Vol.: Posters (Coling '08), pp. 23-26, Aug. 2008.

[6] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '98), pp. 335-336, 1998..

[7] C.L. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Bu¨ ttcher, and I. MacKinnon, "Novelty and Diversity in Information Retrieval Evaluation," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 659-666, 2008.