ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A Novel Approach to Ranking On Data Manifold With Sink Points

A. Mahesh. M.E. [1], G.R.Sumithra[2]

ABSTRACT: Ranking of documents has various applications in data mining, information retrieval and natural language processing. Many approaches are proposed to rank documents according to their measure of importance or relevance. In addition to importance and relevance of ranked documents, diversity is also recognized as an important criterion in ranking. Top ranked documents in traditional approaches contain redundant information, which is not desired by users. In order to address diversity, relevance, novelty and importance of ranked documents, a new novel approach named Manifold Ranking with Sink Points is proposed. The Manifold ranking process finds the most important and relevant data objects very efficiently. The ranked objects are converted to sink points in data manifold and the redundant objects are prevented from receiving a high rank. The Manifold Ranking with Sink Points approach has good convergence property and also satisfies optimization explanation. MRSP is applied on two applications: query recommendation and update summarization, where diversity is of important concern in ranking. Experimental results present that MRSP has strong empirical performance than traditional approaches like MMR and DivRank.

KEY WORDS—Diversity in Ranking, Manifold Ranking with Sink Points.

I. INTRODUCTION

Ranking is a very important problem in natural language processing, data mining and information retrieval. Generally the ranking problem is explained as follows: Given a set of data objects, the ranking function arranges the objects in the set according to their degrees of importance, relevance or preferences. But, the mass of ranked documents might contain duplicated and redundant information which is not of great use to users. The redundancy present in top ranked results reduces the chance to satisfy multi-faceted requirements of different users. So, diversity in ranked results is very important in addition to relevance and importance in ranking documents. Diversity in ranking of documents helps the users to handle user requests with explicit queries related to information present in documents stored on web. Diversity serves as a means to get the appropriate information very easily.

Many real-time applications demand diversity in the ranking approaches. In addition to relevance and importance, diversity is also considered as an important criterion in data mining. For instance, in query recommendation, the queries that are recommended by the ranking approaches must capture the different query requirements of different users (i.e.) the queries recommended must also be diverse. The concept of diversity when applied to text summarization, the summarized candidate sentences must be less redundant and must also contain different aspects of information relevant to the query. The concept of diversity is studied extensively in the recent years. Many approaches have been proposed to address the diversity issue by many domains such as cluster based centroids, maximum marginal relevance and subtopic diversity. But these approaches seldom consider relevance and diversity in a unified way.

The manifold ranking process makes use of manifold sink points. These sink points are the data objects that contain minimum and fixed ranking scores. The ranking scores of data objects closer to the sink points are naturally included in the ranking process which is usually based on intrinsic manifold. In this way, the diversity and importance as well as relevance are captured during the manifold ranking process.

II. RANKING ON DATA DOCUMENTS

There are many traditional ranking approaches available that ranks objects according to their degrees of importance or relevance. In addition to these criteria, diversity of information content is also an important criterion in ranking of documents. Furthermore, the user's needs might be multi-faceted or ambiguous. The redundancy in top ranked results will reduce the chance to satisfy different users. Thus, it is important to reduce redundancy in these top search results. Ranking of documents by servers is considered as an important criterion in applications like data mining, natural language processing, information retrieval and computational biology.

The ranking process gives high ranks to nodes close to the query and also to those sentences in documents that contain high centrality. Ranked documents using traditional approaches contain highly redundant and also duplicated information which is not desirable by users. Also, the needs of users in information retrieval are multi-faceted or ambiguous. So, the redundancy of top ranked results minimizes the chance to satisfy different user's needs.

For instance, the information that the user is expecting from searching experience may not be relevant to the query and it may contain repeatedly same information. So, users with multi-faceted queries are not contended with the searching using traditional approaches. The ranked data according to traditional approaches gives high ranks to documents that are visited the maximum number of times and do not consider the diversity of information present in it. But MRSP approach addresses the concepts of diversity and relevance in a unique way. The redundant objects are prevented from receiving a high rank. As a result, diversity, relevance and importance are captured during the process.

III. THE MRSP APPROACH

A new approach called Manifold ranking with sink points (MRSPs) is proposed, that addresses the aspects of diversification of information, relevance as well as importance in ranking. This approach satisfies optimization explanation on two application tasks: update summarization and query recommendation. Update summarization summarizes the up-to-date information contained in the new document set given a past document. Also, query recommendation is to provide alternative queries to help users search and improve the usability of search engines. Diversity is of great concern both in update summarization and query recommendation. The vast and diverse information present in ranked documents is grouped together and overall summarized information is presented to the users. This is achieved by turning ranked objects into sink points.

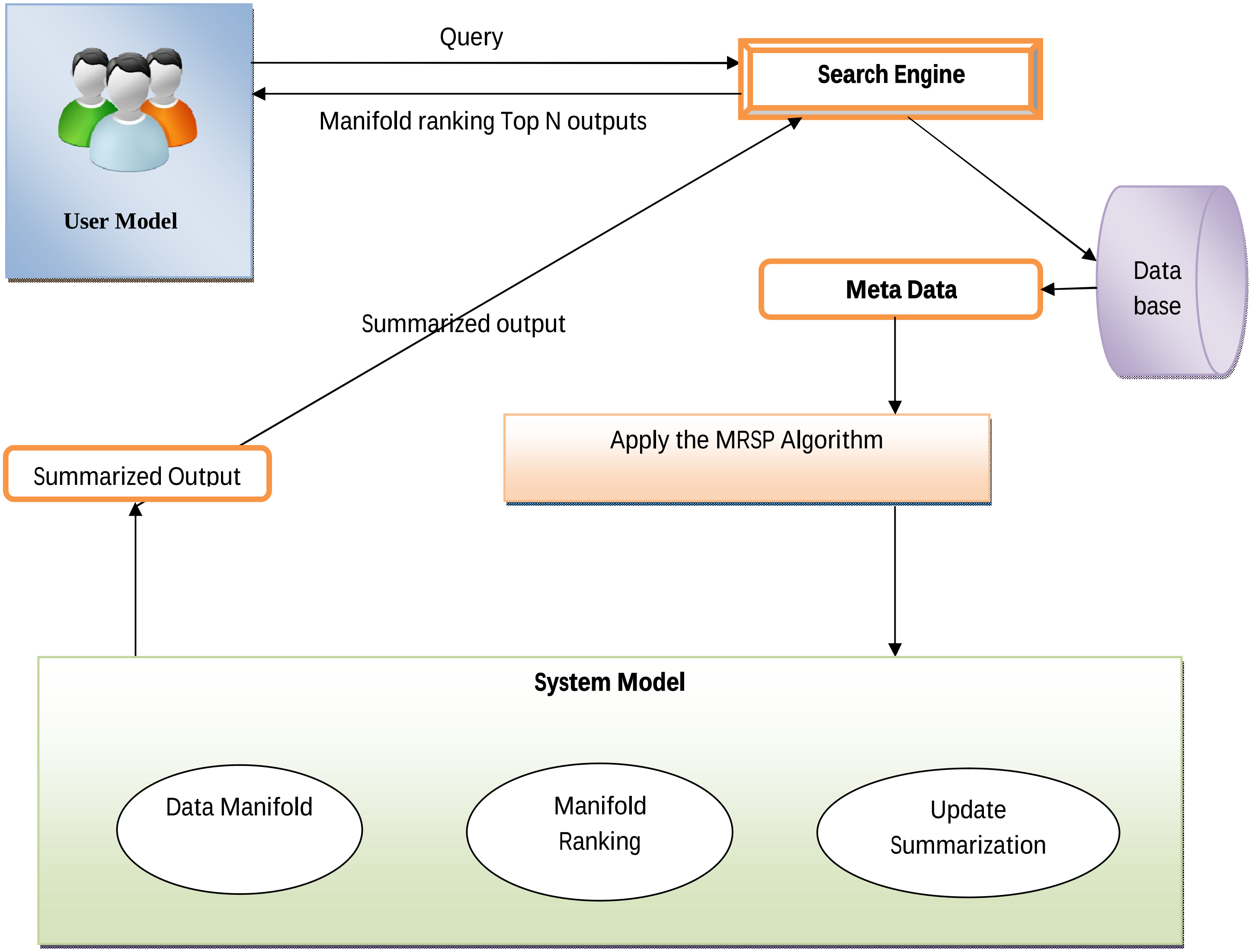

FIGURE-THE NOVEL MANIFOLD RANKING WITH SINK POINTS

THE NOVEL MRSP ALGORITHM

The novel MRSP algorithm works as follows:

1. Initialize the set of sink points _s as empty.

2. Form the affinity matrix W for the data manifold, where Wij = sim(xi, xj) if there is an edge linking xi and xj . Note that sim(xi, xj) is the similarity between objects xi and xj .

3. Symmetrically normalize W as S = D−1/2WD−1/2 in which D is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of W.

4. Repeat the following steps if |_s| < K:

   (a) Iterate f(t + 1) = _SIf f(t) + (1 − _)y until convergence, where 0 ≤ _ < 1, and If is an indicator matrix which is a diagonal matrix with its (i, i)-element equal to 0 if xi ∈ _s and 1 otherwise.
   (b) Let f∗ i denote the limit of the sequence {fi(t)}. Rank points xi ∈ _r according to their ranking scores f i (largest ranked first).
   (c) Pick the top ranked point xm. Turn xm into a new sink point by moving it from _r to _s.

5. Return the sink points in the order that they were selected into _s from _r.

IV. COMPONENETS OF MRSP

The novel MRSP model consists of the following components.

a. DATA MANIFOLD

The data manifold framework is mainly based on two criteria. First, nearby sentences are considered to have close ranking score. Second one, data in the same structure is likely to contain almost same ranking scores. The data manifold process is intuitively explained as follows: a network is constructed with predefined weights for the nodes. The data and query points are considered as nodes of the graph. An edge exists between the nodes of the graph, if the two nodes contain close ranking scores.

b. MANIFOLD RANKING

The manifold ranking process is the next phase of the Data Manifold process and it gives high ranks to the nodes that are closer to the queries on the data representation process. The nodes propagate the rankings scores to their nearest neighbors through weighted network. The propagation process is continued until a global state is obtained. In this way, relevance and importance are well balanced in the manifold ranking.

c. UPDATE SUMMARIZATION

Update summarization summarizes the complete information contained into a new document given a set of documents. There are two kinds of update summarization approach. First one is abstractive summarization and other summarization technique is extractive approach. Abstractive summarization applies some deep natural language processing techniques to compress the sentences present in the sentences to reorganize the sentences, in order to produce a summary of text. The extractive approach composes the summary by extracting the most representative sentences from target documents.

d. QUERY RECOMMENDATION

The task of query recommendation is to provide alternative queries to help users improve the usability if search engines query recommendation mainly focuses on measuring similarity and query log data is used to implement this technique.

V. CONCLUSION

The novel MRSP approach addresses diversity as well as relevance and importance in ranking. MRSP uses a manifold ranking process over the data manifold, which can naturally find the most relevant and important data objects present in a document. MRSP can effectively prevent redundant objects from receiving a high rank. The novel MRSP approach solves the ambiguous requirements of different queries given to the search engine and generates highly relevant query recommendations and update summarization.

REFERENCES

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.
2. J. Allan, R. Gupta, and V. Khandelwal. Temporal summariesof news topics. In SIGIR '01: Proceedings of the 24th annual
3. International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 10–18, New York, NY, USA, 2001.ACM.
4. D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 407–416, 2000.
5. P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In Proceedings of the 2009
6. Workshop on Web Search Click Data, WSCD '09, pages 56–63, New York, NY, USA, 2009. ACM.
7. F. Boudin, M. El-B`eze, and J.-M. Torres-Moreno. A scalable MMR approach to sentence scoring for multi-document update summarization. In Coling 2008: Companion volume: Posters, pages 23–26, Manchester, UK, August 2008.
8. J. Carbonell and J. Goldstein. The use of MMR, diversity based re-ranking for reordering documents and producing summaries. pages 335–336, New York, NY, USA, 1998.
9. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 659–666, 2008.
10. H. T. Dang and K. Owczarzak. Overview of the tac 2009 summarization track (draft). In Proceedings of the Second Text Analysis Conference (TAC2009), 2009.
11. P. G. Doyle and. Snell, J. Laurie (James Laurie). Random walks and electric networks / by Peter G. Doyle, J. Laurie Snell. [Washington, D.C.] : Mathematical Association of America, 1984. Includes index.
12. P. Du, J. Guo, J. Zhang, and X. Cheng. Manifold ranking with sink points for update summarization. In CIKM '10: Proceeding of the 19th ACM conference on Information and knowledge management, Toronto, Canada, 2010. ACM.
13. G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res., 22(1):457–479, 2004.
14. T. H. Haveliwala. Topic-sensitive pagerank. In Proceedings of the 11th international conference on World Wide Web, WWW '02, Pages 517–526, New York, NY, USA, 2002. ACM.
15. K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst., 20(4):422–446, 2002.
16. K. Knight and D. Marcu. Statistics-based summarization – step one: Sentence compression. In Proceedings of the Seventeenth
17. National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 703–710. AAAI Press / the MIT Press, 2000.
18. Y. Lan, T.-Y. Liu, Z. Ma, and H. Li. Generalization analysis of list wise learning-to-rank algorithms. In ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, pages 577–584, New York, NY, USA, 2009. ACM.