Efficient Information Retrieval By Using Multi-Modality Manifold Ranking Based On Syntactic/Semantic Measurement

J. Saranya M.Phil1, P. Ilango, M.C.A, PGDCA, M.Phil 2

1 Mphil Scholar, Department of Computer Science, Gobi Arts and Science College, Gobi, Tamilnadu, India
2 Associate professor, Department of Computer Science, Gobi Arts and Science College, Gobi.

Abstract— Ranking is the major problem in variety of applications like information retrieval (IR), Data mining (DM) and natural language processing (NLP). To rank the objects according to their importance multiplicity has also been identified as a important criterion. Manifold Ranking with Sink Points (MRSP) is one of the novel approaches to conquer this problem. This approach uses a manifold ranking procedure data manifold to determine the most pertinent and significant data objects to rank the data objects. The measurement of the semantic/syntactic similarity between the terms and multi-manifold ranking are not supported in the existing system. While considering the semantic and syntactic based measuring also important to improve the information retrieval (IR) result than the normal keyword based results. Compared to normal keyword based results the semantic and syntactic based measuring also important to improve the information retrieval (IR). Proposed system uses a Multi-Modality Manifold Ranking (MMM) which contains multiple data manifolds, where each and every data is constructed using distinct keywords. It employs syntactic and shallow semantic kernels to estimate the importance between the terms. The tree kernel functions are exploit for multi-modality manifold ranking framework. In addition to syntactic and semantic information can develop the performance of the multimodality manifold-ranking algorithm.

Keywords—Manifold ranking with sink points, update summarization, query recommendation, Multi-manifold ranking.

I. INTRODUCTION

The diversity of applications of ranking is Information Retrieval (IR), data mining (DM) and natural language processing (NLP). To reduce the information retrieval problems like document recovering, collaborative filtering, sentiment analysis, computational advertising that is online ad placement ranking is the necessary part. In the collection of data objects the ranking model selects the data objects in accordance with their significance. Even though a collection of pertinent objects may contain unnecessary data, duplicate information which is unwanted to the users. So, the redundancy in top ranked results will decrease the possibility to satisfy different users. The information need for the user will not be fulfilled due to the ranking model.

To address the problem of diversity and relevance we use a novel approach called Manifold Ranking with Sink Points (MRSP). By using the manifold ranking process in this approach, we can easily identify the most relevant data objects. During the manifold ranking process, we initiate the manifold sink points which the objects whose ranking scores are fixed at the minimum score like zero. During the ranking process, based on the intrinsic manifold the ranking scores of other objects close to the sink points. By turning ranked objects into sink points in the data manifold, we can efficiently avoid redundant objects from receiving a high rank.

Due to the ranking process we can capture diversity as well as relevance and importance. In this manifold ranking process the measurement of the semantic/syntactic similarity between the terms and multi-manifold ranking are not supported. Compared to normal keyword based results the semantic and syntactic based measuring also significant to develop the information retrieval (IR).

In order to compute the semantic/syntactic similarity between the terms we introduce an innovative approach called multi-manifold ranking. By means of the tree structure which we call Semantic Tree (ST), we calculate the semantic similarity between the terms. Computing the numeral of common subtrees between two trees by using tree kernel (TK) function. Such subtrees are topic to the restraint that their nodes are taken with all or none of the children. In the semantic tree, influence is displace with the majority significant word-often referred to as the semantic head. To recognize the semantic head in the argument, firstly we look noun, then verb, then adjective, then adverb. If none of these is there, we take the first word of the disagreement as the semantic head. These decrease the data sparseness with respect to a typical cosine measure representation used in the basic ranking model.

The major steps of the work as follows:

1. First step the manifold ranking process is that first addresses the diversity as well as relevance in addition to importance now ranking in a combined way.

2. To rank the objects according to their significance Manifold Ranking with Sink Points (MRSP) is used.
3. This approach uses a manifold ranking procedure to data manifold to establish the most pertinent and significant data objects to rank the data objects.

4. In order to improve the information retrieval (IR) result than the normal keyword based results semantic and syntactic based measurement is used that is called Multi-modality manifold ranking.

5. Finally measure the performance of the system α—NDCG and Intent coverage comparison.

II. LITERATURE REVIEW

Xiaojun Wan et.al [1] proposed an automated multi-document summarization that is generally used to provide crisp topic description about a cluster of documents and facilitate users to browse the document cluster. To rank sentences or passages, graph-ranking based algorithms have been proposed. Before relating the TimedTextRank algorithm, we first introduce the original Text Rank algorithm. Text Rank algorithm creates the relationships between sentences and selects sentences according to the “votes” from their neighboring sentences, which is related to Page Rank and HITS.

Paolo Boldi et.al [2] proposed query recommendations are typically queries related to the original one, and they are usually attained by analyzing the query logs, for instance, identifying recommendations by clustering of queries, or by identifying frequent re-phrasings [2]. In this method there is a problem of indicating related queries issued by other users and query expansion methods to create artificial queries.

Furu Wei et.al [3] suggested query-sensitive mutual reinforcement chain (Qs-MRC) approach for document summarization. The explosion of the WWW has brought with it a vast hoard of information. With the availability of the number of individual documents, it is impossibile to anyone for read and understands. Automatic document summarization presents an efficient means to handle such an exponentially increased set of information and to sustain information seeking and condensing goals. Instinctively, MR principle is sound and applicable. Since a document is always structured into meaningful text units such as paragraph, sentence, phrase and word in turn, the affiliation relation or the affinity relation between sentence and term can successfully mutually reinforce the importance of each other.

Gune s Erkan et.al [4] proposed stochastic graph-based method for computing relative significance of textual units for Natural Language Processing. The relative significance of textual units for Natural Language Processing is computed in stochastic graph-based method. We test the technique on the problem of Text Summarization (TS). Extractive TS relies on the concept of sentence salience to find the most significant sentences in a document or set of documents. For computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences, we consider a new approach called LexRank.

Kalervo Jarvelin et.al [5] proposed Cumulated Gain-Based Evaluation of IR Techniques for the information retrieval. In the information retrieval process due to the Collection of data objects, the users are dissatisfied because all documents are not of equal relevance to their users. So, in order to overcome this problem, we identified and ranked for highly relevant documents for presentation. To retrieve highly relevant documents, we develop IR techniques in this direction it is necessary to develop evaluation approaches and methods that gives credit to IR methods. This method can be done based on graded relevance judgments. The computation of the cumulative gain so that the user obtains by verifying the retrieval result up to a particular rank position.

Rakesh Agrawal et.al [6] suggested systematic approach to diversifying results that intend to reduce the risk of dissatisfaction of the average user. Web search has become the most important method for people to accomplish their information needs, whereby users usually denote their information needs by providing a few keywords. The significance of result diversification has been predictable on information retrieval. The basic principle is that the relevance of a set of documents depends not only on the individual relevance of its members, but also on how they relate to one another. Particularly, we presume that there exists taxonomy of information, and model user intents at the topical level of the taxonomy.

Rodrygo L. T. Santos et.al [7] proposed novel probabilistic framework for Web search result diversification. Web user’s underlying information need is not clearly specified from the initial query; a valuable approach is to diversify the results retrieved for this query. For Web search result diversification, we establish a novel probabilistic framework. In particular, we diversify a document ranking by estimating how well a given document satisfies each uncovered aspect and the extent to which different aspects are satisfied by the ranking as a whole. Moreover, by simulating an upper-bound query reformulation mechanism from official TREC data, we illustrate useful insights regarding the effectiveness of the query reformulations generated by the different WSEs in promoting diversity.

Qiaozhu Mei et.al [8] proposed a novel query suggestion algorithm based on ranking queries with the hitting time on a large scale bipartite graph. The massive growth of web information has not only created a vital challenge for search engine companies to handle large scale data, but also enlarged the complexity for a user to handle his information need. It has become progressively more complicated for the user to create a concise and accurate query to present the search need. It is a common practice for a search engine to give some types of query suggestions for reducing the trouble of the users.

Jian Guo et.al [9] suggested a ranking algorithm called DivRank, based on a toughened random walk in an information network. Many recovery and mining tasks are
disturbed with finding the most important and/or pertinent items from a large set of data. Top ranked web pages are offered to the users of a search engine. DivRank, based on a reinforced random walk in an information network. These models automatically stabilize the prestige and the diversity of the top ranked vertices in a righteous way.

Xueqi Cheng et al. [10] proposed a new extractive approach based on manifold ranking with sink points for update summarization. There has been an ever-increasing interest in text summarization with the enormous explosion in the amount of data on the Web. Update summarization intend to generate a summary over a topic-related multi-document dataset based on the assumption that the user has already read a set of earlier documents of the same topic. To overcome this problem, we use manifold ranking with sink point’s technique for update summarization.

III. MANIFOLD RANKING WITH SINK POINTS AND SEMANTIC/SYNTACTIC BASED MULTIMODALITY MANIFOLD RANKING

Manifold Ranking with Sink Points (MRSPs) is one of the significant techniques, to undertake diversity as well as relevance and importance in ranking. It undertakes all the data as well as query objects are present points sampled after a low dimensional manifold also leverages a manifold ranking process to tackle relevance also importance. In this technique, we initiate the manifold sink points in which the ranking scores of the data objects are fixed at the minimum score (zero in our case) during the manifold ranking process. Hence, the sink points will not once extent any rank toward their neighbors. Impulsively, we can predict the sink points such as the “black holes” taking place the manifold, everywhere ranking scores spreading to them will be engrossed as well as no ranking scores would discharge from them. Our overall algorithm monitors an iterative structure is defined in MRSP. At each iteration, we procedure manifold ranking to find more relevant points. Then, we chance the ranked points into sink points; fill in scores, also repeat. By turning ranked objects interested in sink points taking place data manifold, we can successfully check redundant objects from receiving a high rank.

A. MSRP algorithm

For ranking the relevance data points from the user query with best result, first elected the top ranked point \( x_1 \), it is defined as the most significant and relevance of the data to the neighborhood data point that is \( x_0 \) and then call our MRSP algorithm. It can well execute the data points that closer to the nearest point that is sink points while ranking the result. Thus, points in the center group were boosted up. Similarly, if we turn the novel top ranked point \( x_i \) into a sink point, it will punish its nearby points in the center collection and make the points in the left group surface and it can also focus to the diversity of the data points as well as significance and importance in a mutual manner. Let \( \mathcal{X} = \mathcal{X}_q \cup \mathcal{X}_s \cup \mathcal{X}_r \in \mathbb{R}^m \) denote a set of data points over the manifold, \( \mathcal{X}_q = \{x_1, \ldots, x_q\} \) where denotes a set of query points, \( \mathcal{X}_s = \{x_1, \ldots, x_s\} \) denotes a set of sink points, and \( \mathcal{X}_r = \{x_1, \ldots, x_r\} \) denotes the set of points to be ranked called free points. Let \( f: \mathcal{X} \rightarrow \mathbb{R} \) denote a ranking function which assigns a ranking score \( f_i \) to each point \( x_i \).

Then can view \( f \) as a vector \( f = [f_1; \ldots; f_s]^T \), where \( N = q + s + r \). \( x_i \) is a query also define \( y = [y_1; \ldots; y_N]^T \) and \( y_1 = 0 \) otherwise. Suppose only top-\( K \) ranked data points are needed to be expanded, the MRSP algorithm works as follows:

1. Initialize the position of sink point \( \mathcal{X}_s \) as vacant.
2. Form the affinity matrix \( \mathcal{W} \) for the data manifold, where \( \mathcal{W}_i = \text{sim}(x_i; x_j) \) if there is an edge linking \( x_i \) and \( x_j \). It is defined as the similarity between the data objects in the sink points.
3. Symmetrically regularize \( \mathcal{W} \) as \( \mathcal{W} = \mathcal{D}^{-1/2} \mathcal{W} \mathcal{D}^{-1/2} \) in which \( \mathcal{D} \) is a diagonal matrix with its \((i,i)\) element equal to the sum of the \( i \)th row of \( \mathcal{W} \).
4. Repeat the following steps if \( |\mathcal{X}_s| < K \):
   4.1 Let \( f = \alpha SUf + (1 - \alpha)y \) until convergence, where \( 0 \leq \alpha \leq 1 \) and \( f \) is indicator matrix and it symbolize as diagonal matrix with its \((i,i)\) element equal to zero. If \( x_i \epsilon \mathcal{X}_s \) and one otherwise.
   4.2 Let \( f_i^* \) denotes the limit of the sequence \( \{f_i(t)\} \) and rank points \( x_i \epsilon \mathcal{X}_r \) based on the ranking scores \( f_i^* \).
   4.3 Pick the first top ranked point \( x_m \) into a new sink point by moving the result to \( \mathcal{X}_r \) to \( \mathcal{X}_s \).
   5. Return the sink points in the order that they were selected into \( \mathcal{X}_s \) from \( \mathcal{X}_r \).

B. Refined MRSP

The data sink points with the ranking algorithm for beyond step solve the optimization problem and can be recognized as a Dirichlet problem. Consequently, the computational efficiency of MRSP can be further enhanced. In this algorithm once the sink points are chosen and get sinked, need to reorganize the matrix \( \mathcal{S} \). This will result in a recalculation of the matrix which is computationally expensive. Fortunately, we only need to acquire \( \Omega = (1 - \alpha \mathcal{S})^{-1} \) once before the first iteration, because the reorganization of matrix \( \mathcal{S} \) has the same effect as the reorganization of matrix \( \mathcal{S} \). In step 5, matrix \( \Omega = \Omega_1 \) is initialized by organizing group sink points into \( \Omega_{11} \) and others into \( \Omega_{22} \). If the set of sink points is empty, \( \Omega_{22} = \Omega \) which means the refined algorithm disintegrate into the traditional manifold. The ranking algorithm at each iteration mark the top ranked object as a new sink point and move it from the group of free points to the group of sink points by reorganizing matrix \( \mathcal{S} \). Then the object is to be elected next will distribute different information from that of previously preferred. With little number of query points in most real scenarios, the calculation in step 6 can be very inexpensive. In this way, our advanced MRSP algorithm is able to tackle the problem of diversity in ranking very efficiently.
Refined MRSP algorithm

1. Calculate the similarity values \( \text{sim}(x_i, x_j) \) of each pair of the data points \( x_i \) and \( x_j \).
2. Connect any two objects with an edge if their similarity values exceed 0. We define the affinity matrix \( W \) by \( W_{ij} = \text{sim}(x_i, x_j) \) if there is an edge linking \( x_i \) and \( x_j \). Let \( W_{ii} = 0 \) to avoid self-loops in the graph.
3. Symmetrically regularize \( W \) 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. Calculate \( \Omega = (1 - \alpha s)^{-1} \) where \( 0 \leq \alpha < 1 \).
5. Attain the submatrices \( \Omega_{11}, \Omega_{12}, \Omega_{21}, \Omega_{22} \) from \( \Omega \) based on free points and query points and 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 highest score \( f^*_m \) as a new sink point.
8. If the predefined number of sink points \( K \) is not reached got to step 5.
9. Arrival the sink points in the arrange that they get obvious as sink points.

C. Shallow semantic and Multimodality MR

Multimodality manifold ranking (MMMR) is based on the syntactic and semantic actions. In our project we projected a Syntactic/semantic based MMR tree kernel (TK) functions compute the numeral of ordinary subtrees among two trees. Such subtrees are subject to the restriction that their nodes are full by means of every one or not any of the family they have in the innovative tree. Although, this clarification of subtrees makes the TK function suitable for syntactic trees but at the similar time make it not well suitable for the semantic trees (ST). TK purpose is that the production of two estimated nodes has to be identical to allow the match of additional descendants. This means that common substructures cannot be composed by a node with only some of its children as an effective ST manifestation would necessitate. Scheming the Shallow Semantic Tree Kernel (SSTK) which permits matching portion of a ST in the data manifold with sink points. The SSTK is based on two ideas: primary it changes the ST by addition SLOT nodes. These offer accommodation dispute labels in an exact order i.e. it provide a fixed numerical of slots, possibly filled with null influence, that educate all possible predicate arguments. Leaf nodes are packed with the wildcard character * excluding they might instead include extra information. After the construction of syntactic and semantic trees the next task is to measure the similarity between the trees. For this every tree \( T \) is represented by an \( m \) dimensional vector.

\[
V(T) = V_1(T), V_2(T) \ldots \ldots, V_m(T)
\]

In which the \( i \)-th element is the number of occurrences of the \( i \)-th tree fragment in tree \( T \). The tree kernel of two trees \( T_1 \) and \( T_2 \) is actually the inner product of \( V(T_1) \) and \( V(T_2) \):

\[
\text{TK}(T_1 : T_2) = V(T_1) : V(T_2)
\]

The calculation of the polynomial time is as follows:

- If the productions at \( n_1 \) and \( n_2 \) are different terms then \( C(n_1 ; n_2) = 0 \).
- If the productions at \( n_1 \) and \( n_2 \) are the same, and \( n_1 \) and \( n_2 \) are pre-terminals, then \( C(n_1 ; n_2) = 1 \).
- Else if the productions at \( n_1 \) and \( n_2 \) are not pre-terminals.

\[
C(n_1 : n_2) = \sum_{j=1}^{\text{terminals}} 1 + (\text{ch}(n_1 ; j) : \text{ch}(n_2 ; j))
\]

Where \( \text{ch}(n_1 ; j) : \text{ch}(n_2 ; j) \) are the child nodes of the tree that represents semantic terms of the manifold data. \( \text{TK}(T_1 ; T_2) \) is the semantics similarity between terms.

IV. EXPERIMENTAL RESULTS

Lastly in this section manifold ranking with sink points and Multi-Modality manifold ranking is compared. The corresponding results of the manifold ranking with sink points and Multi-Modality manifold ranking is measured with \( \alpha \)-nDCG Comparison and Intent-Coverage are shown in Figure 1 and Figure 2.

A. \( \alpha \)-normalized Discounted Cumulative Gain (\( \alpha \)-nDCG)

The \( \alpha \)-nDCG, which rewards diversity in ranking, is a new version of the nDCG, the normalized Discounted Cumulative Gain measure. When \( \alpha = 0 \), the \( \alpha \)-nDCG measure corresponds to the standard nDCG, and when \( \alpha \) is closer to 1, the diversity is rewarded more in the metric. The key difference between \( \alpha \)-nDCG and nDCG is that they use different gain value. For each suggestion, the gain value \( G(k) \) of \( \alpha \)-nDCG is distinct as,

\[
G(k) = \sum_{i=1}^{k} j_i(k)(1 - \alpha)^{G_i(k-1)}
\]

Where \( G_i(k-1) \) is the numeral of applicable recommendations found within the top \( k \)-1 recommendations for intent \( i \), \( j_i(k) \) is a binary variable signifying whether the recommendation at rank \( k \) belongs to intent \( i \) or not, and \( l \) is the total number of unique intents for each test query. The calculation of \( \alpha \)-nDCG
B. Intent-Coverage

The Intent-Coverage actions the amount of single intents covered by the top k recommended queries for each examination query. Because each intent represents particular user information requires, superior Intent-Coverage indicates larger probability to satisfy different users. Intent-Coverage is dissimilar from the multiplicity measure used in automatic evaluation, because only relevant recommendations to the test query will be measured in Intent-Coverage. As a result Intent-Coverage can better reflect the diversity excellence of recommendations than the diversity compute in automatic evaluation. The Intent-Coverage is properly defined as

\[
\text{Intent - coverage}(k) = \frac{1}{I} \sum_{i=1}^{I} B_i(k)
\]

Where \( B_i(k) \) is a binary variable representing whether the intent \( i \) is institute within the top \( k \) recommendation or not, and \( I \) is the full amount numeral of unique intents for each test query.

V. CONCLUSIONS AND FUTURE WORK

MRSP is one of an innovative approach which deals with diversity as well as relevance and significance in ranking. By using a manifold ranking procedure over the data manifold, which can evidently recognize the most appropriate and significant objects. Meanwhile, by rotating ranked objects interested in sink points on data manifold, it can proficiently prevent redundant objects beginning receiving a high rank. Proposed system predetermined the syntactic and semantic information for measuring the data points with different sink points in the multi-modality manifold ranking algorithm for topic-listening carefully multi-document summarization and report that adding syntactic and/or semantic information on top of the standard semantic calculate improve the presentation. This experiment for query recommendation also show that our Shallow semantic Multimodality manifold ranking can effectively generate diverse and highly relevant query recommendations. In future work we sustain a topic oriented manifold ranking with data sink points and also query based suggestion are given to user to explore the information proficiently.

REFERENCES