Review on Enhanced User Centric Similarity Search Using Sigmoid Coefficient

Mrs. Pallavi V. Ghaste[1], Mr.Vivek T. Patil[2], Saad Omair[3], Shivraj D. Mule[4], Mokshda Odak[5], Vyankatesh A. Shivnikar[6]
Department of Computer Engineering
D.Y.Patil College of Engineering Akurdi-Pune
Email: khude.pallavi@gmail.com[1] , vvkpatil300@gmail.com [2] , saadomair888@gmail.com[3] ,shivrajmule99@gmail.com[4] , mokshdaodak@gmail.com[5], vshivnikar@gmail.com[6]

Abstract- Sigmoid coefficient is variation of Jaccard coefficient which is used in similarity calculation. The similarity calculation is the measure of how much alike two data objects are. In computer field similarity plays a fundamental role in information processing, clustering and data analysis. Similarity calculation has application in many areas such as Artificial Intelligence, Recommendation System and Many Business Operations. Some of the important techniques that are used for finding similarity between objects are Euclidean distance, Cosine similarity and k-nearest neighbor measure. These techniques find the similarity between products ignoring user rating or preferences. User preferences play an important role in market analysis. Similarity between two objects are more accurate and effective if it is calculate by considering product attribute value as well as user preferences. Jaccard Coefficient is used to calculate similarity. Sigmoid coefficient is improvement of Jaccard coefficient which gives more accuracy. In order to improve performance of user centric search use sigmoid coefficient with considering user preferences.

1. INTRODUCTION

Database operations, particularly large database operations such as market data analysis, scientific data analysis and research, web searching and so on are very complicated in many real life databases. Finding similarity between objects plays an important role in clustering as well as in data analysis. Existing methods for finding similarity between objects are completely based only on the values of attributes of objects. Various similarity finding measures have been proposed for similarity calculations between objects. For instance it is used to find pages or documents with similar words over the web [1] or in order to detect customers with abnormal behavior based on the products they buy [2]. Estimation of the similarity between objects is a one of fundamental operation in data management. In the literature many different similarity metrics have been proposed for evaluating the similarity between two data items, such as the Euclidean distance and the cosine similarity. Similarity computations can be performed for the detection of similar conversations and comments between users of the social networks (i.e., comments on Facebook, tweets on Twitter) [3]. In order to perform such kind of similarity computations, a query type, termed a reverse top-k query is used [3].

Top-k query is used to rank top k product based on user preferences and reverse top-k query result into set of customers for which particular product is present in their top k set. In contrast to a top-k query that returns the k products with the best score for a specific customer, the result of a reverse top-k query is the set of customers for whom a given product belongs to their top-k set [4].

All these methods estimate similarity measures between objects based on only values of attributes. In reality, similarity values are computed more accurately and more generally when priority values of attributes are taken into consideration in addition to the values of attributes. These types of requirements are very useful in business applications and many critical database operations. For example, in production management attribute values and opinions of customers are both very useful for ranking the products based on the customers‟ preferences. This is a complementary user-centric approach for similarity computation, which takes into account users' preferences. Top-k query is used to rank top k product based on user preferences and reverse top-k query result into set of customers for which particular product is present in their top k set. And then similarity calculation between products is done using Jaccard Coefficient on reverse top-k set of products [4]. If you want to estimate the similarity between products based on the preferences their customers have expressed for them. A common way to rank products for a customer is to execute a top-k query that assigns scores to each product. New linear functions based technique is proposed for product clustering with respect to opinions of customers. Linear function computes similarity values using both values of attributes of the product and the respective priorities or opinions of values of those attributes.

The major motivation is by giving user preferences maximum priority, to design an efficient recommendation system which would basically give suggestions according to the users which have same behavior. Algorithms such as Top-K query, Reverse Top-K query are used for the implementation of user centric search. These algorithms require a coefficient for the similarity calculation. The referred research paper uses Jaccard coefficient for the similarity calculation between two objects by considering user preferences. Sigmoid coefficient is improvement of Jaccard coefficient which gives more accuracy. We are using sigmoid coefficient with considering user preferences to improve performance of user centric search. In the end, to validate the result i.e. the similar products shortlisted, we will use θ-similarity and m-nearest neighbour algorithm.

2. RELATED WORK

In User centric similarity search [4], Konstantinos Georgoulas and Akrivi Vlachou used Jaccard coefficient for the purpose of finding similarity between products. To create a framework which makes use of user-centric approach for similarity computation and capitalizes on rankings of products based on user preferences to discover similar products. To compute similarity they use θ-similarity and m-nearest neighbour queries to validate the result. General operation of showing the similarity between products is done by ignoring user preferences. Instead products are examined in a feature space based on their attributes and similarity is computed via traditional distance metrics on that space. In this paper, rankings of the products are based on the opinions of their customers in order to map the products in a user-centric space where similarity calculations are performed.

Nick Roussopoulos and Stephen Kelley used nearest neighbour search in geographic information system in which user can point or select specific location on the screen and request the system to find the five nearest point to it and also it is used when the user is not familiar with the layout of the special objects[5].

Badrul Sarwar and George Karypis used Item-based Collaborative Filtering, they analysed different item-based recommendation generation algorithms. In this different techniques for computing item-item similarities (e.g., item-item correlation vs. cosine similarity between item vectors) and different techniques for obtaining recommendations from them (e.g., weighted sum vs. regression model). Finally, they experimentally evaluated the results and compared them to the basic k-nearest neighbour approach. Experiments suggest that item- based algorithms provide dramatically better performance than user-based algorithms, while at the same time providing better quality than the best available user-based algorithms [6].

Dr. S. Aquter Babu proposed that in market basket analysis user ranking of products is very important in addition to the values of attributes of objects. According to him Similarity based comparison between objects play a very important role in many business operations such as ranking of objects with respect to the preferences of customers, finding a set of top-k objects in ranking order, and finding a set of k-nearest neighbour objects in ranking order and so on present study proposes a new similarity finding measure between objects. This measure computes weighted sums of values of attributes and priority values of respective values of attributes. These weighted sums are computed using a linear function formula. Finally a new clustering technique is proposed for clustering market basket analysis products using newly proposed similarity search measure between two objects new clustering technique based on new similarity finding measure is very useful in many real time applications and in many very large database operations including query execution [7].

3. METHODOLOGY

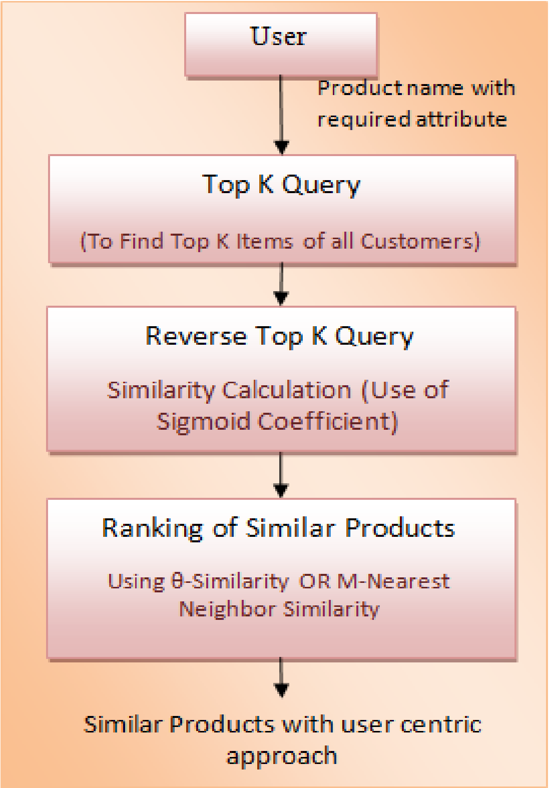
User
Product name with required attribute
Top K Query
(To Find Top K Items of all Customers)
Reverse Top K Query
Similarity Calculation (Use of Sigmoid Coefficient)
Ranking of Similar Products
Using θ-Similarity OR M-Nearest Neighbor Similarity
Similar Products with user centric approach

Fig. 1. User Centric Similarity Search Methodology

In order to calculate the similarity between two products in user centric space that considers both attribute value as well as user preferences we have to calculate score of product. We are using linear function to calculate score given as

$$\sum_{i=1}^{n} P[i] * W[i]$$

Where P[i] is value of ith attribute of product P and W[i] user rating for that attribute.

Top K-query returns the Top most K product for a particular customer. Reverse Top K-query returns list of customers for a particular product. In user centric similarity search, similarity calculation is going to be performed using sigmoid coefficient.

Formula for Sigmoid coefficient is

$$\mathrm{sim\ s}(O1, O2) = \frac{e^{\mathrm{sim\ j}(O1,O2)} - 1}{e^{\mathrm{sim\ j}(O1,O2)} + 1}$$

Where sim j(O1,O2) is Jaccard coefficient that can be calculated as

$$\mathrm{sim\ j}(O1, O2) = \frac{CF(O1, O2)}{DF(O1) + DF(O2) + CF(O1, O2)}$$

4. FRAMEWORK DESING

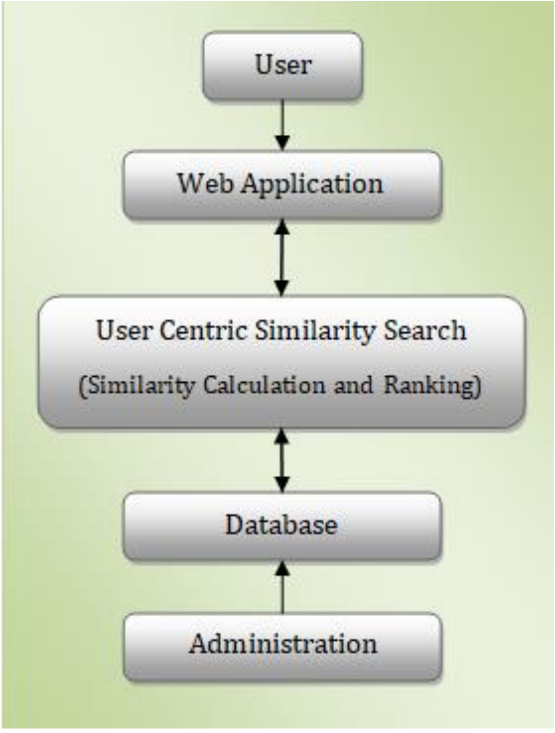
User
Web Application
User Centric Similarity Search
(Similarity Calculation and Ranking)
Database
Administration

Fig: 2 System architecture

Suggested client driven user centric recommendation system has above architecture design.

User Interface of the application is nothing but JSP pages. Application is based on gathering data by tracking customers activates and using it to perform Top K query and Reverse Top K query in order to calculate the similarity. Clients and administrators are two unique actors of the system related with the application. Administrator can add new items or products. Admin will also perform data base related operations. Clients will search the product. They can also rate and view products.

5. DISCUSSION

There is a work on user centric similarity framework which makes use of Jaccard coefficient on reverse top k query outcomes. In this work we are using one modifications of Jaccard coefficient called sigmoid coefficient. Sigmoid coefficient gives more performance than Jaccard coefficient in feature based similarity calculation. We are using sigmoid coefficient in user centric space to improve performance.

6. REFERENCES

[1] A.Rajaraman and J. D. Ullman, Mining of Massive Datasets. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[2] K. Georgoulas and Y. Kotidis, "Towards enabling outlier detection in large, high dimensional data warehouses," in Proc. Scientific Statistical Database Manage. 2012, pp. 591–594.

[3] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in Proc. 3rd ACM Int. Conf. Web Search Data Mining, 2010, pp. 291–300.

[4] Konstantinos Georgoulas, Akrivi Vlachou, Christos Doulkeridis, and Yannis Kotidis "User-Centric Similarity Search" IEEE Transaction on Knowledge and Data Engineering

[5] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl "Item Based Collaborative Filtering Recommendation Algorithms" Army HPC Research Center Department of Computer Science and Engineering University of Minnesota,

[6] Dr. S. Aquter Babu "Product Data Clustering using Weighted Similarity Measure" International Journal of Emerging Trends in Science and Technology

[7] Silvia Likavec, Ilaria Lombardi _ and Federica Cana "How to improve Jaccard's feature-based similarity measure" Department of Information, University of Torino, Cursor Svizzera, Italy