RESEARCH ARTICLE OPEN ACCESS

Cosine Similarity in Clustering With Collaborative Filtering For Service Recommendation

Reshma M Batule, Prof. Dr. S. A. Itkar
Department of Computer Science and Enienering
Savitribai Phule Pune University
Pune –India

ABSTRACT

Different services on the web are available in form of unstructured, semi structured and structured form. So the service recommendation is not done properly therefore the system is developed collaborative filtering with clustering and similarity calculation with cosine similarity. In the Clustering a document are clusters by their similarity calculation with cosine similarity calculation and after that collaborative filtering is applied. There are different similarity calculation algorithms but cosine is giving better results than the Jaccard coefficient.

Keywords:- **Clusters**, Clustering, Jaccard Coefficient, Cosine Similarity, Recommendation, Collaborative filtering.

I. INTRODUCTION

1.1 CLUSTERING

Clustering is a set of physical or abstract objects into classes of similar objects. Similar objects are in a one cluster and dissimilar objects in another cluster. Clustering partitions large amount of data into small parts called as segmentation.

1.2 Hierarchical Method

Hierarchical methods organize the data into tree structure. There are two types agglomerative and divisive.

a) Agglomerative Hierarchical Clustering

It is a bottom up strategy. Clusters size is increased after addition of objects into clusters. It will satisfy condition when all similar objects are placed in one cluster. Divisive clustering is opposite of agglomerative.

1.3 Model Based Clustering

Model based clustering fits the data into mathematical model. These methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. These are of three types- Expectation-Maximization, Conceptual clustering and a neural network approach to clustering.

1.4 Memory Based Collaborative Filtering

User rating is used to calculate similarity or weight between users and items. [6] Make predictions or recommendations according to those calculated similarity values. Similarity values are based on common items and therefore are unreliable when data are sparse. The common items are therefore few.

a) User Based Collaborative Filtering

It predicts user's interest in particular item based on rating information from similar user's profiles. Ratings by more similar users contribute to more predicting the item rating. Set of similar users can be identified by employing a threshold.

b) Item Based Collaborative Filtering

It is same as user based collaborative filtering but works on item rating. Unknown rating of the item can be predicted by averaging the ratings of other similar items. Item ratings are calculated by Pearson correlation coefficient.

1.4 Content Based Filtering

It recommends items based on the comparison between content of the items or user profiles which is also mentioned as cognitive filtering. Content of each item is expressed as set of descriptors or terms, typically words that appear in the document.

1.5 Collaborative Filtering

It is filtering technique which automatically analyzes the data which user could not analyze. This is most widely used technique nowadays.

II. LITERATURE SURVEY

User's characteristics are calculated by similarity algorithm and clustering is applied on the similarities. It will assemble the web visiting message data of users on the basis of similarities of users. Mittal, [7] projected to obtain the predictions for a user by first minimizing the size of item set, the user needed to explore. Movies are partitioned based on the genre requested by the user using k-means clustering algorithm. High-dimensional parameter free, divisive hierarchical clustering algorithm, Simon, [8] It uses implicit feedback on the basis of past purchases of users to find out the similarities between the user. Products of high interest are taken into one cluster. Implicit feedback may not give correct results so domain ontology is created on the basis of semantic interoperatibility. Pham, [10] projected the neighborhoods of the users in social network is determined by applying network clustering technique, and then provide the traditional CF algorithms to produce the recommendations. This work is relying upon on social relationships between users. Li, [11] projected to include multidimensional clustering into a collaborative filtering recommendation model. Background data in the form of user and item profiles was composed and clustered using the projected algorithm in the first stage. Then the poor clusters with analogous features were deleted while the appropriate clusters were further picked based on cluster pruning. At the third stage, an item prediction was made by operating a weighted average of deviations from the neighbor's mean. Such an approach was likely to trade-off on increasing the variety of recommendations while preserving the accuracy of recommendations.

Thomas, [12] proposed collaborative filtering based on weighted co-clustering algorithm. User and item neighborhoods are simultaneously produced via co-clustering and generate predictions based on the average ratings of the co-clusters. The entry of new users, items and ratings is handled by using an incremental co-clustering algorithm. J. Kelleher, [13] proposed a collaborative recommender that uses a user-based model to predict user ratings for specified items. The model generates summary rating information derived from a hierarchical clustering of the users. Its accuracy is good and coverage is maximal. Proposed algorithm is very efficient; predictions can be made in time that grows independently of the number of ratings and items. Rashid, [14] projected ClustKnn approach, a simple and intuitive algorithm that is well suited for large data sets. The projected method first compresses data tremendously by building a straightforward but efficient clustering model. Recommendations are then generated quickly by using a simple Nearest Neighbor-based approach. ClustKnn provides very good recommendation accuracy. Sarwar, [15] proposed a new approach in improving the scalability of recommender systems by using clustering techniques. Experiments suggest that clustering based neighborhood provides comparable prediction quality as the basic CF approach. Author uses a variant of K- means clustering algorithm called the bisecting K-means clustering algorithm. This algorithm is fast and tends to produce clusters of relatively uniform size, which results in good cluster quality.

Zhirao, [16] proposed Community-based collaborative filtering algorithm, its idea is that the users who belong to the same community have the same interests, who do not belong to the same community do not have the same interests, which Data-Providing service in terms of vectors is described by, Zhou, [9] which considers the composite relation between input, output, and semantic relations between them. Refined fuzzy C-means algorithm is applied to cluster the vectors. The capability of service search engine was enhanced narrows the scope of the neighbors. To a certain extent, it solves the problem of data sparseness.

Tseng, [17] proposed Default voting scheme using the cloud model which represents the user's global preference that is computed from users' past ratings to ameliorate the sparsity problem and preferences more accurately and reduce the data sparsity. ZHANG, [18] Considers the user's level of consumption, using the association rule mining formalized the competitive relationship between goods; using the time-based Bayesian probability formalize the complementary relationship between commodities, and through these relationship between the two commodities matches the user's requiring preferences and price preferences into the item sets of user evaluation.

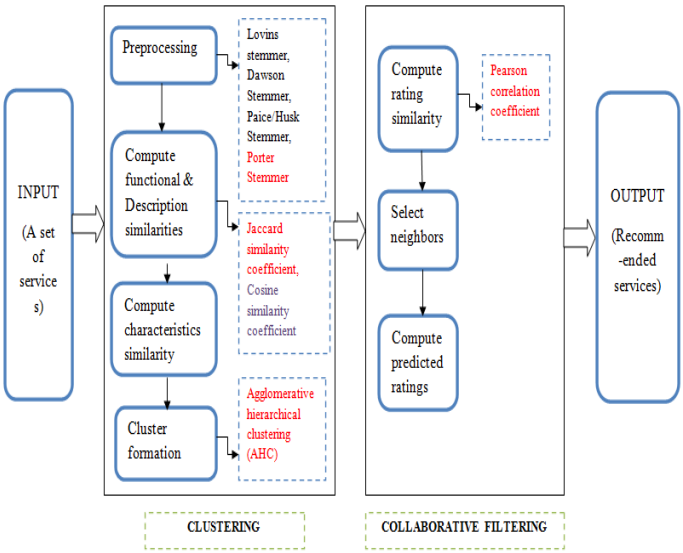
III. SYSTEM FLOW

The system flow shows that the input is user-item matrix it is given for the clustering. The similar services are processed by collaborative filtering and the output is recommended services.

A. Processing Steps

1. Clustering

I. Stem Word: - Developers can use different names to describe the similar services. But it influences to the description similarity. Therefore description words should be uniformed before measurement of description. Morphological similar words are clubbed together under the assumption that they are semantically similar. To stem word porter stemmer algorithm is used.



II. Compute Description Similarity And Functionality Similarity: - Description similarity and functionality similarity are both computed by Jaccard similarity coefficient (JSC) which is a statistical measure of similarity between samples cardinality of their intersection divided by the cardinality of their union.

III. Compute Characteristic Similarity:-Characteristic similarity between two services is computed by weighted sum of description and function similarities.

IV. Cluster Services: - Agglomerative hierarchical clustering is used to calculate the clusters. It creates the tree structure clusters.

2. Collaborative Filtering

Item-based collaborative filtering algorithms have been widely used in many real world applications such as at Amazon.com. It is divided into three main steps, i.e. compute rating similarities, select neighbours and recommend services.

I. Compute Rating Similarity: - Rating similarity items is a time consuming but critical step in item-based CF algorithms. Common rating similarity measures include the Pearson correlation coefficient (PCC).The basic intuition behind PCC measure is to give a high similarity score for two items that tend to be rated the same by many users. PCC which is the preferred choice in most major systems was found to perform better than cosine vector similarity. Therefore, PCC is applied to compute rating similarity between each pair of services in ClubCF.

II. Select Neighbours: - Based on the enhanced rating similarities between services, the neighbours of a target service are determined.

III. Compute Predicted Rating: - For an active user for whom predictions are being made, whether a target Service is worth recommending depends on its Predicted rating.

B. Algorithmic Aspect.

I. Cosine Similarity algorithm

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents.

Input-User and item ratings

Output-similarities between user and item

Steps-

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

1. Get the dot product of vectors a and b
2. Multiply magnitude a and magnitude b
3. Divide the dot product of vectors a and b by the product of magnitude a and magnitude b.

II. Hierarchical Clustering algorithm

Input- Services with similar users and item.

Output- Cluster levels

Steps-

1. Start with n clusters containing one object
2. Find the most similar pair of clusters Ci and Cj from the proximity matrix and merge them into a single cluster
3. Update the proximity matrix (reduce its order by one, by replacing the individual clusters with the merged cluster) 4. Repeat steps (2) and (3) until a single cluster is obtained (i.e. N-1 times)

IV. MATHEMATICAL MODEL

Let, S is the System having Input, Functions and Output.
S = {I, P, O}
Where,

I -> Input,
O -> Output,
P -> Processing.

- I=User-item matrix
- For the system, there are two processing steps P= {P_1, P_2}
- P1= {Clustering}
  1. Stem Word:-Stemming is done using Porter Stemmer algorithm.
  2. Calculate Functionality Similarity and Description Similarity using Jaccard similarity Coefficient.
     Let S_t and S_j are two services, D_sim and F_sim are Description and Function similarities respectively.

$$D\_Sim(S_t, S_j) = \left[\frac{D_t \cap D_j}{D_t \cup D_j} \right]$$

$$F\_Sim(S_t, S_j) = \left[\frac{F_t \cap F_j}{F_t \cup F_j} \right]$$

  3. Calculate Characteristics Similarity. Let C_sim be the Characteristic similarity, α and β are the weight of Description and Function similarities respectively.

$$C\_Sim(S_t, S_j) = \alpha \times D\_Sim(S_t, S_j) + \beta \times F\_Sim(S_t, S_j)$$

  4. Cluster Services:-agglomerative hierarchical clustering is used.

- P2={Collaborative Filtering}
  5. Compute Rating Similarity:-Common rating similarity measures include the Pearson correlation coefficient (PCC).
  6. Select Neighbours.
  7. Compute Predicted Rating for an active user for whom predictions are being made, whether a target service is worth recommending depends on its predicted rating.
- O= {Recommended Services}

V. EXPERIMENTAL RESULTS

This section presents the performance of the Jaccard similarity coefficient algorithm and cosine similarity. The results are obtained by using the dataset from website Programmable Web.

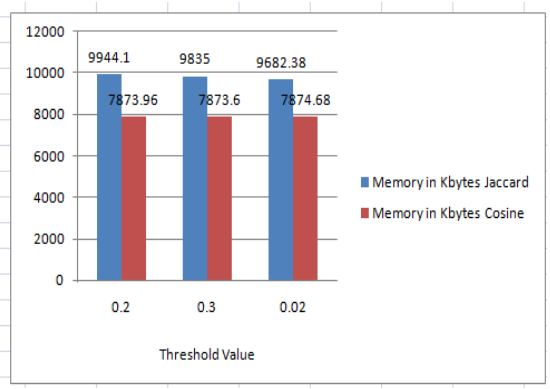


Fig.1. Memory in Kbytes used by Jaccard and Cosine for Threshold values.

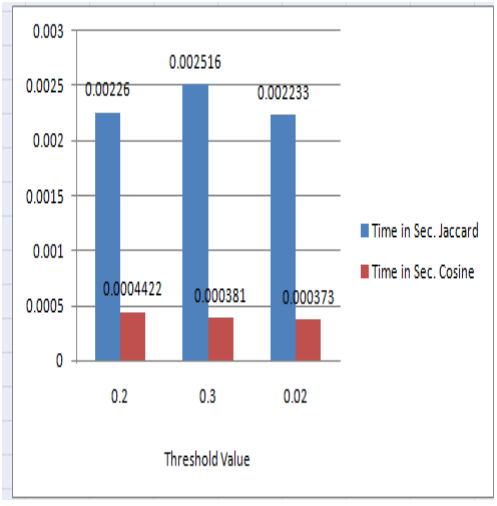


Fig.2. Execution Time in sec. used by Jaccard

and Cosine for Threshold values.

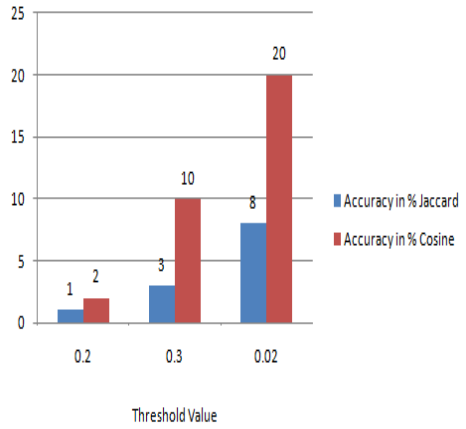


Fig.3. Accuracy of Jaccard and Cosine for Threshold values.

When different threshold values are used in Jaccard and cosine similarity, Cosine takes less time for similarity calculation and less memory to store the similarities.

Cosine gives more accuracy than the Jaccard. When different cluster sizes are taken then also cosine is taking less time. When different cluster values are taken for the comparison then also it is showing better results for time.

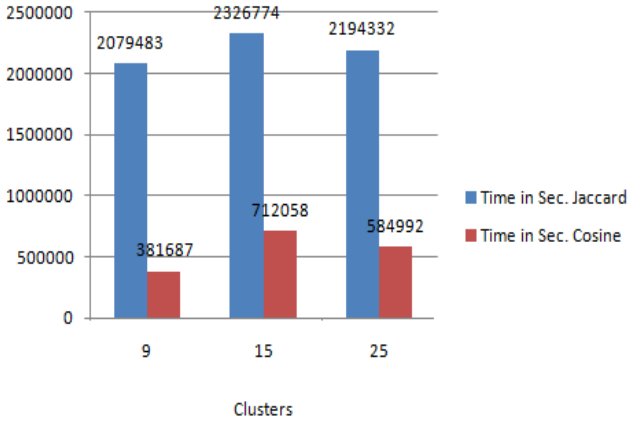


Fig.4. Execution Time in sec. used by Jaccard and Cosine for cluster values.

VI. CONCLUSION

In this paper, I presented a cosine similarity algorithm instead of Jaccard coefficient in collaborative filtering with clustering technique and it take less time than the Jaccard and less memory. It is used to store large amount of data. When the different cluster sizes are taken it takes less time for similarity calculation.

REFERENCES

[1] Rong hu, Wanchun dou and Jianxun liu , ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application, IEEE transaction on Emerging Topics in Computing, vol .2, no. 3, pp. 302313, Sep. 2014.

[2] Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude, Data providing services clustering and management for facilitating service discovery and replacement, IEEE Trans. Autom. Sci. Eng., vol. 10, no. 4, pp. 1311-1146, Oct. 2013.

[3] Ankur Narang, Abhinav Srivastava, Naga Praveen, Kumar Katta, High Performance On Online Distributed Collaborative Filtering, 2012 IEEE 12th International Conference on Data Mining, pp.549-558.

[4] Manh Cuong Pham, Yiwei Cao, Ralf Klamma, Matthias Jarke, A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis, Journal of Universal Computer Science, vol. 17, no. 4(2011) , pp.583-604.

[5] Manh S. Kanimozhi, Effective Constraint based Clustering Approach for Collaborative Filtering Recommendation using Social Network Analysis, Bonfring International Journal of Data Mining, Vol. 1, pp.12-17, Dec. 2011.

[6] Xiaoyuan Su, Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", Hindawi Publishing Corporation Advances in Artificial Intelligence Volume 2009, Article ID 421425, 19 pages.

[7] N. Mittal, R. Nayak, M. C. Govil, and K. C. Jain, "Recommender system framework using clustering and collaborative filtering," in Proc. 3rd Int. Conf. Emerging Trends Eng. Technol., Nov. 2010, pp. 555558.

[8] R. D. Simon, X. Tengke, and W. Shengrui, "Combining collaborative filtering and clustering for implicit recommender system, " in Proc. IEEE 27th Int. Conf. Adv. Inf. Netw. Appl., Mar. 2013, pp. 748755.

[9] Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude, "Data providing services clustering and management for facilitating service discovery and

replacement," IEEE Trans. Autom. Sci. Eng., vol. 10, no. 4, pp. 116, Oct. 2013.

[10] M. C.Pham, Y. Cao, R. Klamma, and M. Jarke, "A clustering approach for collaborative filtering recommendation using social network analysis," J. Univ. Comput. Sci., vol. 17, no. 4, pp. 583604, Apr. 2011.

[11] X. Li and T. Murata, "Using multidimensional clustering based collaborative filtering approach improving recommendation diversity," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., Dec. 2012, pp. 169174.

[12] George Thomas, Srujana Merugu, "A scalable collaborative filtering framework based on co-clustering.", In Proceedings of the IEEE ICDM Conference. 2005.

[13] Jerome Kelleher, Derek Bridge, "RecTree Centroid: An Accurate, Scalable Collaborative Recommender", In Procs. of the Fourteenth Irish Conference on Artificial Intelligence and Cognitive Science, pages 89–94, 2003.

[14] Al Mamunur Rashid, Shyong K. Lam, George Karypis, John Riedl, "ClustKNN: A Highly Scalable Hybrid Model& Memory Based CF Algorithm", WEBKDD '06, August 20, 2006, Philadelphia, Pennsylvania, USA, Copyright 2006 ACM.

[15] Badrul M. Sarwar, George Karypis, Joseph Konstan, John Riedl, "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering", Proceedings of the Fifth International Conference on Computer and Information Technology, 2002.

[16] Jiang Zhirao, "Based on Java Technology System and Implement the Personalized Recommendations of the system", Jilin: Jilin University, 2011.

[17] Kuo-Cheng Tseng, Chein-Shung Hwang, Yi-Ching Su, "Using Cloud Model for Default Voting in Collaborative Filtering", Journal of Convergence Information Technology (JCIT) Volume6, Number12, December 2011.

[18] ZHANG Yao, FENG Yu-qiang, "Hybrid Recommendation method IN Sparse Datasets: Combining content analysis and collaborative filtering", International Journal of Digital Content Technology and its Applications (JDCTA) Volume6, Number10,June 2012.