A Survey and Analysis on Recommendation System Algorithms

Tahira Mahboob[1]
Assistant Professor

Fatima Akhtar[2]

Moquaddus Asif [3]

Nitasha Siddique[4]

Memoona Khanum[5]

[1, 2 ,3, 4] Department of Software Engineering
[5] Department of Computer Science
[1, 2 ,3, 4, 5] Fatima Jinnah Women University, Pakistan

ABSTRACT-***The paper is comprehensive survey of methodologies and techniques used for recommendation systems. Recommendation shows significant part in many fields and has become a great attention in the field of research.Important factor for the searching and recommendation system is to recognize and recognizeuser's personalized first choice from their historic rating behaviors. Recommender systems are now common in the research public, where many suggestions or recommendations are provided on the basis of algorithms. These algorithms usually implement differently in several fields and domains. Without properly defined algorithms systems are useless, therefore, it is essential from the research viewpoint, as well as from a real-world view, to be able to decide on an algorithm that exactly matches the domain and the interest of user. In this research paper we emphasized on graphical model based, supervised, semi-supervised and unsupervised algorithms that are necessary part of recommendation system.***

Key words: *context-aware semi-supervised co-training method commonly known as (CSEL), Graph Mode, Bayes rule, expectation Maximization, Collaborative Filterin, supervised learning, unsupervised learning, semi-supervised learning, data clusering, Frequent Term-Based Text Clustering , ROC, Precision, accuracy etc*

I. INTRODUCTION

Information overloaded on the searching and recommendation systems has created huge challenges for users to get more accurate searched and recommended results at the same time under same one system. Several recommender systems employing different algorithms, approaches and methods are used to address some challenges. This research paper emphasizes on graphical model that offers a standard data demonstration and support different recommendation methods and algorithms. Furthermore models and algorithms needed for Bayesian Recommender Systems are discussed that describes how Bayesian methods can be applied to recommendation systems to make optimalrecommendations. A challenging task for recommendation and searching system is to improve the accuracy for the new items for new users. In recommendation system usually data is analyzed about particular item and interactions between users and items are found as a result. This paper focuses on algorithms that are used by researchers and engineers to improve the behavior of recommendation systems.

II-GRAPHICAL MODEL BASED, SUPERVISED, SEMI-SUPERVISED AND UNSUPERVISED TECHNIQUES AND ALGORITHMS

2.1 Addressing Cold Start in Recommender Systems: A Semi-supervised Co-training Algorithm by Mi Zhang, Jie Tang, Xuchen Zhang, Xiangyang Xue

Recommender systems are today's important part in various present applications that expose the user to a large collection of items. These systems recommend user's preference items that it thinks the user will prefer. In this research paper author proposed an algorithm i.e. context-aware semi-supervised co-training method commonly known as (CSEL) to solve cold start problem. To boost the recommendations for user's author further suggest a semi-supervised ensemble learning algorithm. Author [2] focuses on the efficiency and accuracy of recommendations for new articles and new users. Recommender produces error which increases quickly as the popularity of item decreases. It is observed that the most common error of the most unpopular items usually doubles as compared to popular items. This problem is called cold start problem which most of the system suffers. Author suggested a solution by combining content and collaborative data under a single frame based on Bayes classifier. CSEL and context-aware model

provide more precise and accurate predictions. It comprises of three main steps i.e. constructing multiple regresses, co training, assembling the results. In constructing multiple regresses concept of bagging is introduced which presents the learning algorithm with a training set that consists of a sample of k training examples drawn randomly with replacement from the original training set.

2.2 A Graph Model for E-Commerce Recommender Systems by Zan Huang, Wingyan Chung and Hsinchun Chen

In this research paper author describes some challenges that are faced by customers and system. Customers often experience difficulty in getting searching and recommended results at the same time for required product. Large amount of data handling problem arises due to which system fails to show recommendations for particular item. In this research paper author gets solution to this problem by proposing a recommendation system that analyze data that is utilized by the user. Different recommendation approaches are discussed i.e. neighborhood formation, association rule mining, machine learning techniques etc. To overcome two major problems i.e. presenting diverse information and flexibility of the system to incorporate different recommendations author proposed a graphical model that contains nodes of customer & products and respective links of transactions & limitations. There exist three types of links between nodes. Similarity between the products is captured by the link created between nodes. Based on these links and nodes recommendations can be produced by few strong predicted directed paths joining many users. Based on the original graph, a social network based graph is then produced which generated recommendations by comparing both the graphs. In this way author describes the graphical model on basis of which many recommendation methods can be developed.

2.3 A Research Paper Recommender System by Bela Gippi, Joran Beel, Christian Hentschel

This paper focuses on the recommendation techniques and algorithms that produce more accurate and efficient searched and recommended results. Author [3] presents hybrid system that combines two techniques i.e. content based and collaborative techniques. This hybrid system has potential to improve the problem of finding related possible research articles and papers. In this way system will combine citation analysis, explicit ratings, and author analysis and source analysis methods to produce better results. System combines the existing theories with new concepts in order to produce better recommender system. Many weaknesses and drawbacks become obsolete by this hybrid technique. Appropriate algorithms are applied on the set of five inputs such as script/text, references, authors, sources, ratings or documents to receive relevant recommended results. To deal with explicit ratings, the system generated relevant ratings by monitoring 22 actions of user (View Document Details, Download, View Related Documents, and Follow Recommendations etc.); in this way system will able to improve user's own recommendation accuracy. Research paper lacks some non-technical aspects like privacy and security of the ratings and recommendations.

2.4 Review Recommendation with Graphical Model and EM algorithm by Richong Zhang, Thomas Tran

Nowadays due to the evolution of internet technology and business, many web sites are providing services by demanding users to leave reviews. Author [4] describes many challenges for user/reviewers that how they can make best use of online reviews. Common method of rating is done by using number of stars. This type of star rating scale often combines feelings and opinions of user therefore making difficult for user to have real semantic reviews. In this research paper author proposed recommendation technique that uses the probability density and exploit graphical model and expectation Maximization algorithm. For searching purpose search engines are good tools for searching, but it encounters a problem of huge data set which should be solved by any recommendation system which limits the searching by giving recommended results. Mathematical calculations regarding Bayes rule are formulated. Detailed graphical model is represented by author. Respective nodes and links are created in graphical model. Mathematical and experimental results show that the graphical model will effectively calculate the review's helpfulness. This model will be helpful for getting recommended results.

2.5 Research Paper Recommender System Evaluation: A Quantitative Literature Survey by Joeran Beel, Stefan Langer

Recommender systems are becoming popular day by day. They are becoming important part of applications. In the domain of information technology article searching and recommendation system are necessary applications which keep record of user's preferences in the field of research. More the

recommendation algorithms and techniques are offered the more important their evaluation becomes to decide the best technique. In this way strengths and weakness of each approach can be analyzed and determined. Author [5] basically describes the evaluation criteria and different methods under which recommender system get evaluated. Parameters are such as accuracy, user stratification, satisfaction of recommendation provider, which contribute to the evaluation of recommender systems.

2.6 Bayesian Recommender Systems: Models and Algorithms (Shengbo Guo)

This research paper describes Bayesian methods and techniques to make optimal recommendations. Author discusses dimensions of recommendation systems such as preference elicitation, set based recommendations and matchmaking. All the three dimensions mentioned by author suffer from the problem of uncertainty. Author refers solution in Bayesian approach. Author describes Bayesian approach to be very flexible and robust. These techniques and approaches are introduces for complex models and system in order to achieve highly flexible and robust results for required systems. Such recommender system minimizes user's efforts for searching desired items. Various recommender systems have been introduced which solves the problem of uncertainty. Graphical models and Bayesian network are used for modeling different parameters. Research paper includes some issues that are associated with uncertainty using Bayesian approaches.

2.7 Relational Graphical Models for Collaborative Filtering and Recommendation of Computational Workflow Components by William H.Hsu

In this research paper authorwilliamH.Hsu[7] focuses on graphical models which are used to represent rational data in computation method such as myGrid. The objective is to provide users a system which combines collaborative filtering (CF) algorithm. The basic aim of the research paper is to improveand develop such techniques that discover relational and constraint models. The emphasis is on statistical evaluation and algorithms that are necessary for knowing and implementing graphical models. These techniques include Bayesian networks and decision networks which are applied to large variety of problems. Main contribution of this research is the combination of algorithms in order to learn the structure of graphical models and existing techniques. Author describes few systems such as ResearchIndex / CiteSeer that have some restrictions that hinder their direct application.

2.8 Semi-Supervised Learning for Personalized Web Recommender System by Tingshao Zhu

This paper explains that handling large amount of data is often time consuming and difficult for a user to find relevant preference items. Author describes the affective recommendation system that handles large volume of data and helps the user in finding relevant data with their own interest. Such a recommendation system should be designed that generates personalized recommendations according to user preference and interests. To overcome the problem of limitation of labeled data author introduced semi-supervised learning which includes two major steps i.e. training and prediction. Training process make use of all labeled data to infer general prediction function whereas the later process make use of general function to infer labels for unlabeled data. Semi-supervised approach is slightly different from labeled and unlabeled data and makes recommendation in a step. The relationship between labels and unlabeled data is beneficial for making more accurate predictions therefore author suggested using semi-supervised approach for recommendation systems. Preliminary results using semi-supervised prediction model shows that this approach is better than the other models.

2.9 Data clustering: 50 years beyond K-means (By Anil K. Jain)

One of the most fundamental modes of understanding and learning is to organizing data into sensible groupings such that common scheme of scientific classification. The Cluster analysis is the formal study of algorithms and methods for clustering, objects and, groupings. Clusters that are formed based on measured or assumed intrinsic characteristics or similarity. The category labels that tag objects with prior identifiers do not use by cluster analysis such that class labels. The lack of category information differentiates data clustering (unsupervised learning) from discriminate analysis (supervised learning) or classification. The basic purpose of clustering is to discover structure in data. The K-mean is the most simple and popular clustering algorithm. K-mean designs a general purpose clustering algorithm and the ill-posed problem of clustering.

2.10 *Feature Selection for Unsupervised Learning by Jennifer G. Dy and Carla E. Brodley*

For performing feature subset selection for unsupervised learning the wrapper framework is used. The automated feature subset selection algorithm for unlabeled data involved two issues. The first issue is the need for finding the number of cluster with feature selection. The second one is the need for normalizing the bias of feature selection criteria based on dimension. In feature search the need for finding the number of clusters is identified. The proof is provided for the biases of ML and scatter separable based on dimension. As compared to fixing k to be the true number of classes, the feature subset selection process generates better result. This is because of two reasons first is the number of class is not always equal to the Gaussian cluster. The second reason is different feature subset contain different number of clusters. The experiment shows that ML and scatter separable are biased in some way with respect to dimension. So for the chosen feature selection criterion a normalization scheme is needed. These biases can be removed by cross projection criterion normalization scheme. Although the wrapper framework examined by using FSSEM, the feature normalization schemes, the search method, feature selection criteria can be applied to any clustering method. An appropriate search method, clustering and feature selection criteria can be choose depending on the application.

2.11Frequent Term-Based Text Clustering By Florian Beil and Martin Ester

To structure massive sets of text or hypertext documents, text clustering methods can be used. The special problems of text clustering such that understandability of the cluster description, very large size of the databases, and very high dimensionality of the data usually not addressed by text clustering. The proposed novel approach for text clustering uses frequent item (term). The association rule mining is used to discover frequent set effectively. Measure the mutual overlap of frequent sets with relative to the sets of supporting documents in cluster based on frequent term sets. The two algorithms FTC and HFTC are proposed for frequent term-based text clustering. The HFTC create hierarchical clustering and FTC creates flat clustering. The experimental results based on web documents and classical text documents specify that clusters obtained from proposed algorithms comparable quality significantly more efficiently the other art text clustering algorithms.

2.12A Frequent Concepts Based Document Clustering Algorithm By Rekha Baghel and Dr. Renu Dhir

The paper presents that documents are clustered based on frequent concepts. The Frequent Concepts based document clustering technique is a clustering algorithm. Rather than frequent items used in traditional text mining techniques, FCDC work with frequent concepts. The documents are treated as beg of word in traditional clustering algorithm and it neglect the important relationships between words like synonyms. The semantic relationship is utilizes between the word to create concepts. Low dimensional feature vector are created by using Word Net ontology. This allows developing effective clustering algorithm. Cluster text documents having common concepts uses a hierarchical approach.The comparison between traditional clustering algorithm and FCDC shows that FCDC is more accurate, scalable and effective. So even documents do not contain common words our proposed algorithm is able to group documents in the same cluster.

III- ANALYSIS

In Table-1 essential parameters are discussed below along with their respective meanings and possible values i.e Yes: Y, No: N, Not Discussed: ND. Detail analysis of parameters is mentioned in the Table-2 Necessary parameters such as accuracy in terms of prediction, performance, efficiency, user satisfaction, robustness, and satisfaction of recommender provider, effortless and timing constraints are discussed in table 1. Almost all algorithms are efficient such that system utilizes minimum time and memory. According to MiZhang [2] prediction should be accurate so that precise results are achieved. JoeranBeel[3] and Stefan Langer[3] focus on significant factor that contributes to a good recommender system is its capacity to provide satisfaction to the user. Whenever user search for an article system should be capable of showing related recommendations in this way user satisfaction can be achieved. They further emphasized on the satisfaction of the Recommendation Provider. While addressing cold start MiZhang [2] describes diversity is an essential property to achieve good performance of recommendation systems. ShengboGuo[2] also discussed diversity factor as the important part of recommender systems. They observed inverse relationship between diversity and accuracy. Almost all authors discussed timing constraints in their research papers. Each algorithm or method is associated with specific time under which it will produce best results

Table 1.Evaluation Parameters for Article Recommendation Systems

Serial#	Parameters	Meanings	Possible Values
1	***Efficiency***	A level of performance that describes a process that uses the lowest amount of inputs to create the greatest amount of outputs in minimum time & memory	Yes, No, Not Discussed
2	***Accuracy in terms of prediction***	Accuracy of algorithms should be maintained. Error rate should be minimized.	Yes, No, Not Discussed
3	***User stratification***	User specifications and needs must be satisfied	Yes, No, Not Discussed
4	***Automatable***	Methods describes are automatable which reduces manual work	Yes, No, Not Discussed
5	***Robustness***	Specification are may not be covered and but appropriate performance of a system	Yes, No, Not Discussed
6	***Integration***	Integration of system allows combination of 2 concepts such that system is capable of producing better results.	Yes, No, Not Discussed
7	***Flexible***	Flexibility refers to designs that can adapt when external changes occur.	Yes, No, Not Discussed
8	***Performance***	Backtracking or recovery process defines the performance of the system	Yes, No, Not Discussed
9	***Satisfaction of the Recommendation Provider***	User should be satisfied with recommended results. Relevant information should be provided in order to win out user satisfaction.	Yes, No, Not Discussed
10	***Diversity***	The concept of diversity encompasses acceptance and unique.	Yes, No, Not Discussed
11	***Timing constraint***	Appropriate timing is associated with every algorithm	Yes, No, Not Discussed
12	***Effortless***	The quality of a system that makes the user to use it easily	Yes, No, Not Discussed
13	***Optimization***	Optimization is the process of modifying a system to make some features of it work more efficiently or use fewer resources	Yes, No, Not Discussed
14	***Error rate***	It measures the total number of incorrect predictions against the total number of predictions	Yes, No, Not Discussed
15	***Precision***	It is defined where datasets are much unbalanced.	Yes, No, Not Discussed
16	***Recall***	It is the proportion of the number of data items that system selected as the positive	Yes, No, Not Discussed
17	***F1-Score***	For optimization F1 score combines both recall and precision with equal importance into a one parameter.	Yes, No, Not Discussed
18	***Receiver Operating Characteristic (ROC) graph***	It is technique to organize, visualize, and select classifiers that depend on their performance in 2D space.	Yes, No, Not Discussed

Note: Yes/Y, No/N, Not Discussed/ND

Table 2.Analysis of existing techniques for Article Recommendation Systems

Serial #	Authors	(ROC)	F1-Score	Recall	Precision	Error rate	Accuracy	Optimization	Effortless	Timing constraint	Satisfaction of the Recommendation	Diversity	performance	Flexible	Integration	Robustness	Automatable	User stratification	Accuracy in terms of	Efficiency
1	Mi Zhang	ND	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
2	Zan Huang, Wingyan Chung	ND	Y	N	Y	Y	Y	N	Y	Y	N	N	N	N	Y	Y	N	Y	Y	Y
3	BelaGippi, JoranBeel,Jan 2009	N	ND	Y	N	Y	Y	N	Y	Y	N	Y	N	Y	Y		Y	N	Y	Y
4	Richong Zhang, Thomas Tran,April 2010	N	Y	Y	Y	Y	N	N	Y	N	Y	Y	Y	N	Y	Y	Y	Y	Y	N
5	JoeranBeel, Stefan Langer	N	N	Y	N	Y	Y	Y	Y	N	Y	N	Y	N	Y	N	Y	N	Y	Y
6	ShengboGuo, Oct 2011	Y	Y	ND	Y	N	Y	N	Y	Y	N	Y	Y	Y	Y	Y	N	Y	N	Y
7	William H.Hsu	ND	N	Y	Y	N	N	Y	Y	ND	Y	Y	Y	Y	Y	N	Y	Y	N	Y
8	A .K .Jain	N	N	Y	N	Y	Y	Y	N	Y	N	Y	Y	Y	N	Y	N	Y	Y	Y
9	J.G.Dy and C.E.Brodley	N	N	Y	Y	Y	N	Y	N	Y	Y	Y	N	Y	N	Y	Y	Y	N	Y
10	F.Beil, M.Ester and X.Xu	N	Y	Y	ND	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	ND	Y	Y	Y
11	Rekha Baghel and Dr. Renu Dhir	N	Y	Y	ND	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	ND	Y	Y	Y
12	Z. Huang, W. Chung and H. Chen."	N	ND	N	N	Y	Y	Y	Y	Y	ND	Y	Y	N	Y	Y	N	Y	Y	N

IV- CONCLUSION

From our survey research for the algorithms for recommendation system we conclude that majority of techniques will produce effective results and will minimize those factors which lead to the better generation of recommended results. However there are some techniques which lack tool support and their automation along with their application in other domains. Therefore we suggest that any approach or method that is used in recommendation system must be able to give accurate, efficient results. Main approaches of graphical models and supervised semi supervised and unsupervised are discussed. Further more benefits and drawbacks of algorithms are presented by authors to compare the performance of each algorithm. It is concluded that user can have better recommended results if two or more related algorithms are combined. Graphical models and Bayesian models are emerging algorithms for making recommendations more personalized.

V- REFERENCES

[1] S. Guo. "Bayesian Recommender Systems: Models and Algorithms." PhD thesis, Australian National University, Australia, 2011.

[2] M. Zhang, J. Tang, X. Zhang and X. Xue."Addressing Cold Start in Recommender Systems: A Semi-supervised Co-training Algorithm." in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 73-82.

[3] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger and A. Nurnberger." Research Paper Recommender System Evaluation: A Quantitative Literature Survey." in Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, 2013, pp.15-22.

[4] R. Zhang and T. Tran. "Review Recommendation with Graphical Model and EM Algorithm." in *Proceedings of the 19th international conference on World wide web*, 2010, pp.1219-1220.

[5] B. Gipp, J. Beel and C. Hentschel."Scienstein: A Research Paper Recommender System." in *Proceedings of the International Conference on Emerging Trends in Computing*, 2009, pp. 309-315.

[6] Z. Huang, W. Chung and H. Chen."A graph model for E-commerce recommender systems."*Journal of the American Society for Information Science and Technology*, vol. 55, pp. 259-274, February 2004.

[7] W. H. Hsu. "Relational Graphical Models for Collaborative Filtering and Recommendation of Computational Workflow Components." in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Multi-Agent Information Retrieval and Recommender*, 2005, pp. 18-22.

[8] Tingshao Zhu. "Semi-supervised learning for personalizedWeb recommender system."in*Graduate University of Chinese Academy of Sciences*,Revised manuscript received 24 March 2010.

[9] A .K .Jain. "Data clustering: 50 years beyond K-means." *Elsevier B.V*,2009. Available:http://webcache.googleusercontent.com/search?q=cache:HShPxKwW_HkJ:citeseerx.ist.psu.edu/viewdoc/download%3Fdoi%3D10.1.1.332.3907%26rep%3Drep1%26type%3Dpdf+&cd=1&hl=en&ct=clnk&gl=pk

[10]J.G.Dy and C.E.Brodley."Feature Selection for Unsupervised Learning."*Journal of Machine Learning Research 5*,2004. Available:www.jmlr.org/papers/volume5/dy04a/dy04a.pdf

[11]F.Beil, M.Ester and X.Xu."Frequent Term-Based Text Clustering."*ACM*,2002. Available: www.jmlr.org/papers/volume5/dy04a/dy04a.pdf

[12] R.Baghel and R.Dhir. "A Frequent Concepts Based Document Clustering Algorithm." *International Journal of Computer Applications*, Volume 4 , No.5, July 2010. Available : ijcaonline.org/volume4/number5/pxc3871171.pdf