International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Research Paper Recommender System Evaluation Using Coverage

Shraddha B. Shinde[1], Mrs. M. A. Potey[2]

[1]Department of Computer Engineering
D.Y.Patil College of Engineering & Research Center
Akurdi, Pune, India
shinde.shraddha97@gmail.com
[2]Department of Computer Engineering
D.Y.Patil College of Engineering & Research Center
Akurdi, Pune, India
mapotey@gmail.com

Abstract - *Recommendation systems(RS) support users and developers of various computer and software systems to overcome information overload, perform information discovery tasks and approximate computation, among others. Recommender systems research is frequently based on comparisons of predictive accuracy: the better the evaluation scores, the better the recommender. However, it is difficult to compare results from different recommender systems due to the many options in design and implementation of an evaluation strategy. Additionally, algorithmic implementations can separate from the standard formulation due to manual tuning and modifications that work better in some situations. It have been compared common recommendation algorithms as implemented in three popular recommendation frameworks. We evaluate the quality of recommender systems, most approaches only focus on the predictive accuracy of these systems. Recent works suggest that beyond accuracy there is a variety of other metrics that should be considered when evaluating a RS. This paper reviews a range of evaluation metrics and measures as well as some approaches used for evaluating recommendation systems. Analysis shows that large differences in recommendation accuracy across frameworks and strategies. we are developing the recommender system for research papers using coverage.*

Key Words: **Recommender System, Research Paper Recommender System, Evaluation, Metrics, Coverage.**

1. INTRODUCTION

Recommender Systems (RSs) or recommendation systems (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that user would give to an item. Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles and search queries. Recommender Systems (RSs) can now be found in many modern applications that expose the user to a huge collections of items. This systems typically helps to provide the user with a list of recommended items they might prefer, or supply guesses of how much the user might prefer each item. These systems help users to decide on appropriate items, and ease the task of finding preferred items in the collection.

Recommender Systems are designed to suggest users the items that best fit the user needs and preferences. Recommender systems typically produce a list of recommendations in one of two ways - through collaborative or content-based filtering. Among Recommendation Systems techniques, the two most popular categories are content-based filtering and collaborative filtering and their hybridization is called hybrid approach. Knowledge based is the third category of Recommender Systems. These recommender systems deal with two types of entities, users and items. Recommendation process are entirely based on the input (rating, user profile) provided by the visitors or users.

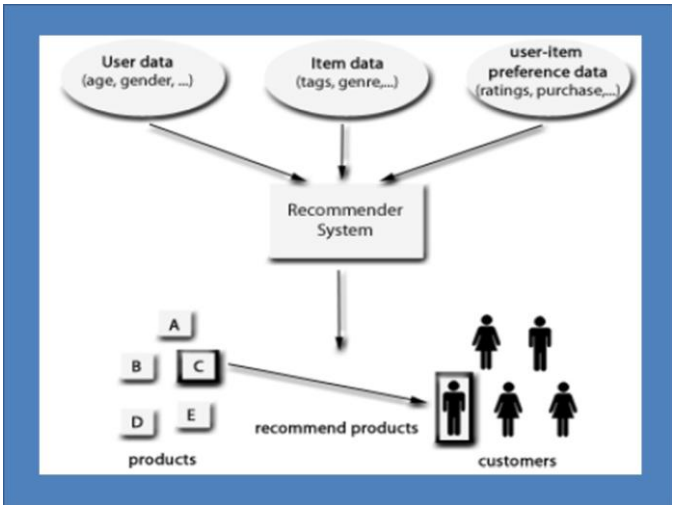


Fig-1: Recommendation Process[4]

1.1 Research Paper Recommender System

Recommender systems for research papers are becoming increasingly popular. In the past 14 years, over 170 research articles, patents, web pages, etc. Were

published in this field. Interpolating from the numbers of published articles in the year, 30 new publications were estimated to appear in 2013 (Figure 2)[1]. Recommender systems for research articles are useful applications, which for instance help researchers keep track of their research field. The more recommendation approaches are proposed, the more important their evaluation becomes to determine the best approaches and their individual strengths and weaknesses.

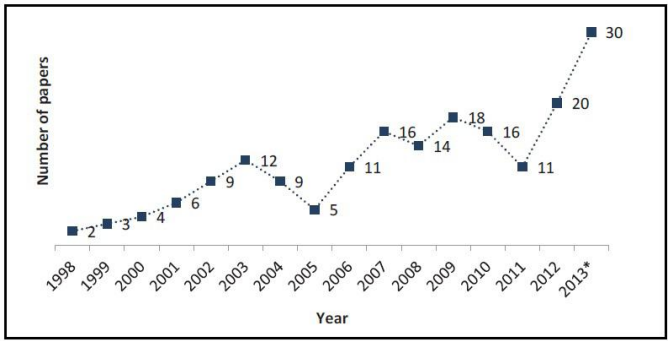


Fig-2: Published papers per year[1]

1.2 Evaluation Metrics

Evaluating recommender systems requires a definition of what constitutes a good recommender system, and how this should be measured. There is mostly consensus on what makes a good recommender system and on the methods to evaluate recommender systems. The more recommendation approaches are proposed, the more important their evaluation becomes to determine the best performing approaches and their individual strengths and weaknesses. To determining the 'best' recommender approach there are three main evaluation methods, namely user studies, online evaluations, and offline evaluations to measure recommender systems quality.

In user studies, users explicitly rate recommendations generated by different algorithms and the algorithm with the highest average rating is considered the best algorithm. It is important to note that user studies measure user satisfaction at the time of recommendation. Users do not measure the accuracy of a recommender system because users do not know, at the time of the rating, whether a given recommendation really was the most relevant. In online evaluations, recommendations are shown to real users of the system during their session. Offline evaluations use pre-compiled offline datasets from which some information has been removed. Subsequently, the recommender algorithms are analyzed on their ability to recommend the missing information. There are three types of offline datasets, which is define as (1) true-offline datasets, (2) user-offline dataset, and (3) expert-offline datasets.

Many evaluation metrics have been suggested for comparing recommendation algorithms and, most researchers who suggest new recommendation algorithms also compare the performance of the new algorithm to a set of existing approaches. Such evaluations are typically performed by applying some evaluation metric that provides a ranking of the candidate algorithms (usually using numeric scores). Most frequently used metrics for evaluating recommendation approaches for correctness in each scenario.

a] Predicting User Ratings:

If the recommendations produced are intended to predict how users rate items of interest then Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) metrics are often used. When calculating RMSE, the difference between actual user ratings and predicted ratings should be determined. MAE, on the other hand, measures the average absolute deviation of predicted ratings from user ratings.

b] Ranking Items:

Ranking measures are used when an ordered list of recommendations is presented to users according to the their preferences. This order can be as the most important, or `most relevant', items at the top and the `least relevant' items at the bottom. When checking for correctness of ranking measures, if a reference ranking (benchmark) is available, the correctness of the ranking can be measured by Normalized Distance based Performance Measure (NDPM). The value returned by NDPM is between 0 and 1 with any acceptable ranking having a distance of 0. A frequently used metric for measuring ranking correctness, considering item ranking position, is Normalized Discounted Cumulative Gain (NDCG). It is calculated based on measuring Discounted Cumulative Gain (DCG) and then comparing that to the ideal ranking. DCG measures the correctness of a ranked list based on the relevance of items discounted by their position in the list.

c] Recommending Interesting Items:

If a recommendation system is providing the items that users may like to use, a common approach to evaluate it is to use classification metrics like precision, recall, accuracy and false positive rate and these metrics have been used excessively across different domains.

When Recommendation systems make recommendations by searching available information spaces

then Coverage is used. Coverage refers to the proportion of available information (items, users) that recommendations can be made for. Coverage usually refers to catalogue coverage (item-space coverage) or prediction coverage (user-space coverage). Catalogue coverage is the proportion of available items that the recommendation system recommends to users. Prediction coverage, on the other hand, refers to the proportion of users or user interactions that the recommendation system is able to generate predictions for. Therefore, if the set of items recommended to a user over a particular recommendation session is Sr and Sa is the set of all available items, catalogue coverage can be calculated by the following formula:

$$\text{Catalogue Coverage} = \frac{S_r}{S_a}$$

Similar to catalogue coverage, prediction coverage can be calculated by measuring the proportion of users that prediction can be made for (Sp) to a set of available users (Su).

$$\text{Prediction Coverage} = \frac{S_p}{S_u}$$

2. LITERATURE SURVEY

The literature on Research Paper Recommender System evaluation offers a large variety of evaluation metrics and using this evaluation metrics to evaluate the performance of recommender systems.

Joeran Beel et.al.[1] presents a quantitative literature survey on Research Paper Recommender System Evaluation. Over 80 approaches for academic literature recommendation exist today. The approaches were introduced and evaluated in more than 170 research articles, as well as patents, presentations and blogs. These approaches were reviewed and found most evaluations to contain major shortcomings. They have concluded that it is currently not possible to determine which recommendation approaches for academic literature are the most promising. However, there is little value in the existence of more than 80 approaches if the best performing approaches are unknown.

In the last sixteen years, more than 200 research articles were published about research-paper recommender systems and Bela Gipp et.al.[2] have reviewed these articles and present some descriptive statistics in this paper, as well as a discussion about the major advancements and shortcomings and an overview of the most common recommendation concepts and approaches. They found that more than half of the recommendation approaches applied content-based filtering (55%). Collaborative filtering was applied by only 18% of the reviewed approaches, and graph-based recommendations by 16%. Other recommendation concepts included stereotyping, item-centric recommendations, and hybrid recommendations. The content-based filtering approaches mainly utilized papers that the users had authored, tagged, browsed, or downloaded. TF-IDF was the most frequently applied weighting scheme. They have concluded that several actions could improve the research landscape: developing a common evaluation framework, agreement on the information to include in research papers, a stronger focus on non-accuracy aspects and user modeling, a platform for researchers to exchange information, and an open-source framework that bundles the available recommendation approaches.

Alan Said et.al.[3] have compared the common recommendation algorithms as implemented in three popular recommendation frameworks. To provide a fair comparison, they have completed the control of the evaluation dimensions being benchmarked: dataset, data splitting, evaluation strategies, and metrics. They have also include results using the internal evaluation mechanisms of these frameworks.

Iman Avazpour et.al.[4] have reviewed a range of evaluation metrics and measures as well as some approaches used for evaluating recommendation systems. The metrics presented in this paper are grouped under sixteen different dimensions, e.g., correctness, novelty, coverage. They have reviewed these metrics according to the dimensions to which they correspond. A brief overview of approaches to comprehensive evaluation using collections of recommendation system dimensions and associated metrics is presented. They also provide suggestions for key future research and practice directions.

Mouzhi Ge et.al.[5] have focussed on two crucial metrics in RS evaluation: coverage and serendipity. Based on a literature review, they first discussed both measurement methods as well as the trade-off between good coverage and serendipity and then analyze the role of coverage and serendipity as indicators of recommendation quality, present novel ways of how they can be measured and discussed how to interpret the obtained measurements. Overall, they argue that their new ways of measuring these concepts re ect the quality impression perceived by the user in a better way than previous metrics thus leading to enhanced user satisfaction.

Asela Gunawardana and Guy Shani[13] have been suggested many evaluation metrics for comparing recommendation algorithms. The decision on the proper evaluation metric is often critical, as each metric may favour a different algorithm. In this paper they reviewed the proper construction of offline experiments for deciding on the most appropriate algorithm and discussed three important tasks of recommender systems, and classify a set of appropriate well known evaluation metrics for each task. they demonstrate how using an improper evaluation metric can lead to the selection of an improper algorithm for the task of interest. They also discussed other important considerations when designing offline experiments.

Alejandro Bellogin et.al.[9] have compared five experimental methodologies, and their experiments with

three state-of-the-art recommenders, four of the evaluation methodologies are consistent with each other and differ from error metrics, in terms of the comparative recommenders performance measurements.

Elena Gaudioso et.al.[15] have proposed a such framework, attempting to extract the essential features of recommender systems. In this framework, the most essential feature is the objective of the recommender system. Next, they will shows that a new metric emerges naturally from this framework. Finally, they have comparing the properties of this new metric with the traditional ones and then evaluate the whole range of recommender systems with this single metric.

Joost de Wit[27] have presented three types of prediction strategies that can be used to predict a user's rating for an item that is not yet rated by him and also focused on how the performance of recommender systems is currently evaluated as described in the literature. This evaluation is important in order to compare different recommender systems and to be able to improve recommender system performance.

3. PROPOSED SYSTEM

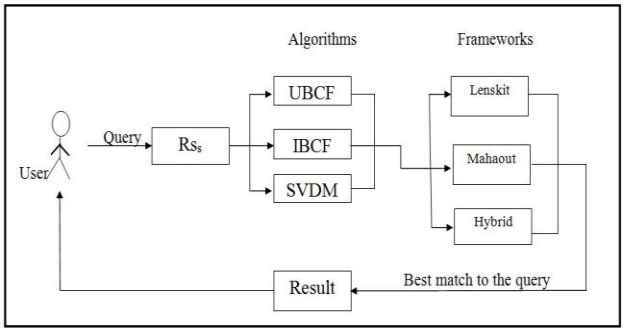


Fig-3: Proposed System Architecture

Where,
RSs-Recommendation Systems.
UB CF-User-Based Collaborative Filtering Algorithm.
IB CF-Item-Based Collaborative Filtering Algorithm.
SVD MF-SVD Based Matrix Factorization Algorithm.

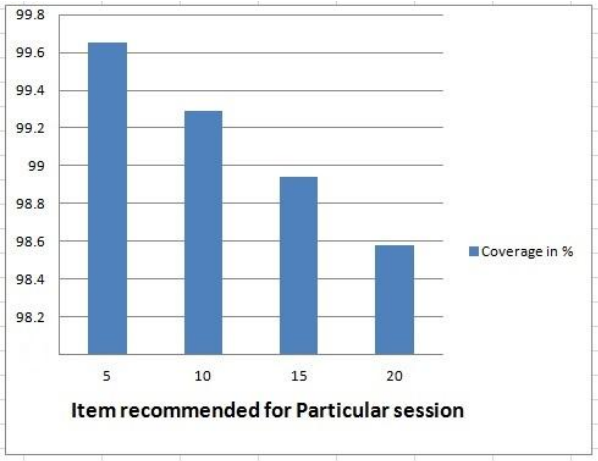
Fig. 3 shows the proposed system architecture. In this architecture user first gives query to the Recommendation System (RS), then Recommendation System produced results using three common state-of-the-art methods from the recommender systems literature namely user-based collaborative filtering (CF), item-based CF and matrix factorization. It have been compared common recommendation algorithms as implemented in three popular recommendation frameworks. Then find which algorithm gives best match results to the user query and return results to the user.

4. RESULT AND DISCUSSION

For Evaluating the Recommendation System many evaluation metrics have been used like precision, recall and F-measure. We are evaluating the Recommendation System using Coverage Parameter. Table 1 shows the result of Coverage as per the items recommended in particular session.

Table 1: Performance of Coverage in Recommendation System.

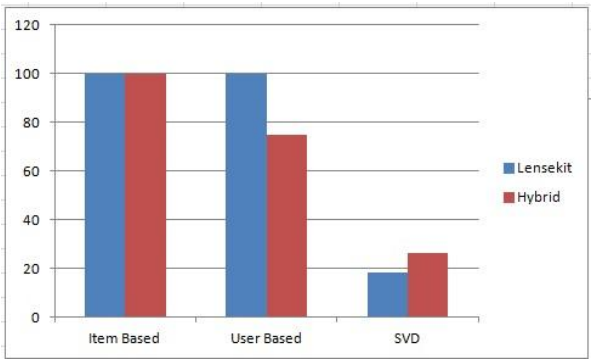
Item Recommended for particular session	Value (%)
5	99.65
10	99.29
15	98.94
20	98.58



Recommendation System produced results using three common state-of-the-art methods from the recommender systems literature namely user-based collaborative filtering (CF), item-based CF and matrix factorization. It have been compared common recommendation algorithms as implemented in three popular recommendation frameworks such as Lenskit, Mahout and MyMediaLite using DBLP dataset. We are proposed new Hybrid framework which contains combination of both Lenskit and Mahout framework. Below graph shows performance of Hybrid against Lenskit. Our analysis shows Hybrid framework gives better performance than other.

Table 2: Framework Accuracy in Recommendation System.

Algorithm	Lenskit	Hybrid
Item Based	100	100
User Based	100	74.99
SVD	18.18	26.086



5. CONCLUSION AND FUTURE SCOPE

We studied a range of common metrics used for the evaluation of recommendation systems in software engineering. Based on a review of current literature, derived a set of dimensions that are used to evaluate individual recommendation systems or in comparing it against the current state of the art. We used coverage, F-Measure metrics to improvise the performance of recommender system for Research paper. Evaluation results shows the disparity between three common recommendation frameworks. Even though the frameworks implement similar algorithms, there exist large differences in the reported recommendation quality. It have been evaluated three types of recommendation algorithms (user-based and item-based CF and SVD-based matrix factorization) using popular and publicly available DBLP dataset. Finally, the content of this report can be used for understanding the evaluation criteria for recommendation systems and this can be improve the decisions when selecting a specific recommendation system for a software development project.

In future the system performance can be enhanced by using new metrics rather than existing metrics for evaluating the recommender system.

ACKNOWLEDGEMENT

We would like to thank the publishers, researchers and teachers for their guidance. We would also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to friends and family members for their support.

REFERENCES

[1] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nurnberger. Research paper recommender system evaluation: a quantitative literature survey. In Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, pages 15-22. ACM, 2013.

[2] Bela Gipp, Joran Beel, and Christian Hentschel. Scienstein: A research paper recommender system. In Proceedings of the international conference on Emerging trends in computing (ICETiC'09), pages 309-315, 2009.

[3] Alan Said and Alejandro Bellogin. Comparative recommender system evaluation: benchmarking recommendation frameworks. In Proceedings of the 8th ACM Conference on Recommender systems, pages 129-136. ACM, 2014.

[4] Iman Avazpour, Teerat Pitakrat, Lars Grunske, and John Grundy. Dimensions and metrics for evaluating recommendation systems. In Recommendation Systems in Software Engineering, pages 245-273. Springer, 2014.

[5] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In Proceedings of the fourth ACM conference on Recommender systems, pages 257-260. ACM, 2010.

[6] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6):734-749, 2005.

[7] L Anitha, M Kavitha Devi, and P Anjali Devi. A review on recommender system. International Journal of Computer Applications, 82(3), 2013.

[8] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nurnberger, and Bela Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, pages 7-14. ACM, 2013.

[9] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In Proceedings of the fifth ACM conference on Recommender systems, pages 333-336. ACM, 2011.

[10] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Papadopoulos, and Roberto Turrin. Comparative evaluation of recommender system quality. In CHI'11 Extended Abstracts on Human Factors in Computing Systems, pages 1927-1932. ACM, 2011.

[11] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. User-centric vs. system-centric evaluation of recommender systems. In Human-Computer

Interaction-INTERACT 2013, pages 334-351. Springer, 2013.

[12] Francois Fouss and Marco Saerens. Evaluating performance of recommender systems: An experimental comparison. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, volume 1, pages 735-738. IEEE, 2008.

[13] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. The Journal of Machine Learning Research, 10:2935-2962, 2009.

[14] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1):5-53, 2004.

[15] Felix Hern_andez del Olmo and Elena Gaudioso. Evaluation of recommender systems: A new approach. Expert Systems with Applications, 35(3):790-804, 2008.

[16] FO Isinkaye, YO Folajimi, and BA Ojokoh. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3):261-273, 2015.

[17] Daniel Kluver and Joseph A Konstan. Evaluating recommender behavior for new users. In Proceedings of the 8th ACM Conference on Recommender systems, pages 121-128. ACM, 2014.

[18] David W McDonald. Evaluating expertise recommendations. In Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, pages 214-223. ACM, 2001.

[19] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In CHI'06 extended abstracts on Human factors in computing systems, pages 1097-1101. ACM, 2006.

[20] Deuk Hee Park, Hyea Kyeong Kim, Jae Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. A review and classi_cation of recommender systems research. School of Management, KyungHee University, Seoul, Korea, IPEDR, 5, 2011.

[21] RVVSV Prasad and V Valli Kumari. A categorical review of recommender systems. International Journal of Distributed and Parallel Systems, 3(5):73, 2012.

[22] Paul Resnick and Hal R Varian. Recommender systems. Communications of the ACM, 40(3):56-58, 1997.

[23] Giancarlo Ruffo and Rossano Schifanella. Evaluating peer-to-peer recommender systems that exploit spontaneous affinities. In Proceedings of the 2007 ACM symposium on Applied computing, pages 1574-1578. ACM, 2007.

[24] Gunnar Schroder, Maik Thiele, and Wolfgang Lehner. Setting goals and choosing metrics for recommender system evaluations. In In CEUR Workshop Proc., volume 811, pages 78-85, 2011.

[25] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In Recommender systems handbook, pages 257-297. Springer, 2011.

[26] Lei Shi. Trading-off among accuracy, similarity, diversity, and long-tail: a graph-based recommendation approach. In Proceedings of the 7th ACM conference on Recommender systems, pages 57-64. ACM, 2013.

[27] Joost Wit. Evaluating recommender systems: an evaluation framework to predict user satisfaction for recommender systems in an electronic programme guide context. 2008.

[28] Zied Zaier, Robert Godin, and Luc Faucher. Evaluating recommender systems. In Automated solutions for Cross Media Content and Multi-channel Distribution, 2008. AXMEDIS'08. International Conference on, pages 211-217. IEEE, 2008.

[29] Markus Zanker, Matthias Fuchs, Wolfram H□opken, Mario Tuta, and Nina M□uller. Evaluating recommender systems in tourism a case study from austria. Information and communication technologies in tourism 2008, pages 24-34, 2008.

[30] https://snap.stanford.edu/data/com-DBLP.html