



A Semantic Approach to Cold Start Problem in Personalized Recommendation Using Hybrid Rating Prediction Method

Kruti Jani	Asst. Prof. SCPPGICA, Ahmedabad, Research Student(1627), HNGU, Patan
Dr. V.M. Chavda	Principal, NPCCSM, Kadi

ABSTRACT	<p>Recommendation System with Semantic Web has changed the traditional approaches suffer from various shortcomings like sparsity, cold start, shilling effect and limited content analysis. Semantic Web technology semantically interlinked data or concepts which may help to know more about particular item or its category, by leveraging knowledge of Linked Open Data Cloud which consists of domain specific and cross domain interconnected datasets. LOD serves as a good source of knowledge if it can be combined with the other small datasets that are specifically designed for that domain. This research approach proposes a hybrid recommender as a rating predictor which utilizes DBpedia Knowledge Base to address cold start problem in traditional personalized recommendation.</p>
-----------------	--

KEYWORDS	Semantic Web, recommendation, DBpedia, Linked Open Data Cloud
-----------------	---

1. INTRODUCTION

Personalization and Recommendation systems are a solution to the problem of content overload, especially in large information systems. Due to that, need of personalization and filtering systems is growing rapidly [1,2]. To address the requirement of effective web navigation, web sites provide personalized recommendations to the end users. Recommendation systems constitute a specific type of information filtering technique that attempt to present items according to the interest expressed by a user. In general, there exist two basic types of recommendation techniques, namely content based filtering and collaborative filtering [2]. Content-based filtering methods examine items previously favored by the actual user, collaborative filtering computes recommendations based on the information about similar items or users [3].

Semantic Web technology (SW) [6, 8] provides semantically interlinked data or concepts which may help to know more about one particular item or its category. LOD [2,7] serves as a good source of knowledge if it can be combined with the other small datasets that are specifically designed for that domain.

This approach combine both techniques into a hybrid approach, where supplementary content features extracted from Linked Open Dataset (DBpedia[4] and LinkedMDB[5]) employed to improve the accuracy of collaborative filtering. This research approach proposes a hybrid recommender which can be used as a rating predictor as well as utilizes DBpedia Knowledge Base to address cold start, sparsity and limited content analysis challenges in traditional personalized recommendation. by using a preprocessing technique for sparsity removal.

2. RELATED WORK

Since 90's research on Recommender System is in progress [1]. Yet, traditional RSs suffer from various limitations like sparsity, cold start, shilling effect. etc. It also bears from Content/Feature generation problem i.e. appropriate feature collection/extraction[7,8]. The beginning of Semantic Web [6,8] provides software agents that are capable to extract more meaningful, linked. Researcher in integrated this Semantic Systems in Recommender System.

While Personalization and Recommendation systems solve the problem of content overload, Content based RS suffered mainly from over generalization problem Collaborative Filtering RS have to face the cold start problem [1, 5]. Hybridization

of both former RS improves of all the stated problems [9]. In 2010, Heitman et. Al. [2] proposed collaborative method for integration of FOAF information with item information about DBpedia Ontology. Ostuni et. al [3] combine both the rating information and item features by learning the path connected to a particular user and his/her liked items along with its features. In proposed work uses DBpedia and LinkedMDB for movie's information to generate features for content based RS. To remove the sparsity problem it also used method in [3,4] to fill the sparse matrix and address the problem of the cold-start using neighborhood liking. To check the accuracy of the method proposed methods compares it to the existing LOD based RSs.

3. PROBLEM FORMULATION

Cold start challenge in collaborative and content based filtering is a situation in which there are not enough ratings for items in order to generate recommendations. In The content-based approach the user has to rate item either explicitly (by querying the user) or implicitly (by observing the user's behavior). In both cases, the cold start problem would imply that the user has to dedicate an amount of effort before the system can start providing any intelligent recommendations. The collaborative filtering approach fails to consider items which no one in the community has rated previously. Adopting a **hybrid approach** between content-based and collaborative filtering reduce the cold start problem where New items (which have not yet received any ratings from the community) would be assigned a rating automatically, based on the ratings assigned by the community to other similar items. Item similarity would be determined according to the items' content-based characteristics.

A proposed new model of personalized recommender System by leverage the current emerging Semantic Web technology and deals with the current challenge of cold start and sparsity in the Recommender Systems. Semantic Web technology (SW) provides semantically interlinked data or concepts which may help to know more about one particular item or its category. LOD serves as a good source of knowledge if it can be combined with the other small datasets that are specifically designed for that domain.

4. HYBRID RATING PREDICTION

In proposed idea for generalized hybrid Recommender system with LOD dataset, the base recommenders follow two user based, item based collaborative filtering strategies and different content-based strategies using various extracted feature sets from DBpedia.

Content information forwarded by the contributing recommender should be in an appropriate data format to be processed by the actual collaborative recommender and information loss through data transformation should be avoided. The following figure displays designated hybrid architecture:

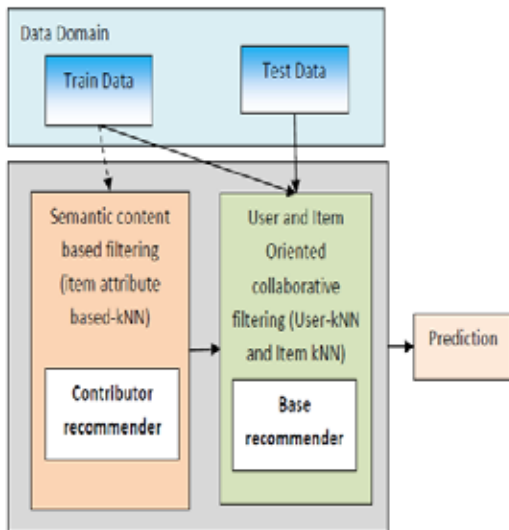


Figure 4.1: Proposed hybrid recommender system

In the first step of this hybrid model, content based method apply item attribute k-NN algorithm and extracts item and its related information from DBpedia [4], LinkedMDB [5] knowledge base. Due to lack of rating data in LOD cloud, MovieLens [6] rating dataset is mapped with semantic movie features of LinkedMDB [5] and DBpedia [4]. The newly generated content feature matrix is normalized with subtractive and multiplicative method and high dimensionality is removed through SVD based Matrix factorization technique. The second step collaborate with former step by applying user based k-NN and item based k-NN collaborative algorithm to compute semantic similarities and predict unknown user-item ratings and item recommendation. Experiments show incorporating item features from LOD cloud (DBpedia[4], LinkedMDB[5]) and weighted hybrid rating predictor and recommender improve result in limited content analysis, cold start and sparsity problem.

Two User and Item based collaborative and semantic content based filtering is combined and predicts the unknown rating. The item (movie) profile expansion is a done by using semantic features retrieved from DBpedia and LinkedMDB knowledge base in semantic content based filtering technique. Collaborative filtering technique applied on extended rating matrix and find item-user similarity for rating prediction.

5. IMPLEMENTATION

For the implementation of the proposed approach, the Recommendation Extension [9] integrated with LOD extension [1,2,9] provided by Rapidminer5.3 are used. The LOD Extension works with datasets that provide a SPARQL endpoint, i.e., a web service which delivers data using the query language SPARQL. A list of such datasets is available at datahub.io. Furthermore, datasets not offering SPARQL endpoints may also be used via dereferencing URIs. Data collected from Linked Open Data sources used as input to the mining process, e.g., reading data from the LinkedMDB and DBpedia Linked Open Data endpoint and running analysis on the data with RapidMiner. The fundamental rating prediction algorithm conceives k-NN classification and cosine similarity for the base recommenders' implementations.

The first phase of feature retrieval process deals with possible knowledge sources from the Cloud of linked Data that we can use to enrich dataset. The secondary dataset used

for experiments was DBpedia, one of the most successful initiatives developed based on the Linked Open Data (LOD) principles. Here system used the English version of DBpedia 3.9. The Linked Movie Database (LinkedMDB) is movie-related database represented in RDF which can be queried using SPARQL.

The second phase of rating prediction describes a dataset from MovieLens, which serves as basis for proposed LOD based hybrid rating predictor. Rating prediction implements SVD based matrix factorization and collaborates different item based and user based k-NN algorithm with content feature extracted by content filtering in first phase of feature retrieval.

6. EVALUATION AND RESULTS

6.1 Evaluation metrics

Proposed hybrid recommender system estimates user ratings and actual values are compared with test dataset (MovieLens) [6]. Proposed method employs RMSE as performance measure. The rating prediction accuracy of the evaluated methods was assessed using RMSE (root mean square error) that puts more weight on larger errors. RMSE is calculated by measuring the average square error between predictions and the actual ratings for all items in the test set:

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{(u,i)} (p_{ui} - r_{ui})^2} \quad (1)$$

6.2 NNH-SVD evaluation

This NNH-SVD approach reduce the dimension of the rating/feature information by means of singular value decomposition.

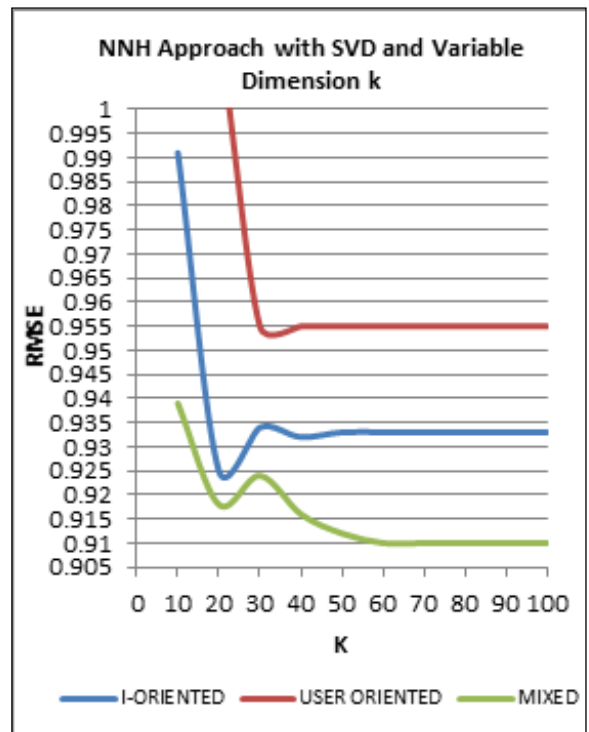


Figure 6.1: NNH Approach with SVD and Variable Dimension k

As a dimension of k = 20 is the optimal adjustment for the item-oriented implementation, and a dimension of k = 10 is the most favorable setting for the user-oriented as well as mixed approach.

6.3 Finding the Optimal Value of Neighborhoods Size (k)

A varied the number of neighbors and corresponding RMSE for an active user, from 0 to 100 and is computed. The results are shown in Fig.6.2.

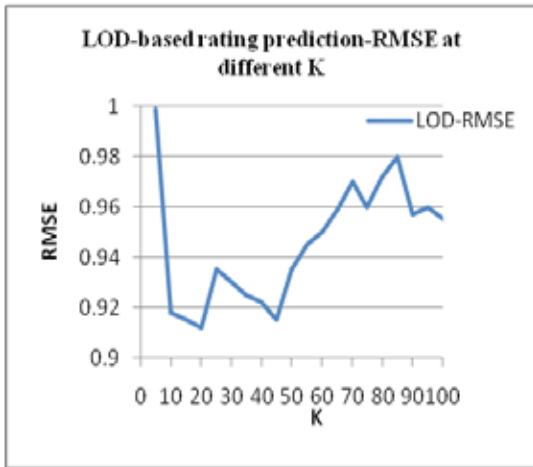


Figure 6.2 LOD based rating prediction RMSE at different k

Table 6.1 and figure 6.3 shows different technique compared with proposed method: user-based CF [1,2], item-based CF [1,2], item attribute-based, user attribute-based, a hybrid recommendation algorithm proposed in [8].

Table 6.1 RMSE evaluation at different k for various models

Recommendation technique		RMSE
Hybrid	LOD based	0.892
	Non LOD based	1.057
Collaborative	User-based CF	1.021
	Item based CF	1.009
Content	User attribute (Content)	1.037
	Item attribute (Content)	1.057

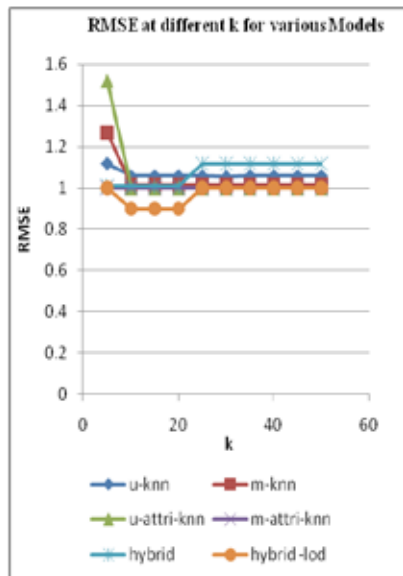


Figure 6.3 RMSE result for different model with various k

7. CONCLUSION

In this paper hybrid recommender combined collaborative and content-based filtering techniques to predict rating using Link Open dataset. This hybrid recommender is implemented by means of the familiar MovieLens rating dataset. Supplementary movie information was retrieved from the online archive, LinkedMDB and DBpedia. By means of various experiments, it demonstrated that the extracted content features are beneficial to the prediction accuracy of our hybrid recommendation system using Linked open data.

8. REFERENCES

- [1] Gediminas Adomavicius, Alexander Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, vol. 17. 2005.
- [2] Heitmann, B., Hayes C. Using linked data to build open, collaborative recommender systems. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010.
- [3] R. Mirizzi, T. D. Noia. A. Ragone, V. C. Ostuni, E. D.Sciascia, Movie Recommendation with DBpedia. IIR 2012.
- [4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia – a crystallization point for the web of data", Web Semant., 7:154-165, September 2009.
- [5] O. Hassanzadeh and M. P. Consens, "Linked Movie Data Base", In Proceedings of the WWW2009 Workshop on Linked Data on the Web, April 2009.
- [6] MovieLens Dataset, USA, 200, University Minnesota.
- [7] Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock. 2002, "Methods and metrics for cold-start recommendations", In the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland: ACM.
- [8] Burke, R. 2002, "Hybrid recommender systems: Survey and experiments", User Modeling and User-Adapted Interaction 12(4):331–370.
- [9] Mihel i , M., Antulov-Fantulin, N., Bošnjak, M., Šmuc, T., Extending RapidMiner with recommender systems algorithms, RapidMiner Community Meeting and Conference 2012, Budapest, Hungary