A Review on Similarity Measures for Cold-starting Problem in Collaborative Filtering

Nguyen Thu Giang
tgn3@illinois.edu

1 Introduction

Facing the problem of huge information available commonly found when searching for relevant documents, movies or even purchases on online stores, users could easily become overloaded and disoriented. Recommendation systems are developed to increase utility of users by suggesting to users relevant items to reduce search time and streamline information load.

Collaborative filtering ("CF") is one of the most prominent approaches in designing recommendation systems. CF provides suggestions to users based on items rated by others similar users, or items similar to what the respective user has rated. CF includes two methods – memory-based (or neighbourhood-based) and model-based. Memory-based method relies on the idea that users would find items highly rated by similar users relevant. Model-based, on the other hand, takes in users' attributes and behaviours and try to make a prediction on the rating of an item. While model-based method could be faster in prediction time since the model is trained prior to application (Liu et al., 2014), memory-based method are preferred due to its simplicity and the ability to respond immediately after receiving a new feedback from a user (Patra et al., 2015).

Memory-based approach utilises traditional similarity measures such as cosine similarity, Pearson coefficient and other variants. This makes the approach perform poorly when there is insufficient of co-rated items. Cold starting problem refers to the challenges of recommending new items or to new users due to the lack of data about them (Ahn, 2016) given that users might only have a limited time interacting with the site. This could be a common problem for e-commerce website where a high portion of users could be new or inactive.

This review outlines the development of different similarity measures to tackle this problem. The structure of the review as follows. In section 2, we discuss the use of similarity measures in memory-based CF. Traditional and state-of-the-art similarity measures are provided in section 3. We conclude the review in Section 4 with possible future research directions.

2 Memory-based Approach

First introduced by the GroupLens Usenet article (Resnick et al., 1994), memory-based approach has gained popularity with the commercial electronic retailers. The memory-based algorithm aims to predict the rating of the i-th item using the similarity information of the u-th user (user-based) or the similarity information of the kth item (item-based).

The formula to predict rating of item *i* by user U_u using the user-based approach is as follows.

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{k=1}^{K} s(U_u, U_k) \times (r_{ki} - \bar{r}_{k_u})}{\sum_{k=1}^{K} |s(U_u, U_k)|}$$

$\bar{r}_u$ is the average ratings made by users U_u; $s(U_u, U_k)$ is the similarity measure of user U_u and neighbour U_k; $\bar{r}_{k_u}$ is the average rating made by neighbour U_k; and r_{ki} is the rating made by neighbour U_k on item *i*.

The prediction formula using the item-based approach is as follows.

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{k=1}^{K} s(I_i, I_k) \times (r_{uk} - \bar{r}_k)}{\sum_{k=1}^{K} |s(I_i, I_k)|}$$

$\bar{r}_i$ is the average ratings by all users on item I_i; $s(I_u, I_k)$ is the similarity measure of item I_i and neighbour item I_k; $\bar{r}_k$ is the average rating made all users on the neighbour item I_k; and r_{uk} is the rating made by the active user on the neighbour item I_k.

Similarity measure, as presented in the two formula above, is clearly a crucial step of the method-based algorithm.

3 Review of Similarity Measures

3.1 Traditional similarity measures

Table 1 presents the similarity measure frequently utilised in memory based CF

Measure	Definition/Formula	Limitations
Cosine (Salton and McGill, 1986)	$s(U,V) = \frac{\sum_{I \in I'} (r_{UI})(r_{VI})}{\sqrt{\sum_{I \in I'} r_{UI}^2 \sum_{I \in I'} r_{VI}^2}}$ r_{UI} is the rating made by user *U* on item *I* and *I'* is the set of co-rated item.	[1], [2], [3], [4]
Adjusted cosine (Sarwar et al, 2001)	$s(U,V) = \frac{\sum_{U \in U'} (r_{UI} - \bar{r}_I)(r_{VI} - \bar{r}_J)}{\sqrt{(r_{UI} - \bar{r}_I)^2 (r_{VI} - \bar{r}_J)^2}}$ r_{UI} is the rating by user *U* on item *I*, $\bar{r}_I$ is the average rating of item *I* and *U'* is the set of co-rated item.	[3]

Pearson Correlation ("PC") (Ekstrand et al, 2011)	$s(U,V) = \frac{\sum_{I \in I'}(r_{UI} - \bar{r}_U)(r_{VI} - \bar{r}_V)}{\sqrt{\sum_{I \in I'}(r_{UI} - \bar{r}_U)^2 \sqrt{\sum_{I \in I'}(r_{VI} - \bar{r}_V)^2}}}$ r_{UI} is the rating by user U on item I, $\bar{r}_I$ is the average rating of item I and U' is the set of co-rated item.	[1], [2], [3], [4]
Constrained Pearson Correlation (Shardanand and Maes, 1995) ("CPC")	$s(U,V) = \frac{\sum_{I \in I'}(r_{UI} - r_{med})(r_{VI} - r_{med})}{\sqrt{\sum_{I \in I'}(r_{UI} - r_{med})^2 \sum_{I \in I'}(r_{VI} - r_{med})^2}}$ r_{UI} is the rating by user U on item I, r_{med} is the median rating and I' is the set of co-rated item.	[1]
Mean squared difference ("MSD") (Shardanand and Maes, 1995)	$s(U,V) = 1 - \frac{\sum_{I \in I'}(r_{UI} - r_{VI})^2}{\lvert I' \rvert}$ r_{UI} is the rating by user U on item I, and I' is the set of co-rated item.	[5]
Jaccard	$s(U,V) = \frac{\lvert I_U \cap I_V \rvert}{\lvert I_U \cup I_V \rvert}$ I_U is the set of items rated by user U.	[7]
Jaccard and Mean-squared-difference ("JMSD") (Bobadilla et al., 2011)	$s(U,V) = s_{MSD}(U,V) \times s_{Jac}(U,V)$ $s_{MSD}(U,V)$ and $s_{Jac}(U,V)$ are similarity in MSD and Jaccard respectively.	[6], [8]

Many major drawbacks are found in the traditional similarity measures (Ahn, 2008; Liu et al., 2014; Patra et al., 2015).

[1] *Few co-rated items*: Under data sparsity, these measures could not calculate the similarity between users or items;

[2] *Only one co-rated items*: Either the similarity measure cannot be calculated (e.g., PC) or yield result as perfect similarity of 1 (e.g., cosine);

[3] *Low similarity regardless of similarity ratings, or high similarity regardless of the difference between users;*

[4] *Flat rating or rating on the same line* (e.g., $\langle 1, 1, 1 \rangle$): Either the similarity measure cannot be calculated (e.g., PC) or yield result as perfect similarity of 1 (e.g, cosine);

[5] *Ignore the proportion of common ratings between two users*: The measures do not differentiate pairs of users with higher proportion of common ratings to be more similar;

[6] *Utilisation of ratings:* The measures do not take into consideration of all ratings of the pair of users;

[7] *Ignores absolute ratings:* Contrary to problem in 6, the measures only consider common ratings of users without considering absolute value of rating, making it very difficult to differentiate users.

[8] *Lack of global information:* The measures only utilise local information of the item and ignore other ratings globally for the items.

When coupled with data sparsity problem (cold-starting problem), the above limitation is further magnified due to higher occurrences from the lack of sufficient ratings and sub-optimal usage of data.

3.2 State-of-the-art similarity measures

We now discuss three new similarity measures that aim to tackle the cold-starting problem.

3.2.1 Proximity-Impact-Popularity ("PIP")

PIP is a heuristic based measure that captures three important factors – proximity, impact and popularity between a pair of ratings on the same item (Ahn, 2008).

$$s(U_u, U_k) = \sum_{j \in Cu,k} PIP(r_{uj}, r_{kj})$$

$$= \sum_{j \in Cu,k} Proximity\ (r_{uj}, r_{kj}) \times Impact(r_{uj}, r_{kj}) \times Popularity(r_{uj}, r_{kj})$$

- Popularity is a simple arithmetic difference between two ratings, and a penalty would be imposed if the two ratings are on different side of the medians (in "disagreement);
- Impact considers how strongly an item is liked or disliked by an user. The higher the strength of preference, the more credible the ratings;
- Popularity rewards ratings further away from the average rating of a co-rated item as items closer to average rating might be based merely on shared preference commonly found in more popular items (e.g, a blockbuster movie).

PIP is able to outperform traditional similarity measures in presence of sparse data. However, it is not able to utilise both local and global information of the ratings, and still lacks strict mathematical foundations being a heuristic measure. PIP also does not perform as well as traditional similarity measures on denser datasets, and Ahn (2008) proposed a hybrid approach of combining PIP and PC to leverage on the strength of both measures.

3.2.2 New heuristic similarity measure ("NHSM")

NHSM proposed by Liu et al. (2014) extends the idea of PIP by the following methods

$$sim(u,v)^{NHSM} = sim(u,v)^{PSS} . sim(u,v)^{Jaccard} . sim(u,v)^{URP}$$

- Utilise a non-linear function $sim(u,v)^{PSS}$ in calculating proximity, impact and popularity to amplify positive factors and restrain negative factors;
- Utilise Jaccard similarity $sim(u,v)^{Jaccard}$ to punish a small proportion of common ratings;
- Utilise mean and variance of user preference $sim(u,v)^{URP}$ to consider global ratings of users.

NHSM, however, only compute co-rated items, and non co-rated items are neglected. In experiment with extremely sparse dataset Epinions outlined by Liu et al. (2014), NHSM does not show very large difference in performance compared to other methods. NHSM did not outperform cosine similarity measure in tasks of providing top N recommendations for one of the dataset in the paper.

3.2.3 Bhattacharyya Coefficient in Collaborative Filtering ("BCF")

Patra et al. (2015) proposed another measure utilising the Bhattacharyya measure that could compute similarity based on all ratings and does not depend on co-rated items. Bhattacharyya measure could achieve such task by computing the similarity based on the densities of the distributions of the two item *i* and *j* vectors.

$$BC(i,j) = \sum_{h=1}^{H} \sqrt{(\hat{p}_{ih})(\hat{p}_{jh})}$$

H is the number of bins and $\hat{p}_{ih} = \frac{count\ of\ h}{count\ of\ i}$ where count of *i* is the number of users rated on item *i* and count of *h* is the number of users rating item *i* with value *h*. Hence, $\sum_{h=1}^{H} \hat{p}_{ih} = \sum_{h=1}^{H} \hat{p}_{jh} = 1$.

The similarity of user *U* and *V* also has Jaccard similarity added to give rewards to common rated items by both users. $loc(r_{Ui}, r_{Vj})$ is a correlation measure of rating made on item *i* by user *U* and rating made on item *j* by user *V* in order to add in the local similarity of the ratings that $BC(i,j)$ disregards.

$$s(U,V) = Jacc(U,V) + \sum_{i \in I_U} \sum_{j \in I_V} BC(i,j) loc(r_{Ui}, r_{Vj})$$

The BCF measure is able to take in both local and global information of the ratings, as well as utilising all absolute ratings. Experiments on different datasets of 4 sparsity levels presented by Patra et al. (2015) show that BCF measure could outperform PIP, NHSM and other traditional measures.

3.2 Other methods

Apart from developing a different similarity measure, data smoothing (Anand et al., 2011), principle component analysis and singular value decomposition (Sarwar et al., 2000) are also proposed as possible methods to fill in missing ratings in order to solve the cold starting problem.

4 Conclusion

This review presents the cold-starting problem in collaborative filtering and feasible new similarity measures to tackle the issue. The performance of different similarity measures still vary depending on different characteristics of the tasks and application settings, as well as depending on different metrics utilised to measure the CF performance. Hybrid approach by combining different measures to leverage on the strengths of different measures to different levels of data sparsity could be further researched.

Reference

Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information sciences, 178(1), 37-51.

Anand, D., & Bharadwaj, K. K. (2011). Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities. *Expert systems with applications*, *38*(5), 5101-5109.

Bobadilla, J., Serradilla, F., & Bernal, J. (2010). A new collaborative filtering metric that improves the behavior of recommender systems. *Knowledge-Based Systems*, *23*(6), 520-528.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). *Collaborative filtering recommender systems*. Now Publishers Inc.

Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. Knowledge-based systems, 56, 156-166.

Patra, B. K., Launonen, R., Ollikainen, V., & Nandi, S. (2015). A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. *Knowledge-Based Systems*, *82*, 163-177.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994, October). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186).

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). *Application of dimensionality reduction in recommender system-a case study*. Minnesota Univ Minneapolis Dept of Computer Science.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295).

Shardanand, U., & Maes, P. (1995, May). Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 210-217).