PipeCF: A Scalable DHT-based Collaborative Filtering Recommendation System*

XIE Bo, HAN Peng, YANG Fan, SHEN Ruimin
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
{ bxie, phan, fyang, rmshen}@sjtu.edu.cn

ABSTRACT

Collaborative Filtering (CF) technique has proved to be one of the most successful techniques in recommendation systems in recent years. However, traditional centralized CF system has suffered from its shortage in scalability as their calculation complexity increases quickly both in time and space when the record in user database increases. In this paper, we propose a decentralized CF algorithm, called PipeCF, based on distributed hash table (DHT) method. We also propose two novel approaches to improve the scalability and prediction accuracy of DHT-based CF algorithm. The experimental data show that our DHT-based CF system has better prediction accuracy, efficiency and scalability than traditional CF systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*information filtering*; H.5.3 [**Information Interfaces and Presentation (1.7)**]: Group and Organization Interfaces–*collaborative computing;* C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*distributed applications*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Collaborative Filtering, Distributed Hash Table

1. INTRODUCTION

Since Goldberg et al [1] published the first account of using collaborative filtering (CF) for information filtering; CF has proved to be one of the most successful techniques in recommendation systems such as GroupLens [4]. However, all these systems have used a centralized memory-based CF algorithm which suffered from scalability as it need to calculate similarity between the active users and all other users.

One way to solve this problem is to use a model-based algorithm [3]; however, these approaches also need complex calculation when compiling models. Another method is to implement CF in a decentralized way [6].

In this paper, we propose a novel distributed hash table (DHT) based technique to implement efficient user database management and retrieval and construct a decentralized CF recommender system based on it; we also propose two novel approaches: significance refinement (SR) and unanimous amplification (UA), to improve the performance of our DHT-based CF algorithm.

2. DHT-BASED CF APPROACH

2.1 Architecture of DHT-based CF System

The neighbor choosing strategy of DHT-based CF algorithm is based on the heuristic that people with similar interests at least rate one item with similar votes. So we only select similar users in the subset in which users have same <ITEM_ID, VOTE> tuple. The key idea of our algorithm is hashing every user for every rated item. In our DHT-based CF system, both the maintenance of user database and the complex computation task of making prediction are done in a decentralized way. Each user keeps his votes locally. The system generates a unique key for each particular <ITEM_ID, VOTE> tuple of each user by hashing it. So each user will have M keys while M is the number of items he has rated. These keys are then used to construct a DHT overlay network [7]. When a user wants to look up other similar users which have the same particular <ITEM_ID, VOTE> tuple, it can fetch them from DHT overlay network efficiently. So with the DHT overlay network, all the users in the CF system are connected together and can find their wanted similar neighbors efficiently through a DHT routing algorithm [2]. Figure 1 gives the architecture of our DHT-based CF system.

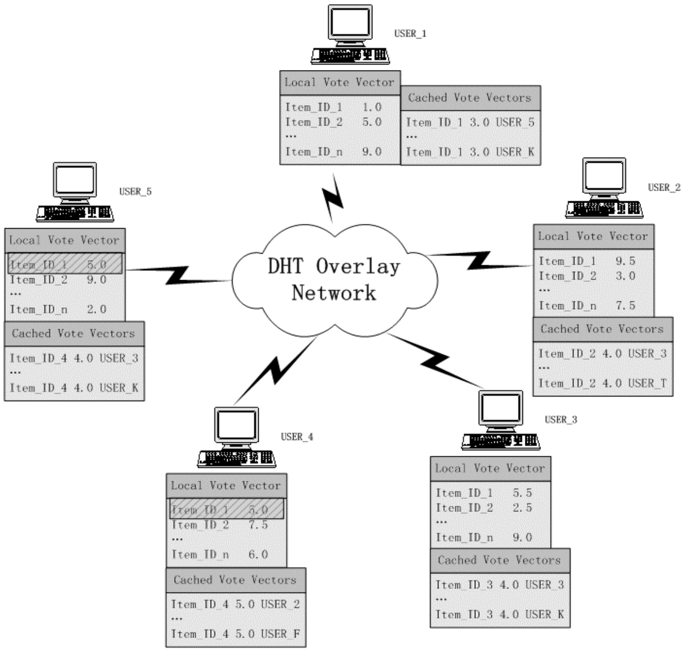

Figure 1 Architecture of DHT-based CF System

* Supported by the National Natural Science Foundation of China under Grant No. 60372078.

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM 1-58113-912-8/04/0005.

2.2 Extensions to Memory-based Algorithm

2.2.1 *Significance Refinement (SR)*

As Breese presented in [3] by the term *inverse user frequency,* universally liked items are not as useful as less common items in capturing similarity. So we introduce a new concept *significance refinement* (SR) which reduce the returned user number of the basic DHT-based CF algorithm by limiting the number of returned users for each <ITEM_ID, VOTE> tuple. We term the algorithm improved by SR as *Return K* which means “for every item, the DHT-based CF algorithm returns no more than *K* users with same <ITEM_ID, VOTE> tuple”. By doing so, we reduce the calculation complexity from $O(M^2N)$ to $O(N^2)$, where M is the user number and N is the item number.

2.2.2 *Unanimous Amplification (UA)*

Enlightened by the method of case amplification [3] which emphasizes the contribution of the most similar users to the prediction by amplifying the weights close to 1, we argue that we should give special award to the users who rated some items with the same vote by amplify their weights, which we term *Unanimous Amplification*. We transform the estimated weights as follows:

$$w'_{a,i} = \begin{cases} w_{a,i} & N_{a,i} = 0 \\ w_{a,i} \cdot \alpha & 0 < N_{a,i} \le \gamma \\ w_{a,i} \cdot \beta & N_{a,i} > \gamma \end{cases} \qquad (1)$$

Where $N_{a,i}$ denotes the number of items which user *a* and user *i* have the same votes. A typical value for α for our experiments is 2.0, β is 4.0, and γ is 4. Experimental result shows that UA approach improves the prediction accuracy of both the traditional and DHT-based CF algorithms.

3. EXPERIMENTAL EVALUATION

We use EachMovie data set [5] to evaluate the performance of improved algorithm. We use Mean Absolute Error (MAE), a statistical accuracy metrics, to report prediction experiments for it is most commonly used and easy to understand:

$$MAE = \frac{\sum_{a \in T} | v_{a,j} - p_{a,j} |}{|T|} \qquad (2)$$

Where $v_{a,j}$ is the rating given to item *j* by user *a*, $p_{a,j}$ is the predicted value of user a on item *j*, *T* is the test set, |*T*| is the size of the test set.

We select 2000 users and choose one user as active user per time and the remainder users as his candidate neighbors, because every user only make self’s recommendation locally. We use the mean prediction accuracy of all the 2000 users as the system's prediction accuracy. For every user’s recommendation calculation, our tests are performed using 80% of the user’s ratings for training, with the remainder for testing.

We compare the prediction accuracy of traditional CF algorithm and DHT-based CF algorithm while we apply both top-all and top-100 user selection on them. The results are shown as Figure 2. We can see that the DHT-based algorithm has better prediction accuracy than the traditional CF algorithm.

4. CONCLUSION

In this paper, we propose a novel distributed hash table (DHT) based technique to implement efficient user database management and retrieval in decentralized CF system. Then we propose a heuristic algorithm to fetch similar users from DHT overlay network and do recommendation locally. Finally, we propose two novel approaches: significance refinement (SR) and unanimous amplification (UA) to improve the performance of our DHT-based CF algorithm. The experimental data show that our DHT-based CF system has better prediction accuracy, efficiency and scalability than traditional CF systems.

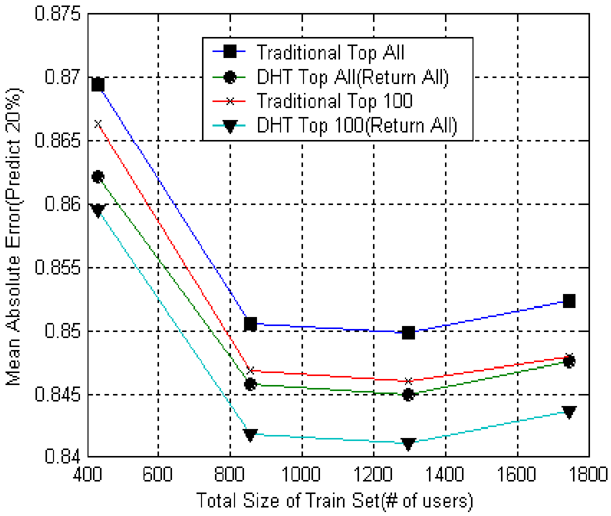

Figure 2. DHT-based CF vs. Traditional CF

REFERENCES

[1] David Goldberg, Using collaborative filtering to weave an information tapestry, Communications of the ACM, v.35 n.12, p.61-70

[2] S. Ratnasamy, A scalable content-addressable network. SIGCOMM, Aug, 2001

[3] Breese, Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 43-52.

[4] Paul Resnick, GroupLens: an open architecture for collaborative filtering of netnews, Proceedings of the 1994 ACM conference on Computer supported cooperative work, p.175-186

[5] Eachmovie collaborative filtering data set, 1997. http://research.compaq.com/SRC/eachmovie

[6] Amund Tveit. Peer-to-peer based Recommendations for Mobile Commerce. Proceedings of the First International Mobile Commerce Workshop, ACM Press, pp. 26-29