ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

SimCo: A Novel Similarity Based Collaborative Filtering In Recommender Systems

Noor Basha

Assistant Professor, Dept. of CSE., Vemana Institute of technology, Koramangala, Bengaluru, India.

ABSTRACT: Inspired by the arrival of Collaborative filtering in the recommender systems became an eminent technology. Among the similitude trait of the users, the dependency evolution is discovered and preserved for the similar items. This dependency form is derived from the resemblance level obtained between the user communities. The survival work has been carried out to solve the issues like data sparsity, inaccuracy and big-error in prediction. In this paper, we made an attempt to form a dependency relation between the users by their phenomenal level of the user's resemblance. We formulated the resemblance and comfort based on user phenomenal collaborative recommendation filtering technique, named SimCo (Similarity based Collaborative Filtering). Here we established a typical feature based CF that it evolutes the similitude among the neighbors in its communities. The experimental analysis is carried out in amazon.com which is high traffic website using Recommender systems. We tried to achieve a novel patent searching by minimizing the big-error predictions, data sparsity is lessened without compromising the accuracy and better usability of patent search.

KEYWORDS: Collaborative filtering; Recommender systems; Data Sparsity; Degree of Similarity and Dependency relation

I. INTRODUCTION

Survival recommender systems made an evolution to create dependency structure between the users by their preferences over items or products. Thus by estimating the preferences among the users established a recommender module in communities. The intention of the recommender systems is to assist the dynamic user by choosing the same preferences [1]. This type of framework create a wide applications in e-commerce, subscription based services and also support a wide variety of the field that is based on the personalized opinions. Thus it can also refer as the 'opinion mining'. In real world environmental data, a variety of choices according to the user's mind level are predefined in order to assist the upcoming grouping of the items or products. In this way, the Recommender system plays a vital role in the field of Data Mining.

This Recommender system (RS) is sorted out as Content based or Collaborative based approaches [2]. The past information is acquired for the similar items to user preferences are known as the content based approaches. From the past information such as description of the items are obtained to generate the user model for demanding the interest among the groups. Another approach CF design a model based on the accumulative preferences of the items and their accessibility level of the users on item descriptions. It opens out as two forms: Memory and Model based approaches.
In Memory based technique [3], the previous ratings given by the user is taken into account to generate future rating predictions over items [4][5]. The hidden predicting value of the dynamic user u for an item i is estimated as the cumulative points obtained by the user's ratings similar to the u for the same item i. This cumulative may be of average, weights and distance among the users. It can also obtain by similarity approaches such as finding correlation, cosine based, and frequency based or case amplification. Discovering similarities between the users have to be modulated effectively for the dynamic users [6].

The model based CF employs the ratings container to study a model and then predicts the rating level [3]. When compared to the memory based approaches, this approach reported a greater level of accuracy with some

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

delimitation. The delimitations such as: a) Computationally very expensive, b) Instead of choosing the rank method, it prefers the rating of an item.

The proposed work carried out as follows:

- To cluster all items into several item groups. For example, we can cluster all books into "Engineering books," "Medical books," and so on.
- To form a user group corresponding to each item group (i.e., a set of users who like items of a particular item group), with all users possessing different similitude between the user groups.
- To build a user-similitude matrix and measure users' similarities based on users' comfort degrees among the groups in order to select the 'neighbors of the users.
- Then, we predict the unknown rating of a user on an item based on the ratings of the "neighbors" of at user on the item.
- To propose an error-correction technique to suggest similar terms for the query keywords and return answers of the similar terms.
- To help users formulate high-quality queries, as user's type in keywords, we suggest keywords that are topically relevant to the query keywords.

The paper is organized as follows: Section II describes the survival work carried out in CF. Section III depicts our novel framework designed for CF. Section IV discusses the experimental analysis using datasets. Lastly, the paper concludes in Section V.

II. LITERATURE SURVEY

A. *Model view and Similitude*

The degree of similarity is attributed as the desirability level of the items modeled in concepts. Under the object of the concept, the common dimension that is imparted by all the objects is considered for explaining the novel concepts [4]. In model view of concepts, a concept is symbolized by the general dimension used by all objects are grouped as the concepts. The general dimension information contains both the relevant and irrelevant dimension of the objects [3]. The concept model imparts the general dimensionality of all objects which is effectively used for prescribing the similarity of the instances. Consider an instance "can-fly" which is probably related to the concept 'bird'. So, similar groups that can fly will be evaluated as the most similarity that cannot fly [7]. Though it opted to be best for model view to act as 'insight', it almost fails in demonstrating as whole in some situations like 'Animal' which can't be used for the co-occurring relations among the instances.

Vanpaemel et al [8] suggested a framework to extend the model view. They introduced the abstractions for each instances of the concepts. The objects that contain the similar abstraction are used for instantiation process. Arguably, the matching dimension of objects is also considered. A combined model has been developed by the Vanpaemel et al to enhance the prototype and exemplar model.

Barsalou [9] discovered two component, Central tendency and frequency of abstraction to minimize the power of objects. The family similarity obtained for an object is known as Central tendency. The object of the member is similar to the object of the other members forms a same concept. The cluster similarity depends on the frequency of the instantiation of an object estimate the preferences of the users. The objects that possess higher frequency in a concept create impressive user preferences.

Rifqi [10] investigated to provide work on object typicality in large datasets and it was also extended by Lesot et al [11]. Depending on the similarity of an object, the items form a category for each item. Au Yeung and Leung [12] investigated this research with the use of ontologies that calculates the property vector and prototype vector of this concept. From this study, we infer that no study discussed with the concept of Collaborative Filtering.

- **Recommender Systems**

Many works has been carried out on Recommender systems that dealt with the creation of profiles using the ratings. The aim of the recommender systems is to create users to discover the items of their demands. Now we shall

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

discuss the recommendation methods by several approaches such as Collaborative Filtering, Content Based and Hybrid methods [13][14].

- Content Based Systems:

This system was developed based on the past information saved by the users. The future items will be classified based on the past information generated by the users. Most demanding items are often searched by the users are used for the content based systems. The similar items are categorized or rated highly will be considered by the user. The user profiles are very important to create a content based model. Many learning approaches have been discovered to construct profiles. The text categorization method in LIBRA systems was developed by Mooney and Roy [15]. A detailed approach about CB systems was further studied by Pazzani and Billsus [16].

- Collaborative Filtering

The aim of CF approach for the user to mine the user profiles based on their demanding opinions about an item will form recommender systems. The group of nearest neighbors who imparts the similar opinions comes under the user based CF approach [2]. Based on the predicted ratings, the unpredicted items are rated. The objective of the item based approach is to predict the ratings by the correlation existing between the users and its search. Sarwar et al [17] discussed various techniques to measure the item ratings. The top-N recommendation algorithm was suggested by Deshpande and Karypis [18]. They found similarity between the users by the Pearson correlation coefficient, cosine based similarity and vector space similarity.

B. Hybrid Recommender Systems

Several hybrid recommender systems incorporated the approach of collaborative and content based methods which eliminates the demerits of the content based and collaborative systems. The linear combinations of ratings encountered by the user predictions of both collaborative and CB methods. The other approach such as rating matrix is created and developed by Melville et al [19]. Ma et al [20] proposed the hybrid approach of item based and user based CF. Some recent works has been carried out to discover an approach to avoid the over fitting problems of user ratings. An implicit feedback was designed by Hu et al [21]. Zhang et al [22] framed a matrix factorization method for the regression analysis. Lee et al [24] focused to design rank based Graph entities for random walks for multidimensional recommendations. Zhou et al [25] framed a solution for cold start problem by exploring functional matrix factorization.

III. PROPOSED SYSTEM

The proposed work carried out as follows:

- To cluster all items into several item groups. For example, we can cluster all books into “Engineering books,” “Medical books,” and so on.
- To form a user group corresponding to each item group (i.e., a set of users who like items of a particular item group), with all users possessing different similitude between the user groups.
- To build a user-similitude matrix and measure users’ similarities based on users’ comfort degrees among the groups in order to select the ‘neighbors of the users.
- Then, we predict the unknown rating of a user on an item based on the ratings of the “neighbors” of at user on the item.
- To propose an error-correction technique to suggest similar terms for the query keywords and return answers of the similar terms.
- To help users formulate high-quality queries, as user’s type in keywords, we suggest keywords that are topically relevant to the query keywords.

A. *Recommendation Algorithms*

The traditional recommendation algorithms works on discovering a set of users and their set of items that conjoint at any point of rating level. In this module, the similar items are aggregated. The item that are already purchased or rated is eliminated and the rest items are allocated to the users. This action is performed by two algorithms such as collaborative filtering and cluster models. The algorithms such as search based methods and own item- to- item

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

collaborative filtering which avoids the similar users. The similar items are encountered by the user's purchased and rated items. Atlast, the similar items are groups and recommend them.

B. *Traditional Collaborative filtering*

Consider a customer as N- dimensional vector of items where N is the number of distinct catalog items. The items are categorized as the positive and negative vector. The positive vector contains the positively purchased or rated items whereas negative vector contains the negatively purchased or rated items. The most demanded items are discovered by multiplying the vector components to the inverse frequency (the inverse of the number of customers who have purchased or rated the item) to predict the well known items. The component vector is in form of sparse. The cosine similarity between two vectors is used to find similarity between the customers A and B.

- Cluster models

The cluster model divides the customer into several segments and addresses this task as a classification problem. Each segment contributes most similar customers to assign the user. The purchases and ratings of the customer in each segment contribute to be a recommendation. The cluster or segment is formed by the most similar customers using a similarity metric.

- Item to Item collaborative filtering

The best instance for item-to item collaborative filtering is the Amazon.com which is a high traffic creating websites. In that websites, click on "Your recommendations" that forwards its customer to webpage where they can find the recommendations for their items and subject are, rate the recommended products, rate their previous purchases, and see why items are being recommended. Our algorithm works on the massive data sets and generates the high quality recommendation system in real time.

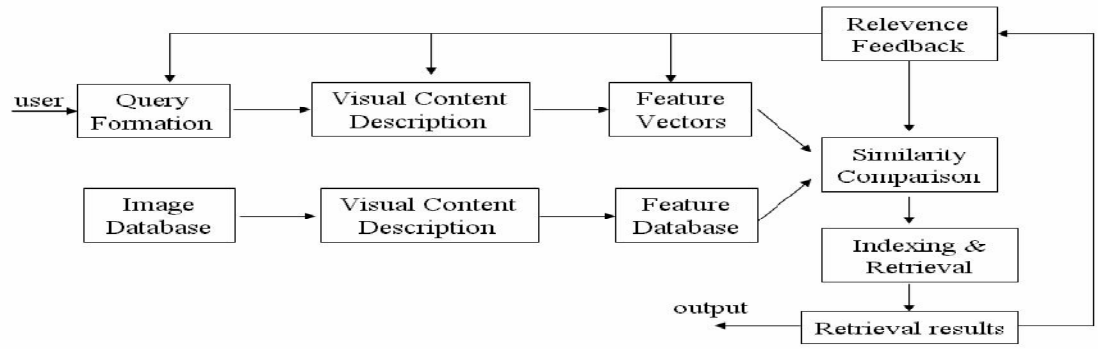

Fig.3.1 Proposed Block Diagram for Recommendation system.

The algorithm preliminaries such as:

A. *Signal:*

A signal is a function that acts as an alarm signal depending on some variable. The signal may be of one-dimensional (time), two dimensional (coordinates in a plane), three-dimensional (describing an object in space) and higher- dimensional. A scalar function may be sufficient to describe a monochromatic image, while vector functions are to represent, for example, color images consisting of three component colors.

B. *Image Functions:*

An image acts as container that contains a sequence stream function of two or three variables. The two variables x and y are the co-ordinates in a plane whereas a third variable is of time t when an image is modified. The function values are obtained by the brightness at image points. The functional value represents the physical quantities such as temperature, pressure distribution, and distance from the observer etc. the brightness function is used to represent the process of image formation. It is combined with variants optical quantities to disseminate the complex process of image production. Arguably, the human eye retina or a TV camera is of 2D, so this function is called as intensity of an image. The 2D intensity image is the projection view of the 3D scene. When 3D objects are transformed

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

into the camera plane by projection view, the information mishaps due to one-to one transformation. The lost information is recovered by the geometric functions and other problem comprehends the image brightness. The light source emitted is created by the energy distribution as C(x, y, t, λ) where (x, y) are co-ordinates in a plane, time t and wavelength λ. The brightness f depends on the energy distribution C and spectral sensitivity and wavelength is estimated by:

$$f(x,y,t) = \int_0^\infty C(x,y,t,\lambda)S(\lambda)d\lambda \quad (3.1)$$

In a color or multispectral image, the image is represented by a real vector function f as:

$$f(x,y,t) = (f_1(x,\mathrm{y},\mathrm{t}),\ f_2(x,\mathrm{y},\mathrm{t}),.....f_n(x,\mathrm{y},\mathrm{t})) \quad (3.2)$$

A monochromatic static image is represented by a continuous image function f(x,y) whose arguments are two co-ordinates in the plane. A monochromatic static image is represented by a continuous image function f(x,y) whose arguments are two co-ordinates in the plane. Computerized image processing uses digital image functions which are usually represented by matrices, so co-ordinates are integer numbers. The customary orientation of co-ordinates in an image is in the normal Cartesian fashion (horizontal x axis, vertical y axis), although the (row, column) orientation used in matrices is also quite often used in digital image processing. The range of image function values is also limited; by convention, in monochromatic images the lowest value corresponds to black and the highest to white. Brightness values bounded by these limits are gray levels. Computerized image processing uses digital image functions which are usually represented by matrices, so co-ordinates are integer numbers. The customary orientation of co-ordinates in an image is in the normal Cartesian fashion (horizontal x axis, vertical y axis), although the (row, column) orientation used in matrices is also quite often used in digital image processing.

c) Image Digitization

The image is denoted as a continuous function f(x, y) of the two co-ordinates in a plane. This function is sampled into a matrix with M rows and N columns. It's an integer value. The continuous range of the image function f(x, y) is split into K intervals is known as image quantization. The greater the sampling (Larger M and N) and quantization (larger K), the better is the approximation of continuous image function f(x, y).

IV. RESULTS AND DISCUSSIONS

The data are collected from the commercial social network site named, Amazon.com which contains interaction between users. Each record in a data symbolizes the contact by a tuple, identity of the sender's contact, identity of the receiver's contact and an indicator that indicates whether the communication is successful (Positive or negative response). It contains a training set sensed by one week period and test set on its subsequent week is collected. The dataset is summarized in Table I.

	# interactions	#positive	# negative
Training set	188255	54754	133501
Test set	199083	56677	142406

Table I: Dataset Description

For each active user, there will be one learning model. To learn an instance, a pair of users recommend by the CF. The pair is generated by considering the target user as from positive response and another target user as from negative response. The indicator indicated as integer value 0 for 1st user pair and 1 for 2nd user in a pair. The prediction evaluation metrics is identified by following metrics:

- Coverage metrics
- Statistical accuracy metrics

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

- Decision support accuracy metrics

A. *Coverage metrics:*

It validates the number of items to provide recommendations for the system. It is the percentage of the customer-product pairs for recommender systems.

B. *Statistical accuracy metrics:*

It validates the accuracy using the numerical recommendation. Statistical accuracy metrics evaluate the accuracy score in the test dataset. This is estimated by the Mean Absolute Error, Root Mean Squared Error and correlation between ratings and predictions. Here we employed Mean Absolute Error (MAE) approach is used.

C. Decision Support Accuracy metrics:

It validates how the search engine effectively retrieves the information to the user to display the high quality products. It operates on the binary value as 0 for bad and 1 for good. Based on the point scale obtained, the decision is generated.

The recommendation evaluation metrics is also estimated by Precision and Recall function. The top N is the customer most preferred and rated. The Recall function is given by:

$$\mathrm{Re}\,call = \frac{sizeofhitset}{sizeoftestset} = \frac{test \cap topN}{test} \quad (4.1)$$

Precision function is given by:

$$\mathrm{Pr}\,ecision = \frac{sizeofhitset}{sizeoftopNset} = \frac{test \cap topN}{N} \quad (4.2)$$

In nature, both Precision and Recall function is proportional to each other. So, we need to measure the F metric to predict the quality level. The F metric is given by:

$$F_1 = \frac{2 * \mathrm{Re}\,call * \mathrm{Pr}\,ecision}{(\mathrm{Re}\,call + \mathrm{Pr}\,ecision)} \quad (4.3)$$

The user preferences' rating is depicted as:

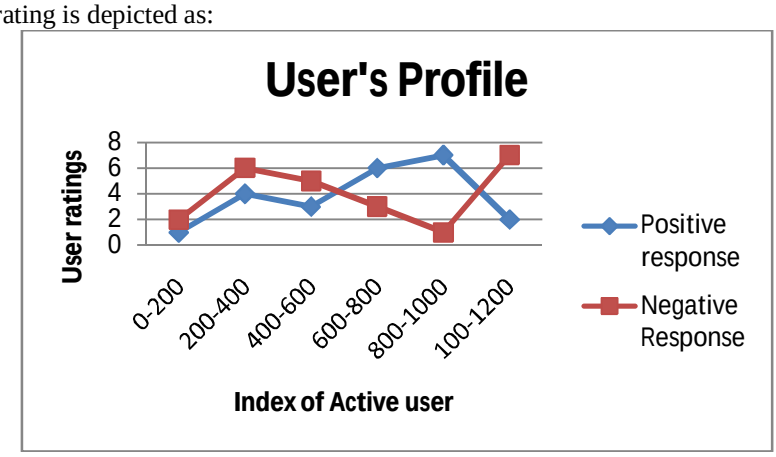

Fig.4.1Generating user's profile

The user ratings and calculated in the above figure the redline in the graph indicated the negative response and the blue line indicates the positive response, based on the response given by the users the graph is depicted above.

The cumulative rating for user preferences is given by: Based on the cumulative response verses number of active users the graph is depicted.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

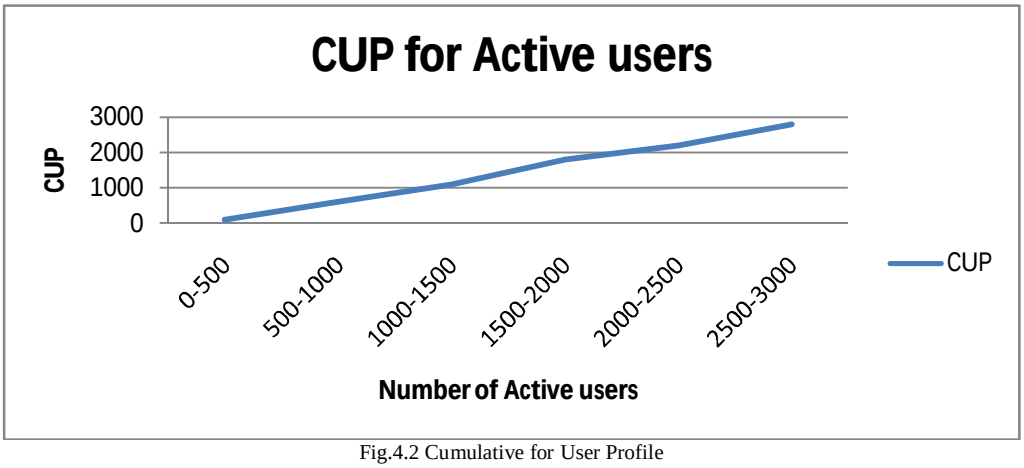

Fig.4.2 Cumulative for User Profile

The F metric is calculated as: F-matric predicts the quality level, the below graph is calculated using the equation 4.3 and the graph is depicted below.

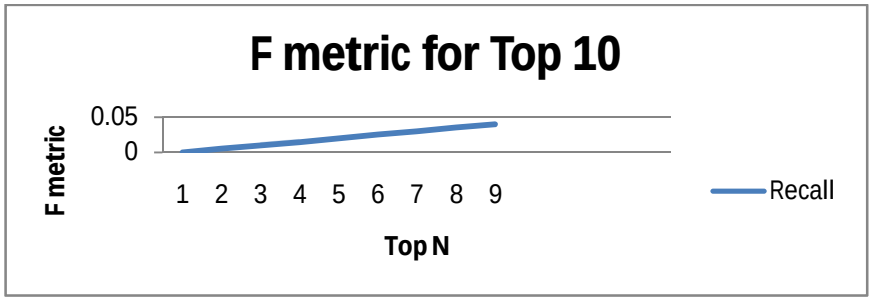

Fig.4.3 Recall function

The coverage metric is calculated as:It validates the number of items to provide recommendations for the system. It is the percentage of the customer-product pairs for recommender systems.

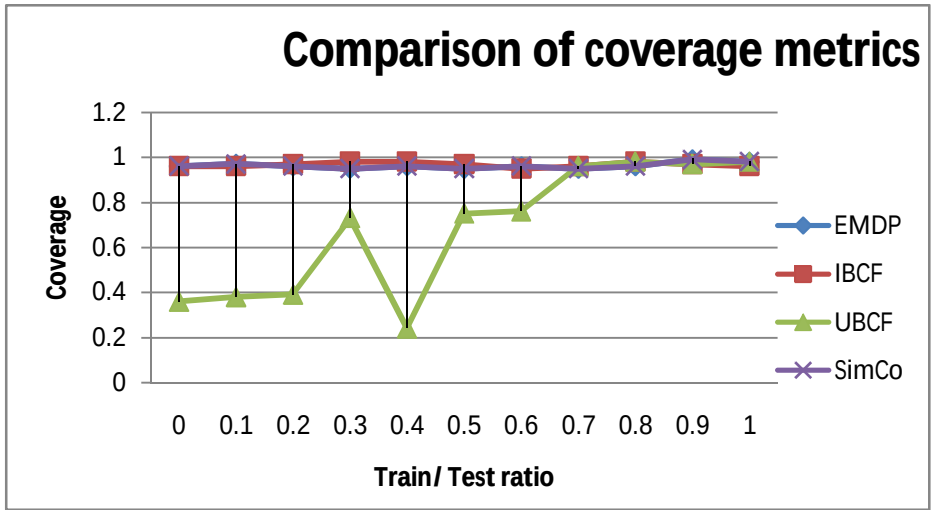

Fig. 4.4 Coverage Metrics

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

V. CONCLUSION

The personalized shopping experience for each customer is very vital. Recommender algorithms acts as a good way for most of the e-Commerce sites like amazon.com, flipkart.com etc. This algorithm is very scalable for the large datasets under time constraint. When compared to other algorithms, the item to item collaborative filtering able to meet our previous mentioned challenges such as predictions accuracy, relevant patents, big error predictions and user similarity. The unique feature in SimCo that it prefers the 'neighbors' by the past user preferences. The degree of similarity is employed to predict the user preferences and its ratings. This type of the methods will be great helpful in the real life applications that are processed in offline. As a future work, we can also make an attempt to try some parallelization techniques for processing the online and offline applications.

REFERENCES

1. Z. Huang, H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," ACM Trans. Information Systems, vol. 22, no. 1, pp. 116- 142, 2004.
2. G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.
3. K.M. Galotti, Cognitive Psychology In and Out of the Laboratory, third ed. Wadsworth, 2004.
4. G.L. Murphy, The Big Book of Concepts. MIT Press, 2002.
5. L.W. Barsalou, Cognitive Psychology: An Overview for Cognitive Scientists. Lawrence Erlbaum Assoc., 1992.
6. S. Schiffer and S. Steele, Cognition and Representation. Westview Press, 1988.
7. D.L. Medin and E.E. Smith, "Concepts and Concept Formation,"Ann. Rev. of Psychology, vol. 35, pp. 113-138, 1984.
8. W. Vanpaemel, G. Storms, and B. Ons, "A Varying Abstraction Model for Categorization," Proc. Cognitive Science Conf. (CogSci '05), pp. 2277-2282, 2005.
9. L.W. Barsalou, "Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories," J. Experimental Psychology: Learning, Memory, and Cognition, vol. 11, no. 4, pp. 629-654, Oct. 1985.
10. M. Rifqi, "Constructing Prototypes from Large Databases," Proc. Int'l Conf. Information Processing and Management of Uncertainty (IPMU '96), pp. 301-306, 1996.
11.] M.-J. Lesot, L. Mouillet, and B. Bouchon-Meunier, "Fuzzy Prototypes Based on Typicality Degrees," Proc. Int'l Conf. Eighth Fuzzy Days '04, 2005.
12. C.M.A. Yeung and H.F. Leung, "Ontology with Likeliness and Typicality of Objects in Concepts," Proc. 25th Int'l Conf. Conceptual Modeling, pp. 98-111, 2006.
13. H. Ma, I. King, and M.R. Lyu, "Effective Missing Data Prediction for Collaborative Filtering," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 39-46, 2007.
14. J. Wang, A.P. de Vries, and M.J.T. Reinders, "Unifying User- Based and Item-Based Collaborative Filtering Approaches by Similarity Fusion," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 501-508, 2006.
15. R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," Proc. Fifth ACM Conf. Digital Libraries (DL '00), pp. 195-204, 2000.
16. M.J. Pazzani and D. Billsus, "Content-Based Recommendation Systems," The Adaptive Web: Methods and Strategies of Web Personalization, pp. 325-341, Springer-Verlag, 2007.
17. B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-Based Collaborative Filtering Recommendation Algorithms," Proc. 10th Int'l Conf. World Wide Web (WWW '01), pp. 285-295, 2001.
18. M. Deshpande and G. Karypis, "Item-Based Top-N Recommendation Algorithms," ACM Trans. Information Systems, vol. 22, no. 1, pp. 143-177, 2004.
19. Y. Hu, Y. Koren, and C. Volinsky, "Collaborative Filtering for Implicit Feedback Datasets," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 263-272, 2008.
20. A. Umyarov and A. Tuzhilin, "Improving Collaborative Filtering Recommendations Using External Data," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 618-627, 2008.
21. Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," Computer, vol. 42, no. 8, pp. 30-37, Aug. 2008.
22. L. Zhang, D. Agarwal, and B.-C. Chen, "Generalizing Matrix Factorization through Flexible Regression Priors," Proc. Fifth ACM Conf. Recommender Systems (RecSys '11), pp. 13-20, 2011.
23. L. Backstrom and J. Leskovec, "Supervised Random Walks: Predicting and Recommending Links in Social Networks," Proc. Fourth ACM Int'l Conf. Web Search and Data Mining (WSDM '11), pp. 635-644, 2011.
24. S. Lee, S.-i. Song, M. Kahng, D. Lee, and S.-G. Lee, "Random Walk Based Entity Ranking on Graph for Multidimensional Recommendation," Proc. Fifth ACM Conf. Recommender Systems (RecSys '11), pp. 93-100, 2011.
25. K. Zhou, S.-H. Yang, and H. Zha, "Functional Matrix Factorizations for Cold-Start Recommendation," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '11), pp. 315-324, 2011.