ISSN: 2321-7782 (Online)
Volume 1, Issue 6, November 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

A Survey on Recommender System

Ramya Laurraine. U[1]
PG student
Department of Computer Science and Engg.
Velammal Engineering College
Chennai - India

M.Shanmugasundaram[2]
Assistant Professor
Department of Computer Science and Engg.
Velammal Engineering College
Chennai - India

Abstract: ***Recommender systems provide useful recommendations to a collection of users for items or products that might be of concern or interest to them. Its design depends on the data characteristics and the domain for which it will be applicable. The applications where recommender systems are used range from electronic products, books, movies, websites to even news articles and so on. The recommendations deal with predicting user's ratings that are yet to be consumed by users, based on the items already rated by them and recommend the top N items with the highest predicted ratings. Many algorithms try to improve the accuracy in predicting the top N items thereby improving the recommendation quality. This paper presents an overview of recommender systems and explores the various approaches to recommender systems.***

Keywords: ***Big Information Retrieval, Content Based Recommender, Collaborative Filtering, Cosine Similarity, Implicit Feedback, Item-Item Collaborative Filtering.***

I. INTRODUCTION

Nowadays, everyone has access to the internet, which means there is high potential for products to be sold online. This is where businesses take the advantage of selling their products through e-commerce sites. These sites need a business tool to let them efficiently market their products. This tool is nothing but the recommender systems. The motive of the system is not merely to increase the sales but to provide the customer with a wide view of products in the market and analyze the customer preferences for better recommendations. There are many e-commerce sites like Amazon, eBay, Netflix, last. fm which uses the inbuilt recommender system. E-commerce sites recommend products based on best sellers of a site, the demographic of the customer, past buying preferences of the customer.

Recommender Systems are a type of information filtering systems that try to predict the ratings that a user would give to an item and recommends the items that match the user's interests. Recommender systems are applied in a variety of applications like recommending electronic products, books, movies, music, restaurants, vacation trips and so on.

Recommender systems work as personalized or non-personalized systems. Non-personalized recommender systems recommend products to users based on other user opinions on the products. The recommendations are not personal to the user, so each individual gets the same recommendation. A personalized system evaluates the user's interests and recommends items based on them. The recommendation approaches can be broadly classified into two categories:

• Content-based filtering

• Collaborative filtering

Content-based Filtering: This approach recommends items that match with the attributes of the user or are similar in contents of the items the user has liked in the past.

Collaboration Filtering: This approach recommends items based on tastes and preferences of similar people/similar items. Amazon.com uses an algorithm based on item-based collaborative filtering to make their recommendations [1].

Recommendation systems are evolving such that it tries to improve its accuracy in recommending items that exactly match the user's needs. The huge amount of information on the Internet and the increase in web users are some of the key challenges in recommender systems. These recommender systems are improving in the algorithms they use to produce accurate recommendations while handling the huge number of products and participants in the system.

II. Architecture of a Recommender System

There are three steps to making a recommendation system (Fig. 1):

STEP 1: This step deals with learning the user's interest. The user profile of the customer is built on implicit/explicit feedback. Implicit feedback is captured from the user's browsing behaviour, clicks, buying items and dwell time and so on. Explicit feedback is got from explicit ratings, votes, likes/dislikes. In content based filtering the items are modelled as a vector of attributes or keyword tags and the user preferences are built on those attributes.

STEP 2: Once the ratings are collected and formed as a user rating matrix, the prediction of the value of the Unrated items are done by information filtering techniques such as content based filtering or collaborative filtering or even a hybrid approach.

STEP 3: Recommendation ranking takes the estimated rating value of the items, ranks them given a threshold and recommends the items to the active user.

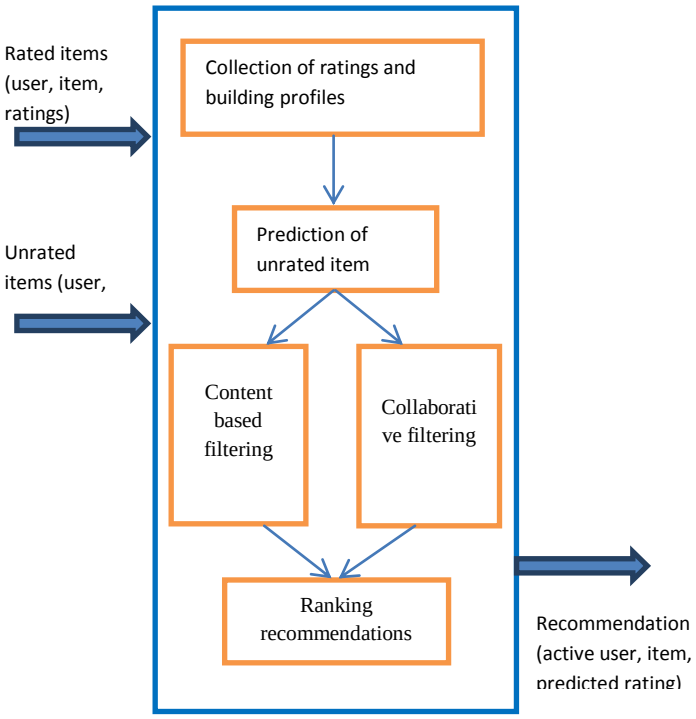

Fig. 1 A simple architecture of a recommender system

III. Approaches to a Recommender System

There are three main approaches to a recommender system.

- Content Based Recommenders
- Collaborative Filtering Recommenders
- Hybrid Approach

A. Content Based Recommenders

Content Based Filtering refers to algorithms that try to recommend items that are similar to those a user liked in the past. The key is to model the items to relevant attributes and model the user preferences by these attributes. Much research in this area has focused on recommending items with textual content such as web-pages, books, and movies. It compares the content of the item with the content that interests the user. The content of each item is represented as a set of terms, basically the words that occur in a document. The user profile is also represented by the same terms and dynamically changed based on the content of items which have been seen by the user. Similarity measures are then used to match the user profile with the item profile and matched items are recommended to the customer.

An example of content based system (fig. 2)'The Krakatoa chronicle' [6], which was the first web newspaper, a highly interactive and a personalized newspaper on the World Wide Web. It creates columns of news collected from several news sites dynamically. Each article is converted to term vectors by TF-IDF (Term Frequency – Inverse Document Frequency) weighting. User profiles are also maintained same as news articles vectors. Based on the feedback from the user, the weight of the user vector is adjusted. This system uses both explicit feedback as well implicit feedbacks from the user.

Fig. 2 Krakatoa Chronicle - An example of content based recommender [6]

Relevance feedback [11] is a useful feature of information retrieval systems, where the feedback from the initial results of the query is used to improve the subsequent results. Explicit feedback is where the users have to manually assign explicit ratings, votes, likes etc. Implicit feedback is inferred from user behavior, such as time spent on viewing an item, clicks on an item, scrolling and so on.

Newsweeder [2] a netnews-filtering system recommends news articles to the user based on the user profile learned from the ratings of news articles by the user. It uses the learning method: TF-IDF weighting, where documents in each rating category are modeled into TF-IDF word vectors. Once the item profile and user profile is created, the profiles are matched using cosine similarity. Profiles which have measured close to 1 are considered to be very similar and items with measures close to 0 are considered as dissimilar items.

1) TF-IDF (Term frequency-Inverse Document Frequency): The TF-IDF is an information retrieval technique to obtain the importance of a word in a related document.

TF-IDF = tf × idf

Where, tf (term frequency) represents the frequency of a term within the document. The larger the frequency the greater the TF-IDF weight and idf (inverse document frequency) represents how rare is the term in the collection of documents.

2) Cosine Similarity: A measure [3] to compute the similarity between two vectors, and measures the cosine of the angle between them. The equation for cosine similarity is given below:

$$Cosine(A,B) = \frac{\sum_{i-1}^{n} A_i * B_i}{\sqrt{\sum_{i-1}^{n}(A_i)^2} * \sqrt{\sum_{i-1}^{n}(B_i)^2}}$$

Where, A and B are two vectors which represents two items in the m dimensional space. The numerator represents the dot product between the two vectors and the denominator is a product of their Euclidean distance [10]. The cosine of the angle between the vectors that represent two items is their similarity. The smaller the angle, the closer are the two vectors.

Other sophisticated methods such as Bayesian classifier, nearest neighbor, neural networks [4] called as model based approaches are also used to estimate the probability of a document being liked. Web pages are represented as their feature vectors and given as input to the classifier, which labels these pages with the relevant classes and later matches it to the interests of the users. This system defines a software agent called syskill & webert that learns from the user feedback and tries to recommend other pages. In this system, experiments were conducted using different classification algorithms and accuracy was noted. Bayesian classifier seemed to perform well than the other algorithms.

Some drawbacks of content based filtering is the overspecialization, where items recommended strictly match to the user profile which narrows down the interests of the user. Another is the New User problem, where items cannot be recommended to new users who do not have a profile set. So, accurate recommendations cannot be given to new users.

B. Case Based Recommender

It is a form of content based recommender [7]. It is best suited for product recommendation domains where each product is described in terms of a well defined set of features (e.g., price, colour, brand, etc.). Case-based reasoning system uses a database of past experiences used in solving problems as its problem-solving expertise. Each case consists of a specification part which defines the problem and a solution part, giving a solution to the problem. Whenever a new problem emerges, the cases from the database are analyzed and the specification which is similar to the current problem is retrieved along with its solution to solve the new problem. Fig. 3 shows a simple form of the case based recommendation system. Based on the target user's query, using the similarity knowledge, matching products are retrieved from the case base, ranked and given as a recommendation to the user.

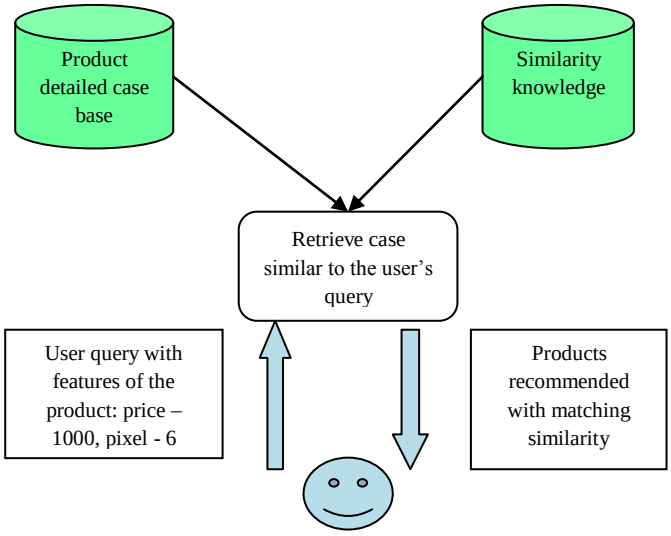

Fig. 3 Case based recommendation system [7]

C. Collaborative Filtering

In general, Collaborative Filtering (CF) is an information filtering process using techniques involving collaboration among multiple agents, data sources, etc. Applications of collaborative filtering involve very large data sets. These techniques have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration [13], environmental sensing over large areas, financial data integrating many financial sources, electronic commerce and web applications where the focus is on user data. In recommender system, collaborative filtering is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users (collaborating). Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors and preferences and predicting what users will like based on their similarity to other users. It works by collecting user feedback in the form of ratings for items in a given domain and exploiting similarities in rating behavior among several users in determining how to recommend an item.

The collaborative filtering process is shown in fig. 4. The user ratings on items are represented as a ratings matrix. Each rating is within a numerical scale, say 1-5. The CF algorithm uses this rating matrix as the input and performs prediction for the unrated items. Once predictions are done, the items are ranked and recommended to the user.

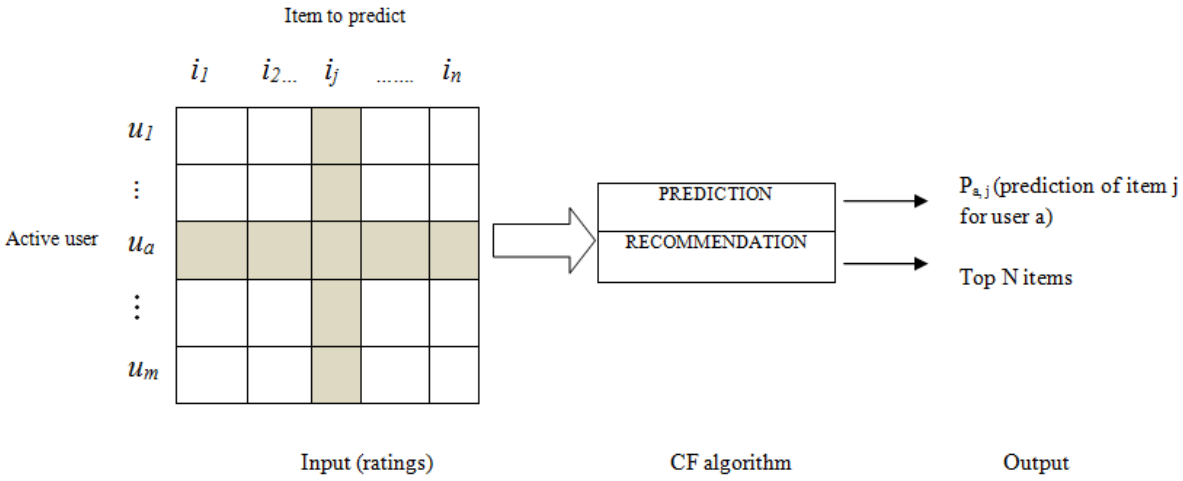

Fig. 4 Collaborative Filtering Process [3]

There are two types of collaborative filtering:

1. User-User Collaborative Filtering
2. Item-Item Collaborative Filtering

1) User-User Collaborative Filtering: This approach is also called as Neighborhood based approach. In this approach, a subset of users is selected based on their similarity to the active user. The selected user ratings are obtained, using which a prediction for the active user is made by calculating a weighted average of these ratings. These types of techniques use the similarity measures to find similarity between users and predict the ratings of the unseen items based on the similarity and recommend these items ranked from highest priority to lowest priority. Some of the similarity measures used is cosine similarity, Pearson correlation measure and so on. Figure 5 shows an approach to User Based CF using KNN algorithm.

User based algorithms are efficient as they provide recommendations that tries to match items related to the user. But these algorithms suffer from certain drawbacks. The number of users and items in major e-commerce website are very large. Most of the users however, rate only a small portion of the total items available. Even very popular items are rated by only a few of the total number of users. This means that the user-item matrix is very sparse because of which it is possible that the similarity between two users cannot be defined thus making the algorithm incapable of giving accurate recommendations. This is one of the

drawbacks called the data sparsity problem which is common in collaborative filtering. Another problem is scalability, where finding optimal clusters of users over large datasets is impractical. Most user-based recommenders use various forms of greedy cluster generation algorithms such as Lazy learner k-nearest neighbours [9]. These cluster generation algorithms require a lot of computations that grow linearly with the numbers of users and cannot be precomputed because users and items are changing over time in the database. So user-based algorithm does not scale well and are not suited for large databases of users and items.

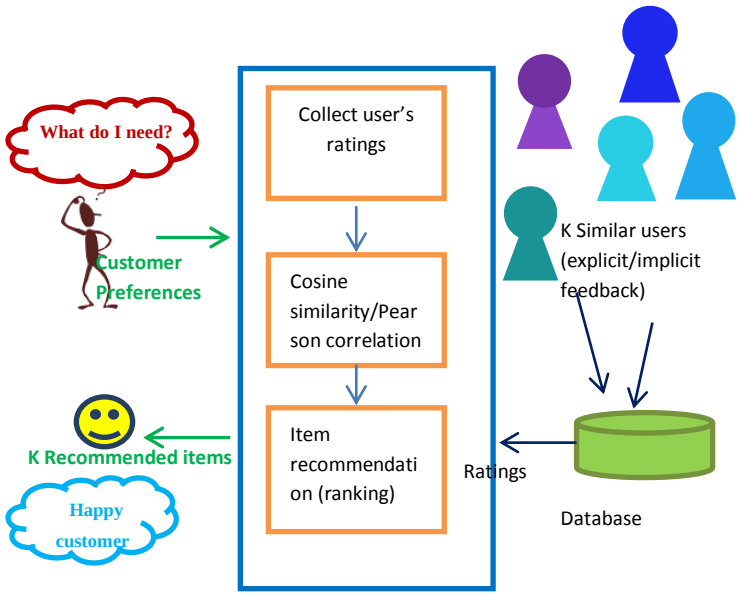

Fig. 5 A general approach to User-based KNN Collaborative filtering

2) Item-Item Collaborative Filtering: The scalability problem is well overcome by this approach. Item-Item collaborative filtering, looks into the set of items the active user has rated & computes how similar they are to the target item. From these results, the k most similar items are selected. A weighted average on the target user's ratings on the most similar items is done to make the prediction. Similarity between the items is computed using adjusted cosine similarity [3], correlation based similarity and other measures. One famous recommender system that uses item centric collaborative filtering is the one inbuilt in Amazon shopping website. Amazon.com uses recommendation algorithms to personalize its Web site for each customer's interests. It has customized its browsing page to concentrate on customers who return for more purchases. Figure 6 shows one of the recommendations used in Amazon.com. It builds a product-to-product matrix and computes the similarity between each pair [1]. Then finally it matches with the user data and recommends the items.

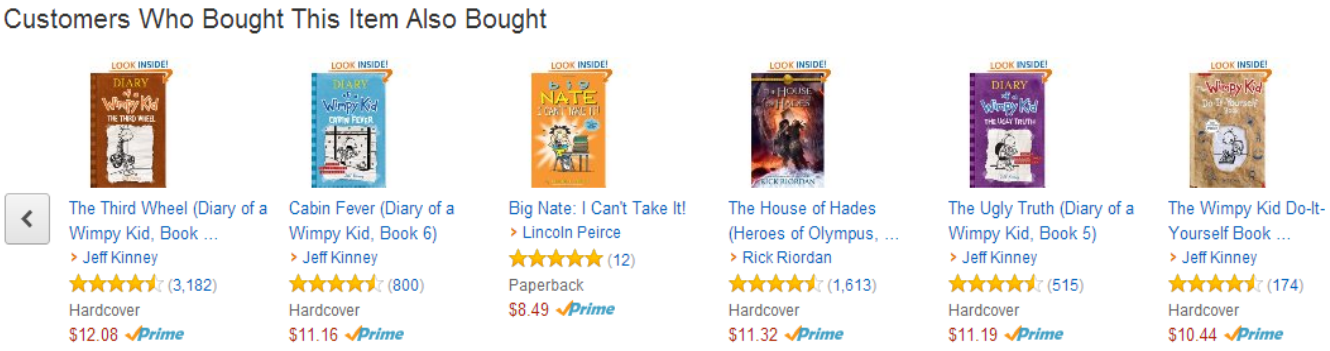

Fig. 6 Amazon.com recommendation [12]

Slope one is a family of item based collaborative filtering algorithms introduced in a research paper by Daniel Lemire and Anna Maclachlan in the year 2005. The slope one predictors are of the f (x) = x+b. It finds the average difference between the ratings of one item and another for users who rated both. These algorithms are efficient and easy to implement. They also support both online queries and dynamic updates, making them good candidates for real-world systems [8]. One of the drawbacks of

slope one algorithm is that sometimes the prediction of certain items would be biased towards largely rated items. To overcome this, the weighted slope one algorithm was defined. The prediction of item j by user u of the weighted slope one is given below:

$$p(u)_j = \frac{\sum_{i \in S(u)-j}(dev_{ij} + u_i)c_{ij}}{\sum_{i \in S(u)-j} c_{ij}}$$

Where dev_{ij} represents the average deviation of item i with respect to target item j, C_{ij} represents the number of users who rated both item i and item j, U_i is the target user's rating of item i. Its accuracy was measured by conducting experiments taking other CF based measures like Bias from Mean, adjusted cosine similarity, Pearson correlation. MAE (Mean Average Error) was used as the evaluation metric to obtain the accuracy of the algorithm. It was found that the basic SLOPE ONE scheme had a higher accuracy than bias from mean. However, Weighted slope one showed a relative accuracy of 1.98 vs. 1.94 with Pearson correlation and a better accuracy of 1.98 vs. 2.09 to adjust the cosine similarity on EachMovie dataset [8].

One problem with the item-item CF algorithm is the early rater problem. Collaborative filtering systems are mostly unable to provide recommendations for new items since there are no user ratings for these new items on which to base a prediction. Even if these items start receiving ratings, it will take some time before the item has enough ratings in order to use them to make accurate recommendations. Likewise, recommendations will also be inaccurate for new users who have rated few items.

D. Hybrid Approaches

Combination of collaborative filtering and content-based filtering as hybrid approach is becoming a way of information filtering. The hybrid approach suggests that using both methods it is possible to overcome each other's shortcoming and make the recommendation process better. Fab [5] a recommendation system combines both content and collaborative technique. It enables users to sift through the huge amount of information in the World Wide Web. It uses a framework involving separate agents to handle the individual user profile as well as the collective topic profile. The collaborative approach shares the high rated user profiles with other users and the content based approach tracks the user's interest to match web pages. Relevance feedback is used to modify the user profile as well as the topic profile. This system was able to overcome the two scaling problems, which is the increase in the number of users and the documents to be matched. Experimental results using Fab-hybrid system achieved high accuracy.

IV. CONCLUSION

Recommendation systems have significantly grown by using various techniques to boom the industrial growth. They are powerful tools for marketing of products by e-commerce sites to increase their sales. This paper presents a survey on the various techniques and how each of it has its own benefits and limitations. The content based recommenders are suitable for recommending products with descriptions. New item problem in collaborative filtering is overcome by content based filtering. Collaborative filtering provides good recommendations by considering the preferences of other users. User based CF algorithm are not as scalable as Item based CF algorithm. Hybrid approaches combine content based features and collaborative features to overcome the drawbacks of each other. This paper also discussed how various applications like Amazon.com uses the recommendation techniques. Performance evaluation of collaborative filtering methods was also discussed which used MAE as the accuracy metric. Thus it can be seen how recommender systems are evolving in its techniques to try to improve its accuracy to provide useful and meaningful recommendations to the user.

References

1. G. Linden, B. Smith and J. York," Amazon.com Recommendations: Item-to-Item Collaborative Filtering", IEEE Internet Computing, pp. 76-80, Jan. /Feb, 2003.
2. K. Lang, "Newsweeder: Learning to Filter Netnews", Proc. 12th Int.Conf. Machine Learning, 1995.
3. B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", Proc. 10th Int. WWW Conf., 2001.

4. M. Pazzani and D. Billsus,"Learning and Revising User Profiles: The Identification of Interesting Web Sites", Machine Learning, Kluwer Academic Publishers, vol. 27, pp. 313-331, 1997.
5. Balabanovic, M. and Shoham, Y.," Fab – content based, collaborative recommendation", Communications of the ACM, vol. 40, pp. 66-72, March, 1997.
6. Kamba, T., Bharat and K. & Albers, M. C.,"The Krakatoa Chronicle – an Interactive, Personalized, Newspaper on the Web", In Proc. of the 4th Int. WWW Conf., pp. 159–170, 1995.
7. Smyth, B "Case-based Recommendation", In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, vol. 4321, pp. 342–376. Springer, Heidelberg, 2007.
8. Lemire, D., and Maclachlan, "A. Slope one predictors for online rating based collaborative filtering", In Proceedings of SIAM Data Mining Conference, 2005.
9. D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, and W. Nzoukou. "A Survey Paper on Recommender Systems," arXiv:1006.5278, Dec. 2010.
10. Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze , Introduction to Information Retrieval [online] .Available: http://nlp.stanford.edu/IR-book/html/htmledition/dot-products-1.html
11. Relevance Feedback, [online], available: http://en.wikipedia.org/wiki/Relevance_feedback. accessed on- 23/11/2013
12. Amazon (2003) [online], Available: http://www.amazon.com, accessed on 23/11/2013.
13. Collaborative Filtering [online], Available: http://en.wikipedia.org/wiki/Collaborative_filtering, accessed on 23/11/2013.

AUTHOR(S) PROFILE

Miss. Ramya Laurraine.U, Pursuing Master of Engineering in Velammal Engineering College, Affiliated to Anna University, Chennai. Received Bachelor of Engineering Degree in Computer Science and Engineering in 2006. Currently working on a project in Recommender System.

Mr. M. Shanmugasundaram, completed his B.E in Computer Science and Engineering from Madurai Kamaraj University in 1988 and M.E in Computer Science and Engineering from College of Engineering Guindy, Anna University in the year 1992. He has two decades of professional experience - both in IT industry and in academia. Currently, he is working as Faculty in the Department of Computer Science and Engineering at Velammal Engineering College, Chennai, Tamil Nadu, India. His research interests include Data Mining, Social Network Mining and Data Mining Applications in Disaster Management.

