

Methods of Mining the Data from Big Data and Social Networks Based on Recommender System

P.Priyanga

Department of Computer Science, Thassim Beevi Abdul Kader College for Women, Kilakarai-17.
Email: p.priyanga19@gmail.com

Dr.A.R.Nadira Banu Kamal

Department of Computer Science, Thassim Beevi Abdul Kader College for Women, Kilakarai-17.
Email: nadirakamal@gmail.com

ABSTRACT

Data has become a vital part of every action of offline based or online based activities. By use of some technological advancement like internet, Smartphone, social network, mobile computing, fine tuning of ubiquitous/ pervasive computing, the size of data grows exponentially day by day. Big data technology is sure to soon knock on the door of every enterprise, organization, and domain. Organizations data is one of the organization's assets. Because of daily operational activity, data will grow up. More data means the organization need more effort to choose which data is important to process become information. Big data is not big if we know how to use it. Mining is the analysis of the data with the indent to discover the gems of hidden information in the vast quantity of data that has been captured in the normal course of running the business. At present most of the enterprises realize the significance of using voluminous data to take better decisions. Even though a larger amount of data gives a better output, there will be a challenging task for processing the data. Mining the information helps organizations to make knowledge driven decisions. It applies many computational techniques available. Recommender System is one of the most information filtering techniques for nowadays. Recommender systems look for to predict the "rating" or "preference" that a user would give to a particular item. This paper gives the general idea about the methods used to filtering or mining the information from big data and social networks based on recommender systems.

Keywords – Big Data, Challenges and Issues, Data Mining, Recommended System, Social Network.

1. Introduction

Recommender Systems (RS) typically apply techniques and methodologies from other neighboring areas – such as Human Computer Interaction (HCI) or Information Retrieval (IR). However, most of these systems bear in their core an algorithm that can be understood as a particular instance of a Data Mining (DM) technique. The data mining methods that are most commonly used in RS: classification, clustering and association rule discovery. There are much variety of different methods are available in mining of the data or information based on the recommended system. There is a lot of data available in our offline activities and online activities. We have a more technology and computing power processing the data but the objective of the processing is more or less same, to know about the data, explain about the data and finally predict the data. Big data is not only the much data the association of Data Mining, Science, Artificial Intelligence, Parallel Processing, Storages and So on. It is a high Volume, Variety, Velocity, veracity and Value of the information.

Big Data is becoming more popular with the internet technology development, which means the data sets whose size is beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time. There are many recommendation algorithms, such as collaborative filter algorithm [3],

social recommendation algorithm [4], content based filtering algorithm [5], and so on.

Online social networks facilitate connections between people based on their shared interests, values, membership in particular groups (i.e., friends, professional colleagues, Students, Researchers, etc. They make it easier for people to find and communicate with individuals who are in their networks using the Web as the interface.

Social networking on social media websites involves the use of the internet to connect users with their friends, family and acquaintances. Social media websites are not necessarily about meeting new people online, although this does happen. Instead, they are primarily about connecting with friends, family and acquaintances you already have in real life.

Table 1: Bits and Bytes of data

Byte	8 bits	One Grain Of Rice	 Hobbyist
Kilobyte	1024 bytes	Cup Of Rice	 Desktop
Megabyte	1024 Kilobyte	8 Bags Of Rice	
Gigabyte	1024 Megabyte	3 Semi Trucks	 Internet
Terabyte	1024 Gigabyte	2 Container Ships	

Petabyte	1024 Terabyte	Blankets Manhattan	
Exabyte	1024 Petabyte	Blankets west coast states	 Big Data
Zettabyte	1024 Exabyte	Fills the Pacific Ocean	
Yottabyte	1024 Zettabyte	A Earth Size Rice Ball!	 Future?

Online shopping is the process whereby consumers directly buy goods or services from a seller in real-time, without an intermediary service, over the Internet. It is a form of electronic commerce. Online shopping is one of the social network area, the peoples are connected by feedback and recommendation of particular product or items.

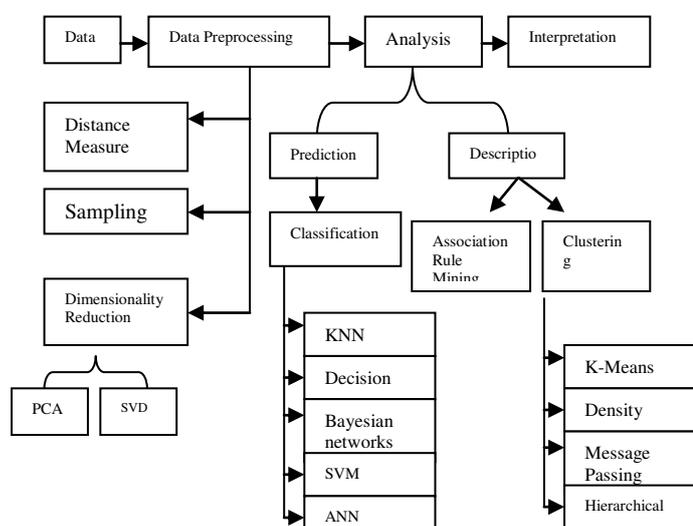


Fig 1: Data Mining Methods for Recommender Systems

2. Data Mining Methods for Big Data

Extract relevant business information from large data sets, many methods have been established that are based on identifying important relationships, patterns, and trends. These methods can also be used for statistical processes.

- **Outlier detection:** Extreme values that stand out from the rest of data are known as outliers.
- **Cluster analysis:** Clusters refer to a group of objects that, in one way or another, are similar to one another. The goal of this analysis is to segment unstructured data. To this end, algorithms are used to search for similarities in the structures of large data sets, in order to identify new clusters.
- **Classification:** While cluster analyses primarily aim to identify new groups, classification involves making use of predefined classes. Allocating these occurs with the help of matching characteristics from the data set. A decision tree presents a common method for automatically classifying data. For each node, a property of the

object is called up. The presence of this property determines the choice of the following node.

- **Association analysis:** An association analysis aims to identify relationships within data sets that can be formulated as inference rules. When it comes to e-commerce, these data mining methods can be used in order to identify the correlation of individual products within shopping carts along the pattern of 'if product A is bought, then product B will also be bought'.
- **Regression analysis:** Regressions analyses help create models that explain dependent variables through various independent variables.

One of the best methods for turning raw data into useful information is Map Reduce method. Map Reduce is a method for taking a large data set and performing computations on it across multiple computers, in parallel. In essence, Map Reduce consists of two parts. The Map function does sorting and filtering, taking data and placing it inside of categories so that it can be analyzed. The Reduce function provides a summary of this data by combining it all together.

2.1 Tools used for analyse the big data.

Perhaps the most influential and established tool for analyzing big data is known as Apache Hadoop. Apache Hadoop is a framework for storing and processing data in a large scale, and it is completely open source. Hadoop is broken into four main parts:

- The **Hadoop Distributed File System (HDFS)**, which is a distributed file system designed for very high aggregate bandwidth;
- **YARN**, a platform for managing Hadoop's resources and scheduling programs which will run on the Hadoop infrastructure;
- **MapReduce**, sorting and filtering, taking data and placing it inside of categories, a model for doing big data processing;
- And a common set of libraries for other modules to use.

2.1.2. Data Ingestion Strategy and Acquisition

Data ingestion in the Hadoop world means ELT (Extract, Load and Transform) as opposed to ETL (Extract, Transform and Load) in case of traditional warehouses.

- Determine the frequency at which data would be ingested from each source
- Data append replace ,Pre-Processing

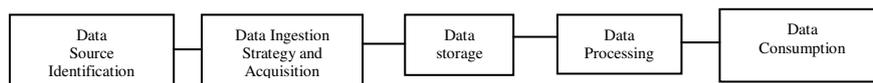


Fig 2: planning the Logical layers of Big Data architecture

- Segregate the data sources based on mode of ingestion – Batch or real-time

2.1.3. Data Storage

Store large amounts of data of any type and should be able to scale on need basis. We should also consider the number of IOPS (Input output operations per second) that it can provide. Hadoop distributed file system is the most commonly used storage framework in BigData world, others are the NoSql data stores – MongoDB, HBase, Cassandra etc. One of the salient features of Hadoop storage is its capability to scale, self-manage and self-heal.

There are 2 kinds of analytical requirements that storage can support:

- **Synchronous** – Data is analysed in real-time or near real-time, the storage should be optimized for low latency.
- **Asynchronous** – Data is captured, recorded and analysed in batch.

Things to consider while planning storage methodology:

- Type of data (Historical or Incremental)
- Format of data (structured, semi structured and unstructured)
- Compression requirements
- Frequency of incoming data
- Query pattern on the data
- Consumers of the data

2.1.4. Data Process

The Processing methodology is driven by business requirements. It can be categorized into Batch, real-time or Hybrid based on the SLA.

- **Batch Processing** – Batch is collecting the input for a specified interval of time and running transformations on it in a scheduled way. Historical data load is a typical batch operation
Technology Used: Map Reduce, Hive, Pig
- **Real-time Processing** – Real-time processing involves running transformations as and when data is acquired. Technology Used: Impala, Spark, spark SQL, Tez, Apache Drill

- **Hybrid Processing** – It's a combination of both batch and real-time processing needs. Best example would be lambda architecture.

2.1.5. Data Consumption

This part consumes the output provided by processing layer. Different users like administrator, Business users, vendor, partners etc. can consume data in different format. Output of analysis can be consumed by recommendation engine or business processes can be triggered based on the analysis.

Different forms of data consumption are:

- **Export Data sets** – There can be requirements for third party data set generation. Data sets can be generated using hive export or directly from HDFS.
- **Reporting and visualization** – Different reporting and visualization tool scan connect to Hadoop using JDBC/ODBC connectivity to hive.
- **Data Exploration** – Data scientist can build models and perform deep exploration in a sandbox environment. Sandbox can be a separate cluster (Recommended approach) or a separate schema within same cluster that contains subset of actual data.
- **Adhoc Querying** – Adhoc or Interactive querying can be supported by using Hive, Impala or spark SQL.

And finally, the key things to remember in designing Big Data Architecture are:

- **Dynamics of use case:** There a number of scenarios as illustrated in the article which need to be considered while designing the architecture – form and frequency of data, Type of data, Type of processing and analytics required.
- **Myriad of technologies:** Proliferation of tools in the market has led to a lot of confusion around what to use and when, there are multiple technologies offering similar features and claiming to be better than the others.

3. Social Network Analysis

Social network analysis includes visual and formal analysis of human relationship. Web and pages existing in

it are an example of social network. Actually, pages can be considered as nodes and links among them as edge among these nodes. On the other side, as new generation of webs appeared and considering their main factors i.e. weblogs and wikis, the importance of social networks in web is now higher. Recently, much attraction and interest has been observed in social network analysis among data mining groups. Its main motivation is exploitation, recognition and awareness of values collected concerned with users' social behavior in on-line environments. Data mining techniques when analyzing social network data especially for massive data collections are considered as useful which are not controllable through traditional methods.

In geographical based systems, user's preferences, its required information or dimensions related to the content used e.g. day of the week, weather, time, user's activity and transfer tool are not presented. Therefore, user becomes afloat in suggestions that may even be of no interest for him/her and there is a possibility that this suggestion result in dissatisfaction for user. The problem gets worse in mobile devices due to limitations like; small screen, limited input and things like that [2].

4. Recommender System

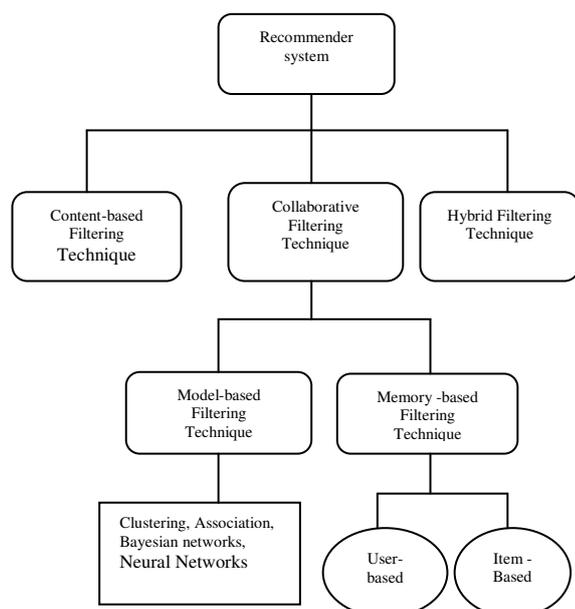


Fig 3: Recommendation techniques

Recommender Systems make use of different sources of information for providing users with predictions and recommendations of items. They try to balance factors like accuracy, novelty, disparity and stability in the recommendations [1]. A typical Recommendation system cannot do its job without sufficient data and big data supplies plenty of user data such as past purchases, browsing history, and feedback for the Recommendation systems to provide relevant and effective recommendations. In a nutshell, even the most advanced

Recommenders cannot be effective without big data. Social recommender systems are a combination of social data on web like; user's social networks and spatial information. Because user's information include personal information and interests in social network sites, considering user's current location and the information existing in social network data base, it is possible to provide user with a suitable suggestion. Through this method users' interaction decreases and they can acquire their favorite information and services.

4.1. Collaborative Filtering

Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. Many algorithms have been used in measuring user similarity or item similarity in recommender systems. For example, the k-nearest neighbour (k-NN) approach and the Pearson Correlation.

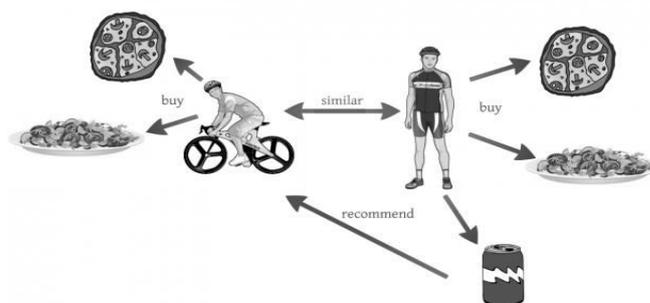


Fig 4: Collaborative Filtering

4.2. Content Based Filtering

Content-based filtering methods are based on a description of the item and a profile of the user's preference. In a content-based recommendation system, keywords are used to describe the items; beside, a user profile is built to indicate the type of item this user likes. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach has its roots in information retrieval and information filtering research.

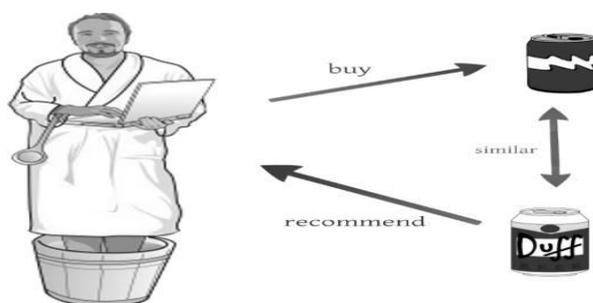


Fig 5: Content Based Filtering

4.3. Hybrid Recommendation Systems

Recent research has demonstrated that a hybrid approach, combining collaborative filtering and content-based filtering could be more effective in some cases. These methods can also be used to overcome some of the common problems in recommendation systems such as cold start and the sparsity problem.

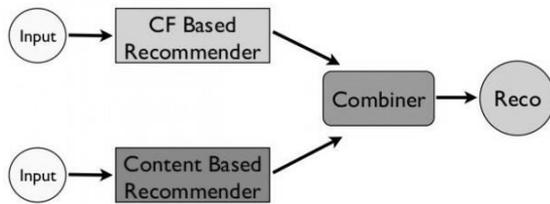


Fig 6: Hybrid Recommendation Systems

5. Issues and Challenges

5.1. Benefits of Recommendation system

- Recommendation systems are based on actual user behavior i.e. objective reality.
- Recommendation systems are great for discovery.
- Recommendation systems are effective tools for personalization.
- Recommendation systems are always up-to-date.
- Most of the organizational maintenance of a site is keeping the navigation system in line with the users' changing needs. Based on user activity, the system recommends navigation options to the user.
- Recommendation systems are intensive, database-driven applications that are difficult to set up and get running.

5.2. Limits of Recommendation systems

- Recommenders **depend totally on data** and their hirers must constantly supply them with large volumes of data. That is why; smaller firms are more disadvantaged than the bigger firms such as Google and Amazon.
- Recommenders may **find it difficult to exactly identify user choice patterns** if the user preferences tend to vary quickly, as in fashion. Recommenders depend a lot on historic data but that may not be suitable for certain product niches.
- Recommenders face problems with **unpredictable items**.

5.3. Recommender System Challenges and Issues

- Cold start
- Sparsity
- Trust
- Privacy
- Scalability

5.4. Social Network Challenges and Issues

- Privacy
- Terms of Agreements
- Losing Authenticity
- Unbelievable worthy member data
- Dishonesty
- Misuse
- Access for those disabilities
- Security Concerns
- Slower response rate
- Loss in work place activity
- Spamming
- Commercial advertising on Social Media
- Deception
- Difficult to monetize
- Lacks of metrics hard to measure
- Integrity risk

5.5. Big Data Challenges and Issues

- Privacy Security and
- Analytical challenges
- Infrastructure fault
- Data access and sharing of information
- Skill requirements
- Issues related to characteristics
- Storage and processing issues
- Technical Challenges
- Technical challenges
 - Fault tolerance
 - Quality of data
 - Scalability
 - Heterogeneous data
- Issues related to characteristics
 - Data volume, data velocity, data variety, data value, data complexity

6. Conclusion

In this paper we presented some of the general ideas of mining methods for filtering the data from big data and social network based on recommender system. Social network and big data information are very vast that it can't be closed using hand. Recommendation systems are not only filtering the data, but also it gave some decision making power for all the offline and online activities of day by day life.

REFERENCES

- [1] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey Universidad Politécnica de Madrid, Ctra. De Valencia, Km. 7, 28031 Madrid, Spain, journal homepage: www.elsevier.com/locate/knossys, *Knowledge-Based Systems* 46 (2013) 109–132.
- [2] Fatemeh Khoshnood, Mehregan Mahdavi and Maedeh Kiani sarkaleh, Designing a Recommender System Based on Social Networks and Location Based Services, *International Journal of Managing Information Technology (IJMIT)* Vol.4, No.4, November 2012.
- [3] G. F. Sun, L. Wu, and Q. Liu, et al., "Recommendations based on collaborative filtering by exploiting sequential behaviors", *Journal of Software*, Vol. 24 No. 11, pp.2721-2733, 2013.

- [4] L. Guo, J. Ma, and Z. M. Chen, et al., “Incorporating item relations for social recommendation”, *Chinese Journal of Computers*, Vol. 37, No. 1, pp. 218-228, 2014.
- [5] Y. R. Wang, M. Chen, and H. H. Wang, et al., “A content-based filtering algorithm for scientific literature recommendation”, *Computer Technology and Development*, Vol .21, No. 2, pp. 66-69, 2011.
- [6] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation, *Egyptian Informatics Journal* (2015) 16, 261 – 273.
- [7] Soanpet .Sree Lakshmi ,Dr.T.Adi Lakshmi, Recommendation Systems:Issues and challenges, *International Journal of Computer Science and Information Technologies*, Vol. 5 (4) , 2014, 5771-5772.
- [8] Daniar Asanov, Algorithms and Methods in Recommender Systems.
- [9] Jaseena K.U. And Julie M. David Issues Challenges,And Solutions:Big Data Mining, Computer Science & Information Technology (Cs & It) , Pp. 131–140, 2014.
- [10]Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol, P.B (Eds) Handbook of Data Mining Methods for Recommender Systems (2). (New York-Springer.)

Websites:

<https://haifengl.github.io/bigdata/>
<https://opensource.com/resources/big-data>
http://en.wikipedia.org/wiki/Social_network_service
<http://accan.org.au/tip-sheets/introduction-to-social-networking>
<http://www.kdnuggets.com/2015/10/big-data-recommendation-systems-change-lives.htm>
<https://www.1and1.com/digitalguide/online-marketing/web-analytics/data-mining-analysis-methods-for-big-data/>