Science-related Articles Recommendation System from Big Data

Mie Khine Oo, Myo Kay Khaing
University of Computer Studies, Yangon
mikhineoo@ucsy.edu.mm, myokaykhaing5@gmail.com

Abstract

Under the explosive increase of global data, the term Big data is mainly used to describe enormous datasets. With the availability of increasingly large quantities of digital information, it is becoming more difficult for researchers to extract and find relevant articles pertinent to their interests. In this system, we propose an approach to discover and recommend the desired articles by combining collaborative filtering (CF) with topic modeling. Correlated Topic Model (CTM) is used for modeling topics. Our approach not only considers the interactions between users through collaborative filtering but also learns the properties of items involved through topic modeling to improve recommendation. In order to handle a large dataset, a Big data analytics tool Hadoop is used to perform processing over distributed clusters. The proposed approach learns the accuracy of the recommendation.

Keywords – *Big data, Collaborative filtering, Topic modeling, Recommendation System*.

1. Introduction

Big data is mainly used to describe massive, heterogeneous, and unstructured digital content that is difficult to process using traditional data analysis tools and techniques. Compared with traditional datasets, Big data typically includes masses of unstructured data that need more analysis [9]. With the rapid growth of various Big data, users find it challenging to locate information that is relevant to their purposes. Although this growth has allowed researchers to access more scientific information, they need new ways to manage the massive influx of information.

Recommender systems are information filtering systems that can solve the information overload problem by filtering crucial information from large amounts of data. The goal is to generate meaningful recommendations to users according to users' preferences, interests, or observed behavior about item to improve user satisfaction.

There are basically two basic approaches for making recommendations. A recommendation system could be a content-based approach or a collaborative filtering (CF) approach. Content-based recommender system matches the user profile and some specific characteristics of an item while CF recommender system filters information based on the collaboration of users or the similarity between items [13]. For large data sets, CF methods give better accuracy and performance. However, it can suffer from cold-start, scalability, and sparsity problems.

Because Big data produces large data size, the performance may be slow if recommendation has been done in a single system. To handle this slowness problem, a distributed environment is needed to increase the computation of recommendation. An open-source Big data processing platform Hadoop provides a reliable,

scalable, and manageable distributed environment. Hadoop provides distributed processing in less amounts of time by processing MapReduce implementation [5]. Therefore, the proposed system has been modeled on Hadoop platform. The purpose of this paper is to alleviate the CF problems, to develop a collaborative filtering recommender model, and to perform parallel computation to deal with Big data using Hadoop.

The main contribution of this paper is proposing a Big data recommendation system that can able to recommend unseen and latent articles to the user by the use of correlated topic modeling. Another contribution is item-based collaborative filtering model on increasing volume of Big data that can able to perform online recommendation system within acceptable response time.

The rest of the paper has been arranged in different sections. Section 2 describes the related researches in this area. In section 3, we defines about Big data, and its characteristics, and challenges that are now facing. Section 4 explains about the background theory. Section 5 presents the brief description of the proposed approach for Big data on Hadoop platform. Then section 6 focus on conclusion and further extensions.

2. Related Works

The most related works are the implementations of collaborative filtering (CF) algorithms used in recommendation systems in terms of Hadoop platform that utilizes MapReduce framework. Recommendation with Big data is still an ongoing research direction.

Zhao et al [16] proposed an implementation of user-based CF algorithm based on Hadoop. Their approach is scalable and showed that they take the improved performance. But the Hadoop platform cannot reduce the recommendation response time, and the calculation process of MapReduce is very resource-consuming.

Pandey et al [10] proposed a CF recommendation system which combines the results of user-based and item-based CF based on a threshold value. The system is modeled on Hadoop to solve the scalability problem. To perform CF on Hadoop, an open-source java library, Apache Mahout is used to enhance accuracy of the recommendation.

Vinodhini et al [15] proposed a hybrid recommendation system that recommend books to the user by combining the two approaches, CF and content-based filtering. They considered both the ratings of the user and the item's feature. The dataset used is Big data so that a Big data analysis tool Hadoop is used. They showed that the proposed system is reliable, fault tolerant, adaptive, and more accurate than existing recommender systems.

Recently there are few works have been done on topic modeling with Big data.

Hansmann et al [5] derived dimensions for Big data from existing definitions of Big data. These dimensions are validated and enriched with a two-step topic models. The first step validates the derived dimensions on all of the identified publications. The second step enriches the individual dimensions on dimension-specific publications. The results are assigned to a generic data analysis process to clarify the Big data research.

Chen et al [2] proposed a lifelong topic model (LTM) to automatically and dynamically mine the prior knowledge to generate more coherent topics in order to solve the problem of incoherent topics generation. The model can also be deal with Big data by dividing a Big dataset into a number of small datasets from multiple domains. The LTM can be implemented in

MapReduce to get better topic quality and to reduce the execution time.

3. Big Data

Big data is a collection of data sets that is so large and complex that it becomes difficult to process using traditional data processing approaches. Big data comes from different sources and that data can be structured, semi-structured, and unstructured. Structured data has formal schema and data model as well as similar formats and predefined lengths, and are generated by users or automatic data generators, including computers and sensors. Unstructured data does not have a predefined data model. Satellite images, videos, and social media data are examples of unstructured data. Semi-structured data need not to have a predefined length or type, it lacks strict data model structure [17]. Many definitions are evolved for big data, for example, TechAmerica Foundation defines Big data as follows [14]:

"Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information."

Similarly, O'Reilly defines Big data as [3]: "Big data is data that exceeds the processing capacity of conventional database systems. The data is huge and massive, moves at a very high speed, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it."

3.1. Characteristics of Big Data

Big data is described by using the following characteristics: Volume, Variety, Velocity, Veracity, and Value, or the five Vs [8].

- Volume refers to the size of the data set that is growing enormously. It is the vast amounts of data generated every second.
- Variety refers to the different formats and types of data that Big data can comprise. This data can be structured as well as unstructured.
- Velocity refers to the increasing speed (rate) how fast the data is generated and processed to meet the challenges of Big data. Big data processing can be batch, near-time, real-time, and streams.
- Veracity refers to the quality, understandability or trustworthiness of data, i.e., how much the data can be trusted when the reliability of its source is given.
- Value is the most important of Big data. It means a worth that a company can gain from employing Big data. The value lies in detailed analysis of given accurate data, and the information and insights it provides.

In [9], there are two more Vs to represent Big data.

- Variability refers to data whose structure changes constantly.
- Visualization is a way of presenting the data in a manner that's understandable and accessible. It is one of the challenges of Big data.

3.2. Issues & Challenges with Big Data

There are many research challenges in the field of Big data. Nowadays, the development of information technology brings huge challenges to data acquisition, storage, management, and processing of the explosive growing Big data. The key challenges in the analytics of Big data are listed below [4] [12]:

- Distributed data source: Big data comprises various data sets which come from different

sources, and has heterogeneity in data type, structure, and locations.

- Unstructured nature of data: Almost 80% of Big data are unstructured data, and therefore this data is needed to transform to a proper structured data to make efficient processing.
- Time evolving data: Big data is changing constantly or increasing dramatically over time. So effective algorithms and methods are required to adapt with this time evolving data.
- Analytics architecture: The architecture of the analytics system should be able to perform a scalable and speedy processing, and respond with a timely manner.
- Data compression: Datasets can contain redundant data, and they take more space to store. To reduce the storage cost of the system, data compression is needed.
- Privacy: Handling sensitive and private data plays an important role in Big data. Suitable protection measures are required when transferring data between various sources.

4. Background Theory

Today, there are huge amounts of online documents are available and easily accessed by online users. So it needs a better way to manage these online documents by using new techniques and tools that can automatically organize and search documents. This section gives the theory background of CTM and two types of CF.

4.1. Correlated Topic Model (CTM)

Topic modeling techniques are used to find patterns of words in documents, and these documents are mixtures of topics, where a topic is a probability distribution over words [1]. The four methods of topic modeling are: Latent semantic analysis, Probabilistic latent semantic analysis, Latent Dirichlet Allocation (LDA), and Correlated topic model (CTM).

CTM builds on LDA model which assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. A limitation of LDA is inability to model topic correlation because of independence assumption. CTM models the correlation between latent topics. Each document exhibits multiple topics in different proportions. The key of CTM is the logistic normal distribution. The logistic normal cannot conjugate to the multinomial, which adds complexity to the variational inference process. But it provides a more expressive document model and more robust to overfitting [6].

Specifically, the correlated topic model assumes that an N-word document arises from the following generative process. Given topic distributions for K topics $\beta_{1:K}$, a K-vector μ and a $K \times K$ covariance matrix $\sum$:

1. Draw $\eta_d \mid \{\mu,\sum\} \sim \mathrm{N}(\mu,\sum)$ where η_d is a natural parameterization of the multinomial for d^{th} document, $\{\mu,\sum\}$ is a K-dimensional mean and covariance matrix, and $\mathrm{N}(\mu,\sum)$ is a K-1 dimensional Normal distribution.
2. For $n \in \{1,...,N_d\}$ where N_d is the number of words associated with document d,
   a) Draw $z_{d,n} \mid \eta_d$ from $Mult(f(\eta_d))$ where $z_{d,n}$ is a topic assignment associated with n^{th} word and d^{th} document.
   b) Draw $w_{d,n} \mid \{z_{d,n},\beta_{1:K}\}$ from $Mult(\beta z_{d,n})$ where $w_{d,n}$ is a n^{th} word in the d^{th} document.

Let $f(\eta)$ maps a natural parameterization of the topic proportions to the mean parameterization,

$$f(\eta)=\frac{\exp\{\eta\}}{\sum_i \exp\{\eta i\}} \quad (1)$$

The CTM models the same type of data as LDA and only differs in the first step of the generative process.

4.2. Collaborative Filtering (CF)

Collaborative filtering is a machine learning algorithm which is widely used for recommendation purposes. A big problem of CF is the scalability problem, which occurs when the volume of the dataset is very large. In order to alleviate this problem, the CF algorithm is implemented on cloud computing platform [7].

CF technique works by building a database (user-item matrix) of preferences for items by users. It then matches the users with relevant interests by calculating similarities between them to make recommendations. Such users build a group called neighborhood. In such a way, a user gets recommendations to those items that he has not rated before but that were already positively rated by users in his neighborhood [13].

There are two types of CF techniques. Memory-based CF are not as fast and scalable as model-based CF especially in processing very large datasets. Memory-based algorithms use the whole dataset to compute their recommendation. They use similarity measures to select users or items that are similar to the active user. Then, the recommendation is calculated from the ratings of these neighbors. The idea of memory-based recommendation systems is to calculate and use the similarities between users or items and use them as weights to predict a rating for a user and an item.

Model-based CF involves building a model based on the dataset of ratings. This technique analyzes the user-item matrix to identify relations between items, and uses these relations to compare the list of recommendations. The idea is to extract some information from the dataset, and use that information to build a model to make recommendations without having to use the complete dataset every time. This approach increases both speed and scalability, and resolves the sparsity problem. A model-based system allows trimming of the model to make the system more scalable. It can limit the number of similar entities (users or items) that are stored for each entity [7].

5. Proposed Approach

The idea of the proposed approach is to develop a recommendation system using collaborative filtering to recommend scientific articles to the users with increased accuracy in terms of correlated topic modeling on Big data. Our approach considers two phases. The former is a model building phase, which involves preprocessing, extracting topics, building utility matrix, and building recommendation model. The latter is a recommendation phase, which uses the pre-computed model to recommend what kind of items the user may prefer.

For the system to process Big data, we use Hadoop, which is an open-source platform to distribute processing. Firstly, the user sends the request to the recommender system, the system calculates the data by using the recommendation model, then the system sets back the recommender results to the user, and then the results will be displayed on the interface screen.

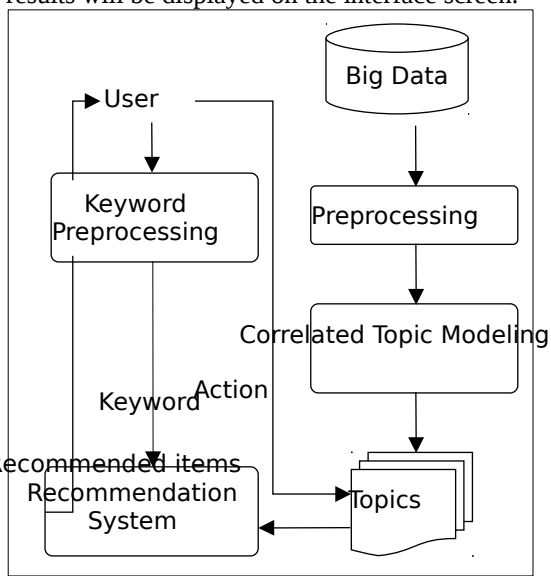

Figure 1. Design of the proposed approach

5.1. Model Building Phase

5.1.1. Preprocessing

The large volume of data can demand preprocessing tasks for filtering data. Preprocessing is necessary if analytics is to yield trustworthy and useful results, and to speed up the recommendation. Most raw data, especially Big data, are not suitable for human consumption, but the information derived from the data is. The input to the preprocessing step is a large documents collection from a large number of domains, which is called the Big data. Preprocessing reduces the number of files that need to be subjected to detailed analysis by checking the extensions of files included in the collection. The output is a large set of documents containing only Portable Document Format (pdf) files.

5.1.2. Extracting Topics

After preprocessing, topics are extracted from documents by utilizing correlated topic modeling (CTM) technique. The idea of correlated topic modeling is that every document is about several topics and that each word in the document can be associated with one of these topics. The topic representation of articles allows to make meaningful recommendations of articles before anyone has rated them. Once a set of topics have been extracted, where each topic contains the lists of the most probable words, the optimal number of topics must be evaluated to contain only the top coherence topics.

5.1.3. Building Utility Matrix

In this step, a utility matrix is constructed to build the recommendation model. The utility matrix, also known as users' preferences matrix, offers known information about the degree to which a user likes an item. Without a utility matrix, it is almost impossible to recommend items. The utility matrix is discovered by implicitly monitoring the actions of user on the system – that is, to log the actions of user, which is known as the implicit rating. The preference of each user is monitored by the system in a way that a user clicked the item only if he has interest in it. Therefore, the value of 1 means that the user clicked the item. And 0 means that the user has not viewed the item. This sort of rating system has only one value (0 or 1). After collecting preferences of the user, item-based CF approach is used to produce recommendations for users.

5.1.4. Building Recommendation Model

This step computes the similarity between items and then select the most similar items. There are many techniques to compute similarity between items. In this approach, the Pearson correlation is used to compute similarity between two items. The co-rated cases are isolated to make the correlation computation accurate. The co-rated case means that both items are rated by the users. The correlation similarity between items i and j is given by

(2)

$$¿(i,j)=\frac{\sum_{u\in U}\left(R_{u,i}-\hat{R}_i\right)\left(R_{u,j}-\hat{R}_j\right)}{\sqrt{\sum_{u\in U}\left(R_{u,i}-\hat{R}_i\right)^2}\sqrt{\sum_{u\in U}\left(R_{u,j}-\hat{R}_j\right)^2}}$$

where U is the set of users who both rated i and j, $R_{u,i}$ is the rating of user u on item i, and $\hat{R}_i$ is the average rating of the i^{th} item [11].

5.2. Recommendation Phase

To perform recommendation, the user's keyword to be searched is entered to the recommendation system. Recommendation of items is computed by taking a weighted average of the target user's ratings on these similar items. Each rating is weighted by the corresponding similarity $s_{i,j}$ between items i and j. The prediction of on an item i for a user u is given by (3)

$$P_{u,i}=\frac{¿\ s_{i,N}\vee ¿\ ¿\ ¿\ \sum_{all\ similar\ items,N}\ ¿\ \left(s_{i,N}*R_{u,N}\right)}{\sum_{all\ similar\ items,N}\ ¿}$$

where N is the similar item, $s_{i,N}$ is the similarity between i and N, and $R_{u,N}$ is the rating of user u on N [11]. The output of the recommendation is a list of top items along with their top preferred articles that the user will like the most.

6. Conclusion and Future Works

Big data is often heterogeneous, inter-related and untrustworthy. Due to the emergence of Big data, it becomes a difficult task to recommend the preferences for users. Thus, CF algorithms face the problem of large dataset and data sparsity in utility matrix, which describes the relation between users and items. In this paper, we have proposed a collaborative filtering method for recommending science-related articles. The proposed approach is intended to provide some advantages. First, it can be used to solve the scalability problem because the computation of recommendation is implemented in MapReduce framework that process large amounts of data in parallel. Second, the use of model-based collaborative filtering technique reduces the recommendation response time because model-based technique do not need to utilize the whole dataset every time to compute recommendations. Finally, correlated topic modeling technique able to recommend unseen and latent articles to the user, and thus increases the recommendation accuracy. There are still many further works that are needed to be done for better recommendation. Improving the recommendation accuracy on the sparse data will be one of our future works.

References

[1] Blei, D. M. and Lafferty J. D., "A correlated topic model of Science", The Annals of Applied Statistics, Volume 1, No. 1, 2007.

[2] Chen, Z. and Liu, B., "Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data", Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014.

[3] Dumbill, E., "What is Big data?" January 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html

[4] Fan, W. and Bifet, A., "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2, 2013.

[5] Hansmann, T. and Niemeyer, P., "Big Data – Characterizing an Emerging Research Field using Topic Models", International Joint Conferences on WI and IAT, IEEE, 2014.

[6] Huang, J. and Malisiewicz, T., “Correlated Topic Model Details”.
[7] Isinkaye, F. O., Folajimi, Y. O. and Ojokoh, B. A., “Recommendation systems: Principles, methods and evaluation”, Egyptian Informatics Journal, 2015.
[8] Marr, B., “Big Data: The 5 Vs Everyone Must Know”, March 2014. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.html
[9] McNulty, E., “Understanding Big Data: The Seven V’s”, May 2014. http://dataconomy.com/seven-vs-big-data.html
[10] Pandey, S. and Kumar, T. S., “Customization of Recommendation System using Collaborative Filtering Algorithm on Cloud using Mahout”, International Journal of Research in Engineering and Technology, May 2014.
[11] Sarwar, B., Karypid, G., Konstan, J. and Riedl, J., “Item-based Collaborative Filtering Recommendation Algorithms”, WWW10, May 2001.
[12] Singh, N., Garg, N. and Mittal, V., “Big Data – insights, motivation and challenges”, International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December 2013.
[13] Su, X. and Khoshgoftaar, T. M., “A Survey of Collaborative Filtering Techniques”, Advances in Artificial Intelligence, Volume 2009.
[14] TechAmerica Foundation’s Federal Big Data Commission, “Demystifying Big data: A practical guide to transforming the business of Government”, 2012.
[15] Vinodhini, S., Rajalakshmi, V. and Govindarajalu, B., “Building Personalised Recommendation System with Big Data and Hadoop MapReduce”, International Journal of Research in Engineering and Technology, April 2014.
[16] Zhao, Z. D. and Shang, M. S., “User-based Collaborative Filtering Recommendation Algorithms on Hadoop”, Third International Conference on Knowledge Discovery and Data Mining, January 2010.
[17] Zikopoulos, P. and Eaton, C., “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGraw-Hill Osborne Media, 2011.