Avrupa Bilim ve Teknoloji Dergisi
Cilt. 4, No. 7, S. 51-57, Haziran 2016
© Telif hakkı EJOSAT'a aittir
Derleme Makalesi



www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Vol. 4, No. 7, pp. 51-57, June 2016
Copyright © EJOSAT
Review Article

Distributed Recommender Systems with Sentiment Analysis

Yeliz Yengi[1], Sevinç İlhan Omurca[1]

[1]Department of Engineering, Faculty of Engineering-Computer, Kocaeli University, Kocaeli, Turkey, yelizyengi@gmail.com, silhan@kocaeli.edu.tr

Abstract

The aim of this study is to realize a recommender system (RS) for big data application by using sentiment analysis instead of user ratings. By becoming widespread of e-commerce systems through internet, too much user data has became available. So traditional storage systems remained incapable and the stored data is divided. Nowadays we collect lots of reviews from users and feedback on e-commerce web sites therefore the importance of increasing big data analysis technology, which increases the need of big calculation. In this study, we report the performance improvement by adding the natural language processing steps to the classical recommender system.

Keywords: recommender systems, big data, distributed systems, sentiment analysis

Büyük Veride Tavsiye Sistemlerini Duygu Analizi ile Desteklemek

Bu çalışmanın amacı kullanıcı puanlama temelli tavsiye sistemlerinin, kullanıcı puanları yerine duygu analizinden elde edilen değerler ile büyük veri üzerinden gerçeklenmesidir. Internet üzerinden e-ticaret sistemlerinin yaygınlaşması ile çok fazla kullanıcı verisi oluşması, alışılmış depolama sistemlerinin artık yeterli gelmemesi ve verinin bölünmesi durumunu oluşturmuştur. Ancak dağıtık dosya sistemleri teknolojileri ile veri bütünlüğünü sağlamak mümkündür. Bu veri üzerinden makine öğrenmesi algoritmalarının çalıştırılması ve sonuçların değerlendirilmesine büyük ihtiyaç duyulmaktadır. Bu çalışmada tavsiye sistemlerinin dağıtık veri üzerinden değerlendirilmesinde, doğal dil işleme adımlarının sisteme eklenmesi ile sağlanan iyileştirme raporlanmıştır.

Anahtar Kelimeler: tavsiye sistemleri, büyük veri, dağıtık dosya sistemleri, duygu analizi

1. Introduction

Big Data analysis is one of the most important issues in data mining research nowadays. Big Data becomes greatly important when we need to process high calculations or the amount of data is greater than the capacity of the system. There are the three features of big data: volume, velocity and variety. Volume indicates the amount of data that needs to be processed. Could it be moved to zeta bytes? Velocity refers to the speed at which data can be processed with minimal error rate. Variety aim to all types of data for example unstructured, semi structured and structured data. Facebook and Twitter logs, text files, word documents etc. are raw form data that it is called unstructured data. Analyzing semi structured data is like of xml files could not be done easily. Structured data is in the form of rows and columns and can be easily retrieved using SQL, the data which has a well-defined schema.

Large user feedback data is being produced every day in various areas such as movies, food, music, electronic and so on. The amount of user data is increasing dramatically due to growth of e-commerce websites. At this point it raises the importance of topics like how could we stored and how to understand all this data. Big data is one of the best solutions of storing that data and has motivated the interest of recommendation systems to understand that data. Reviews have become a very effective way for people to share expectations of product or services, express their opinions, support their products or even educate each other's by publishing textual data. The meaningful and useful information obtained from user reviews by sentiment analysis could be used to improve recommender systems. Researchers have argued that recommendation systems could be very beneficial. By explaining why a product is recommended, help users make better and faster decisions, and help users to buy or enjoy.

The most common way of recommendation is collaborative filtering (CF) methods generate user specific recommendation of an item based on patterns of ratings or usage without need for outside information about both item and user [1].

Incorporating big contextual information of user-product reviews also improves recommendation accuracy by using CF. Recommendation engines are computational intensive, hence and become large data size therefore ideal for Hadoop Distributed File System (HDFS) Platform [2]. Hadoop is a large scale distributed batch processing infrastructure. It allows for the distributed calculation of large data sets called big data through clusters of computers using Map Reduce and HDFS file system. Hadoop can be used with data mining applications also recommendation algorithms by Mahout Apache project. The whole dataset is then transferred to the Hadoop File System and it uses a recommendation algorithm over frameworks.

In this paper, we report our structure of recommendation systems based on sentiment analysis of user review data stored on a distributed file system. We suggest architecture to divide the whole system function into three tasks, sentiment analysis task, storage task and recommendation algorithms task. To conquer the massive data the Hadoop distributed file system is used to store and manage the raw text data and rating of product. As a study we also discussed the implementation of sentiment analysis framework for get ratings information.

The paper is organized as following: Section 2 provides an overview on existing big data recommendation systems, sentiment analysis frameworks. Section 3 proposed structure of big data and other frameworks. Section 4 is an experiment of setup and evaluation. Section 5 reports the experiments comparison on different dataset to validate effectiveness of the proposed methods. Section 6 includes conclusion and discusses future works.

2. Related work

A lot of recommendation systems have been developed in both academia and industry, shown in recent studies mostly over big data technology. In [3] the authors propose a personalized recommendation; users' preferences change and improve over time. Developed models for such notions of user evolution, in order to assess which best captures the dynamics present in product rating data. Their model to capture similarities between users approximately 15 million reviews rating get from various sources including beers, wines, gourmet foods, and movies. In [4] proposed a keyword-aware service of recommendation method, called KASR. The method provides keywords are used to demonstrate users' preferences for user based Collaborative Filtering algorithm are accepted to generate convenient recommendations. To develop the scalability and efficiency of KASR in "Big Data" environment which is implemented it on a Map Reduce framework in Hadoop platform.

Results demonstrate that KASR method substantially improves the accuracy and scalability of service recommender systems than existing approaches. In [5] propose a Bayesian-inference based recommendation system that personalized recommender systems to improve accurate by using social network area. Bayesian network inferences the measure of conditional probability distribution that using rating similarity between friends in the way it calculates the most probable recommendation. In [6] available an active web service recommendation based on active users usage history. The prototype was able to manage large scale experiments over real world data in distributed structures. Results show that system has a better performance than other based on collaborative filtering web services recommendation approaches.

In this work [7] proposed a probabilistic personalized travel recommendation model were travel photo logs as well as people attributes. People feature and pictures are used in order to effective data mining with demographics for travel landmarks based on Bayesian learning model. In this way greatly useful personalized travel recommendation systems. In [8] proposed a recommender system for video systems that transferred on the web or broadcast, large-scale events in the context, which has been tested with the Olympic Games. The recommender approached audiovisual consumption and was not related on the number of users, working only on the client side. Specific video fragment cannot be recommended using this work. In [9] proposed a framework for cloud services for personalized quality of service ranking prediction by using other user's historical usage experiences for defining ranking approach. The experimental results shows that approach better performance than existing rating based approaches.

In our study we investigate the rating based on the approach that the users can obtain sentiment analysis ranking prediction over data stored on distributed file system. We suggest architecture to divide the whole system function into three tasks, sentiment analysis task, storage task and recommendation algorithms task. To conquer the massive data the Hadoop distributed file system are used to store and manage the raw text data and rating of product. As a study we also discussed the implementation of sentiment analysis framework for get ratings information.

3. Overview about Models and Tools

The definition and modelling of an architecture dedicated to the activities of analysis of big data using mahout ML library for created data sets by using sentiment analysis:

3.1 Hadoop distributed file system (HDFS)

To store large amounts of data on a file system designed that across multi nodes of commodity hardware. Architecture has master slave relation of data nodes which every node store part of data and report back to master node [10].

3.2 MapReduce

The Map Reduce approach in machine learning performs batch learning to build learning model read training data set in its entirety. Inconvenience of this batch model is speed and computational resources. In normally read data sets from HDFS to the mapper as a set of key value match with batch oriented workflow. The result will be key and their associated values list that is written to the disk. Example of classification task, key value match may be a filename or list of instances, the mapper would be associated class with a list of each instance [10].

3.3 Mahout

Mahout is well-known tools for large scale ML. In May 2015, Mahout 0.10.1 was used in project for recommender systems [11].

3.4 Sentiment analysis

Sentiment analysis for get the raking have use annotations for tokenization and sentence separation, although several of the baselines use parsing and tagging as well were processed with state-of-the-art NLP tool that Stanford CoreNLP tools (version 3.4.1) [12].

Lexicon features

Lexicon of words marked with their prior polarity: positive, negative, or neutral. We create three label based on the presence of any words from the lexicon [13].

All technology and tools are based on our architecture Figure 1 shown more detailed.

Sentiment Analysis of Review

Stanford CoreNLP supplies a set of natural language analysis tools. It is able to give the base forms of words, their parts of

speech, whether they are names of place, product, etc., normalizes special information in text like date, time, and numeric quantities, and give forming the structure of sentences in terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, show opinion of text, etc.

There are few toolkits for naturel language analysis,

Stanford CoreNLP is awareness proven, and a main purpose is trying to analyze the attributes to get real opinion of sentences [14].

Toolkit return results like that very bad, bad, naturel, good, very good (-2,-1,0,1,2) (Table 1).

Table 1. CoreNLP sentiment analysis sample

Phrase	Sentiment Value
The film has an excellent rhythm that gives you but a few minutes to recover from all the action before plunging you right into another battle.	Very Good(2)
Product did not come in a protective packaging, I hope it's not broken internally	Bad(-1)
This camera came up with a great resolution and low price.	Very Good(2)

Lexicon Based Sentiment Analysis

Lexicon-based approach to extracting sentiment from words uses dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation (Table 2). To capture main subject matter the opinion in text's need to process text for assigning a positive or negative label [15].

Table 2. Example of words in the dictionaries

Word	Sentiment Orientation Value
disaster	–5
disgust	-4
loathing	-3
execration	-3
generate	-2
lag	-1
specification	1
inspire	2
afflatus	2
ingratiate	3
pleasure	4

Rating Interface

Amazon review and Movie Lens datasets person expressing a preference (a 0–5 star rating), reviews is three point labeled < 3 result in preference score of 0, rating = 3 score of 1, rating 3 < score of 2. Sentiment analysis word level or phrase level (Stanford NLP) are return negative or positive value then we addition that value at last our preference, if < 0 result in sentiment score of 0, if = 0 score of 1, 0 < score of 2.



Figure 1. Architecture of distributed recommender system

Recommender System

Recommender systems predict the 'rating' or 'preference' that a user would give to an item kind of a subclass of information filtering system. The recommender systems concept was introduced to deal with retrieve the most relevant information of users and the challenges of information overload, to scan through the large information sets [16].

Collabrative Filtering:

Methods are based on collecting and analyzing a large amount of information on user's behaviors, activities or preferences and predicting what users will like based on their similarity to other users. The input to collaborative filtering consists of user, item and rating to build recommendations using any of the following ways:

a) User Based Recommendations: User based recommendations are computed based on users with identical characteristics.

Item Based Recommendations: Item based recommendations are computed based on similar items.

1. Matrix Factorization with ALS:

Mahout offers Alternate Least Square algorithm for matrix factorization. Matrix factorization is a dimensionality reduction technique that factorizes a matrix into a product of matrices. User-item matrix, V can be factorized into one m X k. User-feature matrix, U and an n X k. Item-feature matrix, M, m X n. Mahout ALS recommender engine can be used for large data sets that spread over many machines in HDFS environment.

We have used open source libraries of Apache Mahout for our experiment. Mahout is an open source project to provide free implementations of scalable and distributed machine learning algorithms for recommender systems. It provides both non-distributed and distributed (Map-Reduce) algorithms for recommendation. In this work, the distributed algorithms of Mahout have been used.

Mahout has several algorithms for similarity measures, neighborhood computation and evaluation. We are going to use the distributed algorithms for user-based and matrix factorization with ALS recommender in Mahout [17].

4. Experiments

4.1 Experimental Setup

System Configuration: We carried out the assessment using following configuration:

Table 3. System and Other Configuration

Processor	Intel(R) Xeon(R) CPU E3-1231 v3 3.40GHz
Ram	4 GB Per Data Node and Name Node
Operating System	Linux Ubuntu 14.04 LTS Per Data Node and Name Node
JVM	JRE1.7.0_79
Mahout	Apache Mahout 0.11.1
Hadoop	Apache Hadoop 2.6.0

4.2 Datasets

The datasets we consider are summarized in Table 4. Each of our datasets was obtained from Amazon product data. Also for comparison movie data sets had used Movie Lens 1M and 10M data sets. [18, 19].

4.3 Evalution

As actual preferences of items by a user don't exist for items, which have not been already rated by the user, a simulation technique needs to be used for the evaluation. A small part of real data 10% (with preference values) is set aside as test data. Another part of real data 90% is used as training data. A recommendation system is asked to estimate the preference values for the test data and the results are compared with actual preference values to measure the quality of recommendation. A score can be generated for a recommender from evaluation. Root-mean-square (RMSE) between estimated and actual preferences of the differences for calculation of scores. Lower score is better as that indicates that estimates are closer to actual preference values.

Table 4. Summarized data sets

Reviews	Number of reviews	Number of users	Number of products	Timespan
Movie	7.911.684	889.176	253.059	Aug 1997 - Oct 2012
Fine Food	568.454	256.059	74.258	Oct 1999 - Oct 2012
Electronics	9.000.000	4.201.732	476.005	Jun 1995 - Mar 2013
Digital Music	836.016	478.244	266.457	Jun 1995 - Mar 2013
Movie Lens	1.000.209	6.040	3.900	In 2000
Movie Lens	10.000.054	71.567	10.681	In Jan 2009

5. Qualitative Analysis

In this section, we present the results of performance and quality assessment of sentiment recommendation on large data sets approximately 30 million reviews ratings and sentiment rating data worked on. We have performed this assessment using user-based recommendation of Apache Mahout. Our observation about the performance and quality of different sentiment rating datasets is lower RMSE than traditional star rating with algorithms in recommending items is evaluation scores (RMSE) of different datasets are shown below Figure 2.

As we can see from the table or graph which created datasets using sentiment analysis were decreases of error rate. Sentiment analysis was is change our results, Stanford Core NLP tools give us more detailed analyze of sentences but lexicon analyze is much more effective on all categories of datasets.

Also matrix factorization improves recommender systems over sentiment rating datasets. In Table 5 is clearly seen the improvements over recommendation, besides to growth of data is effect of reducing error rate in clearly seen movie datasets while growing data RMSE value has decrease. In [20] study was using information retrieval over same user reviews, our model gives better results for recommendation. Also in [21] the recommendation is made by using restaurant review through sentiment analysis, our study with distributed architecture get more efficient result.

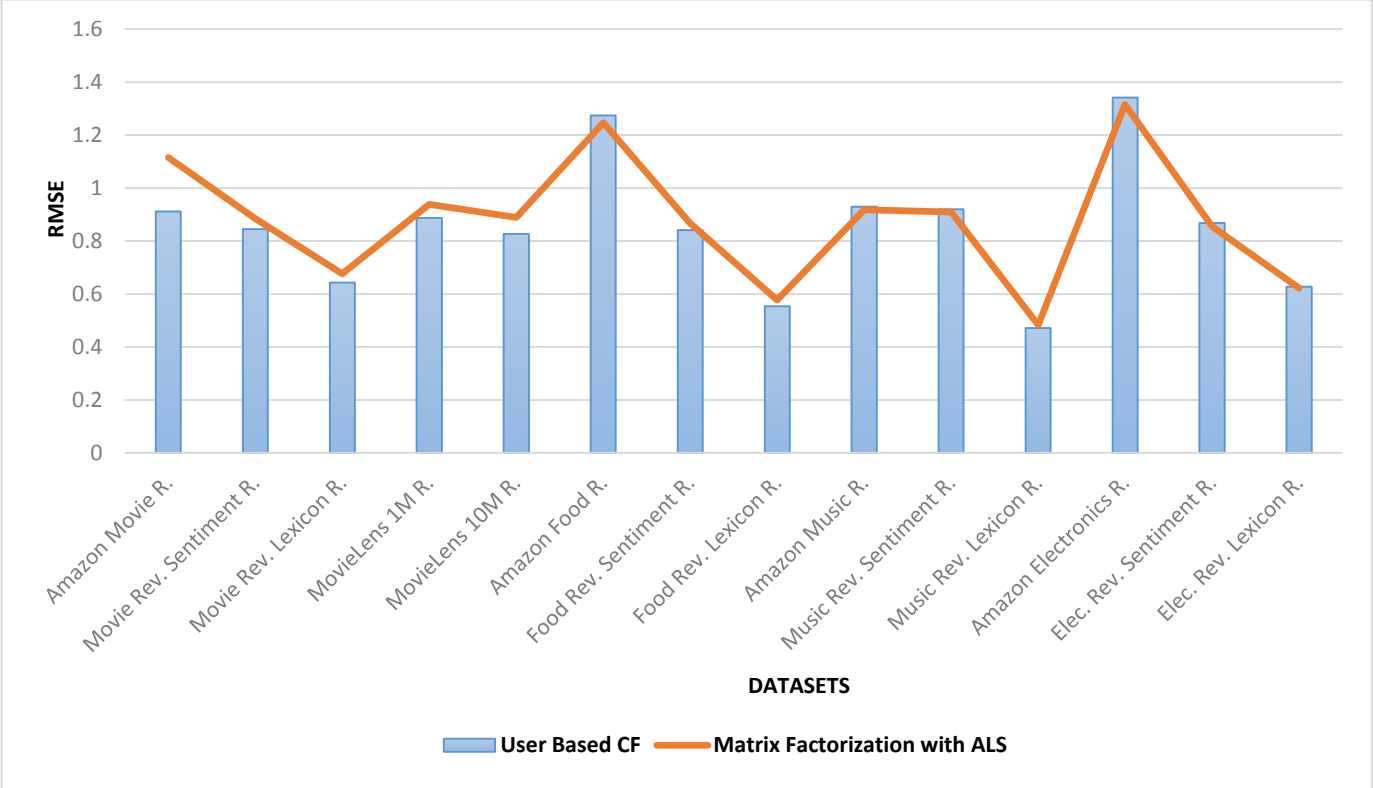


Figure 2. Comparison of recommender algorithm and created sentiment data sets. Table has Rev. Short notation are "Review" and R. short notation are "Rating"

Table 5. Datasets RMSE values have shown in the table.

Categories	Datasets	User Based CF RMSE	Matrix Factorization with ALS RMSE
Movie	Amazon Movie Rating	0,9107	1,11502
	Movie Review Sentiment Rating	0,8445	0,8841
	Movie Review Lexicon Rating	**0,6434**	**0,6757**
	Movie Lens 1M Rating	0,8869	0,9382
	Movie Lens 10M Rating	0,8271	0,8894
Fine Food	Amazon Food Rating	1,2736	1,2468
	Food Review Sentiment Rating	0,8413	0,867
	Food Review Lexicon Rating	**0,5534**	**0,5775**
Music	Amazon Music Rating	0,9299	0,9182
	Music Review Sentiment Rating	0,9201	0,9092
	Music Review Lexicon Rating	**0,471**	**0,4833**
Electronics	Amazon Electronics Rating	1,3418	1,3149
	Elec. Review Sentiment Rating	0,8673	0,8545
	Elec. Rev. Lexicon R.	**0,6275**	**0,6215**

6. Conclusion

Recommendation systems play a very important role in e-commerce, social networking, etc. Similarity measurement is a key aspect in any recommendation system and performance of a recommendation system is highly dependent on the performance of the large data and data quality. In this paper we have assessed the performance and quality of sentiment rating datasets in collaborative filtering and matrix factorization process on Hadoop cluster using Apache Mahout, a library for machine learning algorithms. By using sentiment rating over big data approach, accuracy of recommendation has improved. This approach has scaled well with the Hadoop platform. This paper also shows how we recommend without user rating stars, sentiment analysis help to detect user preferences to use under recommender systems.

Feature work over recommender system but using GPS and sensor data that prefer to working on it. Analysis to user behavior, risks, insurances costs, advice systems, traffic analysis and so on could be show over that king of data.

References

Bell, R. M., & Koren, Y. (2007, October). Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on (pp. 43-52). IEEE, ISO 690.

Zhou, B., Jia, Y., Liu, C., & Zhang, X. (2010, October). A distributed text mining system for online web textual data analysis. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2010 International Conference on (pp. 1-4). IEEE.

McAuley, J. J., & Leskovec, J. (2013, May). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In Proceedings of the 22nd international conference on World Wide Web (pp. 897-908). International World Wide Web Conferences Steering Committee.

S Meng, S., Dou, W., Zhang, X., & Chen, J. (2014). Kasr: A keyword-aware service recommendation method on mapreduce for big data applications. Parallel and Distributed Systems, IEEE Transactions on, 25(12), 3221-3231.

Yang, X., Guo, Y., & Liu, Y. (2013). Bayesian-inference-based recommendation in online social networks. Parallel and Distributed Systems, IEEE Transactions on, 24(4), 642-651.

Kang, G., Liu, J., Tang, M., Liu, X., Cao, B., & Xu, Y. (2012, June). AWSR: Active web service recommendation based on usage history. In Web Services (ICWS), 2012 IEEE 19th International Conference on (pp. 186-193). IEEE.

Chen, Y. Y., Cheng, A. J., & Hsu, W. H. (2013). Travel recommendation by mining people attributes and travel group types from community-contributed photos. Multimedia, IEEE Transactions on, 15(6), 1283-1295.

Sanchez, F., Alduán, M., Alvarez, F., Menéndez, J. M., & Baez, O. (2012). Recommender system for sport videos based on user audiovisual consumption. Multimedia, IEEE Transactions on, 14(6), 1546-1557.

Zheng, Z., Wu, X., Zhang, Y., Lyu, M. R., & Wang, J. (2013). QoS ranking prediction for cloud services. Parallel and Distributed Systems, IEEE Transactions on, 24(6), 1213-1222.

White, T. (2012). Hadoop: The definitive guide. " O'Reilly Media, Inc.".

Anil, R., Dunning, T., & Friedman, E. (2011). Mahout in action (pp. 145-183). Shelter Island: Manning.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations) (pp. 55-60).

Wilson, T.; Wiebe, J.; and Hoffmann, P., (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics 35(3):399–433.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations) (pp. 55-60).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Prem Melville and Vikas Sindhwani, “Recommender Systems”, IBM T.J. Watson Research Center.

Owen S., Anil R., Dunning T. and Friedman E.: “Mahout In Action”, (2012). Manning Publications Co. ISBN 978-1-9351-8268-9

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 43-52). ACM.

McAuley, J., Pandey, R., & Leskovec, J. (2015, August). Inferring networks of substitutable and complementary products. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(pp. 785-794). ACM.

McAuley, J. J., & Leskovec, J. (2013, May). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In Proceedings of the 22nd international conference on World Wide Web (pp. 897-908). International World Wide Web Conferences Steering Committee.

Xing Margaret, F. U., & Xiaocheng, L. I. (2015). From Movie Reviews to Restaurants Recommendation.