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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

A Study on Product Recommendation System vide Machine Learning using Big Data

Priyanka [1], Shalini [2]

P.G. Student, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India [2]

ABSTRACT: Today buyers are presented to an expanding assortment of items and data never observed. This prompts an expanding assorted variety of shoppers' interest, transforming into a test for a dealer to give the correct items as needs be to client inclinations. Recommender frameworks are an instrument to adapt to this test, through item proposal it is conceivable to satisfy clients' needs and desires, helping keeping up faithful clients while drawing in new clients. Anyway the enormous size of value-based databases like Big Data is not a normal volume for retail business and diminishes the effectiveness and nature of suggestions. In this paper or scheme we will elaborate the techniques and suggested framework that joins content-based, synergistic sifting and machine learning strategies utilizing and Naïve Bayes and Big-Data wherein proposes to outperform these troubles. The suggestion calculation begins to get comparable or different gatherings of clients utilizing tweets i.e. miniaturized scale blogging webpage or crowd sourcing platform. Thereafter, next an affiliation rule mining approach dependent on comparable shopping bins of clients of a similar group, in a particular timeframe is executed so as to give increasingly decisive and customized client item proposals dependent on logical inquiry and sentiment mining or conduct investigation utilizing synopsis procedures, content filtering and celebrative filtering for client product recommendation . Therefore client can build the estimation of the deals without losing suggestion precision and adequacy based on recommendations available on twitter.

KEYWORDS: Product Recommendation, Twitter, Big Data, Machine Learning, Naïve Bayes, Linear Regression,

I. INTRODUCTION

Twitter is a social data crowd sourced organized platform or eco-system where short messages or tweets are shared among countless through a straight forward information rendering system exemplary views, reviews, sentiments, behaviours, opinions and recommendations. With a populace of more than 100M clients creating more than 300M tweets every day, Twitter clients can be effectively overpowered by the huge measure of data accessible and the tremendous number of individuals they can connect with. To conquer the above data over-burden issue, recommendations frameworks can be acquainted with assistance clients make the fitting determination. Various researchers have started to think about proposal issues in Twitter yet their works for the most part address singular suggestion errands. There is so far no complete overview for the domain of proposal in Twitter to arrange the current fills in just as to recognize territories that should be additionally considered. The paper hence means to fill this hole by presenting a scientific classification of suggestion assignments in Twitter, and to utilize the scientific categorization to depict the applicable works as of foundation to develop more effective system. This scheme depicts various strategies utilized in the above. At last, it proposes a couple of research bearings for suggestion undertakings in Twitter.

The taxonomy via unique functions or facilitation to users are provided by the eco-system to participate and explore the following taxonomy on the context for recommendations, opinions and reviews :-

1. **Tweet** : This refers to posting a message of up to 280 characters, known as tweets. The content of tweets may vary from users' daily activities to news, views, opinions, advices, recommendations Some messages may also include URLs to web pages or hashtags to relate tweets of similar topics together. Each hashtag is a keyword

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

prefixed by a # symbol. For example, #rafael #namo and #elec2019 have been used to group tweets related to India's context and origin.

2. **Re-Tweet** : This alludes to sending a tweet from another client to the followers. Such re-sharing of tweets is a common instrument in Twitter to diffuse data.
3. **Follow** - This refers to linking to another user and receiving the linked user's tweets after that. The user creating such a link is called the follower and the linked user is known as the follower.
4. **Mention** - One may mention one or more users in a tweet by including in the tweet the mentioned user name(s) prefixed by the @ sign. The mentioned user(s) will subsequently receive the tweet. This is a means for users to gain attention from the other users so as to start new conversations.

Big Data: As the data or information at Twitter repositories are in high volume therefore to produce the recommendation system on the fly needs the best mechanism to store the tweets, re-tweets and related context in comprehensive manner therefore, big data will be the best exemplary model for such mammoth data based on velocity, volume and variety however, below the same is depicted for consideration to inculcate in future schemes. Huge information alludes to information volumes in the scope of exabytes (1018) and past. Such volumes surpass the limit of current on-line stockpiling frameworks and preparing frameworks. Information, data, and learning are being made and gathered at a rate that is quickly moving toward the exabyte/year extend. Be that as it may, its creation and collection are quickening and will approach the zettabyte/year go inside a couple of years. Volume is just a single part of enormous information; different qualities are assortment, speed, esteem, and multifaceted nature. Capacity and information transport are innovation issues, which appear to be resolvable in the close term, however speak to long haul difficulties that require look into and new ideal models. We dissect the issues and difficulties as we start a collective research program into approaches for huge information examination and structure.

The idea of enormous information has been endemic inside software engineering since the most punctual long periods of figuring. "Enormous Data" initially implied the volume of information that couldn't be prepared (productively) by conventional database techniques and instruments. Each time another capacity medium was developed, the measure of information available detonated in light of the fact that it could be effectively gotten to. The first definition concentrated on organized information, yet most analysts and experts have come to understand that the vast majority of the world's data dwells in monstrous, unstructured data, to a great extent as content and symbolism.

The blast of information has not been joined by a comparing new capacity medium. We characterize "Huge Data" as the measure of information just past innovation's capacity to store, oversee and process effectively. These impersonations are just found by a vigorous examination of the information itself, express preparing needs, and the abilities of the instruments (equipment, programming, and techniques) used to investigate it. Similarly as with any new issue, the finish of how to continue may prompt a suggestion that new devices should be manufactured to play out the new errands. As meager as 5 years prior, we were just reasoning of tens to several gigabytes of capacity for our PCs. Today, we are thinking in tens to several terabytes. Along these lines, huge information is a moving target. Put another way, it is that measure of information that is simply past our quick handle, e.g., we need to strive to store it, get to it, oversee it, and process it. The present development rate in the measure of information gathered is amazing. A noteworthy test for IT analysts and professionals is that this development rate is quick surpassing our capacity to both plan suitable frameworks to deal with the information viably and dissect it to remove important significance for basic leadership. In this paper we distinguish basic issues related with information stockpiling, the executives, and handling. To the best of our insight, the exploration writing has not successfully tended to these issues.

Five suggested fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues. Each of these represents a large set of technical research problems in its own right:-

Data Volume: Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

Data Velocity: Data velocity measures the speed of data creation, streaming, and aggregation. eCommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue; it is also an ingest issue (extract-transform-load.

Data Variety: Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

Data Value: Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data.

Complexity: Complexity measures the degree of interconnectedness (possibly very large) and interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all.

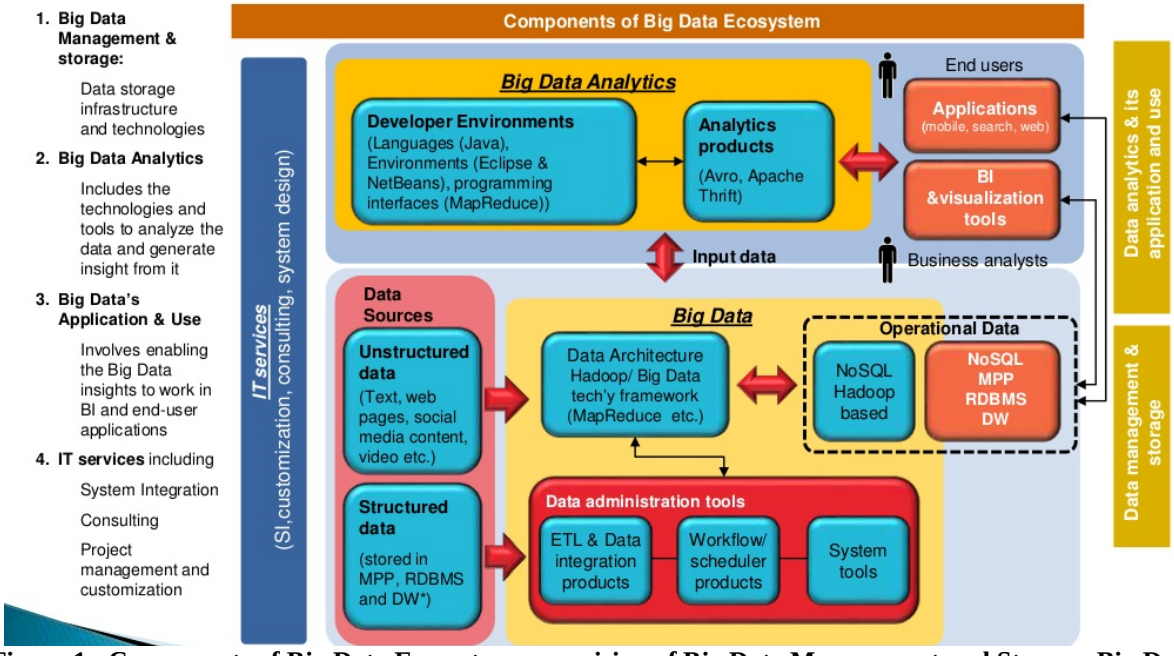

Figure 1: Components of Big Data Ecosystem comprising of Big Data Management and Storage, Big Data Analytics, Big Data Applications along-with its Uses and Last Information Technology Services and Solutions.

Machine Learning : Machine Learning is concerned approach towards algorithm oriented programs that empower the conduct of a computers to be gained from precedents or experience as opposed to directed through guidelines composed by hand. It has useful incentive in numerous application regions of software engineering, for example, on-line networks and advanced libraries. This class is intended to show the viable side of machine learning for applications, for example, mining newsgroup information or building versatile UIs. The accentuation will be on learning the way toward applying machine adapting adequately to an assortment of issues instead of underlining a comprehension of the hypothesis behind what makes machine realizing work. This course does not accept any earlier presentation to machine learning hypothesis or practices some of the approaches depicted to explain some of the methods which can be used for product recommendation system under the machine learning techniques.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

Naive Bayes Naive Bayes is a simple probabilistic classifier based on Bayes' rule. The naive Bayes algorithm builds a probabilistic model by learning the conditional probabilities of each input attribute given a possible value taken by the output attribute. This model is then used to predict an output value when we are given a set of inputs. This is done by applying Bayes' rule on the conditional probability of seeing a possible output value when the attribute values in the given instance are seen together. Before describing the algorithm we first define the Bayes' rule. Bayes' rule states that $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$, where *P(A|B)* is defined as the probability of observing *A* given that *B* occurs. *P(A|B)* is called posterior probability, and *P(B|A)*, *P(A)* and *P(B)* are called prior probabilities. Bayes' theorem gives a relationship between the posterior probability and the prior probability. It allows one to find the probability of observing *A* given *B* when the individual probabilities of *A* and *B* are known, and the probability of observing *B* given *A* is also known. The naive Bayes algorithm uses a set of training examples to classify a new instance given to it using the Bayesian approach. For an instance, the Bayes rule is applied to find the probability of observing each output class given the input attributes and the class that has the highest probability is assigned to the instance. The probability values used are obtained from the counts of attribute values seen in the training set. In our weather example, for a given instance with two input attributes $temp\_A_t$ and $temp\_B_t$, with values *a* and *b* respectively, the value v_{MAP} assigned by the naive Bayes algorithm to the the output attribute $temp\_C_t$ is the one that has the highest probability across all possible values taken by output attribute; this is known as the maximum-a-posteriori (MAP) rule. The probability of the output attribute taking *a* value v_j when the given input attribute values are seen together is given by $P(v_j \mid a,b)$. This probability value as such is difficult to calculate. By applying Bayes theorem on this equation we get $P(v_j \mid a,b) = \frac{P(a,b \mid v_j)P(v_j)}{P(a,b)} = P(a,b \mid v_j)P(v_j)$, where $P(v_j)$ is the probability of observing v_j as the output value, $P(a,b|v_j)$ is the probability of observing input attribute values *a, b* together when output value is v_j. But if the number of input attributes *(a, b, c, d, ...)* is large then we likely will not have enough data to estimate the probability $P(a, b, c, d, ... \mid v_j)$. The naive Bayes algorithm solves this problem by using the assumption of conditional independence for the all the input attributes given the value for the output. This means it assumes that the values taken by an attribute are not dependent on the values of other attributes in the instance for any given output. By applying the conditional independence assumption, the probability of observing an output value for the inputs can be obtained by multiplying the probabilities of individual inputs given the output value. The probability value $P(a, b \mid v_j)$ can then be simplified as $P(a,b \mid v_j) = P(a \mid v_j)P(b \mid v_j)$, where $P(a \mid v_j)$ is the probability of observing the value *a* for the attribute $temp\_A_t$ when output value is v_j. Thus the probability of an output value v_j to be assigned for the given input attributes is $P(v_j \mid a,b) = P(v_j)P(a \mid v_j)P(b \mid v_j)$. Learning in the Naive Bayes algorithm involves finding the probabilities of $P(v_j)$ and $P(a_i|v_j)$ for all possible values taken by the input and output attributes based on the training set provided. $P(v_j)$ is obtained from the ratio of the number of time the value v_j is seen for the output attribute to the total number of instances in the training set. For an attribute at position *i* with value a_i, the probability $P(a_i|v_j)$ is obtained from the number of times a_i is seen in the training set when the output value is v_j. The naive Bayes algorithm requires all attributes in the instance to be discrete. Continuous valued attributes have to be discretized before they can be used. Missing values for an attribute are not allowed, as they can lead to difficulties while calculating the probability values for that attribute. A common approach to deal with missing values is to replace them by a default value for that attribute.

Linear Regression : The Linear Regression algorithm performs standard least squares regression to identify linear relations in the training data. This algorithm gives the best results when there is some linear dependency among the data. It requires the input attributes and target class to be numeric and it does not allow missing attributes values. The algorithm calculates a regression equation to predict the output (*x*) for a set of input attributes $a_1,a_2,.....,a_k$. The equation to calculate the output is expressed in the form of a linear combination of input attributes with each attribute associated with its respective weight $w_0,w_1,....,w_k$, where w_1 is the weight of a_1 and a_0 is always taken as the constant 1. An equation takes the form $x = w_0 + w_1a_1 + ...+ w_ka_k$ .For our weather example the equation learned would take

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

the form $temp\_C_t = w_0 + w_{At-2}\ temp\_A_{t-2} + w_{At-1}\ temp\_A_{t-1} + w_{At}\ temp\_A_t + w_{Bt-2}\ temp\_B_{t-2} + w_{Bt-1}\ temp\_B_{t-1} + w_{Bt}\ temp\_B_t + w_{Ct-2}\ temp\_C_{t-2} + w_{Ct-1}\ temp\_C_{t-1}$,where $temp\_C_t$ is value assigned to the output attribute, and each term on the right hand side is the product of the values of the input attributes and the weight associated with each input. The accuracy of predicting the output by this algorithm can be measured as the absolute difference between the actual output observed and the predicted output as obtained from the regression equation, which is also the error. The weights must be chosen in such as way that they minimize the error. To get better accuracy higher weights must be assigned to those attributes that influence the result the most. A set of training instances is used to update the weights. At the start, the weights can be assigned random values or all set to a constant (such as 0). For the first instance in the training data the predicted output is obtained as $w_0 + w_1 a_1^{(1)} + \ldots\ldots\ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)}$ where the superscript for attributes gives the instance position in the training data. After the predicted outputs for all instances are obtained, the weights are reassigned so as to minimize the sum of squared differences between the actual and predicted outcome. Thus the aim of the weight update process is to minimize

$\sum_{i=1}^{n}\left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)}\right)$,which is the sum of the squared differences between the observed output for the i^{th} training instance ($x^{(i)}$) and the predicted outcome for that training instance obtained from the linear regression equation.

LeastMedSquare: The LeastMedSquare or Least Median Squares of Regression algorithm is a linear regression method that minimizes the median of the squares of the differences from the regression line. The algorithm requires input and output attributes to be continuous, and it does not allow missing attribute values. Standard linear regression is applied to the input attributes to get the predict the output. The predicted output x is obtained as $w_0 + w_1 a_1^{(1)} + \ldots\ldots\ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)}$,where the a_i are input attributes and w_i are the weights associated with them. In the LeastMedSquare algorithm, using the training data, the weights are updated in such a way that they minimize the median of the squares of the difference between the actual output and the predicted outcome using the regression equation. Weights can be initially set to random values or assigned a scalar value. The aim of the weight update process is to determine new weights to minimize $\underset{i}{median}\left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)}\right)$, where i ranges from 1 to the number of instances in the training data that is being used, $x^{(i)}$ is the actual output for the training instance i, and the predicted outcome for that training instance is obtained from the regression equation.

II. RELATED WORK

Previous research on Recommender Systems (RS), especially the continuously popular approach of Collaborative Filtering (CF), has been mostly focusing on the information resource of explicit user numerical ratings or implicit (still numerical) feedbacks. However, the ever-growing availability of textual user reviews has become an important information resource, where a wealth of explicit product attributes/features and user attitudes/sentiments are expressed therein. This information rich resource of textual reviews have clearly exhibited brand-new approaches to solving many of the important problems that have been perplexing the research community for years, such as the paradox of cold-start, the explanation of recommendation, and the automatic generation of user or item profiles. However, it is only recently that the fundamental importance of textual reviews has gained wide recognition, perhaps mainly because of the difficulty in formatting, structuring and analyzing the free-texts. In this research, we stress the importance of incorporating textual reviews for recommendation through phrase-level sentiment analysis, and further investigate the role that the texts play in various important recommendation tasks.[1].

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

Recommender system applied various techniques and prediction algorithm to predict user interest on information, items and services from the tremendous amount of available data on the internet. Recommender systems are now becoming increasingly important to individual users, businesses and specially e-commerce for providing personalized recommendations. Recommender systems have been evaluated and improved in many, often incomparable, ways. In this paper, we review the evaluation and improvement techniques for improving overall performance of recommendation systems and proposing a semantic analysis based approach for clustering based collaborative filtering to improve the coverage of recommendation. The basic algorithm or predictive model we use are – simple linear regression, k-nearest neighbours(kNN), naives bayes, support vector machine. We also review the pearson correlation coefficient algorithm and an associative analysis-based heuristic. The algorithms themselves were implemented from abstract class recommender, which was extended from weka distribution classifier class. The abstract class adds prediction method to the classifier.[2]

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains.[3]

The recommendation systems are widely used to support the users to handle the ever increasing data over the internet efficiently. Recommendation systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. To date a number of recommendation system algorithms have been proposed such as collaborative filtering recommendations, content based recommendations and hybrid approach algorithms. The focus is generally on content based recommendation systems methods which are mainly based on naïve Bayesian machine learning algorithm. In this paper, a survey of techniques is presented which suggest naïve Bayesian algorithm for similarity in recommendation systems.[4]

III. CONCLUSION

A key issue that rises up out of this investigation, requesting future research, alludes to the viability of the proposals produced by such a framework and how this can be enhanced utilizing the client responses to the suggestions as input or feeling. Along these lines the framework won't just utilize acquiring information as info, yet in addition the client responses to the proposals, which is the most essential proportion of its viability. Another pending examination issue is the effective execution of suggestions, which much of the time must be made vide twitter. The utilization of a standard on an item set is a high request intricacy undertaking and the presence of an immense measure of principles may make issue their environment, where the speed of reaction to the client is basic. Proposal frameworks have turned out to be not kidding business apparatuses and are re-forming the universe of businesses. Powerful suggestions are a significant support of the clients and a beneficial support of the retailer. We trust that the future plan will be instilled with present learning around there and open up new points of view utilizing Big Data and Machine Learning to create progressively viable and precise proposed framework for future.

REFERENCES

1. Yongfeng Zhang, Min Zhang and Yiqun Liu, Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation,Eighth ACM International Conference on Web Search and Data Mining, pp. 435 – 440, February 2015.
2. Lovedeep Kaur, Naveen Kumar, A Research on user Recommendation System Based upon Semantic Analysis,International Journals of Advanced Research in Computer Science and Software Engineering ISSN: 2277-128X (Volume-7, Issue-11), November 2017
3. Nada Elgendy and Ahmed Elragal, Big Data Analytics: A Literature Review Paper, P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014. Springer International Publishing Switzerland 2014

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 1, January 2019

4. Meghna Khatri, A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems, Volume 2, Issue 5 International Journal of Advanced Research in Computer Science and Software Engineering, May 2012
5. Stanford large network dataset collection. http://snap.stanford.edu/data/index.html.
6. Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. Proceeding of WSDM 2011, pages 635–644, 2011.
7. Jure Leskovec, Lada A.Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. ACM Transactions on the Web (ACMTWEB), 1(1), 2007. Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-toitem collaborative filtering. IEEE Internet Computing, 7(1):76–80, 2003.
8. Tao Zhou, Jie Ren, Matus Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. Physical Review E, page 76, 2007
9. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and Tweet: Experiments on Recommending Content from Information Streams. In: The 28th International Conference on Human Factors in Computing Systems (2010)
10. Choudhary, A., Hendrix, W., Lee, K., Palsetia, D., Liao, W.K.: Social media evolution of the Egyptian revolution. Communications of ACM 55(5), 74–80 (May 2012)
11. De Francisci Morales, G., Gionis, A., Lucchese, C.: From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation. In: The 5th ACM International Conference on Web Search and Data Mining (2012)
12. Garcia, R., Amatriain, X.: Weighted Content Based Methods for Recommending Connections in Online Social Networks. In: The 2nd ACM Workshop on Recommendation Systems and the Social Web. Barcelona, Spain (June 2010)