International Journal of Data Science and Analysis
2015; 1(1): 1-7
Published online July 31, 2015 (http://www.sciencepublishinggroup.com/j/ijdsa)
doi: 10.11648/j.ijdsa.20150101.11

Survey on Sina Weibo Research Based on Big Data Mining

Ru Wang

School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, PR China

Email address:
ruwang@outlook.com

To cite this article:
Ru Wang. Survey on Sina Weibo Research Based on Big Data Mining. *International Journal of Data Science and Analysis*. Vol. 1, No. 1, 2015, pp. 1-7. doi: 10.11648/j.ijdsa.20150101.11

Abstract: In recent years, with the advances in information communication, Sina Weibo has attracted the attention of scholars in China. The big data analytics platform at Sina Weibo has experienced tremendous growth over the past few years in terms of size, complexity, number of users and variety of use cases. Without a clear description of how the underlying data were collected, stored, cleaned, and analyzed, however, Weibo network analysis and modeling become difficult. To analyze the Weibo data, the structure framework of Weibo need firstly be known, and the composition and characteristics of Weibo data must be understood. Then by comparing different application programming interface (API), the more efficient and convenient method of data collection are found. Moreover, according to the characteristics of Weibo data, quarrying the cleaning methods and strategies provide convenient for the further processing of data. Finally, the integration of big data mining and the properties of Weibo find the most effective method based on large Weibo data, and discuss the future research.

Keywords: Sina Weibo, Big Data, Analytics Platform, API, Data Mining

1. Introduction

The Sina Weibo as a new network service and tools is increasingly integrated into the visual presentation of news and scholarly publishing in the form of hash tags, user comments, and dynamic charts displayed on screen during conference presentations, in television programming, alongside political coverage in newspapers, and on web sites. In each case, the mass-scale Sina data is transformed into statistics, tables, charts and graphs without explanation of how the data are collected, stored, cleaned and analyzed. So the readers are unable to assess the given methodology to the social phenomena. As an increasing amount of everyday social interaction is mediated by Weibo or other social software, their servers actively aggregate vast stores of information about user behavior. So the repositories offer new methodological opportunities that combine character istics of the micro and the macro, in combination with the falling costs of mass storage and parallel computing. The new paradigm "big social data" [1] seem to combine the grand scale and generalizability of methods.

Recent years have witnessed an exciting increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze and store. Consider the Internet data, in 1998 the web pages indexed by Google were around one million, but quickly reached one billion in 2000 and have already exceeded one trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create content freely and amplify the huge Web volume. Since researchers pursue big social opportunities, they face a host of daunting challenges theoretical and ethical as well as technical, and the challenges may not be obvious from the outset. Whereas the reliability and validity of established social scientific methods depend on their transparency, big social data are almost universally produced within closed, commercial organizations. In other words, the stewardship of this unprecedented record of public discourse depends on an infrastructure that is both privately owned and operationally opaque.

The big social data is collected among the many sites.

Because of perceived accessibility of Sina Weibo, it is particularly compelling. In comparison to other social sites like Facebook and YouTube, the message of Weibo are small in size, public by default, numerous and topically diverse. Facebook is largely closed-off to the academic community, and YouTube need a high-bandwidth. With little more than a laptop, an Internet connection, and a few lines of scripting code, researchers can aggregate several million messages in a short period of time using widely-available, low-cost tools. It can be foreseen that it can be foreseen that Internet of things (IoT) applications will raise the scale of data to an

unprecedented level. People connected others by Weibo frequently. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help us analysis people's behavior and make predictions. In section 2, we summarize network services architecture of Weibo. We introduce collection and analysis of Weibo data in Section 3 and elaborate big data mining of Weibo in Section 4. Finally, we give some conclusions in Section 5.

2. Weibo Network Service Architecture

2.1. Data Center Principle

Sina Weibo is a Web2.0 network service essentially. The principle of the system is the same with all Web service, which hosts the network traffic data by http protocol. A complete system is composed of Weibo service, data centers and Weibo client. The network architecture of Weibo service system shows in Figure 1.

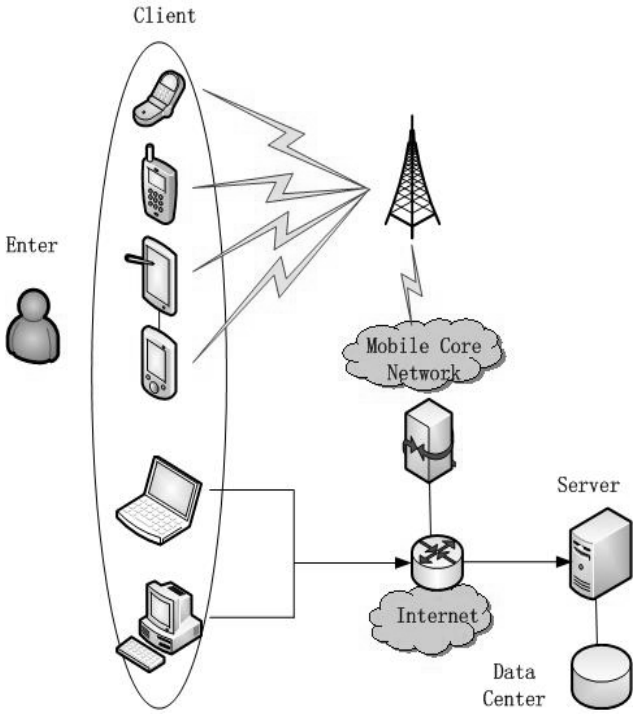

Fig. 1. *The network architecture of Weibo service system.*

In Weibo service system, the function and implementation of server and client are based on Web2.0 services framework, the data center is the focus of Weibo service system. Weibo data center is a huge and complex database system, which stores all personal information of user, following relationship between users and posted messages, etc. In addition, Weibo data center also runs a social network service (SNS) program which calculates the relationship between users and counts Weibo hot topic. Weibo message could widely spread depending on the client relationship maintained by data center.

2.2. Data Characteristics Analysis

In computer networks and communications, the Weibo data characteristics have been concerned in recent years. Many studies are expanded based on the network data measurement and analysis [2]. Weibo data refers to the various types of data stored in the data center, including Weibo user files, customer relationship, messages and hot topics, etc. Many researches are based on the component of Weibo message spreading. The research content is message spreading and member organizations. The purpose is to discover the laws of user, message, hot topic and user relationship in Weibo. In the news propagation, user, message and user relationship directly affect the power of spreading. They interact with each other. The message can spread rapidly among users. The relationship between the components is shown in Figure 2.

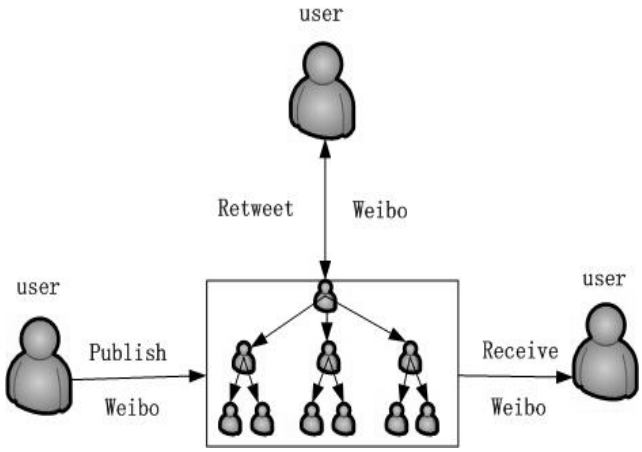

Fig. 2. *The relationship between the components.*

Weibo users are manufacturers, retweeters and receivers.

It is the starting point, the intermediate and end points in information dissemination process. Weibo content directly affect the degree of concern and retweet. Since Weibo play a role in recent mass incidents and emergencies, Weibo monitoring has become an important research topic. Analysis of Weibo data has become crucial. The rules and characteristics of Weibo are found by analyzing message data. The process is divided into two stages: data acquisition and data analysis. In the data acquisition phase, the main task is to get a large number of Weibo data. The data has been fetched by API [3]. Weibo maintains three accessible APIs: the Search API, Rest API, and Streaming API. Each offers a set of different methods for interacting with the data center and constrains the user in different ways. None of the accessible APIs offers interviewers the same degree of access. Recently, Sina, Inc. represents a new class of pay-as-you-go APIs, which has its own features and restrictions. In addition to the many different APIs currently in use, a number of earlier access options are no longer available. So it is difficult to access to the Weibo data. In the data analysis stage, the main task is to extract and analysis the feature of data, and dig out the key features of Weibo. Methods have used

including statistical data analysis, complex network analysis, data classification and mining.

3. Collection and Cleaning of Weibo Data

Data collection requests not only the methodological tool kit required to enable comparison among the many studies drawing on Weibo data, but also a shared set of expectations regarding the identification, aggregation, cleaning, and archiving of messages. In spite of the considerable volume of published work drawing on Weibo data, misinformation abounds regarding the validity of different data collection strategies and their appropriateness for the analytic processes to which they are subjected.

3.1. API Comparing and Selecting

Making sense of the many different Weibo APIs is challenging enough for an individual researcher, and the need for experimental comparison among different APIs has been long noted on various occasions, but the published record remains quite small. The abroad study of APIs in Twitter has made some progress. These studies are worth learning. Using each of the publicly accessible APIs along with two commercial platforms, Mazon and his collaborators attempted to aggregate messages related to the indignados social movement during a protest [4]. Although Mazon's experiment was exploratory, the results clearly demonstrated significant differences among the various APIs that would fundamentally alter the outcome of any analytic inquiry.

In 2012, Oxford Internet Institute conducted an experiment to compare the tweets returned by similar queries of the Search and Streaming APIs [5]. According to Twitter's documentation of the two APIs, search is "focused in relevance and not completeness" and "not all Tweets are indexed" which means that "some Tweets and users may be missing" from the results [6]. These clues indicate that the Search API will return fewer results than the Streaming API, but it does not indicate how the subset will be produced. Gonzalez Bailon found that the Streaming API returned more than four times as many tweets as the Search API in a series of experiments [7]. Nearly all of the Search API results were found in the Streaming API results, with a few unexpected exceptions. While these omissions could be due to interruptions in their connection to the API or other errors, the discrepancies provide further evidence that data from the public APIs are not complete and should not be considered the entirety of all public tweets matching the search criteria. After comparing the two data sets using social network analysis, the researchers determined that the Search API results skewed steeply toward central users and more clustered regions of the network. Conversely, peripheral users were less accurately represented and may have been absent altogether [8]. Search API results are not a random sample of overall Twitter activity. Twitter's internal software plays an editorial role in selecting and yielding tweets according to a set of heuristic algorithms that are not known to outside users. Delivered tweets are also subject to the limitations of the local cache on Twitter's servers [9]. None of the available APIs provide an unfiltered, direct interface to the internal data store of Twitter. As a result, the artifacts of an emerging protest or the activities of users at the periphery may not be represented in data collected from the Search API.

The Streaming API is a publicly accessible interface for third-party software programs to collect data from the Twitter platform. The "filter" method provides external clients with a real-time stream of tweets matching a set of keyword filters [10]. The volume of these results is constrained by an undocumented upper limit. If the access stream is up to 1% of the entire Twitter stream at any point in time, the API sends a single message indicating a running total of the number of tweets that were not sent, or rate limited [11]. Since the most recent connection to the Twitter API was initiated. Twitter is strategically ambiguous about this constraint, and its documentation suggests that users should consider purchasing a subscription to a commercial data provider.

Both Weibo and Twitter transmit tweet data in the machine-readable JSON format, but each structures the data according to a different ontology. The same tweet delivered by the Streaming API and Weibo will require different software to read, take up different amounts of space on the disc, and include different supplementary metadata [12]. The supplementary metadata added by Weibo includes such features as expanded versions of any shortened URLs. Users of the Streaming API must manually add this supplementary information to their data sets. The distinction between these two formats makes visible the intermediary role played by Twitter and Weibo in the production of trace data. These data are not "raw" but are rather shaped according to unspoken criteria regarding what is valuable to know and how it ought to be categorized.

If the network connection is interrupted because the user 's local machine crashes, is turned off, or the API disconnects, there will be a temporal gap in the final data set. For this reason, the demands on local infrastructure are much higher for real-time streams than for asynchronous systems. Ideally, one or more computers will be dedicated to the data collection process and remain connected to the Streaming or Weibo API 24 hours a day. Merely catching tweets as they come across the network connection is just one aspect of the local data apparatus, of course. Additional human and technology resources are required to monitor and manage the incoming stream, clean and categorize new tweets, and maintain an archive of past activity.

3.2. Exploratory Data Analysis

Exploratory data analysis always reveals data quality issues. A large, real-world data set was directly usable without data cleaning. Sometimes there are outright bugs, such as inconsistently formatted messages or values that shouldn't exist. Cleaning data often involves sanity checking. A common checking technique for a service is to make sure the sum of component counts matches the aggregate counts. Another is to compute various frequencies from the raw data

to make sure the numbers seem reasonable. This is surprisingly difficult to identify values that seem suspiciously high or suspiciously low requires experience, since the aggregate behavior of millions of users is frequently counter-intuitive. We have encountered many instances in which we thought that there must have been data collection errors, and only after careful verification of the data generation and import pipeline were we confident that users did really behave in some unexpected manner.

Sanity checking frequently reveals abrupt shifts in the characteristics of the data. For example, in a single day, the prevalence of a particular type of message might decrease or increase by orders of magnitude. This is frequently an artifact of some system change, for example, a new feature just having been rolled out to the public or a service endpoint that has been deprecated in favor of a new system. Weibo has gotten sufficiently complex that it is no longer possible for any single individual to keep track of every product roll out, API change, bug fix release, etc. Thus abrupt changes in data sets appear mysterious until the underlying causes are understood. But understanding the underlying causes often wastes time to require cross-team cooperation. Even when the messages are correct, there are usually a host of outliers caused by non-typical use cases, most often attributable to non-human actors in a human domain. Without discarding these outliers, any subsequent analysis will produce skewed results. Although over time, a data scientist gains experience in data cleaning and the process becomes more common, there are frequently surprises and new situations. We have not yet reached the point where data cleaning can be performed automatically.

4. Big Data Mining of Weibo

Typically, after exploratory data analysis, the data scientist is able to more precisely formulate the problem, cast it within the context of a data mining task, and define metrics for success.

4.1. Big Data Mining

The term of Big Data appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck [13]. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [14]. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2003 in a paper by Diebold [15]. The origin of the term of Big Data is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [16] in his invited talk at the KDD BigMine'12 Workshop presented amazing data numbers about internet usage. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies such as Google, Apple, Facebook, Twitter, Weibo are starting to look carefully to this data to find useful patterns to improve user experience. Alex in Human Dynamics Laboratory at MIT, is doing research in finding patterns in mobile data about what users do, not what they say they do. We need new algorithms, and new tools to deal with all of this data. The three V's of big data management had been given as following [17]:

Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process.

Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more.

Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Nowadays, there are two more V's:

Variability: there are changes in the structure of the data and how users want to interpret that data.

Value: business value that gives organizations competitive advantages, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

Feldman [18] summarizes this in their definition of Big Data as high velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. There are many applications of Big Data, for example the following [19]:

Business: costumer personalization, churn detection.

Technology: reducing process time from hours to seconds.

Health: mining DNA of each person, to discover, monitor and improve health aspects of every one.

Smart-cities: cities focused on the sustainable economic development and high quality of life, with wise management of natural resources.

These applications allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [20].

With a precisely formulated problem in hand, the data scientists can now gather training and test data. In this case, we could use data from a few weeks ago to predict if the user is active today. Now feature extraction and machine learning would be familiar to all data mining researchers and practitioner. Applying domain knowledge, the data scientist would distill potentially tens of terabytes of log data into much more compact sparse feature vectors, and from those, train a classification model. In Weibo or Twitter, this would typically be accomplished via Pig [21] that is compiled into physical plans executed as Hadoop jobs. For a more detailed discussion of our large-scale machine learning infrastructure, we refer the reader to a recent paper [22]. The data scientist would now iteratively refine the classifier using standard practices, such as cross-validation, feature selection, tuning of model parameters, etc. After an appropriate level of effectiveness has been achieved, the classifier might be evaluated using data from today and verifying prediction accuracy a few weeks from now. This ensures that we have not inadvertently given the classifier future information. At

this point, we have achieved a high level of classifier effectiveness by some appropriate metric, on both cross-validated retrospective data and on prospective data in a simulated deployment setting. For the academic researcher, the problem can be considered solved.

4.2. Weibo Data Analysis Method

The evolution of Network analysis is from the initial quantitative analysis [23] and sociological network analysis [24] to the online social network analysis in early 21st century. Many popular online social networks, such as Weibo, Twitter and Facebook have become increasingly popular in recent years. These online social networks usually contain a lot of links and content data. The link data is a graphical structure, which represents the communication between the two entities. The content data contains the text, images and other multimedia data network. The rich content for networks data analysis to unprecedented challenges, but also opportunities.

According to the data-centric view, the research of Weibo can be divided into two categories: structural analysis of the link-based and content-based analysis [25]. The link-based structural analysis has focused on link prediction, community found, social network evolution, social impact analysis and other fields. Weibo can be visualized as a graph. The point of graph corresponds to a person, while the edges represent some correlation between the corresponding persons. Since Weibo network is a dynamic network, new vertices and edges have been added into the graph constantly. Link prediction hopes to predict the possibility of connection between two nodes in future. Many techniques can be used to link prediction, for example, feature-based classification, probabilistic methods and linear algebra. Feature-based classification can select a set of features as a sentinel, and use the existing link information to produce a binary classifier to predict the future situation links [26].Probability method tries to establish the model for the connection probability between fixed-points in Weibo network [27]. Linear algebra mainly calculates the similarity between two points based on reduced rank similarity matrix [28]. Community refers to a sub-graph structure. The density of edges at point is greater, while the density of point is lower in sub-graph structure. It has been proposed and compared methods for testing community. Most of the methods are based on topology and relies on the objective function of the concept of community structures [29]. Du using overlapping communities in real life nature presents a more efficient large-scale social network community detection method [30]. The research of Weibo network aimed at finding laws which could interpret network evolution and deducing model. Some empirical studies have found the proximity bias, geographical restrictions and other factors play an important role in the evolution of Weibo networks [31]. Simultaneously it also proposes many generation methods to help network and system design [32]. Social impact is that individuals change their behavior by the influence of others in network. The strength of social impact depends on many factors, such as the relationship between people, network distance, time effect, network and individual characteristics [33]. Marketing, advertising, recommended, and many other applications can gain benefits with measuring of qualitatively and quantitatively personal influence on others [34]. Typically, if take the content of proliferation between the Weibo networks into account, the performance based structural analysis of the link can be further improved. Thanks to the technology of Web2.0 revolutionary advances, the content of Weibo network has exploded. Researches of content-based refer to media analysis in Weibo network. Weibo media includes text, multimedia, location and comments. Almost all of the research topics of structural analysis, text analysis and multimedia analysis can be interpreted as Weibo media analysis, but Weibo media analysis is facing unprecedented challenges.

Firstly, we need to automatic analyze growing and a large number of Weibo media data within a reasonable period of time. Secondly, social media data contains many noise data. Such as a large number of spam blog in Weibo, there are also trivial tweets in Twitter. Thirdly, the social network is a dynamic network, often frequent changes and updates in a very short time. Weibo media close to the Weibo network, Weibo media analysis inevitably is affected by Weibo network analysis. Weibo network analysis refers to a Weibo networking context, in particular texts and multimedia analysis of the structure features of network. The research of Weibo media analysis is still in its infancy. Text analysis applications of Weibo network include key-word search, classification, clustering, and migration study in heterogeneous network. Keyword Search attempts to use content and link behavior to search [35]. The hidden meaning behind the application is usually linked together with a text document containing similar keywords. In the classification process, the assumed Weibo network node has a label, then spiked node for classification purposes [36]. In the clustering process, the researchers try to determine the set of nodes have similar content, and thus cluster [37].

Given the Weibo network contains a large amount of information of different types of objects linked to each other, such as articles, labels, graphics and video, in a heterogeneous network of transfer learning designs to migrate information and knowledge between different links [38]. In Weibo network multimedia data set a structured form of organization includes a wealth of information content, such as semantic ontology, social interaction, community, media, geographic maps, and multimedia views. The structure analysis of multimedia in Weibo network is also referred to multimedia information networks. Link structure of multimedia information networks which is a logical structure is essential for multimedia in network multimedia. Multimedia information networks in the logical connection structure can be divided into four categories: ontology, social media, personal photo albums and location [39]. Based on the logical connection structure we can further improve the retrieval system [40], recommendation system [41], collaborative tagging systems [42] and other applications.

5. Conclusion

Currently, the research of big data analysis has made a lot of achievements, the characteristic of data in Weibo network has also been relatively clear. But using big data theory to analysis Weibo data is also faced with many new problems. First, we faced with the complex data objects and variety of property types in Weibo network, the complexity for researching the distributed of Weibo data and collaborative association law is a challenge. Secondly, Weibo data from a large number of Internet users, the problem of artificially randomness and high noise arise, so finding a way to simplify the process paradigm becomes necessary. Finally, with the increase of data, Weibo data facing security and privacy issues, it requires that in data processing data can be protected, which is a new challenge in Weibo data processing.

References

[1] L. Manovich, "Trending: the promises and the challenges of big social data," Debates in the digital humanities, pp. 460–475, 2011.

[2] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.

[3] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," Knowledge and Data Engineering, IEEE Transactions on, vol. 23, no. 10, pp. 1498–1512, 2011.

[4] N. Kshetri, "The emerging role of big data in key development issues: Opportunities, challenges, and concerns," Big Data & Society, vol. 1, no. 2, p. 2053951714564227, 2014.

[5] S. Gonza´ lez-Bailo´ n, N. Wang, A. Rivero, J. Borge Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," Social Networks, vol. 38, pp. 16–27, 2014.

[6] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, 2013, pp. 995–1004.

[7] S. Gonza´ lez-Bailo´ n, J. Borge-Holthoefer, and Y. Moreno, "Broad- casters and hidden influentials in online protest diffusion," American Behavioral Scientist, p. 0002764213479371, 2013.

[8] C. R. Shalizi, A. Rinaldo et al., "Consistency under sampling of exponential random graph models," The Annals of Statistics, vol. 41, no. 2, pp. 508–535, 2013.

[9] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Information, communication & society, vol. 15, no. 5, pp. 662–679, 2012.

[10] K. Nahon, J. Hemsley, R. M. Mason, S. Walker, and J. Eckert, "Information flows in events of political unrest," 2013.

[11] A. Bruns and J. E. Burgess, "The use of twitter hashtags in the formation of ad hoc publics," 2011.

[12] A. Bruns and J. Burgess, "Notes towards the scientific study of public communication on twitter," Science and the Internet, pp. 159–169, 2012.

[13] F. X. Diebold, "On the origin (s) and development of the term'big data'," 2012.

[14] S. M. Weiss and N. Indurkhya, Predictive data mining: a practical guide. Morgan Kaufmann, 1998.

[15] F. X. Diebold, "big datadynamic factor models for macroeconomic measurement and forecasting," in Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,(edited by M. Dewatripont, LP Hansen and S. Turnovsky), 2003, pp. 115–122.

[16] U. Fayyad, "Big data analytics: applications and opportunities in on-line predictive modeling," in Keynote Talk. BigMine: BigData Mining Workshop KDD-2012, Beijing, China, 2012.

[17] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," ACM sIGKDD Explorations Newsletter, vol. 14, no. 2, pp. 1–5, 2013.

[18] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering," in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 2013, pp. 1434–1453.

[19] C. C. Aggarwal, Managing and mining sensor data. Springer Science & Business Media, 2013.

[20] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza, "Big data, big business: bridging the gap," in Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. ACM, 2012, pp. 7–11.

[21] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon, "Parallel data processing with MapReduce: a survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11–20, 2012.

[22] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012, pp. 793–804.

[23] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." MIS quarterly, vol. 36, no. 4, pp. 1165–1188, 2012.

[24] C. D. Brummitt, R. M. DSouza, and E. Leicht, "Suppressing cascades of load in interdependent networks," Proceedings of the National Academy of Sciences, vol. 109, no. 12, pp. E680–E689, 2012.

[25] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Community detection with edge content in social media networks," in Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012, pp. 534–545.

[26] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 1046–1054.

[27] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," Access, IEEE, vol. 2, pp. 652–687, 2014.

[28] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," ACM Transactions on the Web (TWEB), vol. 6, no. 2, p. 9, 2012.

[29] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," Knowledge and Information Systems, vol. 42, no. 1, pp. 181–213, 2015.

[30] J. D. Cruz, C. Bothorel, and F. Poulet, "Entropy based community detection in augmented social networks," in Computational aspects of social networks (cason), 2011 international conference on. IEEE, 2011, pp. 163–168.

[31] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location- based online social network: analysis and models," in Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, 2012, pp. 145–158.

[32] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013, pp. 587–596.

[33] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," GeoJournal, vol. 78, no. 2, pp. 319–338, 2013.

[34] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang, "Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships," in Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013, pp. 657–666.

[35] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi, "Online team formation in social networks," in Proceedings of the 21st international conference on World Wide Web. ACM, 2012, pp. 839–848.

[36] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining text data. Springer, 2012, pp. 163–222.

[37] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song, "Evolution of social-attribute networks: measurements, modeling, and implications using google+," in Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, 2012, pp. 131–144.

[38] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of big data on cloud computing: review and open research issues," Information Systems, vol. 47, pp. 98–115, 2015.

[39] K. Fujimoto and T. W. Valente, "Social network influences on adolescent substance use: Disentangling structural equivalence from cohesion," Social Science & Medicine, vol. 74, no. 12, pp. 1952–1960, 2012.

[40] M. Rabbath, P. Sandhaus, and S. Boll, "Multimedia retrieval in social networks for photo book creation," in Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ACM, 2011, p. 72.

[41] S. Shridhar, M. Lakhanpuria, A. Charak, A. Gupta, and S. Shridhar, "Snair: a framework for personalised recommendations based on social network analysis," in Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, 2012, pp. 55–61.

[42] S. Maniu and B. Cautis, "Taagle: efficient, personalized search in collaborative tagging networks," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012, pp. 661–664.