Integration Issues of Big Data Analysis on Social Networks

A V Ivaschenko[1], N Yu Ilyasova[1,2], A A Khorina[1], V A Isayko[1], D N Krupin[3], V A Bolotsky[3], P V Sitnikov[4]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, Molodogvardeyskaya str. 151, Samara, Russia, 443001
[3]IPSI SEC "Open Code", Yarmarochnaya Str. 55, Samara, Russia, 443001
[4]ITMO University, Birzhevaya liniya 14 lit. A, Saint-Petersburg, Russia, 199034

Abstract. Nowadays Social Media becomes one of the major providers of Big Data for analysis of users' behaviour, focus, trends, and deviations. One user can be presented in several social networks by various avatars. Most users have different dynamics of data processing and generation. In order to provide a solution capable to deal with this, there was developed and implemented a software library for integration with a number of social networks. This paper describes the problem, solution architecture and technical details of its implementation supported by the results of simulation and real data analysis for a number of popular social networks.

1. Introduction

The way we deal with the advent of the era of Big Data is crucial. Although this phenomenon has the right take place in conditions of uncertainty form the future, but with increasing automation of data collection and analysis - the number of algorithms that can extract and illustrate large-scale models of human behavior also increases. How do systems conduct this practice, and how do they regulate the flow of data?

The market sees big data like net opportunity: marketers optimizing their proposals, based on market analysis, Wall Street bankers process tons of information about the dynamics of changing rates. Legislation has already been suggested to limit the collection and storage of data, as a rule, about the inviolability of private life.

In recent years, the amount of information formed by business, science and social networks increases in geometric progression. This phenomenon is also called phenomenon known as a data stream.

In business, Valmart's valuation transaction databases estimate the amount of data currently stored in more than 2.5 petabytes of data, including: information on customer behavior and preferences, data about network activity and devices, information on market trends.

As for science, for example, the Large Hadron Collider (LHC) in The European Organization for Nuclear Research produced 13 petabytes of data in 2010.

In addition, the sensor, social networks, mobile data, subscriber data and the location data grow at a frenzied pace. Simultaneously with this growth in the volume of information, data also become more

interconnected. Facebook, for example, is almost completely connected, from 99.91% network to one, large connection component.

Modern social media can be treated as a major source of Big Data that describes the process of users' interaction and various information exchanges. Analysis of this data turns out to become a complicated technical problem: it is required to integrate with multiple social media for data import, associate separate profiles of the same users in different networks, match the facts of their interaction across the real events and derive basic trends and deviations.

To solve this problem there was developed a model of social media user behavior and a based on it software solution that provides capabilities for social networks analysis and simulation.

2. State of the Art

New opportunities of interaction in virtual environments allow Internet users to exchange the ideas immediately. At the same time everybody needs to obtain and process lots of incoming events. Under such informational pressure individuals start prioritizing the most important data, filtering and rejecting everything that is not currently interesting. Such a focus on the current interest instead of importance leads to various imperfections, including the creativity constraints. This process can be described by the modern principles of distributed simulation and decision-making support powered by multi-agent technology [1].

The virtual world of social media should be treated as a complex network of continuously running and co-evolving intelligent agents. Such solutions are based on holons paradigm and bio-inspired approach [2], which requires development of new methods and tools for supporting fundamental mechanisms of self-organization and evolution similar to living organisms (colonies of ants, swarms of bees, etc) [3]. As for the human beings represented by actors or agents, social network user should consider a combination of human and time factors. Interaction of customers and service providers powered by intermediary services generate and can be characterized by a big number of events that form Big Data and require modern technologies for its analysis [4].

Modeling the Internet users' behavior can be based on the modern principles of knowledge representation in the form of Ontologies [5]. These concepts allow formalizing self-organization and semantics, which is advantageous for abstract description of social concepts and their interaction in technical applications [6]. The papers discusses the detection process of uncharacteristic behavior of users [7] and methods classify users [8].

In the context of this paper there should be mentioned the papers on Internet development strategies [9], virtual communities and social networks studies [10-11]. Despite the successful application of mathematical statistics used to cluster and generalize the user's behavior the problem of Big Data analysis of social networks remains open. This happens due to a necessity to personalize user activity models and understand individual features of human behavior.

Our experience in the area of integrated information space development and its users' behavior analysis [12-15] can be used to build a software solution to derive basic trends in social media and provide intelligent functionality for social media big data analysis. The proposed abstract model and solution vision are given below.

3. Abstract Model

Let us present a community of Internet users by u_i, where $i = 1\backslash..N_u$ – a number of users. The activity of users information exchange can be presented by posts, comments or messages p_j, where $j = 1\backslash..N_w$ – an absolute number of an informational object. Post generation is an event

$$g_{i,j} = \left(u_i, p_j, t^0_{i,j}\right). \quad (1)$$

Issue or processing of an information object can be presented by an event $e_{i,j,k}$ that can be characterized by the combination of user, focus, and time:

$$e_{i,j,k} = e\left(p_j, \left(u_i, f_{i.k}, t_{i,j,k}\right)\right) = \{0,1\}, \quad (2)$$

where focus $f_{i,k}$ presents the current user interest and can be described by a tag cloud, which is a set of pairs:

$$f_{i,k} = \left\{ \left(\tau_n, w_{n,k} \right)_{i,k} \right\}, \tag{3}$$

where τ_n is a tag (keyword) with weight $w_{n,k}$.

The sequence of interdependent user focuses represents the evolution of the user's interest.

Each user has own ontology that forms the basis of his perception. It changes with time under the influence of learning and forgetting the information (presented by posts, comments or messages) and can be presented by a chain of contexts:

$$c_{i,m} = \left\{ \left(\tau'_l, w'_{l,m} \right)_{i,m} \right\}. \tag{4}$$

This change is correlated with user focus.

The focus cannot be considered new in order to provide positive perception, and at the same time it is not equal to the context to be able to excite interest.

Considering this correlation let us synchronize the context and focus changes:

$$e'_{i,j,m} = e' \left(p_k, \left(u_i, c_{i.m}, t_{i,j,m} \right) \right) = \{0,1\}. \tag{5}$$

The statements (2) and (5) are Boolean variables, which mean that appearance or perception of a post, comment or message does not guarantee changes in focus and context.

Events (2) and (5) can be used for analysis. One of the possible implementations is presented below. Study of the user's focus and context trends allows identifying tendencies, variations and iterations that form the patterns of user creativity.

In case new informational proposal remain suspended and does not make any effect over the user focus, this means that the user does not see any interest.

Possible reasons are concerned with context: additional education is needed to provoke such interest. On the other side, lots of changes of the user context indicate the search for a stable interest that should be proposed for the user at a certain time.

Context and focus can be also influenced by negative intervention.

In order to manage the user focus there can be generated a series of repeated affections partially covering the actual context and the targeted interest. Such patterns can also be identified applying cross-correlation analysis to the proposed model, which helps identification and resistance to negative informational influence.

4. Solution Architecture

The proposed approach is based on simulation of focus and context. It was implemented in multi agent architecture, which is presented in Figure 1.

Under the bounds of our proposed architecture we provide profile descriptor, post generator and navigator.

These are methods generated and used to simulate real activity of users in the social networks. Post generator used to create posts according to predefined logic.

Navigator is used to process incoming data which is described by network and can be presented like a sorted graph there the nodes are informational objects, for example web sites, documents, posts, comments, and the links are references between these objects.

Each object can refer to several other objects and documents; and the navigator according to all predefined logic decides which link to go.

In addition to navigator and post generator the informational frames are provided under the multi-agent architecture that correspond to a predefined above focus and context concepts. Focus is used to represent the current interest of Internet users.

Context is used to formalize informational space in which the agent performs its negotiating activity. Based on the provided model an algorithm has been developed for social media big data analysis.

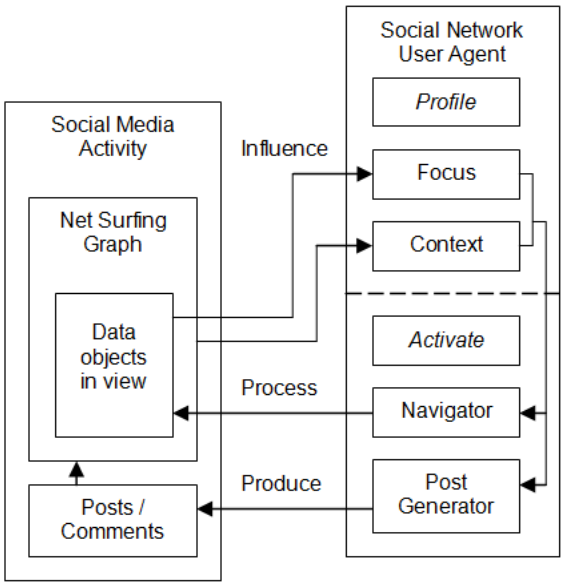

Figure. 1. Solution multi-agent architecture.

The model is used to formalize the social media user and integrate the analytical software with various social media open for data import and analysis.

This algorithm consists of 2 stages:

1. Calculation of the sample frequency vector for all users and development of the standard deviation vector for a variety of users.

You need to select topics and convert them to a view

$$\left(T_i, \frac{1}{N_{i,p}} \right), \tag{6}$$

where p is an hour of publication, $N_{i,t}$ is a number of users that posted on a certain topic in a time period p, T_i is an identificator of a topic.

Then process the received pairs and calculate the amount

$$\left(T_i, \sum_{k=0}^{i} \frac{1}{N_{k,p}} \right) \tag{7}$$

and calculate the standard deviation for each pair σ_j^k.

The obtained values are divided by the period.

2. Calculating the deviation metric for a particular user. You need to select topics and convert them to a key-value view. After that you need to process the data pairs and count the sum of topics with the same key:

$$\begin{cases} 0, \Delta_i \le 3\sigma_i, \\ 1, \Delta_i > 3\sigma_i \end{cases} \tag{8}$$

Count the deviation of a particular user, summarize the deviations of a particular user, divide the sum of the deviations by the number of topics (n) for a particular user and generate the resulting CSV-file in the form of a table with user data and information of standard deviation of this user.

One of the main features of Internet users' activity online that should be considered in the explored scope is mutual influence of contexts and focuses of communicating peers. This factor makes it possible to introduce the control loop: in addition to web content semantic analysis the platform starts to manage the users interest based on focus identification and context feedback.

This information is being collected in social networks and has all necessary details to get actual estimations. Still in this case it is required to provide integration with social networks and the data being processed contain tons of subjective assessments and perceptions.

Online libraries and professional communities are more neutral. For example, Wikipedia enforces various groups of authors to update the articles targeting maximum objectivity. Analysis of this data can help adequate identification of significant trends of consumers' focus identification that can be practically used e.g. for marketing and product placement.

Activation method is used to simulate multi agent activities in real time. The special agent dispatcher will call all the agents by using this activation method and after be activated each time an agent generates the time series period according to some distribution rule. Agent generates time series of navigation calls and post generation methods.

At these stages, we solve the proposed navigation and generation in such a way that we can model post-writers or readers and introduce some specific patterns of online activity, for example, the agent can be more active at night, or we can use some time frame for high / low activity.

Focus and contexts update the results of real agent behavior based on the influents of informational objects. We can generate focus and contexts according to our goals and in case we want agent to behave in a sort of specific way we introduce this control directly and formulate focus and context the agent will do that you want.

This approach allows simulating this influence and is introduced in the system. In this case we need to analyze the focus and context changes of the agent during the period of time. On the basis of analysis we introduce changes in focus by generating informational objects inside this network. This can be done in real systems using the contexts based advertisements. We can generate just the objects with certain informational context, which can be described by tag clouds.

The introduced architecture can be used to simulate online users in social networks and model realistic Internet behavior. In the area of simulation, practical application is generation of cognitive patterns of collective behavior based on self-organization. In this case the agent should be simple and the logic of focus and context should be close to very simple but generic behavior. This logic can correspond to know real users of social networks but it can represent some generalized behavior and the community of agent. Such behavior can be used to study and develop some visual cases.

In another case it is implemented as a sort of a frame, using which the algorithms of syntactical analysis or other large data analysis can look onto the real world of social networks and filter the data for intelligent study.

5. Implementation

To implement the proposed approach there was developed a software solution for social media focus identification based on knowledge discovery and Big Data analysis.

The solution can integrate with various data sources, pick out concepts, generate tag clouds for contexts and focuses and process their changes in time. Solution implementation architecture is presented in Fig. 2. The data imported from social networks is captured in database and can be processed either in real time or in batch mode.

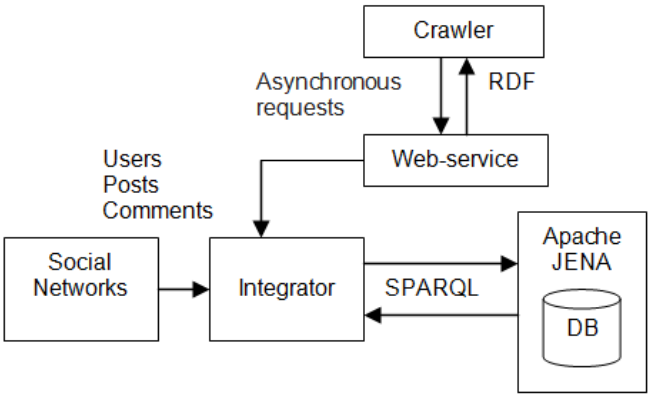

Figure 2. Integration model.

Crawler addresses asynchronously to a web service with requests for data from social networks. After receiving the request, the web service starts processing it. Next, the web service accesses the integrator, which starts downloading the requested data in the form of RDF / XML files, storing the intermediate data received from the single request of the crawler to receive the data by the single block to transfer the already downloaded ones.

Then in the background, i.e. in a mode where there is no need to control the data unloading process, the integrator automatically continues the embedded process and uploads the data to the database and uses Apache JENA to generate RDF / XML files that will be transferred to the first crawler address. The described model, software solution and its implementation was probated and tested using a typical data set derived from a number of social networks. In addition to a real regular result set of social media users' negotiation there was introduced a peak batch of posts generated by an online bot.

Apart from the social media (getting no a prior knowledge of a data structure) the big data analysis algorithms was able to identify the online bot influence. The results are presented in Fig. 3. Gray lines represent the annual trends of users' activity. The peak identified on Aug 15 corresponds the Bot activity and can be easily identified by the agent comparing the behavior of previous periods. The described research results show that the proposed model can be used for online behavior analysis and identification of negative informational influence.

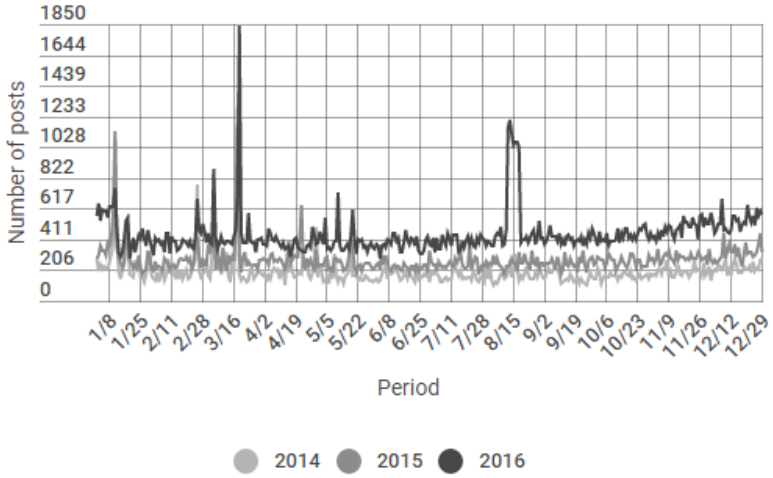

Figure 3. Bot activity identification.

There were processed the data of 32,000 users and their posts for the period of 2014 – 2017. To simulate the intervention there were modeled 50 bot users that automatically perform actions through interfaces intended for people. The given statistics show the distribution of posts throughout the considered time for each year. The horizontal axis, respectively, is temporary, contains the values of t recalculated in step 6 of the abovementioned algorithm.

Each line has 2 similar peaks at the beginning of the year (a detailed analysis showed that such emissions fall on holidays), they are similar to each other throughout the rest of the time. But the curve of 2016 has an unusual outburst (see 8/15), which characterizes the appearance of users' unusual behavior.

This example illustrates that the model and statistically developed patterns of users creativity can be used to identify negative deviations and the attempts to influence using repeated affections.

6. Conclusion

As shown above, the proposed model allows capturing the process of Internet user's activity considering a combination of human and time factors.

7. References

[1] Wooldridge M 2002 *An introduction to multi-agent systems* (Chichester: John Wiley and Sons) p 340

[2] Leitao P 2009 Holonic rationale and self-organization on design of complex evolvable systems *HoloMAS* **LNAI 5696** 1-12
[3] Gorodetskii V 2012 Self-organization and multiagent systems: I. Models of multiagent self-organization *Journal of Computer and Systems Sciences International* **51(2)** 256-281
[4] Bessis N and Dobre C 2014 *Big Data and Internet of Things: A roadmap for smart environments* (Berlin: Springer) p 450
[5] Mouromtsev D, Pavlov D, Emelyanov Y, Morozov A, Razdyakonov D and Galkin M 2015 The simple, web-based tool for visualization and sharing of semantic data and ontologies *CEUR Workshop Proceedings* **1486** 77
[6] *One Internet. Global commission on Internet Governance* 2016 (Access mode: https://www.cigionline.org/initiatives/global-commission-internet-governance) (01.11.2017)
[7] Shatalin R, Fidelman V and Ovchinnikov P 2017 Abnormal behavior detection method for video surveillance applications *Computer Optics* **41(1)** 37-45 DOI: 10.18287/2412-6179-2017-41-1-37-45
[8] Rybintsev A, Konushin V and Konushin A 2015 Consecutive gender and age classification from facial images based on ranked local binary patterns *Computer Optics* **39(5)** 762-769 DOI: 10.18287/0134-2452-2015-39-5-762-769
[9] Balakrishnan H and Deo N 2006 Discovering communities in complex networks *Proceedings of the 44th Annual Southeast Regional Conference* 280-285
[10] Wei W, Joseph K, Liu H and Carley K 2016 Exploring Characteristics of Suspended Users and Network Stability on Twitter *Social Network Analysis and Mining* 6-51
[11] Kadushin C 2012 *Understanding social networks: theories, concepts, and findings* (Oxford: Oxford University Press) p 264
[12] Ivaschenko A 2014 Multi-agent solution for business processes management of 5PL transportation provider *Lecture Notes in Business Information Processing* **170** 110-120
[13] Ivaschenko A, Minaev A and Spodobaev M 2015 Self-mediator software for sensor networks *Proceedings of the 2015 International Siberian Conference on Control and Communications (SIBCON)* 1-4
[14] Ivaschenko A, Lednev A, Diyazitdinova A and Sitnikov P 2016 Agent-based outsourcing solution for agency service management *Lecture Notes in Networks and Systems* **16** 204-215
[15] Protsenko V, Kazanskiy N and Serafimovich P 2015 Real-time analysis of parameters of multiple object detection systems *Computer Optics* **39(4)** 582-591 DOI: 10.18287/0134-2452-2015-39-4-582-591