The analysis of profiles on social networks

V.A. Bakayev[1], A.V. Blagov[1]

[1]Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

Abstract

The article is devoted to the analysis of data and communications on various social networks. The method of search of the profiles belonging to the same users, based on the analysis of the communications and communities which are available for a profile is offered. The program complex realizing this method is created.

Keywords: data mining; social networks; graphs; Label Propagation algorithm; Apache Spark

1. Introduction

Now one of the most urgent directions in information technologies is the analysis of data or Data Mining. The analysis of data represents process of detection of data, suitable for use, in large data sets, often diverse. Usually such data can't be found at traditional viewing and search as communications are too difficult, or because of the excessive volume.

On social networks big data flows are generated (profiles, communications, content are created). Analyzing these data it is possible to obtain a lot of useful information as on various groups, communities and discussions, and on each user separately [1-4].

The great interest in social networks is shown by various commercial organizations using them as the instrument of interaction with audience. Applying specialized services, the companies analyze information on users, their activities and personalize offers for separately taken segments of the target audience, thereby increasing conversion and reducing costs of advertizing campaign.

In the article the method of increase in efficiency of this sort of tools and services which is based on psychology and patterns of human behavior is offered.

The offered method is based on the following facts:

- many Internet users have accounts on several popular social networks at once (VKontakte, Facebook, Instagram and Twitter);
- many users of social networks hide information on themselves from strangers (including information on existence of accounts in other social networks);
- as social networks are a subject of socialization of people, for each user it is possible to allocate at least one community of people it that users of it community are in pairs familiar with each other;
- the person has accounts in different social networks and contacts to the same people.

Is developed the program complex which for realization of a method:

a) analyzes all profiles of target social networks and on the basis of public data finds the accounts belonging to the owner of an initial profile;

b) for users on whose pages there is no information on existence at them of accounts in other social networks finds communications with other social networks on the basis of communities in which the user consists.

2. The object of the study (Model, Process, Device, Sample preparation etc.)

At the first stage the system analyzes all users of social networks VKontakte, Twitter and Instagram and groups them in the following rules:

- in each group there are no more than one profile from each social network;
- all profiles in one group belong to one person.

This problem is solved by means of a program framework of Apache Spark (in particular, superstructures of Spark Streaming intended for stream data processing see fig. 1) and the broker of messages RabbitMQ realizing delivery of basic data in Spark Streaming [5].

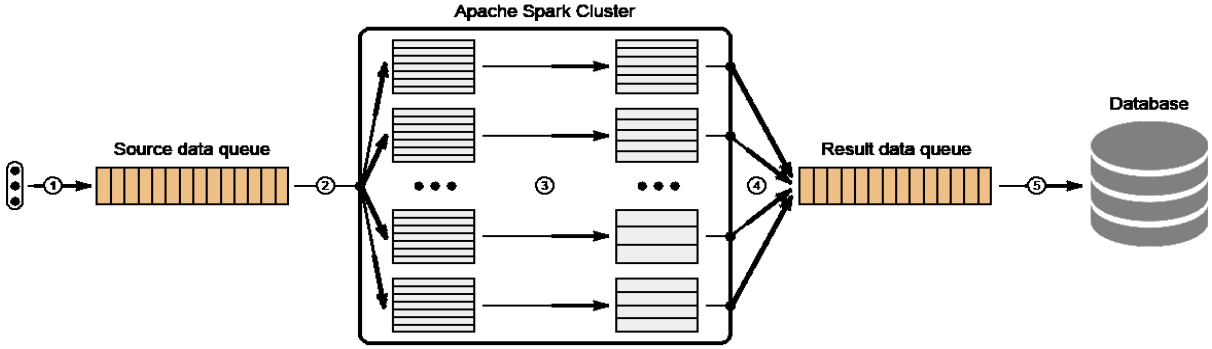

Fig. 1. Architecture of the aggregator of users of social networks (pipeline).

Description of steps:

1. Adding of data from different sources in queue for later processing. Data represent a set of couples (network_id, user_id) containing information on profiles which are required to be analyzed.

2. RDD (Resilient Distributed Dataset) formation by a packing of the basic data which are in queue for increase in productivity.

3. RDD (mapping) conversion. For each couple (network_id, user_id) the algorithm finds and groups profiles on other social networks, and also additional information on the person to whom belongs the initial account. The algorithm is restarted for each found profile until all available information on the user is found. As sources can be: the public information specified on the page of the user (the status, contact information, entries in the film, etc.).

4. Export of data retrieveds from RDD in queue for the subsequent saving.

5. Saving results in NoSQL to the MongoDB database in the form of documents with structure, the reflected in table 1.

The speed of data processing makes about 120-130 profiles a second. For work the Microsoft Azure A2 v2 virtual computer was used (2 kernels, 4 GB of RAM, 20 GB of SSD). Casual users of social network VKontakte (1.000.000 profiles) were analyzed.

Thus, if to assume that speeds of processing of profiles of VKontakte, Instagram and Twitter are equal, we will receive an approximate assessment of time which will be required for the analysis of all users of target social networks:

$$T(n) = \frac{4 * 10^8 + 6 * 10^8 + 13 * 10^8}{120n} \approx 5324\ hours \approx 221\ day, \quad (1)$$

where n - the number of servers in a cluster with a similar configuration.

In case of horizontal scaling of a cluster the linear dependence between the number of servers and processing rate of profiles is watched.

Example of the reference to the table: results of an experiment are reflected in table 1.

Table 1. Data storage structure of profiles.

Name of the field	Type	Description
_id	ObjectId	the document identifier in a collection
vk_id	Int32	the profile identifier in VKontake
facebook_id	Int64	the profile identifier in Facebook network
instagram_id	Int64	the profile identifier in Instagram network
twitter_id	Int64	the profile identifier in Twitter network
other	Object	The additional information (phone number, e-mail address, skype, etc.)

The further task comes down to expansion of the received base by association of profiles by the rules described earlier on which pages bek-links aren't specified other social networks.

3. Methods

Based on a hypothesis that the person consists in the same communities on all social networks which uses, we can establish connection between profiles of the different social networks which are in one community and with a certain probability to assume that they belong to one person.

The purpose of this stage is separation of communities among the people who are in the database received from the previous step. Extension of Apache Spark GraphX which is intended for distributed processing of graphs is for this purpose used.

The algorithm of preliminary data handling:

1. Generation of a graph on the basis of the available data. Peaks represent an entity of "people" and store identifiers of the profiles belonging to the specific user. Connection between peaks is established by the following principle: two peaks are adjacent if profiles of matching social networks are coherent. We will determine edge weight between peaks as:

$$w(A,B) = |\{i: i \in [0, n-1] \wedge \exists\, A_i, B_i \wedge rel(A_i, B_i)\}|, \quad (2)$$

where rel (a, b) - function, which the truth in only case when when profiles an and b are interconnected (the relation of friendship or manifestation of activity is established).

2. The algorithm Label Propagation [6] realized in GraphX API which solves the problem of a clustering is applied to the received graph and finds communities in the graph.

3. For each community RDD with a set of profiles of VKontakte, Facebook, Instagram and Twitter which are connected to one or several members of the initial community is generated.

Thus the data set (dataset) on the basis of which we can speculate about accessory of group of accounts of different social networks (without obvious indication of interrelations) to one person is formed.

Grouping of the common features given on the basis of the analysis is final stage in the solution of an objective. The following procedure is applied to each data set from a dataset:

1. Creation of the full multiple-count graph in which information on profiles of social networks and the potential characterizing probability of their accessory to one person is stored;

Tops of the graph contain information on profiles which is used at their comparison.

For comparison of two profiles the multilayered neural network is used. On an entrance layer of network the vector of dimension 12 containing the following data moves:

- Name ↔ Name'
- max (Name → Username', Name' → Username)

- max (Name → E-mail', Name' → E-mail)
- max (Name → Skype', Name' → Skype)
- Username ↔ Username'
- max (Username → E-mail', Username' → E-mail)
- Username ↔ Skype'
- max (Skype → Username', Skype' → Username)
- max (Skype → E-mail', Skype' → E-mail)
- E-mail ↔ E-mail'
- Phone ↔ Phone'
- Website ↔ Website'

We will determine operations a → by b and a ↔ b:

1.1 Completeness of entry of a into b:

$$a \to b = 1 - \frac{d+r+s}{len(a)} \in [0, 1], \tag{3}$$

where d - the number of operations of removal for transformation a to b;

r - the number of operations of replacement for transformation a to b;

s - the number of operations of a transposition for transformation a to b;

len(x) - function of calculation of length of an argument.

1.2 Comparison of an and b:

$$\forall i \in [1, len(a)], j \in [1, len(b)]\ d[i,j] = 1 - \frac{dist(a[i], b[j])}{len(b[j])} \in [0, 1], \tag{4}$$

$$a \leftrightarrow b = \frac{\sum_{J}^{len(a)} d[i, fit(i)]}{min(len(a), len(b))} \in [0, 1], \tag{5}$$

where dist (a, b) - the function calculating Damerau-Levenstein [7] distance for lines an and b;

fit(i) - the function returning an index of the word of a line b put in compliance to the word a[i].

Operation of comparison doesn't consider a word order. All words of initial lines are in pairs compared, and then, by means of Kuhn-Munkres [8] algorithm, to each word of a line the word of a line b is put in compliance so that the similarity sum on all couples of words was maximum. Also punctuation marks and other symbols aren't considered (except for letters and figures).

Before processing all symbols of entrance data are given to Latin by rules of a transliteration. The summary table of the main alphabets is for this purpose used (Russian, Ukrainian, Bulgarian, Indian, Arab).

4. Results and Discussion

The training and control selections are collected on the basis of primary data. The amount of the training selection ~ 106 couples.

As negative examples both casual couples of profiles, and couples found by means of full text search in different parameters were used (name, username, email, skype).

In generated to the column for each couple of shares the following algorithm is carried out:

1. edges are sorted in decreasing order of scales;
2. edges which weight is less than threshold value are removed or one of incidental tops is already connected with any top of an opposite share.

As a result of these transformations the count in whom everyone a component of connectivity represents group of accounts of different social networks which belong to one person turns out.

Because the person can belong at the same time several community and also if the same group has been created in several communities, then it is possible to believe that accounts of this group really belong to one user.

The data obtained at the last stage register in the same base where primary data are stored. However they aren't used as entrance data for the realized algorithm in view of the unreliability.

5. Conclusion

The processing and data analysis of social networks allows to personalize a product or service for a specific segment of target audience. The program complex received as a result of operation erases boundaries between social networks for different services, allowing them to integrate API and to operate with an entity of "people", but not "profile" that does their operation more effective.

The further research can be continued regarding upgrade of algorithms of aggregation and the analysis of data retrieveds. It is necessary for implementation of the decision of the following tasks:

- the detection of popular social networks and services which users mention on the pages (for example, LinkedIn, Last.fm) and development of parcers for them;
- the analysis of additional information sources on pages of users in VKontakte, Instagram and Twitter (for example, a news feed on Twitter);
- the analysis of pages of users on target social networks and search of additional parameters based on which one user can compare profiles on accessory.

Acknowledgements

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian

Federation within the framework of implementation of the Program for Improving the Samara University Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.

References

[1] Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Computing 2013; 5: 62–69.
[2] Khotilin MI, Blagov AV. Visualization and Cluster Analysis of Social Networks. CEUR Workshop Proceedings 2016; 1638: 843–850.
[3] Protsenko VI, Serafimovich PG, Popov SB, Kazanskiy NL. Software and hardware infrastructure for data stream processing. CEUR Workshop Proceedings 2016; 1638: 782–787. DOI: 10.18287/1613-0073-2016-1638-782-787.
[4] Popov SB. The Big Data methodology in computer vision systems. CEUR Workshop Proceedings 2015; 1490: 420–425. DOI: 10.18287/1613-0073-2015-1490-420-425.
[5] Apache Spark Documentation. Access mode: http://spark.apache.org/docs/2.1.0.
[6] Smetanin N. Fuzzy search in the text and the dictionary. URL: https://habrahabr.ru/post/114997. (in Russian)
[7] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A: Statistical Mechanics and its Applications. 389(7): 1493–1500.
[8] Hungarian algorithm for solving the assignment problem. URL: http://e-maxx.ru/algo/assignment_hungary. (in Russian)