International Journal of Soft Computing 10 (3): 242-246, 2015
ISSN: 1816-9503

The Method of Support Vectors in the Analysis of Social Networks User Profiles

Alexander A. Chumak and Alla G. Kravets
Volgograd State Technical University, Lenin Avenue 28, 400005 Volgograd, Russia

Abstract: Sociological surveys and tests exist for a long period of time and help to explore humanity to identify their interests and desires but we live in a world where almost everyone has their own profile in the social networks where they post pictures, write thoughts, notes and so on. Popular social networks in Russia can be represented as an online environment which is used to facilitate social interactions such as content sharing, points of view, experience and relevant information. In this study, we consider the method of support vectors for analysis of users' profiles in social networks with the help of machine learning. We developed a system that could determine the selected network user opinion about the annexation of Crimea to Russia.

Key words: Social networks, SVM-classifier, social network user opinion analysis, pictures, notes

INTRODUCTION

SVM-classifier implements a binary classification: divides the set of input vectors into two parts positive-negative, us vs. them, yes-no, etc. In applied studies, many different methods are used to collect sociological information. The most common method of sociological research is a survey. With the help of this method about 90% of sociological information is collected in Russia. Sociological survey can be defined as a method of gathering information through oral or written questions of the researchers to the total sample of people (respondents), whose content reflects the investigated problem at the level of empirical indicators, subsequent registration, statistical processing and analysis.

Sociological methods of collecting information include methods such as: Questionnaire (Questioning), it's a survey method in which the respondent fills in the questionnaire/inquirer (profile) themselves. It means the communication between the researcher and the respondent who is the source of the information which is required in questionnaire. Questioning is divided into group and individual.

Interviewing, it's a survey method in which the questionnaire is filled in with the words of the respondent. It means that a verbal interaction (through conversation) of the researcher (interviewer) with the respondent is used as a source of information.

Thus, we can observe that for a long period of time people collect information about others and then the collected data is further analyzed. But, these methods are already outdated, laborious and take a lot of time. Moreover, methods such as postal questionnaire surveys and polls compaction are generally, ignored by users or filtered as spam by mail services.

In 1954, a sociologist from the (Manchester School) Jason Barnes in his work (Classes and meetings in the Norwegian Island Parish) introduces the concept of (social network). Social networking is a social structure which consists of a group of people and the relationships between them, this is not only sites on the internet but it can be any community of people who share common interests (Barnett, 2011).

The first social network appeared on the internet in 1995. It was the American portal odnoklass-niki.com ("classmates"). This project has been very successful that is why in the next few years a huge number of similar sites appeared. But, the official beginning of the life of the social networks is considered to be 2003-2004. In these years, many resources such as: LinkedIn, MySpace and Facebook were launched. LinkedIn was created in order to establish business contacts, the owners of MySpace and Facebook tried to do something which could help to satisfy the human's needs for self-expression. There were analogues of these services In Russia: Facebook, VKontakte and odnoklassniki.com-Odnoklassniki.ru (Barnett, 2011).

Popular social networks in Russia can be represented as an online environment which is used to facilitate social interactions such as content sharing, points of view, experience and relevant information.

Social networks have greatly changed our lives, especially when they appeared on the internet. They have provided us with a huge potential for the development of

Corresponding Author: Alexander A. Chumak, Volgograd State Technical University, Lenin Avenue 28, 400005 Volgograd, Russia

relations, grant access to a wealth of information all over the world. Social networks are highly unidirectional. This means that they can unite people with similar interests, professions, they can be directed to the search of the partner and could unite all the people in general.

The main aim of this research is to show how text classification methods can be used, exactly the method of the supervised support vector machine which was used social networks "VKontakte" and "Odnoklassniki".

Social networks analysis: For the data analysis in social networks following social networks have been selected: "VKontakte" is the most popular internet resource in Russia. In addition to the impressive number of participants in the network, "Vkontakte" can boast of a giant database of media content that attracts new users to make new accounts, create new community that can connects people with their interests.

"Odnoklassniki" (odnoklassniki.ru). It was originally created to find and communicate with odnoklassniki but like other social networks, eventually grew up and has acquired additional functions and features.

Both social networks have individual user's profiles which contain a way of publishing information on their profile (Chumak *et al*., 2013). In "VKontakte" such functionality is a user's "wall", "Odnoklassniki" is a "ribbon".

So we had the idea of analyzing profiles and we decided to analyze the social networks users from Ukraine. The main aim of this research is to verify the possibility of analyzing the social network users (Alexander *et al*., 2014). We developed a system that could determine the selected network user opinion about the annexation of Crimea to Russia.

MATERIALS AND METHODS

Selected social networks support API (Application Programming Interface) methods, including standalone applications. Standalone applications support authentication, based on OAuth 2.0 protocol (Quyen and Kravets, 2014). Thus in order to begin working through your own application with your social network profile, you must be authorized through an API Method and receive an access key that gives you access to the data on the page and users in social networks (Algorithm).

Algorithm; All methods of (VKontakte's) API returns JSON structure in response to the query in the form of:

response: {
count: 1,
items: [{
id: 2943,
first_name: 'Ivan',
last_name: 'Babich',
screen_name: 'antanubis',
photo: 'http://cs307805.vk.me/v307805943/343d/kWYZkr7tCFk.jpg'
}
}

In turn, the methods of «Odnoklassniki»'s API return the data in the structure of JSON too:

{
"uid": "2494A013S3EE",
"birthday": "1901-03-03",
"age": 110,
"first_name": "Name",
"last_name": "Surname",
"name": "Name Surname",
"gender": "male",
"has_email": true,
"pic_1": "http://i113.odnoklassniki.ru/getImage?photoId=93412337&photoType=4",
"pic_2": "http://i342.odnoklassniki.ru/getImage?photoId=93412337&photoType=2"
}

Having an access key you can obtain information from the user's wall. To do this, we use the method "wall.get" which returns a list of records from the wall or the user's community. One of the parameters which is passed to this method is (owner_id) the user's ID or the community's one, the wall of which is necessary to obtain records. To the formed request the server "VKontakte" send structured JSON which will contain a list of entries from the wall of the specified user. One of the parameters will contain JSON Response "text" which will contain the record on the wall itself and which we need to process further.

For "Odnoklassniki" such API Method is "stream.get" which returns the events tape "what's new" for the current user, it means the events of his friends, groups and so on. For events generated by user the function of stream getByAuthor is used.

Text preprocessing: Before moving to the next step, we need to break a sentence into words and to parse the words for this we use a parser called MaltParser.

MaltParser is a system for data-driven dependency parsing which can be used to induce a parsing model from tree data bank and to parse new data using an induced model (Nivre *et al*., 2007).

Parsing in linguistics and computer science is the process of comparison of the linear sequence of tokens (words, tokens) of the natural or formal language with its formal grammar. The result is usually a parse tree (syntax tree). It's usually used in conjunction with lexical analysis (Marmanis and Babenko, 2009).

Social networks analysis method on the base of support vector machine classifier: SVM-classifier implements a binary classification: divides the set of input vectors into two parts, positive-negative, us vs. them, yes-no, etc.

Before we classify the text, we need to present it in the form of a frequency dictionary and train on a set of specially selected training examples which consist of two parts the positive and negative examples. In this case, the approach of supervised training is needed where each training example is a pair of (a training input, the correct answer), then we need to make a vector of pairs in the format:

[label×index]:[value×index×value]

Where:

Label = A value
1 = A positive example
-1 = If the input example is negative

Linearly separable sample is the best one as the SVM-classifier as we have two disjoint classes where objects are described by n-dimensional real vectors:

$$X = R^n,\ Y = \{-1, +1\}$$

In this case, we shall construct a linear threshold classifier:

$$a(x) = \text{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \text{sign}\left((w, x) - w_0\right)$$

Where:

x = $x^1,\ldots, x^n$ is a feature descriptions of object x
vector w = $w^1,\ldots,w^n \in R$ and scalar threshold$w_0 \in R$ is a parameter of the algorithm

Equation (w, x) = w_0 describes a hyperplane that separates the classes in the space R^n. Suppose that a sample $X^l = (x_1, y_1)^l_{i=1}$ is linearly separable. Then, there are parameters w, w_0 in which the functional number of errors:

$$Q(w, w_0) = \sum_{i=1}^{l}\left[y_i\left((w, x_i) - w_0\right) < 0\right]$$

becomes zero. But, then the separating hyperplane is not unique. You can select more of its provisions that implement the same partition of the sample into two classes. The idea of the method is that a reasonable way to dispose of this freedom of choice (Fig. 1).

To make separating hyperplane as far distant from the sampling points as possible, the bandwidth must be maximized. Let x_- and x_+ two training object classes -1 and +1, respectively which are on the boundary of the strip. Then there is the bandwidth:

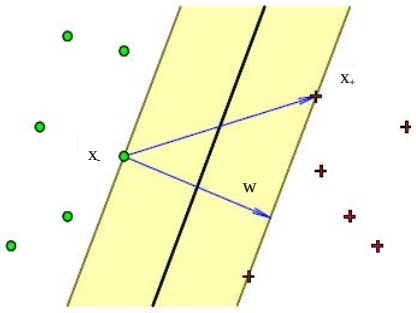

Fig. 1: A linearly separable sample. Educational facilities x_- and x_+ are on the border of the separating the strip. Normal vector w to the separating hyperplane specifies the width of the strip

$$\left[(x_+ - x_-), \frac{w}{\|w\|}\right] = \frac{(w, x_+)(w, x_-)}{\|w\|} = \frac{(w_0 + 1) - (w_0 + 1)}{\|w\|} = \frac{2}{\|w\|}$$

The bandwidth is maximized when the norm of the vector w is minimal. So in the case of linearly separable samples, we obtain a quadratic programming problem: it is required to find the values of the parameters w and w_0 at which the following l inequality constraints and the norm of the vector w is minimal:

$$\begin{cases}(w, w) \to \min; \\ y_i\left((w, x_i) - w_0\right) \geq 1,\ i = 1,\ldots,l\end{cases}$$

RESULTS AND DISCUSSION

Application development and results review: Figure 2 shows the architecture of the application which consists of four main components:

- Vkontakte server: provides access to the profiles of users of the social network "VKontakte"
- Odnoklassniki server: provides access to the users profiles of the social network "Odnoklassniki"
- Model training: SVM Trained Model
- Application module: A basic module which connects the GUI with the algorithmic module that communicates with the server "VKontakte" and "Odnoklassniki" to train the model to obtain the result

This module is divided into two classes: learning models and product testing. To the input of this program

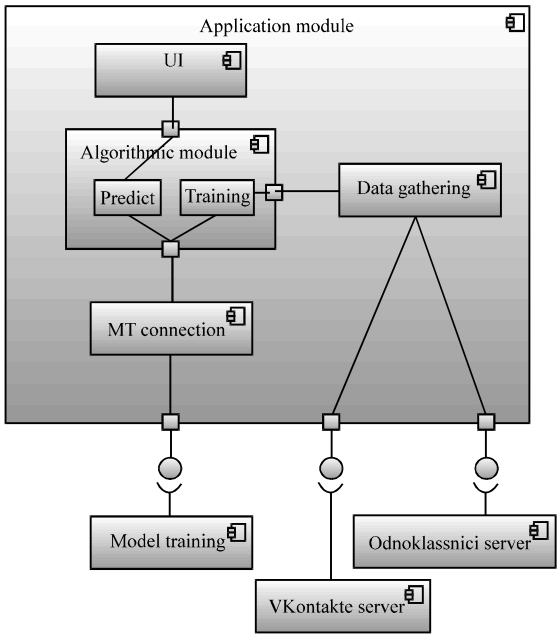

Fig. 2: Module of the application's architecture

Table 1: Examples of phrase

Phrases	Classification
Putin signed a contract and officially took Crimea and Sevastopol to Russia. Our friends! Crimea is RUSSIAN!	+1
While there are Crimean Tatars in Crimea, no Russian will be here. Deal with it!	-1

a supplied link to the page the user is given as defined in this example if a person supports the accession of the Crimea to Russia or not (Table 1).

As we use a supervised support vector machine, firstly, we need to train model which means to give vectors for the training to the classes which were described in this study in the part of the software on output, we get the training model. The partial example of the content model:

0.008636794682939264,0.0,0.0012400751754621
075,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00947192394764
4687,0.04555797088589852,0.043699057798659
485,0.05317963961751165 4,0.0,0.0,0.004089459
638719235,0.0,0.012401333061746971,1.277632
2930456576e-18,0.0005009098665680165,
0.0009649215908873201,-0.000878299512
1716355,1

After we trained the model, we can submit texts in forecasting in order to find out to which class this text belongs. We can create a file that contains records from the walls of users who live in Ukraine and who referred

Table 2: Belonging words to category

Sign	+1	-1
Against	0.15	0.60
Sign	0.44	0.08
And	1.10	0.91
Youth	0.25	0.24
Constitute	0.37	0.22
Spring	0.01	0.01
People	0.14	0.14

the tag "Crimea". Classifier stores all encountered signs as well as the likelihood that the feature is associated with a particular classification. Examples are presented to the classifier one by one. After each example, classifier updates its data by calculating the probability that the text of this category contains a particular word. After the training, we obtain a set of probabilities (Table 2).

Table 2 shows that after learning, associations with different categories of signs are getting stronger or weaker. The word "against" is more likely to "bad", the word "sign" is more likely to "good". Ambiguous signs such as the word "and" have similar probabilities for categories (the word "and" is found in almost any document, regardless of its subject matter). Unlike some other classification methods, there is no need to store the original data as training.

CONCLUSION

In this research, a linear support vector machine which was used for text classification, trained on data submitted from the social network "VKontakte" and tested through "Odnoklassniki". Was considered, so it can be concluded that although, social networks were different, classifier can determine the identity of the text to the class and it is not dependent on the type of social network. Researchers aimed to analyze the users' profiles in a social network "VKontakte" in Ukraine and Russia to develop a system that could determine the selected network user opinion about the annexation of Crimea by Russian Federation. To achieve this goal the methods of data collection, data processing, training and application of machine learning were used. Researchers considered an algorithm to determine an opinion on the annexation of the crimea to Russia. We developed a desktop application for using in practice.

REFERENCES

Alexander, A., Chumak, S.S. Ukustov and A.G. Kravets, 2014. Analysis of user profiles in social networks. Knowledge-based software engineering. Commun. Comput. Info. Sci., 466: 70-76.

Barnett, G.A., 2011. Encyclopedia of Social Networks. SAGE Publications, Inc., USA., Pages: 1112.

Chumak, A.A., S.S. Ukustov, A.G. Kravets and J.F. Voronin, 2013. Social networks message posting support module. World Applied Sci. J., 24: 191-195.

Marmanis, H. and D. Babenko, 2009. Mining Algorithms Internet: Best Practices in Data Collection, Analysis and Data Processing. Manning Publications Co., America.

Nivre, J., J. Hall, J. Nilsson, A. Chanev and G. Eryigit *et al*., 2007. MaltParser: A language-independent system for data-driven dependency parsing. Nat. Language Eng., 13: 95-135.

Quyen, L.X. and A.G. Kravets, 2014. Development of a protocol to ensure the safety of user data in social networks, based on the backes method. Knowledge-based software engineering. Commun. Comput. Inform. Sci., 466: 393-399.