The high-level overview of social media content search engine

A O Savelev[1], A Yu Karpova[2], D V Chaykovskiy[2], A D Vilnin[1], A Yu Kaida[1], S A Kuznetsov[1], L O Igumnov[1] and N G Maksimova[3]

[1]School of Computer Science & Robotics, Tomsk Polytechnic University, Tomsk, Russia
[2]School of Core Engineering Education, Tomsk Polytechnic University, Tomsk, Russia
[3]School of Engineering Entrepreneurship, Tomsk Polytechnic University, Tomsk, Russia
E-mail: ayk13@tpu.ru

Abstract. An increasing amount of social networks users-generated data is the most remarkable research challenge nowadays. Despite the progress in the field of semistructured data processing algorithms creation, even initial data collection could not be treated as issues that have been optimally solved. The paper covers a high-level overview of the automated social media content search system. The proposed structure enables to implement instruments for multisource content extraction tasks as well as supporting of identification processes of new patterns, which describe a certain type of content. Issues of Search engine organization, logically unified extracted data repository and possible content classification techniques with the appropriate knowledge base's application are considered. Under the work, existing approaches and automated web-data extraction methods have been analyzed; social media API's functions and limits, as well as ways of semistructured data storage system organization, have been studied. The planned result's application area is automation and informational support of sociological research based on the social media content analysis techniques namely a content propagation simulation in interconnected groups; social and personal anomy study; clarification of the weak linkage's strength concept.

1. Introduction

The modern sociology which has entered in the age of digitalization focuses on the understanding of how to operate with various digital devices and modern data aggregation and analysis approaches for the human behavior diagnosis in a context of social networking. Modern digital technologies enable sociologists to evaluate the level of digital content influence on the human being: views, perspectives, and beliefs [1]. The "Digital Sociology" is inextricably linked to exploring the Internet space whose scale and speed of growth as well as the level of diffusion into the daily life may cause a serious potential risk to spread the radical views and beliefs across the society [2]. It is important to note that the most vulnerable group for producing social tensions is the youth. The existing sociological approaches that are used by leading social science services are deprecated for the "digital age"[3]. The scientific community had started to pay attention to social networks analysis about ten years ago. This kind of analysis is considered as a multidisciplinary research area that gathers to aggregate, extend and adapt methods of social media content data analysis [4–6]. It was extended through the years, and now, in a context of the research to define the impact of the disruptive content driven to promote radical ideas, has evolved to a milestone where the social media has a great impact. Internet users leave loads of data about its activity surfing on web pages – a digital footprint [7]. The digital footprints analysis based on the aggregated preprocessed data from social networks may let to define a particular behavior of potentially dangerous groups and users that develop view related to extremism. To study the users' behavior in social networks, the set of particular tools created on the basis of artificial intelligence implementation models and methods is required. The core idea is to define the networks as dynamic social groups driven to make social tensions and to be able making a prognosis of the consequences of the actions within the Internet [8]. The solution can be implemented only as a result of the collaboration between sociologists and IT-specialists.

2. The general issue

Today, Internet users that produce disruptive content use modern digital technologies, as well as common users do. This fact makes a new challenge in front of social security services. Tracking the digital footprints of Internet users is still a pressing issue that requires a complex solution that can be used for social media data aggregation, analysis, and visualization. The proposed solution is a particular search engine that able to find a user or a group of users by a context of their posts and comment as well as analyze the activity on a particular page. The aim of the developing engine is to provide a set of tools for defining objects (users or groups) that are producing disruptive content in social media. The general schema of the engine is presented in Figure 1.

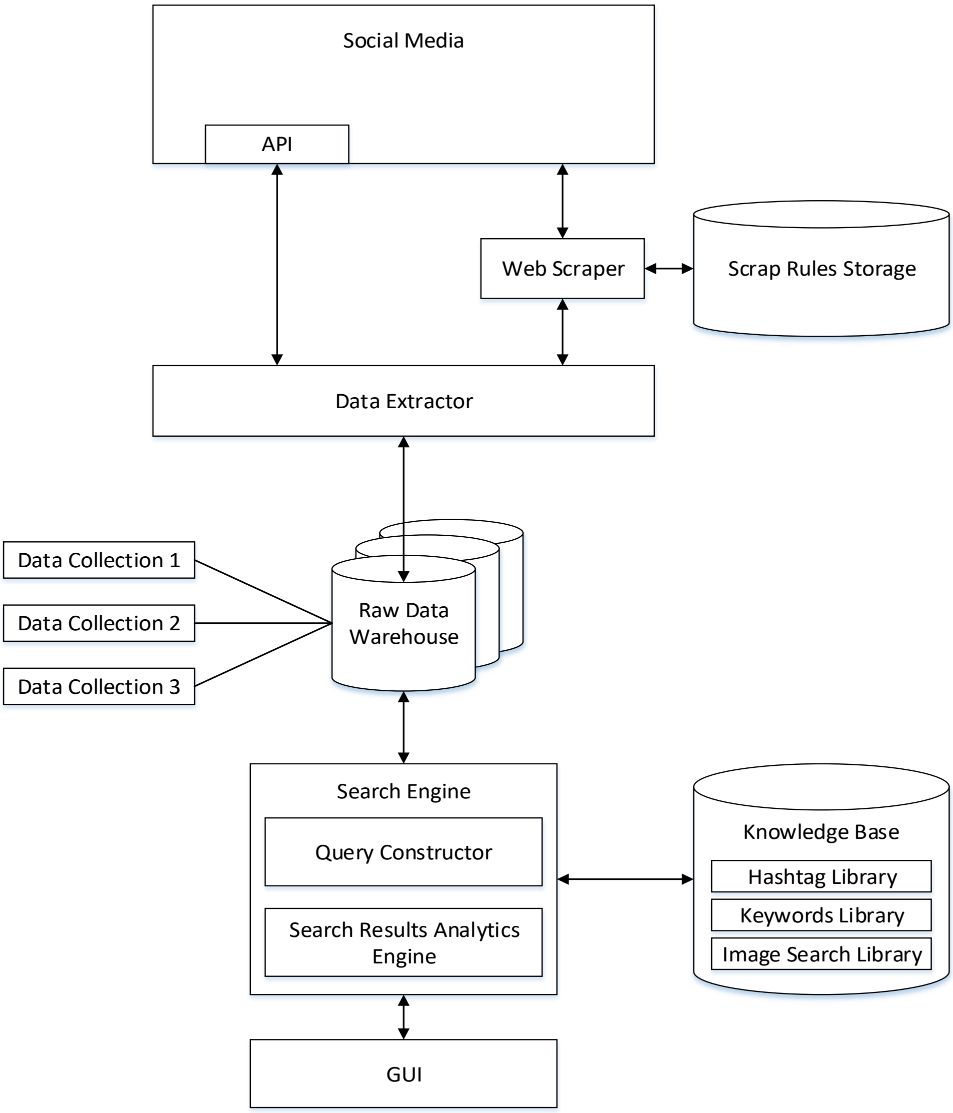

Figure 1. The high-level overview of the search engine.

3. Web scraping engine

Before among solutions for data acquiring from social networks, it is possible to distinguish such services as that could communicate on API (Application Programming Interface) and the rest that could not. The only way to get data from the last ones is a direct way also known as Web scraping.

API is an interface for interaction between a site and external applications including a complete set of classes, procedures, data structures, etc. which is needed for connection. With API an external application could get and use for its purposes a user's public data if user permission is provided, for instance, list of friends, photos, audio and video files.

VKontakte API allows acquiring data from vk.com database over HTTP requests. An applying of such interface avoids the necessity of detailed understanding about internal database structure namely, what tables and field types it consists of. Request's syntax and a type of the response data are strictly specified into the relevant technical documentation.

Despite the benefits of API, one limit still exists: a social network could not provide all details visible while site browsing. It depends on that at first social networks try keeping the privacy of its users and secondly some functions make a too high load for a database [9]. To surmount the limit the web scraping should be used.

Web scraper - an algorithm for collecting data from the network. A web crawler is the very first part of the data extraction process. The crawler is a web robot also known as an Internet bot or bot that systematically scans the World Wide Web with the aim of web spidering. Web crawlers are used by some sites and search engines to update their contents. The procedure of crawling starts with the list of URLs to crawl, these URLs are called seeds URLs. After visiting these seed URLs, crawler identifies all the links that exist on the page and store them into the list of URLs to visit called Crawl Frontier. Now URLs from the crawl frontier are recursively visited. This process goes on until all URLs get visited.

The data needed for analysis process are present on the Internet, this data contains a large amount of information and it is not possible for the analyst to manually take that data from any World Wide Web to perform data extraction. The website data is unstructured, which means it can contain noise or unwanted data. To extract large and unstructured data from the web there are some data extraction techniques and tools that can easily extract large unstructured data and convert them into a meaningful and structured format.

Web data extraction is basically a technique of extracting user-required information from the websites. The extraction process starts with indexing or crawling. It crawls data from the web using a web crawler. In the crawling process, it also transforms unstructured data on the web page into structured data. After crawling, the list of crawled links will be available to us. In these web pages, a lot of junk and useful data is mixed, So we have to extract the useful data and convert that data into needed format. The Data extractor extracts the needed information from the crawled web pages. To separate out the used one from the mixture we target only that part of the page in which information is present. To target that specific part, we will use the CSS Selectors [10].

4. Data warehouse

The aggregated data from social media – extracted using API or Web Scraping method – are related to unstructured or semi-structured data that causes a set of requirements for the data warehouse. The unstructured data is represented as a text that may include the other data fragments for future analysis. At the same time, semi-structured data is represented in some formats that do not require a particular schema (e.g. XML and JSON). Due to the uncertain complexity of analyzed text data, JSON format of files seems as the most suitable ones across serializable formats. Moreover, it is suitable for complex data structures processing.

The raw data warehouse is a key element of the system that keeps raw data aggregated from social media for future analysis. The search engine considers a flexible data model because users can fulfill their own account in social networks partially according to their own decision. Moreover, during the analysis, the number of attributes may be changed because of the results of the previous steps. It is worth noting, however, that the major problem of using database management systems is normalization large volumes of data. The normal form, in a relational data model, considers data representations with the minimal level of superfluous data; however, because of semi-structured form, this way of data representation is a time-consuming task with a high risk of data schema scalability failure or that a large number of empty cells occur. In this case, the normalization process also leads to the enormous growth of the total number of tables, so it is too difficult to define the intersections between analyzed objects (groups or users). NoSQL storages do not use a relational data model (so they may be schemaless), so it rids the requirement to store semi-structured data using relational way.

Today, a huge variety of NoSQL storages exist, so it is essential to define some additional requirements for the system. The developing system is defined as a distributed ones. According to the CAP theorem [11], all data storages can satisfy only two of three requirements at the same time: consistency, availability and partition tolerance. In general, the system consists of a number of standalone independent components where any system element is able to obtain the same set of relevant data – in other words, it is possible to get access to any node and obtain the data without any contradictions. MongoDB has been chosen as a raw data warehouse cause it ensures consistency and partition tolerance. The MongoDB architecture is able to provide distributed data storage using a sharding option. It allows decreasing the load level of a single server operating with the large volumes of data as well as it prevents exhausting CPU capacity. MongoDB storage data on a hard drive source in JSON-like files with a flexible schema. Moreover, MongoDB has a Python distribution containing tools, PyMongo, that makes a process of extraction and analysis faster and easier. PyMongo is able to extract data as JSON files, so then all data can be transfer easily to any other tool for future operations.

5. Knowledge base

The search engine checks the raw data using a knowledge base that contains predefined sets of keywords, hashtags, and images. The main task of the knowledge base is content classification, for the implementation of which statistical correlations are used to build rules for placing objects in specific content classes. The classification issue is the task of recognition, where, for a control sample, the system assigns a new object to a particular category.

Methods based on machine learning to create a training set and build a classifier model, as a rule, require the formation of a dictionary with marked keywords. The tagged (marked) keywords are considered a positive example. The remaining words are considered a negative sample. Next, the relevance of each word of the training text is calculated by comparing the vector of values of various parameters, for example, TF-IDF measures, word lengths, parts of speech, word positions in the headline, word positions in the first paragraph, last paragraph, in literature lists. Further, the difference between the values of the vectors of these parameters for keywords and not important ones (not keywords) is recorded. The next step is to calculate the probability of assigning each word to a group of keywords and setting its threshold, thus learning the model. The extraction of keywords from the new document occurs by calculating the relevance of the words and their probability of attributing to the key following the constructed model [12].

Among the various classification methods based on machine learning, researchers distinguish the naive Bayes classifier and the support vectors as the least complex in their implementation and at the same time the most effective in solving many practical problems.

Naive Bayes classifier

A Bayes classifier is a method that assigns class labels to observations represented by feature vectors. Each attribute independently affects the probability that the observation belongs to the class. For example, an object can be considered an apple if it has a round shape, red color, and a diameter of about 10 cm. The naive Bayes classifier “believes” that each of these attributes independently affects on the probability that this object is an apple, regardless of any possible correlations between the characteristics of color, shape, and size [13].

A Naive Bayes classifier based on a supervised learning method. An additional advantage is the small number of examples required for training.

Support-vector machine

The support vector method is a set of supervised learning algorithms based on the translation of the original vectors of parameter values into a higher dimension space and the search for a separating hyperplane with a maximum gap in this space. This approach involves obtaining some rules that allow distinguishing objects of two classes.

The method of support vectors belongs to the family of linear classifiers used for the problems of classification and regression analysis, one of the main properties of which is the continuous decrease in the empirical classification error. When the method works, the summation is performed not over the entire sample, but only over the reference vectors. This property of sparsity that distinguishes the method of support vectors from other linear delimiters — the Fisher discriminant, logistic regression, and the single-layer perceptron [14].

6. Summary

«Digital sociology» specializing in people’s behavior diagnosis questions, informational content influence to opinions, views, and beliefs forming based on the information presented on the Internet and particularly in social media has completely developed as a scientific direction. Research tasks’ complexity also depends on the availability of automation tools for big amount processing of initial information. The main task considered in the paper is an adaptation of the smart content analysis methods to meet the diagnosis of Informational influence in social media challenges. The automated social media content search system high-level overview has been developed. The proposed overview has several major advantages as follows:

- On the one hand a hybrid approach to the initial data extraction implementing by the combined use of social networks’ API and web-scraping allows to organize a unified data collection from a number of heterogeneous information sources, on the other hand the system sustainability against changes due to API functionality reconsideration by the social media owners or HTTP structure updating is increasing.
- Knowledgebase availability formalizing attribute space for content classification in addition to supportive mechanisms development for the identification of the new attributes on the search module’s results enables to identify semantic communication among groups and communities while absence from cross-references as well as new communities, categories, and users’ interaction language timely tracking.

The presented results are interim ones and have been achieved with a project devoted to the creation of the software tools and new content promotion communicative technologies’ examination method.

Acknowledgments

The reported study was funded by RFBR according to the research project № 19-011-31535.

References

[1] Holt T, Freilich J, Chermak S 2018 *American Journal of Criminal Justice* **44** 83–105
[2] Porta D 2018 *Annual Review of Political Science* **21** 461–474
[3] Ali K, Dong H, Bouguettaya A 2017 *Proceeding of the IEEE International Conference on Web Services (ICWS)* 660–667
[4] Zeng D, Chen H, Lusch R 2010 *IEEE Intelligent Systems* **25** 13–16
[5] Chen X, Madhavan K and Vorvoreanu M 2013 *Proceeding of the International Conference on Cloud and Green Computing* 383–388
[6] Jensen A, Seate A and James P 2018 *Terrorism and Political Violence* 1–24
[7] Gao C and Liu J 2017 *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **47** 171–183
[8] Chung W, Mustaine E and Zeng D 2017 *Proceeding of the IEEE International Conference on Intelligence and Security Informatics (ISI)* 191–193

[9] Sukhanov A and Maratkanov A 2017 *Proceeding of the International Scientific Review of the Problems and Prospects of Modern Science and Education: International Scientific and Practical Conference* 22–25

[10] Parvez M, Tasneem K, Rajendra S *et al* 2018 *Proceeding of the International Conference on Smart City and Emerging Technology (ICSCET)* 1–7

[11] Gilbert S and Lynch N 2002 *ACM SIGACT News* **33** 51–59

[12] Sheremetyeva S, Osminin P and Kokhanova L 2015 *Bulletin of the South Ural State University* **12** 76–81

[13] Kotelnikov E and Klekovkina M 2012 S *Computational Linguistics and Intellectual Technologies* 27–37

[14] Volik A and Murlin A 2014 *Scientific works of KUBSTU* 63–68