Intelligent Instrumentation for Opinion Mining in Social Media

N Yarushkina[1], A Filippov[1], V Moshkin[1], G Guskov[1] and A Romanov[1]

[1] Ulyanovsk State Technical University, Ulyanovsk, Russia

Abstract. The paper presents a developed intelligent tool for Opinion Mining of social media. In addition, the article presents new algorithms to the hybridization of ontological analysis and methods of knowledge engineering with methods of nature language processing (NLP) for extracting the semantic and emotional component of semi-structured and unstructured text resources. These approaches will improve the efficiency of the analysis of social media content-specific data and fuzziness of natural language.

1. Introduction

Active growth of social media audience on the Internet (social networks, forums, blogs and online media) made them a new source of data and knowledge. The specifics of working with social media has several advantages and disadvantages.

Advantages include:

- high speed of access to information;
- a broad audience;
- a wide range of data topics;
- large amount of data.

The disadvantages are:

- large amount of data;
- unstructured presentation of information;
- absence of a single conceptual framework.

A large amount of social media data is both an advantage and a disadvantage at the same time. Monthly in Russian social networks about 30 million unique authors publish 580 billion messages according to statistics for 2017.

However, a large amount of data makes it possible to obtain a large training sets, for machine learning methods and a large statistical sample for social studies.

The monthly billions of unstructured text messages and publications that users leave monthly cannot be processed manually.

There is a need for methods of automated intelligent and sentimental analysis of text data. These methods handle large amounts of data and understand their meaning (Text Mining), determine the sentiment (Opinion Mining) of user messages and publications in a short time [1-5].

Understanding the meaning and sentiment of publications in social media is the most important and complex element of automated text processing [6-11].

Our scientific group has created an intelligent tool for Opinion Mining of social media. This tool includes new approaches to the hybridization of ontological analysis and methods of knowledge engi-

neering with methods of nature language processing (NLP) for extracting the semantic and emotional component of semi-structured and unstructured text resources [12-16].

These approaches will improve the efficiency of the analysis of social media content-specific data and fuzziness of natural language.

2. The architecture of the software system for Opinion Mining social media

Service-oriented approach is the basis of the architecture of the software system for Opinion Mining Social Media (SOM). This approach allows:

1. To increase the overall fault tolerance of the SOM by performing services in different address spaces.
2. To increase the scalability of the SOM by running several instances of services and balancing the load between them.
3. To provide the ability to use different operating systems, programming languages, storage technologies, etc.
4. To reduce the downtime of SOM when making changes, correcting errors, etc.
5. To provide an opportunity to completely replace services while maintaining the interface of interaction with other parts of the SOM.

REST in conjunction with the HTTP protocol [0] is the basis for the organization of the interface for the interaction of SOM services. REST allows a distributed system of any type to have the following properties: performance, extensibility, simplicity, updatability, intelligibility, portability and reliability.

The architecture of SOM is shown in Figure 1.

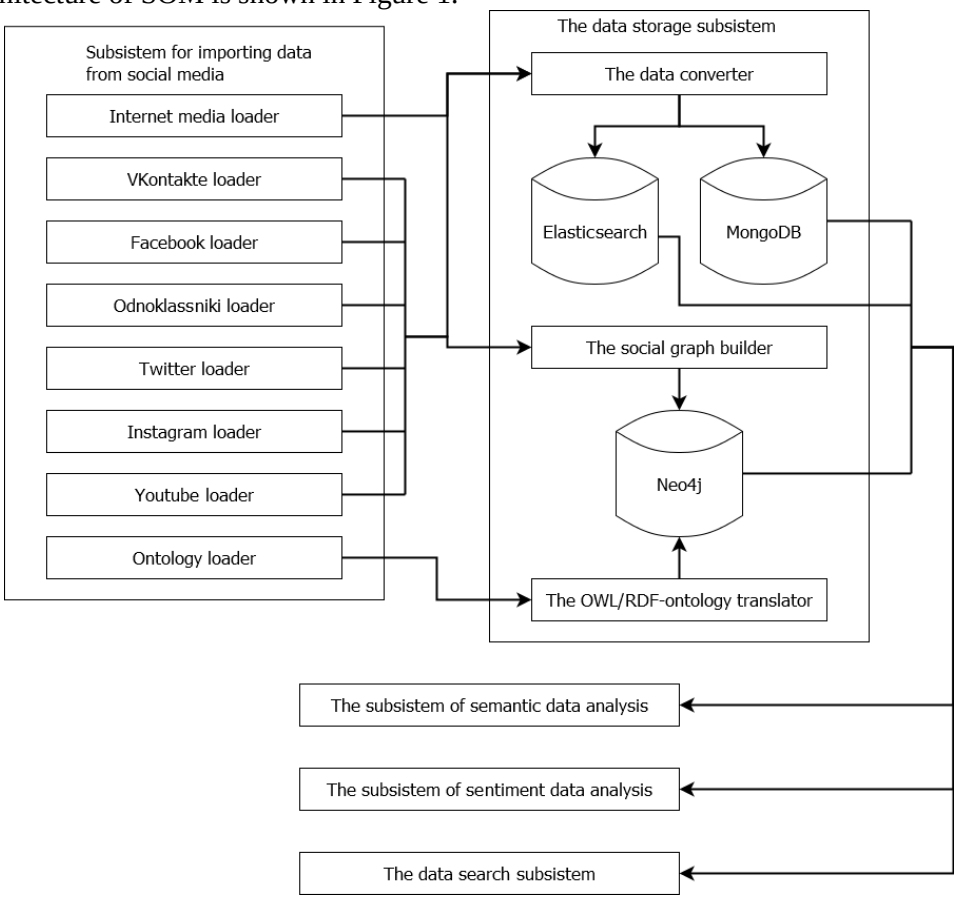

Figure 1. Architectural diagram of the software system for Opinion Mining social media.

The SOM consists of the following subsystems:

1. Subsystem for importing data from social media. This subsystem works with popular Internet services (Vkontakte, Facebook, Odnoklassniki, Twitter, Instagram, Youtube) through the public application programming interface (Public API). The data loader from the Intranet media retrieves data from HTML pages based on rules. You need to create your own rule for each Internet media. The rule should consist of a set of CSS-selectors. The ontology loader loads into the storage subsystem a description of the features of the problem area (PrA) in the form of ontologies in the language RDF or OWL.

2. The data storage subsystem provides the representation of information extracted from social media in a unified structure that is convenient for further processing. The data is stored in the context of users, collections, data sources, versions, etc. As database management systems (DBMS) are used:

- Elasticsearch for indexing and retrieving data [0];
- MongoDB for storing data in JSON format [0];
- Neo4j for storing graphs of social interaction (social graph) and ontology [0].

The data converter converts the data imported from social media into an internal SOM submission. The social graph builder constructs a social graph. The social graph based on the relationship of users and social media communities. The translator OWL/RDF-ontology in the graph translates the ontology into the graph knowledge base [0].

3. The subsystem of semantic data analysis performs preprocessing of text resources. In addition, this subsystem performs statistical and linguistic analysis of text resources.

4. The subsystem of sentimental data analysis determines the attitude of a speaker, writer, or other subject with respect to some topic or emotional reaction to a document, interaction, or event from text.

5. The data search subsystem searches for objects related to a specific task. The task presented in the form of a set of keywords. In this case, the user's query can be extended semantically using an ontology. Ontology contains descriptions of features of the PrA.

2.1. The graph knowledge base and a social graph as data models of SOM

The SOM storage subsystem stores the following kinds of data:

- data extracted from social media;
- description of PrA in the form of a graphical knowledge base;
- social graph that reflects the users and their connections of in social media.

The graph DBMS Neo4j used to store the description of the PrA in the form of a graph knowledge base and a social graph. The main advantages of Neo4j are:

1. Native storage format for graphs.
2. One copy of the DBMS can control graphs with billions of nodes and links.
3. Neo4j can control graphs that do not completely fit into RAM.
4. Graph-oriented query language - Cypher.

The search engine Elasticsearch used to organize data retrieval. The main advantages of Elasticsearch are:

1. Elasticsearch can process petabytes of structured and unstructured data.
2. Using denormalization to increase the search efficiency.
3. Elasticsearch is one of the most popular search engines that is currently used by many large organizations and services such as Wikipedia, The Guardian, StackOverflow, GitHub, etc.

Document-oriented DBMS MongoDB is used to store data extracted from social media. The main advantages of MongoDB are:

1. High performance.
2. Document-oriented query language.
3. Fault tolerance.
4. Scaling.

2.2. Description the main concepts of the Social Media and their relations in knowledge base

The main SOM data model concepts are:

Mass media concept stores information about different social media (VKontakte, Facebook, Twitter, etc.) or news site. The SOM import subsystem downloads data from these social media using their API and from news site by using set of CSS-selectors.

The Person concept is a list of users extracted from social media.

The Person concept has a set of attributes often used in social networks: surname, first name, date of birth, hobbies, education, etc.

The Group concept stores information about communities extracted from social media. The Group concept has a set of attributes often used in social networks: group name, group description, age restrictions, creation date etc.

The Post concept stores information about records in social media. The Post concept has the following attributes: author, title, content, creation date, attachments etc.

The Comment concept stores information about comments in social media. The Comment concept has the following attributes: author, title, content, creation date, attachments etc.

The Attachment concept stores information about the attachments of entries and comments in social media. The Attachment concept has several types and allows you to store the following types of attachments: photos, photo albums, audio, video, links, documents (files), surveys etc. Table 1 shows the correspondence of the social media concepts and SOM concepts.

Table 1. The correspondence of the social media concepts and SOM concepts.

SOM	VKontakte, Facebook, ok.ru	Twitter	Instagram	Youtube	Social media
MassMedia	URL, For example, vk.com	URL	URL	URL	URL
Person	User	User	User	User	-
Group	Group	-	-	-	-
Post	Post	Twit	Photo	Video	News, Article
Comment	Comment	Comment	Comment	Comment	Comment
Attachment	Attachment	Attach-ment	tags, links	Link	Attach-ment

The main concepts of the SOM data model allow storing data downloaded from most existing social media. Unified presentation of SOM data allows efficient processing, analysis and search. The data converter is used to transform data downloaded from social media into the internal representation of the SOM. It is necessary to develop a data converter module for each new Internet resource. The Internet media loader generates the same data representation for all sites. Therefore, the converter for each site separately is not necessary to adapt.

3. Conclusion

Intelligent tool for Opinion Mining social media developed by our research group will allow you to download data from the social network VKontakte and Internet media.

The social graph is formed during the download of data from the social network VKontakte. This social graph contains the following types of relationships: is a friend, is a subscriber, is a relative, is in a relationship, is in the community. The statistical index of text data is formed when data is loaded using the search engine Elasticsearch. The data is converted into the SOM data model concept and stored in MongoDB.

The data search subsystem searches for data by keywords in the context of data sources and concept types: users, communities, entries, comments and attachments. The user's initial search query can be extended during the search based on the graphical knowledge base.

The graph knowledge base is formed during the translation of the ontology in the OWL format into nodes and the relationship of the graph knowledge base.

Further development of the SOM consists of:

1. Development of downloaders for social networks Twitter, Facebook, Instagram, Youtube, ok.ru.
2. Testing the storage subsystem on large amounts of data.
3. Development of a subsystem of sentimental data analysis.
4. Development of a subsystem of semantic data analysis.
5. Finalization of the user interface.

The resulting SOM should improve the effectiveness of analyzing the content of social media taking into account the specifics of data representation and the fuzziness of natural language.

4. References

[1] Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. pp 631-636 (2006).

[2] Gjoka M. et al. Practical recommendations on crawling online social networks //Selected Areas in Communications, IEEE Journal on. Vol. 29. №. 9. pp 1872-1892 (2011).

[3] Boyd D., Ellison N. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication. Vol. 13(1). article 11. (2007)

[4] Pallis G., Zeinalipour-Yazti D., Dikaiakos M.. Online Social Networks: Status and Trends. New Directions in Web Data Management 1, Studies in Computational Intelligence Volume 331, pp 213-234 (2011).

[5] Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner2012-emerging-technologies-hype-cycle-2, last accessed 2018/05/11.

[6] Korshunov A. Tasks and methods for determining the attributes of users of social networks // Proceedings of the 15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" - RCDL'2013

[7] Korshunov A., Beloborodov I., Gomzin A., Chuprina K., Astrakhantsev N., Nedumov J., Turdakov D. Determination of demographic attributes of users of microblogging // Proceedings of the Institute of System Programming of RAS. Vol. 25, 2013 DOI : 10.15514 / ISPRAS-2013-25-10.

[8] Fleuret F. Fast Binary Feature Selection with Conditional Mutual Information // JMLR, 5:1531–1555 (2004).

[9] Crammer K., Dekel O., Keshet J., Shalev-Shwartz S., Singer Y. Online Passive-Aggressive Algorithms // JMLR, 7(Mar): pp 551–585 (2006).

[10] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques. pp 79–86 (2002).

[11] Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the Association for Computational Linguistics. pp 417–424. arXiv: LG/0212032 (2002)

[12] Chetviorkin I., Loukachevitch N. Sentiment Analysis Track at ROMIP-2012. Computer linguistics and intellectual technologies. Computer linguistics and intellectual technologies: Dialogue-2013. Sat. scientific articles volume 2, pp 40-50.

[13] Antonova A., Soloviev A., Using the method of conditional random fields for processing texts in Russian. Computer linguistics and intellectual technologies: Dialogue-2013. Sat. scientific articles / Issue. 12 (19)- Moscow: Publishing house of the RSUH. pp 27-44 (2013).

[14] Pazelskaya A., Soloviev A. Method of definition of emotions in texts in Russian. Computer linguistics and intellectual technologies. Computer linguistics and intellectual technologies: Dialogue-2011. Sat. scientific articles / Issue. 11 (18). Moscow: Publishing House of the RSUH. pp 510-523 (2011).

[15] García-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R.: Retrieving product features and opinions from customer reviews. IEEE Intelligent Systems 28(3), pp 19–27 (2013)

[16] Tarasov D. Deep Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference “Dialogue-2015”. Issue 14(21), Vol.2, pp 65-74 (2015).

[17] Representational state transfer, https://en.wikipedia.org/wiki/ Representational_state_transfer, last accessed 2018/05/11.
[18] The Heart of the Elastic Stack, https://www.elastic.co/products/elasticsearch, last accessed 2018/05/11.
[19] MongoDB. For Giant ideas, https://www.mongodb.com, last accessed 2018/05/11.
[20] Introducing the Neo4j Graph Platform, https://neo4j.com, last accessed 2018/05/11.
[21] Yarushkina N., Filippov A., Moshkin V. Development of the unified technological platform for constructing the domain knowledge base through the context analysis. Communications in Computer and Information Science. 2017. Vol. 754. pp 62-72.

Acknowledgments
This study was supported Ministry of Education and Science of Russia in framework of project № 2.4760.2017/8.9 and by the Russian Foundation for Basic Research (Grants No. 18-47-730035 and 18-37-00450).