Conference Paper

Application of Formal Grammar in Text Mining and Construction of an Ontology

Cunningham, S. and Kanev, A.

This is a paper presented at the th IEEE Int. Conference on Internet Technologies and Applications ITA-17, Wrexham, UK, 12-15 September 2017

Application of Formal Grammar in Text Mining and Construction of an Ontology

Anton Kanev
Department of Information and Control Systems
Bauman Moscow State Technical University
Moscow, Russia
tony.longpoint@gmail.com

Stuart Cunningham
School of Applied Science, Computing and Engineering
Glyndŵr University
Wrexham, UK
s.cunningham@glyndwr.ac.uk

Terekhov Valery
Department of Information and Control Systems
Bauman Moscow State Technical University
Moscow, Russia
terekchow@bmstu.ru

***Abstract*—This work describes an investigation of formal grammar with application to text mining. It is an important area since text is the most widespread type of data and it contains a lot of potentially useful information. Unstructured nature of text requires other methods for its processing, in contrast to other types of data mining. In this work, the authors propose an original approach to text mining by making a parse tree for each sentence using regular grammar and creating an ontology and provide a demonstration of this system being implemented in a constrained scenario. This ontology can be used for different tasks, ranging from expert systems to automatic machine translation. The ontology is a network consisting of concepts linked by relations. The authors developed a new system to implement proposed approach working in different languages.**

Keywords—text mining; text data mining; natural language processing; formal grammar

I. INTRODUCTION

Text mining is a method for the detection of new, non-trivial and potentially useful knowledge [1, 2, 3]. Text mining has some common features with, and uses many techniques from the conventional data mining [4]. Both areas closely overlap with machine learning [5, 6]. But, according to much of the literature surveyed in the domain, text mining is dissimilar to the rest of data mining because of the unstructured nature of the text [1, 4, 7, 8, 9]. It is explained in [3] that in fact text has a structure, but that it is very complicated, making the process of analysis non-trivial. Another reason for this opposition is that 80% of all information is textual [1, 3, 10, 11].

Zhang, Chen and Liu [1] provide a holistic view of the general directions that are available in text mining. The first work in this area has appeared in the 1950s, during the 20th century, and since then has constantly evolved. In their work, they go on to describe several methods for the processing of the text. This was known as the Knowledge Discovery in Text (KDT), a general framework, document warehouse and concept-based mining model. Also in this seminal article they list four main tasks for text mining: categorization, clustering, association rules, the selection and definition of the trend.

Although the direction of text mining is clearly well established [1] there is still on-going controversy, some authors, for example, propose new definitions, concepts and directions. Thus, in [3], the author assumes that almost all modern research activities in text mining are more similar to conventional data mining, using its methods, and are not intended to identify truly new knowledge. They simply translate the text into an intermediate form, because this form is more convenient to work with text as with conventional data. But the text has a much more complex patterns that are lost. Therefore, Sanchez [3] proposes to translate the text into an intermediate form of knowledge, and then to search knowledge, using this form. He intends to use it for deductive and abductive (deductive with preconditions) methods instead of the inductive method used for normal text mining and text data mining. Also he pays attention to the need for background knowledge and creating a proper, correct knowledge.

The authors of the current paper applied the text mining method for construction of the ontology that represents the knowledge consisting in the text. Such approach allows to analyze relations between objects that were mentioned in the text. To evaluate this approach the created ontology was used for the extraction of the objects locations.

A. Text mining applications

As recognized earlier, one of the unique challenges in mining textual data comes from the structural nature of the characters, words, punctuation, and grammar that occurs in text. Since it is necessary to first deal with this challenge, at a high level, the majority of approaches to text mining require breaking the task into several sub-tasks. These tasks predominantly require that the text be first processed to obtain

978-1-5090-4815-1/17/$31.00 ©2017 IEEE

the underlying linguistic structures and sequences, the 'intermediate form' mentioned earlier, and then this organized information can be analyzed using statistical methods, machine learning, and pattern recognition [12].

Areas that use text mining are very broad: evaluating financial risk [13] and insurance [13], as well as medicine [7, 15, 16] and biology [17], education [18] and the analysis of web documents [4] in view of the large number of terms and volumes of newly available information. A 2009 study, for example, [9] gives an overview of a large number of works on text mining carried out in the field of bioinformatics in different directions. Recently, there has been considerable interest in the use of text mining in social media platforms, since much of the information that users post features text-based comments, and these have spanned a variety of application scenarios, ranging from marketing and consumer research to management of public health crises [19, 20, 21, 22]. A particularly interesting aspect relating to text mining, which is worth noting for the purposes of the work discussed in this paper, is that social media posts are often tagged with the user's location, adding a contextual dimension to any information that may be being posted.

B. *Types of text mining*

Contemporary approaches to text mining can be divided into two specific areas [6]: bag-of-words and natural language processing (NLP). The first term includes statistical and machine learning methods. It is appropriate for clustering and categorization. But is not good for question answering and semantic search. To do this, NLP is more suitable. But it is believed that NLP is not a part of text mining. Other work on the subject [9] explains that text mining only uses NLP to perform specific tasks, and that NLP is understood as revealing the value of all the text. In this approach, several important methods are used: parse trees and rule-based approaches, regular expressions [6, 15].

Many articles on the bag-of-words method [1, 4, 7, 18, 23] show that an integral part of the algorithm is the processing of stop-words. In Amarasinghe, Manic and Hruska [23] this stage was given special attention. They emphasized that the removal of the words leads to the loss of some useful information. Therefore, it was proposed that an alternate method, in which the stop-words are considered separately from the key words and the dimension is reduced using a genetic algorithm, be used instead. In [23] experiments were carried out that showed that the accuracy of the algorithm increased by two percent. However, their experiments have been conducted on a fairly small amount of data and there are questions as whether or not the proposed method is effective and, most importantly, can it quickly reduce the dimension of stop-words with a large amount of data?

Another important step in this direction is stemming [18, 23]. It is considered as combining synonyms of the concepts, as well as combining different forms of words as one single value.

Many works apply the Term-Frequency and Inverse Document Frequency (TF-IDF) method [4, 7, 10, 23, 24]. It is based on the calculation of the similarity of two documents on the basis of frequency of matching terms occurring in the two documents. It also takes into account the frequency of the terms in all documents. Simplistically this feature resembles the covariance of two random variables. The frequency of key words appearances in the text is also analyzed in [18].

II. METHODOLOGY

After analysis of the literature, described in the previous section, the authors decided to use NLP for construction of our ontology as this type of text mining allows for the extraction of greater information from text. Our approach follows several key phases:

A. *Sentence parsing*

The first step of the algorithm is the parsing of sentences. We suggest using regular grammars for processing the sentences and making parse trees for them. These rules are not complicated; they stored in the database associated with the implemented system and can be changed during runtime. This feature allows better functions for development, testing and maintenance.

B. *Construction of ontology*

The second step of the proposed approach is the mining of knowledge from the parse trees. The authors intend to use intermediate forms of data in relations and entities to collect this useful data and preprocess it to ensure it is ready for further analysis. Several types of relations are suggested for this purpose, the most popular ones are generalization and aggregation. Generalization is «a kind of» relation; aggregation is «a part of» relation. Using these, and more complicated types of relations, it is possible to save the knowledge into a database and reuse it as part of the next step.

C. *Location extraction*

Several times per day the developed application scans text news looking for information about objects that are identified in the location table of the system database. It separately analyses each occurrence of this text and puts data about the object's popularity, position, activation time and deactivation time into database (Figure 1). The knowledge is mined using regular expressions. Each sentence is parsed using parallel processing and its structure is used to get relationships between terms. All rules are combined into sentences: this is the most complex tier. Each tier, except leaf one, consists of more primitive rules. The application reads sentences word by word from the beginning and chooses rules that are the most appropriate. If several rules could be used on the current step then the parse tree is divided into several trees, which are processed separately. The process finishes when the tree organizes the sentence and all of the words in sentences have been used. In addition to rules the program checks the forms of verbs and nouns. It is necessary to distinguish similar words with different meanings, for example, "people" and "peoples". At the end of this step the program puts parse trees into the collection according to their order in text.

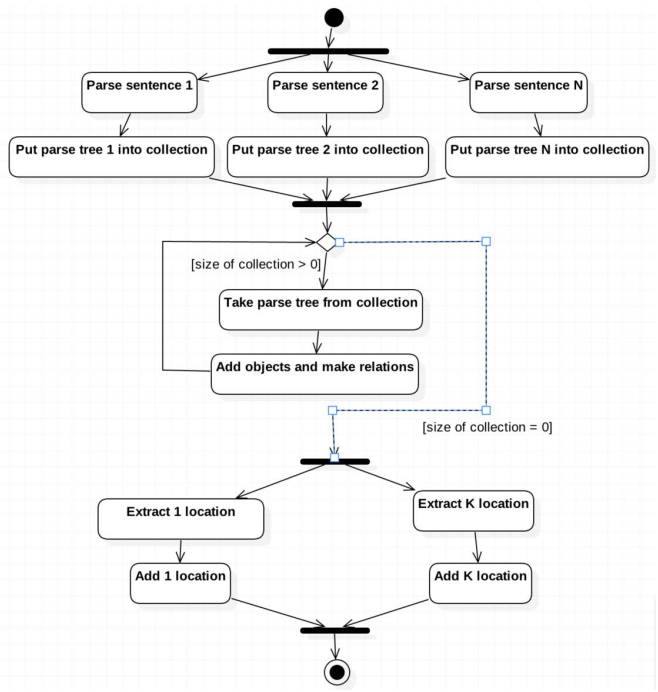


Figure 1. Activity diagram of text mining modules

After constructing parse trees the application begins to mine knowledge. Pronouns are replaced with concepts, which were mentioned earlier in the analyzed text, according to the form and meaning of the word in question. Using the structure of the tree it adds entities (objects) for subject and object of sentence into the database and makes relationships between them. The system uses two main relations: generalization, when one entity is a kind of another one, and aggregation, when one entity is a part of another one. In this step, parse trees are processed sequentially because knowledge from previous sentences may be very important for current one. Nevertheless, the program calls several recursive functions for making relations and searching objects, thus, those functions could be processed in parallel to provide scalability of the whole system.

The last step is an extraction of locations from the acquired knowledge. The system searches for entities, which are connected with locations in the corresponding table. When the program finds the entity in sentence it creates new locations with the same coordinates, but with a new name, and its own period of time. This step is also performed in parallel.

III. Experimental Investigation of the Proposed System

The developed system consists of PostgreSQL database and Java web-service. The database stores rules of the formal grammar and the information about objects locations. The web-service analyses texts and mines data about locations from them.

A. Experiments with parse trees

Experiments were undertaken with an implementation of the described text mining system. This application uses regular grammar to parse sentences and get the structure of these sentences. Every word is an element in the collection of words. Each word has reference to it in the Terminal Index collection and key of terminal element for this word. Several words with the same spelling have references in one Terminal Index element: it simplifies the process of searching terminals and parsing. There are several samples of words and keys of terminals (word - key).

```
"tree" - "noun"
"were" - "to be"
"go" - "to go"
"a" - "article"
```

All terminals have references in the non-terminal collection. This collection keeps lists and orders of terminals or non-terminals in each non-terminal. The final level of this hierarchy is a sentence and this is also the end of the parsing process. Here are two non-terminals and two rules for each element. The first expression means that *An* (adjective noun) could be achieved with a sequence of adjectives plus a noun. The second expression means that *An* or article plus *An* could be combined into a *Nf* (noun fragment):

```
An = noun | adjective+An (1)
Nf = An | article+An (2)
```

What follows is an example of a preliminary study. One sentence "*A cat is in garden.*" was parsed using common regular rules.

During the first step the algorithm deduces that "*a*" is an article and it is a part of *Nf* (noun fragment):

```
Nf - article – a
```

Next the program began to analyzes the second word "cat":

```
- Nf - article - a
- An - noun – cat
```

Than it combines two words into one element (rule 2):

```
- Nf - article - a
    |- An - noun – cat
```

At the next steps the program added the verb "is" and preposition "in" using the same approach:

```
- Sentence - Cnf - Nf - article - a
                     |- An - noun - cat
           |- to be – is
- Sentence - Cnf - Nf - article - a
                     |- An - noun - cat
           |- to be - is
           |- in – in
```

Finally, it combines everything into one sentence with a clear structure, which can be used in following parse process:

```
- Sentence - Cnf - Nf - article - a
                     |- An - noun - cat
           |- to be - is
           |- in - in
           |- Cnf - Nf - An - noun – garden
```

Another example is the following:

```
- Sentence - Cnf - Nf - article - the
                     |- An - noun - dog
           |- to be - is
           |- in - in
           |- Cnf - Nf - An - noun – garden
```

B. An Experiment with the Russian language

This approach is scalable and generalizable and, as such, could be used for different languages and purposes. For

example, the full name of a firm in Russian was parsed into the business entity and name:

```
Публичное Акционерное Общество
"Автоматика."
  - Наим - ОПФ - прил - Публичное
                |- прил - Акционерное
                |- сущ - Общество
          |- назв - "Автоматика"
```

"Публичное Акционерное общество" means private company limited by shares, "Автоматика" is the name of the company, "ОПФ" means business entity, "прил" and "сущ" are reductions for adjective and noun respectively.

C. *Extracting Locations from Text*

Finally, the authors provide several exemplary experiments with target type of sentences that contain information about locations of objects and their connection to past or present tenses. To analyse it the authors used following text:

```
Guernica came to Hermitage yesterday. The
Sistine Madonna arrived at Louvre on Tuesday.
```

These sentences mean that some objects moved into a new place at a particular moment of time. The system subsequently parsed them and constructed following parse trees:

```
- Sentence - proper noun - Guernica
           |- move verb - to come - came
           |- to - to
           |- place - Hermitage
           |- time – yesterday
```

for the first one;

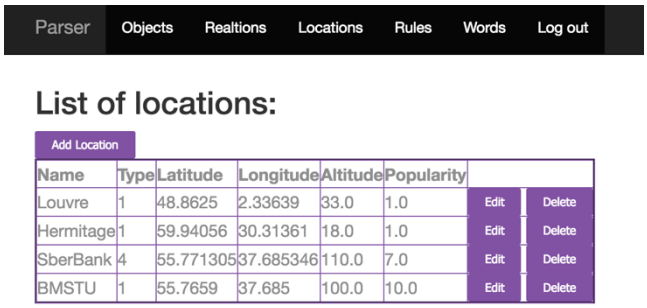
```
- Sentence - Nf - article - the
           |- proper noun - proper noun - Sistine
                          |- proper noun - Madonna
           |- move verb - to arrive - arrived
           |- at - at
           |- place - Louvre
           |- time - on - on
                |- time – Tuesday
```

for the second one.

The system successfully extracted the structure of the sentences. It determined what the subject is, where it should be, and when. The parse trees can be used to connect the location of a place and location of a subject and set a time period for it. These example experiments have positive results, indicating that this text mining approach is appropriate for future systems and worth of deeper investigation.

To mine knowledge from parse trees the system uses several types of relations between objects or entities. The most important among these are generalization and aggregation. As explained, generalization means that one object is a kind of another. For example, in the sentence "*All cats are mammals*" subject "*cats*" is an instance of the object "*mammals*" and it is connected with the object by a generalization relationship. Aggregation means that the object is a part the other one. For example, in the sentence "*The pyramid of Cheops is placed in Egypt*" the subject is part of an object and it is connected with the object using an aggregation relationship. Other, more complicated, relations are similar to these pair and they use the same approach. For example, using another relation type between two sequential states of one object and two aggregation relations between these states and their locations the system represents relocation of this object.

Relations are used for extracting locations. The system uses two operations for this stage. It creates a new location when it finds a new object that is connected with an existing location by aggregate relation. It deactivates locations by editing its end date when it finds a relation that links this location with new place. In the following example (Figure 2) the first two rows represent museums.



Name	Type	Latitude	Longitude	Altitude	Popularity		
Louvre	1	48.8625	2.33639	33.0	1.0	Edit	Delete
Hermitage	1	59.94056	30.31361	18.0	1.0	Edit	Delete
SberBank	4	55.771305	37.685346	110.0	7.0	Edit	Delete
BMSTU	1	55.7659	37.685	100.0	10.0	Edit	Delete

Figure 2. List of locations before text mining

The system reads the sentences "Guernica is in Hermitage. Madonna is in Louvre" and makes two additional records in the table with the same coordinates (Figure 3).

List of locations:

Add Location

Name	Type	Latitude	Longitude	Altitude	Popularity		
Louvre	1	48.8625	2.33639	33.0	1.0	Edit	Delete
Hermitage	1	59.94056	30.31361	18.0	1.0	Edit	Delete
Guernica	0	180.0	151.0	130.0	2.0	Edit	Delete
Madonna	0	100.0	120.0	90.0	1.0	Edit	Delete
SberBank	4	55.771305	37.685346	110.0	7.0	Edit	Delete
BMSTU	1	55.7659	37.685	100.0	10.0	Edit	Delete

Figure 3. List of locations after text mining

IV. Conclusion

The developed system allows for increasing the quantity and quality of extracted data. The function of activation and deactivation of locations by editing their start and end dates makes data temporal and allows to requests for information that is current for the required time; ensuring that data queries are contemporaneous.

To retrieve new data the system employs the NLP approach of text mining. Such an approach requires more system resources but it provides better quality of mined data and is able to extract larger amounts of data. The main algorithm is based on the rules of regular grammar, which are saved in a database and can be changed during runtime.

The mined data are transformed into intermediate form of knowledge. The opportunities of NLP are numerous: the system collects data about different objects. This process generates a lot of relations between many objects that allows

them to be used in future improvements and affording the implementation of new functions of the system. It can be information about kinds of locations and their common features.

The text mining method that was presented in this work showed very good results and it proved that it is appropriated for different languages and a variety of tasks. The presented approach is a complex investigation task that requires further research. First of all, the authors intend to investigate modification of the method by adding new rules for complicated sentences, increasing the amount of relations between objects and respectively mined knowledge, and additional conditions for extracting locations.

REFERENCES

[1] Y. Zhang, M. Chen, L. Liu. A Review on Text Mining. Software Engineering and Service Science International Conference, 2015, pp. 681-685. DOI: 10.1109/ICSESS.2015.7339149

[2] M. Sukanya, S. Biruntha. Techniques on Text Mining. Advanced Communication Control and Computing Technologies International Conference, 2012, pp. 269-271. DOI: 10.1109/ICACCCT.2012.6320784

[3] D. Sanchez, M. Martın-Bautista, I. Blanco, C. Justicia. Text Knowledge Mining: An Alternative to Text Data Mining. Data Mining Workshops International Conference, 2008, pp. 664-672. DOI: 10.1109/ICDMW.2008.57

[4] S. Yin, Y. Qiu, J. Ge. Research and Realization of Text Mining Algorithm on Web. Computational Intelligence and Security Workshops International Conference, 2007, pp. 413-416. DOI: 10.1109/CISW.2007.4425522

[5] I. Witten, E. Frank, M. Hall. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers, 2nd edition, 2005, 560 p. DOI: 10.1186/1475-925X-5-51

[6] H. Mousavi, D. Kerr, M. Iseli, C. Zaniolo. Mining Semantic Structures from Syntactic Structures in Free Text Documents. Semantic Computing International Conference, 2014, pp. 84-91. DOI: 10.1109/ICSC.2014.31

[7] R. Amarasiri, J. Ceddia, D. Alahakoon. Exploratory Data Mining Lead by Text Mining Using a Novel High Dimensional Clustering Algorithm. Machine Learning and Applications Proceedings of the Fourth International Conference, 2005, 6 p. DOI: 10.1109/ICMLA.2005.29

[8] V. Verma, M. Ranjan, P. Mishra. Text Mining and Information Professionals. Role, issues and Challenges. Emerging Trends and Technologies in Liabraries and Information Services International Symposium, 2015, pp. 133-137. DOI: 10.1109/ETTLIS.2015.7048186

[9] Y. Qi, Y. Zhang, M. Song. Text Mining for Bioinformatics: State of the Art Review. Computer Science and Information Technology, 2009, pp. 398-401. DOI: 10.1109/ICCSIT.2009.5234922

[10] C. Silva, B. Ribeiro. Margin-based Active Learning and Background Knowledge in Text Mining. Hybrid Intelligent Systems Fourth International Conference, 2004, pp. 8-13. DOI: 10.1109/ICHIS.2004.70

[11] A. Akilan, M. Phil. Text Mining: Challenges and Future Directions. Electronic and Communication Systems 2nd International Conference, 2015, pp. 1679-1684. DOI: 10.1109/ECS.2015.7124872

[12] K. Wang, Q. Wu, H. Mao, M. Zhou, K. Jiang, X. Zhu, L. Yang, T. Wang, H. Wang. Intelligent Text Mining Based Financial Risk Early Warning System. Information Science and Control Engineering 2nd International Conference, 2015, pp. 279-281, DOI: 10.1109/ICISCE.2015.68

[13] A.H. Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. 1999, Vol. 8, pp. 65-70.

[14] X. Huosong, F. Zhaoyan, P. Liuyan. Chinese Web Text Outlier Mining Based on Domain Knowledge. Intelligent Systems Second WRI Global Congress, 2010, pp. 73-77. DOI: 10.1109/GCIS.2010.66

[15] H. Champion, N. Pizzi, R. Krishnamoorthy. Tactical Clinical Text Mining for Improved Patient Characterization. Big Data International Congress, 2014, pp. 683-690. DOI: 10.1109/BigData.Congress.2014.101

[16] T. Gong, C. Tan, T. Leong, C. Lee, B. Pang, T. Lim, Q. Tian, S. Tang, Z. Zhang. Text Mining in Radiology Reports. Eighth Data Mining International Conference, 2008, pp. 815-820. DOI: 10.1109/ICDM.2008.150

[17] Z. Hu, B. Cohen, L. Hirschman, A. Valencia, H. Liu, M. Giglio, H. Cathy. iProLINK: A Framework for Linking Text Mining with Ontology and Systems Biology. International Conference on Bioinformatics and Biomedicine, 2008, pp. 467-472. DOI: 10.1109/BIBM.2008.73

[18] I. Pinho, D. Epstein, E. Reategui, Y. Corrêa, E. Polonia. The Use of Text Mining to Build a Pedagogical Agent Capable of Mediating Synchronous Online Discussions in the Context of Foreign Language Learning. Frontiers in Education Conference, 2013, pp.393-399. DOI: 10.1109/FIE.2013.6684853

[19] M.M. Mostafa. More than words: Social networks' text mining for consumer brand sentiments, *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241–4251, Aug. 2013.

[20] L. Ampofo, S. Collister, B. O'Loughlin, A. Chadwick. Text mining and social media: When quantitative meets qualitative and software meets people. In P. J. Halfpenny, P.J. and Procter, R. (Eds.), *Innovations in digital research methods*. London, United Kingdom: SAGE Publications, 2015.

[21] A. Sun, M. Lachanski, and F.J. Fabozzi. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction, *International Review of Financial Analysis*, vol. 48, pp. 272–281, Dec. 2016.

[22] A.J. Lazard, E. Scheinfeld, J.M. Bernhardt, G.B. Wilcox, and M. Suran. Detecting themes of public concern: A text mining analysis of the centers for disease control and prevention's Ebola live Twitter chat, *American Journal of Infection Control*, vol. 43, no. 10, pp. 1109–1111, Oct. 2015.

[23] K. Amarasinghe, M. Manic, R. Hruska. Optimal Stop Word Selection for Text Mining in Critical Infrastructure Domain. Resilience Week, 2015, pp. 1-6. DOI: 10.1109/RWEEK.2105.7287440.

[24] D. Dunlavy, T. Shead, E. Stanton. ParaText: Scalable Text Modeling and Analysis. Proceedings of the 19th ACM International Symposium in High Performance Distributed Computing, 2010, pp. 344-347. DOI: 10.1145/1851476.1851526