Hadoop-based Framework for Information Extraction from Social Text

Ferdaous Jenhani, Mohamed Salah Gouider and Lamjed Bensaid
Department of Computing, SMART Laboratory,
Institut Superieur de Gestion de Tunis, Universite de Tunis, Rue de la Liberte, Le Bardo, Tunisia

Keywords: Hadoop, Social Data, Twitter, Information Extraction, Drug Abuse.

Abstract: Social data analysis becomes a real business requirement regarding the frequent use of social media as a new business strategy. However, their volume, velocity and variety are challenging their storage and processing. In a previous contribution (Jenhani et al., 2016a, 2016b), we proposed an events extraction system in which we focused only on data variety and we did not handle volume and velocity dimensions. So, our solution cannot be considered a big data system.
In this work, we port previously proposed system to a parallel and distributed framework in order to reduce the complexity of task and scale up to larger volumes of data continuously growing. We propose two loosely coupled Hadoop clusters for entity recognition and events extraction. In experiments, we carried time test and accuracy test to check the performance of the system on extracting drug abuse behavioral events from 1000000 tweets. Hadoop-based system achieves better performance compared to old system.

1 INTRODUCTION

Nowadays, social media are becoming used for business purposes which created a new source of information rich with knowledge useful to make better decisions in competitive environments. In the last recent years their analysis becomes the interest of research community, data scientists and industrials in order to extract structured information mainly entities, relations and events from user-generated content. In (Jenhani et al., 2016a, 2016b), we combined natural language processing, linguistic rules and machine learning techniques to extract meaningful, structured and coherent information. However, we did not handle data volume and velocity which are main challenges of building big data solution. In fact, used technologies fail to store and process continuously growing volumes of data, on one hand. On the other hand, the proposed solution is complicated since it combines many technologies so that cannot be straightforward applied to data generated in very high frequency and bring real time results. In fact, rapidly ingesting, storing, and processing big data requires a cost-effective infrastructure that can scale with the amount of data and the scope of analysis. Therefore, we port the system to parallel programming model namely Hadoop. Hadoop is the most pragmatic solution for managing huge volume of data generated in high frequency thanks to its parallel programming model and distributed storage. Indeed, we propose a parallel and distributed framework to implement information extraction process and reduce the complexity of the task thanks to Map Reduce paradigm and HDFS storage in distributed clusters. We focus mainly on two information extraction tasks; named entity recognition and events extraction. Regarding the complexity of the handled tasks, we propose two loosely coupled clusters; one for each. In order to check the performance of the system, we carried out run time test by comparing running time of the solution implemented as standard java application and when implemented in Hadoop framework. Data crawling and HDFS load time test and accuracy test by measuring precision, recall and F-measure. Running time and HDFS load linearity face to data volume is not strange in a parallel and distributed system. However, how this affects the accuracy is very important. The system is tested on twitter data for extracting drug abuse behavioral events and have shown improvement of accuracy when the volume of data increases.

The paper is organized as follows; in section two we investigate the various big data solutions for information extraction from social media data. Hadoop, the most famous and reputed big data technology used in our system is presented in section three. The proposed system is composed of two clusters namely the named entity recognition cluster (NER-Hadoop Cluster) and events extraction cluster (EE-Hadoop Clus-

Jenhani F., Salah Gouider M. and Bensaid L.
Hadoop-based Framework for Information Extraction from Social Text.
DOI: 10.5220/0006501402330237
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 233-237
ISBN: 978-989-758-271-4

ter) presented in section four. Implementation and test details are discussed in section five. Finally, we conclude in section six.

2 STATE OF THE ART

Social data are the most important big data type handled and analyzed by data scientists regarding the rich information they embed. However, they are complex since they are unstructured, informal, conversational, voluminous and generated in a very high frequency. In fact, by the increase of data volumes, data scientists and researchers are convinced that natural language processing is no longer possible on a single computer whatever its memory size or processing speed. Therefore, the use of big data technologies based on distributed storage and parallel computing model is the solution to reduce the complexity of the task and achieve analysis goals. Hadoop is the most popular distributed framework and commonly used big data technology by scientists and industrials. It is used in Trendminer system (Preotiuc-Pietro et al., 2012) which is a real time system for social media text analysis based on Hadoop Map Reduce framework in order to decrease the complexity of analyzing terabytes of tweets, improve the performance of pre-processing and extract structured information like named entities. Trendminer without Hadoop processes only 0.51 millions of Tweets in one hour. However, with Hadoop, it processes 7.6 million Tweets for the same period. GeLo (Nesi et al., 2014) is a hybrid system for geographic information extraction from web pages composed of web crawler, geographic information extraction part and a geo-coding module. The crawler of the system is based on Hadoop. Authors relied on the use of distributed and parallel framework for web information retrieval. Unlike KOSHIK framework (Exner et al., 2014) which used Avro for web documents serialization from web sources to HDFS. Hadoop MapReduce engine is used to parallelize natural language processing, annotation and querying of annotated data. NLP tasks including tokenization, pos tagging and parsing are implemented in map reduce jobs. Querying is achieved thanks to Hive and Pig capabilities. In (Prabhakar Benny et al., 2016), authors proposed a system for entity resolution which heart is map reduce functions. Connected to data stream collector and after preprocessing, data is matched to a set of existing rules. If entities match existing rules, the rules are updated and new entities are stored in a data base. If not, sets of blocks of entities are generated and promoted to the map phase for similarity computation and an average similarity is generated which is the input of the reduce function in order to compute a threshold value and make comparison. Tested on Google and Amazon product bench mark data, the system achieved 70% average accuracy. An emerging research is about opinion mining and sentiment analysis in social data especially Twitter. In (Ramesh et al., 2015, Ha et al., 2015) authors used Hadoop data file system for social data storage. Sentiment computation algorithms are divided between map reduce functions to improve processing time and process the large volume of social data thanks to parallel and distributed computing logic. In (Ha et al., 2015), authors used four nodes for Hadoop cluster implementation. Experiments showed a stable memory usage regardless the data size which is a maximum 2.5% and a minimum 0.10%. As well as a stable CPU load especially for master node. The time processing for sentiment analysis increased linearly from 26s to 68s in accordance with the scale of the data sets. Concerning accuracy, sentiment analysis results using Hadoop are close to results obtained with a manual process. In (Sunil et al., 2014), contributors implemented a set of functions for sentiment analysis in Twitter based on a sentiword dictionary. The proposed solution achieved an interesting accuracy average of 72% and important time efficiency thanks to distributed processing which reduced access time and response time eventually. Sarcasm is a type of sentiment expressing negative feelings using positive words. It is hard to detect in textual data. In (Bharti et al., 2016) authors proposed hadoop-based framework for sarcastic content detection in tweets using map reduce programming model. Then a tweet is either, actual positive, actual negative, sarcastic or neutral. Hadoop cluster composed of five nodes is implemented. The use of hadoop showed an interesting decrease of preprocessing and processing time. In (Dhamodaran et al., 2015), Hadoop is used to build an automatic emergency alert system by crawling tweets, storing them in an HDFS and extracting disaster keywords as well as location information using map reduce engine.

3 HADOOP TECHNOLOGY

Hadoop is an open source project managed by Apache software foundation and adopted by many organizations such as Google, Yahoo, Facebook and Amazon. It is the framework that allows distributed processing of large data sets across clusters of computers using simple programming models which allow big problems to be broken down into smaller tasks so that analysis could be easily, quickly and cost effectively. Its hardware architecture is designed to scale up from

single server to thousands of machines, each offering local computation and storage, able to detect failures, and self-adjust to continue to operate without interruption and without data loss thanks to data replication technique. Hadoop has two main components (White, 2011):

- Hadoop Data File System-HDFS: It offers a distributed storage of data in files across clusters of servers. It is a reliable, high-bandwidth, low-cost data storage technique.
- Map Reduce engine: a programming model composed of two user defined functions for parallel processing of data among a set of computers. Hadoop performs its implementation in parallel among distributed set of computers belonging to the same cluster.

4 HADOOP-BASED FRAMEWORK

In this work, we cover two heavy information extraction tasks namely named entity recognition and events extraction. Therefore, we develop two loosely coupled Hadoop clusters; named entity recognition cluster (NER-Cluster hereafter) and events extraction cluster (EE-Cluster hereafter) where the output of the first is the input of the second as shown in figure 1 and explained in subsequent sections:

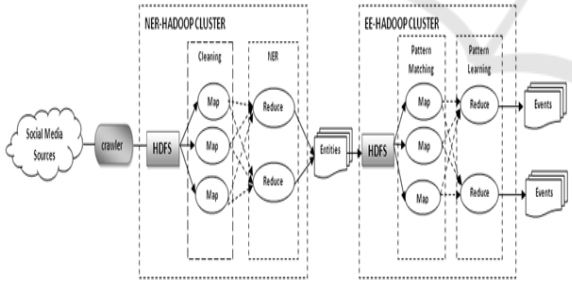

Figure 1: Hadoop-based Information Extraction Framework.

Data collection is performed by a keyword-based crawler developed at the heart of the system in order to collect social data streams from API platforms and load them into HDFS for analysis based on domain keywords.

4.1 NER-Hadoop Cluster

The NER cluster regroups data cleaning and named entity extraction in its map and reduce functions. The input data is load into HDFS and streamed to later

Figure 2: NER-Map Reduce Engine.

functions. In map function we implement data cleaning which includes removing redundancy, punctuations, special characters, URLs and URIs in order to get data ready for analysis. Linguistic processing is implemented in the reduce function. We use Stanford Core NLP (Manning et al., 2014), a natural language processing tool for tokenization, Part of Speech-POS tagging and common named entity recognition such as date, location and person. However, to recognize domain entities such as DRUGS, routes of intake-ROI and effects-EFF, we extend NLP pipeline with a new annotator based on domain dictionary (Jenhani et al., 2016). Social text is plenty with colloquial words, informal data and abbreviations. Therefore we used fuzzy matching used to match words similar for a given threshold to capture the maximum of correct named entities. In the implementation, we match words which are 80% similar. In short, in the reduce phase, the set of tweets are reduced to a set of entities (date, location, person, DRUG, ROI, PATIENT, etc). The final step consists on parsing data in order to identify all syntactic and grammatical relationships between named entities.

4.2 EE-Hadoop Cluster

Events extraction consists on extracting complete, meaningful and coherent information. It represents the n-ary semantic relationship between entities. In order to extract meaningful events with high precision, we proposed a hybrid approach combining linguistic rules with machine learning technique namely classification.

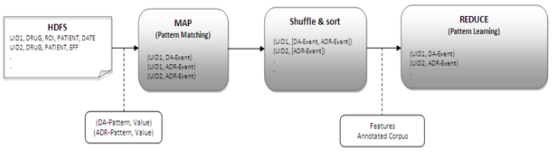

Figure 3: EE-Map Reduce Engine.

We implement this hybrid process in map reduce engine which input is named entities extracted in the first layer. In the map phase we match the set of output entities and semantic predicates with lexico-syntactic patterns. In fact, we give as input to map functions a set of patterns prepared based on linguistc grammar. In our drug abuse case study, we consider drug abuse patterns (DA pattern) and adverse drug

reaction patterns (ADR Pattern). To each set of input entities could correspond more than one pattern. Therefore, in the reduce phase, final patterns will be selected using a classification technique thanks to discriminative features. In the context of our application, features are named entities and triggers (take, play, purchase, need, cause, etc), semantic relationships (Abuser, Substance, Dose, unit, etc) and event tags (Adverse Drug Reaction-ADR, Drug Abuse-DA, NONE). We consider also the sentence length because social media data suffer from short messages syntactically poor but meaningful. Indeed, a given event may be expressed with one keyword in a sentence full of colloquial meaningless words like "marijuana, oh, good experience". We use also window of words mainly the word before subject (WBS) and word after subject or before trigger (WBT). To choose the best classification technique for this problem, we compared a set of classification techniques and the best suited for this task is Support Vector Machine (SVM). We refer the reader to (Jenhani et al., 2016) for more details.

5 IMPLEMENTATION AND EXPERIMENTS

In this section we present implementation details and tests to check the performance of the proposed system compared to previous implementation. The proposed system consists of two parallel servers. In each one we create 3 nodes using pseudo-distributed paradigm. We use eclipse java development environment to which we import Hadoop library for map reduce job programming and HDFS construction. We import also Twitter4j library to build a key-word based crawler. In table 1 we summarize information about social data collected from Twitter Streaming API to run the proposed system:

Table 1: Data Sets Crawling Details.

Data Set	NB Tweets	Extraction period (days)
DS1	504000	30
DS2	300000	23
DS3	196000	12
TOTAL	1000000	65

For performance analysis, we carry out run time test, data crawling and HDFS loading time test and accuracy test on the three data sets having different sizes to see system performance on different data volumes.

5.1 Run Time Test

We compare run time of the system implemented as a standard java program (SJP hereafter) with Hadoop map reduce solution (HDP hereafter) as depicted in table 2.

Table 2: Run Time Test (minutes).

	NER Cluster		EE Cluster		All	
	SJP	HDP	SJP	HDP	SJP	HDP
DS1	125	100	148	102	127	95
DS2	101	89	136	93	113	85
DS3	83	67	106	77	93	69

For named entity recognition, running time is decreased by about 25% using Hadoop. However, events extraction is decreased by 45% where we observe clearly the efficiency of distributed and parallel solutions for complex problems. Events extraction takes much time which is explained by the complexity of treatment for a standard java program. In fact, events extraction combines two heavy processing methods; linguistic pattern matching and classification which is costing process and hard using traditional technologies especially when data volume is large. Nevertheless, using Hadoop map reduce, the process is much more simplified and processing time is improved considerably. The processing time, when all the system is implemented using Hadoop and data is smoothly promoted between layers, is improved compared to results obtained from each layer separately.

5.2 Data Crawling and HDFS Load Time Test

Data crawling and loading time test is an important test to check the performance of Hadoop-based solution. Table 3 shows this test carried on each data set. We observe linear data loading in HDFS while data crawling time is increasing. Thus, whatever data volume, data load in HDFS is stable regarding its big capacity to store larger volumes of data.

Table 3: data crawling and HDFS load time (seconds).

	Crawling Time	HDFS Load Time
DS1	2187	18
DS2	1452	16
DS3	925	12

5.3 Accuracy Test

Table 4 summarizes performance test results carried out on Hadoop-based system with three data sets us-

Table 4: Precision, Recall and F-measure for Accuracy Test.

	DA Events			ADR Events		
	P%	R%	F%	P%	R%	F%
DS1	93.3	93.3	93.3	98.1	97.8	97.8
DS2	91	90.4	90.4	97.1	96.8	96.8
DS3	84.7	83.3	83.9	96.6	95.6	95.8
Avg	89.7	89	89.2	97.3	96.7	96.8

ing SVM classifier to test accuracy of extracting drug abuse events (DA) and adverse drug reactions (ADR). We use three measures of accuracy; precision, recall and F-measure calculated based on the number of positive instances, negative ones, false positives and false negatives.

6 CONCLUSIONS

Social data generated in very high frequency require real time analysis to make the quicker decision in highly competitive environments. However, existing technologies fail to process voluminous, unstructured and dynamic data in real time and return required information for immediate decision making.

In this work, we adopt Hadoop, distributed framework which offers a parallel programming model and effective storage system in distributed clusters. Tested on Twitter data for drug abuse events extraction, Hadoop-based system achieved good performance results.

REFERENCES

Peotiuc-Pietro,D., Samangooei, S., Cohn, T., Gibbins, N., Niranjan, M. (2012). Trendminer: An Architecture for Real Time Analysis of Social Media Text, Association for the Advancement of Artificial Intelligence (www.aaai.org).

Nesi, P., Pantaleo, G., and Tenti, M. (2014). Ge(o)Lo(cator): Geographic Information Extraction from Unstructured Text Data and Web Documents. DISIT - Distributed Systems and Internet Technology Lab, Department of Information Engineering, University of Florence, Italy.

Exner, P. and Nugues, P. (2014). KOSHIK- A Large-scale Distributed Computing Framework for NLP, ICPRAM2014-International Conference on Pattern Recognition Applications and Methods

Ramesh, R., Divya, G., Divya, D., Kurian, M., Vishnuprabha, V. (2015). Big Data Sentiment Analysis using Hadoop, IJIRST International Journal for Innovative Research in Science and Technology— Volume 1 — Issue 11 — April 2015, ISSN (online): 2349-6010

Ha, I., Back, B. and Ahn, B. (2015). MapReduce Functions to Analyze Sentiment Information from Social Big Data, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume.

Dhamodaran, S., Sachin, K. R., and Kumar, R. (2015). Big Data Implementation of Natural Disaster Monitoring and Alerting System in Real Time Social Network using Hadoop Technology, Indian Journal of Science and Technology, Vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79102.

Sunil, B., Mane, B., Sawant, Y., Kazi, S., Shinde, V. (2014). Real Time Sentiment Analysis of Twitter Data Using Hadoop International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 3100

Prabhakar Benny, S., Vasavi, S., Anupriya, P. (2016). Hadoop Framework For Entity Resolution Within High Velocity Streams. International Conference on Computational Modeling and Security (CMS 2016). Procedia Computer Science 85(2016)550 557.

Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. Digital Communications and Networks 2, pp.108121

White, T., second edition, Hadoop: the definitive guide, (2011) , Published by OReilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Jenhani F., Gouider M. S., Ben Said L. (2016) A Hybrid Approach for Drug Abuse Events Extraction from Twitter In 20th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, (ICKIIES16) York, United Kingdom pp. 1032-1040.

Jenhani F., Gouider M. S., Ben Said L. (2016) Lexicon-based System for Drug Abuse Entity Extraction from Twitter In 12th International Conference, BDAS, Ustro, Poland, Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery, Volume 613 of the series. Communications in Computer and Information Science pp. 692-703.

Manning, D., Mihai, C., Bauer, S., Finkel, J., Bethard, J., McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit.