A NoSQL Approach for Aspect Mining of Cultural Heritage Streaming Data

Gerasimos Vonitsanos*, Andreas Kanavos*, Alaa Mohasseb†, Dimitrios Tsolis‡
*Computer Engineering and Informatics Department
University of Patras, Patras, Greece
{mvonitsanos, kanavos}@ceid.upatras.gr
†School of Computing
University of Portsmouth, Portsmouth, UK
alaa.mohasseb@port.ac.uk
‡Department of Cultural Heritage Management and New Technologies
University of Patras, Agrinio, Greece
dtsolis@upatras.gr

***Abstract*—Aspect mining constitutes an essential part of delivering concise and, perhaps more importantly, accurately tailored cultural content. With the advent of social media, there is a data abundance so that analytics can be reliably designed for ultimately providing valuable information towards a given product or service. Naturally representing and efficiently processing a large number of opinions can be implemented with the use of streaming technologies. Big data analytics are especially important in the case of cultural content management where reviews and opinions may be analyzed in order to extract meaningful representations. In this paper, we present a NoSQL database approach for aspect mining of a cultural heritage scenario by taking advantage of Apache Spark streaming architecture.**

***Index Terms*—Apache Cassandra, Apache Spark Streaming, Big Data Analytics, Cultural Heritage Management, Knowledge Representation, Topic Modeling Tweet, Stream Analysis**

I. Introduction

The excessively increasing abundance of data challenges the research community in processing and analyzing it, which will enable the extraction of important knowledge. The management of cultural content has emerged as a major industry driver, where a collection of services such as guided visits to monuments are constantly on high demand in both digital and physical markets around the globe. The still rapidly expanding digital world is currently playing a prominent role in the promotion of cultural content, contributing thus to the preservation and rediscovery of cultural heritage.

In the meantime, as data grows, cloud computing evolves. Frameworks like Hadoop, Apache Spark, Apache Storm and distributed data storages like HDFS and HBase are becoming popular, as they are engineered in a way that makes the process of very large amounts of data almost effortless. Such systems evolve in many aspects, and as a result, libraries, like Spark's MLlib that make the use of Machine Learning techniques possible in the cloud, are introduced.

The need for automatic categorization and extraction of topics is increasing every day with the rise of the digital content in all domains. Aspect extraction can be considered as a major problem in knowledge discovery. There are two types of methods, those that extract aspects without categorizing them and those that extract aspects and categorize them using unsupervised topic modeling; the categorization concerns synonymous aspects in the same category.

In Big Data environment, there are two types of data processing engines, namely batch, and streaming engines. Batch processing deals with the handling of massive volume of data while streaming with processing data of high velocity. Several platforms and tools have been developed to support big data environments, of which the most widely known is Hadoop, which utilizes MapReduce batch processing. It is obvious that some applications require near real time analysis and this type of analysis can be supported by streaming engines, such as Storm and Spark Streaming, which are fault tolerant and guarantee message delivery [6].

In this paper, we propose a real time data analytics platform to delve into cultural content by harnessing Big data analytics in a distributed environment. A framework to render advanced data analytics, i.e., Apache Spark, along with a NoSQL database for handling large portions of data, i.e., Apache Cassandra, provide highly available service with no single point of failure for the aspect mining of cultural heritage content. This framework will be oriented towards efficient content delivery, and it is shown how it can be applied to a dataset consisting of different topics in order for meaningful opinions about historical sites and monuments to be obtained. In order to detect conversations related to the event under consideration, we applied LDA (Latent Dirichlet Allocation) based probabilistic system to discover latent topics.

The research is organized as follows: Section II explains the related work, while Section III depicts the cluster-computing framework Apache Spark Streaming, the distributed NoSQL database management system as well as the Topic Modeling algorithm, which was utilized in our proposed system. In addition, our proposed model is introduced and further analyzed in Section IV. Furthermore, in Section V, the evaluation of the experiments is presented. Finally, Section VI focuses on the main conclusions extracted from the study.

II. Related Work

Topic detection and extraction is concerned with detecting trending topics and extract titles or sets of keywords representing these topics. For text categorization, a work has been presented in [21], where authors use topic modeling for clustering news stories into several k topics; this process is entitled unsupervised learning with automatic topic labeling. Topic modeling reflects the thematic structure of the collection of documents by treating data as observations, which are derived from a generative probabilistic process that comprises hidden variables for documents.

Aspect Extraction is a field with academic interest that is proved by the extensive and growing available literature. Traditional methods are categorized into two basic models, pLSA [10] and LDA [4]. Previous works introduce the discovery of both global and local aspects [19], the extraction of key phrases [5], the multi-aspect rating [22], the aspect summarization [13] as well as the attitudes modeling [17].

Researchers have tried to generate "meaningful" and "specific" topics/aspects. Works [3] and [16] used document label information in a supervised setting. Authors in [11] relied on user feedback during Gibbs sampling iterations. Authors in [1] used another approach (DF-LDA) by introducing must-link and cannot-link constraints as Dirichlet Forest priors. Authors in [14] present two novel statistical models in order to initially extract and in following categorize aspect terms in an automatic way, given some user categories.

III. Preliminaries

A. Twitter

The platform that is being studied in this work is Twitter. It is a platform for uploading posts, exchanging messages between users, and modifying their private profiles according to their needs. Founded in 2006, Twitter is a service which allows users to share 140-character posts. Twitter has 1.3 billion accounts and 3 million monthly active users. The amount of tweets sent per day achieves value equal to 500 million. Twitter provides streaming APIs, which can be connected with a streaming of tweets. It has thus resulted in hosting massive datasets of information whose data are gaining increasing interest.

B. Streaming

The gigantic amount of data created by thousands of sensors and the shipment of those data records simultaneously are defined as streaming data. These data require processing on a record-by-record basis to draw valuable information. The analytics can be filtered, correlated, sampled, or aggregated. This analysis is in a form useful for various business and consumer aspects. For example, the industry can track most popular products among consumers based on corresponding comments and likes on social streams or can track sentiments based on some incidents, and take timely intuitive measures. Over the time, stream processing algorithms are applied to further refine the insights.

C. Spark Streaming

Streaming APIs[1] give access to world-wide twitter users' posts, including tweet itself, tweet language, location where the account was opened, and place from which the tweet was sent, images/videos, etc. Earlier computational models for distributed streaming were less consistent and low leveled while lacked fault recovery. A new programming model, known as discretized streams (D-streams) that provide efficient fault recovery, better consistency, and high-level functional APIs [24], was introduced.

Spark Streaming[2] converts live input stream into batches which are later processed by Spark engine to produce output in batches. D-streams are a high-level abstraction provided by Spark Streaming, while the latter allows parallel processing of data streams by connecting to multiple data streams [12].

Discretized streams are a sequence of partitioned datasets (RDDs) that are immutable and allow deterministic operations to produce new D-streams. The computations executed by D-streams are considered as a series of deterministic and stateless tasks. Across different tasks, states are represented as Resilient Distributed Datasets (RDDs), which are fault-tolerant data structures [23]. Streaming computations are done as a series of deterministic batch computations on discrete time intervals.

D. NoSQL Databases

Nowadays, it is stated that NoSQL databases have a lot more to offer than just offering solutions to scale problems. The real meaning of this generic term NoSQL is that this type does not follow the principles of the traditional relational databases. Instead, while maintaining their principles, they can handle efficiently the output of modern web scale databases like Twitter, Facebook, Google, etc. Due to their difference with the traditional databases, they are superior in terms of scalability and availability problems. As stated in [9], NoSQL databases define a filter for exactly determining the databases that fulfill these requirements. In particular, they provide the following advantages:

- The data representation is schema-less and there is no need to define a certain structure from the beginning since new fields at run-time can be added.
- The speed even with a small amount of data can be processed in milliseconds rather than hundreds of milliseconds.
- The elasticity of the applications is considered due to particular scalability features.
- The development time is reduced freeing developers from having to deal with complex SQL queries and joins as the data are collated from different tables into a new view.

E. Apache Cassandra

Apache Cassandra[3] is an open source NoSQL database, which is extensively scalable. For this reason, it is ideal for

[1] https://developer.twitter.com/en/docs.html

[2] https://spark.apache.org/streaming/

[3] http://cassandra.apache.org/

managing huge amounts of data in different data centers and the cloud as well. Some of its characteristics are the provision of availability that is continuous, linear scalability, as well as the simplicity in operating on multiple servers without any single point of failure. The data model of Cassandra is exceptionally flexible and provides swift response times.

Its design was based on the premise that system/hardware failures always happen, and this fact results in a peer-to-peer distributed system. All nodes are the same whereas the master node as well as the name nodes are missing. The data are distributed among all cluster nodes and the copying as well as sharing procedures are automatic and transparent. Cassandra also provides an advanced custom replication which saves copies of the data on all nodes participating in a Cassandra ring. So, if a node is shut down, then one or more copies of the information that the node has will be available to another cluster node. Cassandra provides linear scaling capability, which means that the overall capacity of the system can be quickly increased by adding new nodes to the network thus deeming it excessively efficient.

F. Topic Modeling

In our proposed work we want to take into consideration the verification of whether all the posts discuss the specific topic. Topic modeling considers a document as a "bag-of-topics" representation, and its purpose is to cluster each term in each post into a relevant topic. Variations of different probabilistic topic models [2], [15] have been proposed, and LDA [4] is considered to be a well known method.

Concretely, the LDA model extracts the most common topics discussed that are represented by the words most frequently used, by simply taking as input a group of documents. The input is a term-document matrix, and the output is composed of two distributions, namely document-topic distribution θ and topic-word distribution ϕ. Expectation-Maximization (EM) [8] and Gibbs Sampling [7] algorithms were proposed to derive the distributions of θ and ϕ. In this paper, we use the Gibbs Sampling based LDA. In this approach, one of the most significant steps is to update each topic assignments individually for each term in every document according to the probabilities calculated using Equation 1.

$$\mathbb{P}(z_i = k | z_{\_i}, w, \alpha, \beta) \propto \frac{(n^{-i}_{(k,m,\cdot)} + \alpha)(n^{-i}_{(k,\cdot,w_i)} + \beta)}{n^{-i}_{(k,\cdot,\cdot)} + V\beta} \quad (1)$$

where $z_i = k$ shows that the i_{th} term in a document is assigned to topic k, $z_{\_i}$ signifies all the assignments of topic except the i_{th} term, $n^{-i}_{(k,m,\cdot)}$ is the number of times that the document d contains the topic k, $n^{-i}_{(k,\cdot,w_i)}$ is the number of times that term v is assigned to topic k, V represents the size of the vocabulary as well as α and β are hyper-parameters for the document-topic distribution and topic-word distribution, respectively.

The number of the Gibbs sampling iterations performed for every term in the corpus is N; after this component, the document-topic θ and topic-word ϕ distributions are estimated using following Equations 2 and 3 respectively.

$$\hat{\theta}_{m,k} = \frac{n_{(k,m,\cdot)} + \alpha}{K\alpha + \sum_{k=1}^{K} n_{(k,m,\cdot)}} \quad (2)$$

$$\hat{\phi}_{k,v} = \frac{n_{(k,\cdot,v)} + \beta}{V\beta + \sum_{v=1}^{V} n_{(k,\cdot,v)}} \quad (3)$$

IV. Implementation

A. Model Overview

We propose a system that consists of two main components, which are data collection and data analysis. The data collection module is developed to crawl the tweets from Twitter using Apache Spark Streaming and in following to store the tweets into Cassandra, a NoSQL database for scalability and scheme less data storage purpose. After the storing procedure takes place, the system mainly performs an online aspect mining methodology in order to answer the following question: what are the topics discussed by people online to help us understand people's interests?

A similar method was presented in [20], where a NoSQL database approach for modeling heterogeneous and semi-structured information by integrating Apache Spark with Apache Cassandra was depicted. Authors focus on a model capable of predicting the relationship between tourist arrivals and nights spent in Greece.

B. Dataset

To simplify management, access, as well as aggregation, data are stored in individual collections organized by day. It is evident that the relational database is struggling in handling large amount of unstructured data [18]. Our experiments were conducted on a cluster with 20 nodes, where each node is equipped with quad-core Intel(R) Core(TM) i5-2400 CPU@3.10 GHz with 4GB RAM.

To ensure that only relevant tweets are considered, we filter all captured posts with keywords that are relevant to Cultural Heritage in the domain of Greece. These keywords are related to different heritages, activities as well as specific tourist destinations. The filtered data set resulted in $5,000$ out of $35,000$ tweets from $01/03/2019$ to $30/03/2019$. In following Figure 1, the hashtags studied are presented.

V. Evaluation

One of the trickier tasks about LDA model is to specify the number of topics to be generated. Our goal is to find topics associated with each post. Hence we obtain the probabilistic Latent Semantic Analysis as a topic distribution. It is obvious that Latent Dirichlet Allocation could be also used as a clustering algorithm, as it groups together words with similar meaning and assigns them to topics that LDA generates.

To evaluate our system, we conducted a user study in which results from our approach were compared to those from users. Human annotators were used in order to read the downloaded posts and specify whether the extracted aspects are correctly identified. 10 students associated with the University of Patras manually classified the tweets and their corresponding aspects, and in following, we compared the system results to the users'

Fig. 1. Hashtags: #Acropolis, #Parthenon, #Epidaurus, #Mycenae, #Meteora, #Vergina, #Knossos, #Thermopylae, #Diolkos, #HerodesAtticus, #Ancientgreece, #Ancientathens

responses. These responses were used as gold standard for the evaluation of the system's performance.

The percentages of corrected identified tweets are presented in following Table I. We can observe that our proposed methodology seems to achieve notable accuracy as 65% of aspects are correctly identified, whereas 20% of aspects cannot be considered as correct. The objective of this experiment is to examine whether our approach extracts and categorizes correctly aspects from a real time data analytics platform.

TABLE I
PERCENTAGES OF ASPECTS EXTRACTED FROM DOWNLOADED TWEETS

Category	Percentage of Aspects
Correctly Identified	65%
In-between Identified	15%
Wrongly Identified	20%

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have developed a real time data analytics platform for analyzing cultural content by harnessing Big data analytics in a distributed environment. The proposed concept relies on distributed architectures and algorithms which are able in real time to manage, process, and evaluate high volume and high velocity of hybrid data. The proposed framework is based on a NoSQL scheme for the aspect mining of cultural heritage content as it takes advantage of Apache Spark streaming architecture as well as Apache Cassandra NoSQL database. Finally, we applied it to a dataset consisting of different topics about historical sites and monuments.

The proposed framework can be extended in a number of ways. Initially, as cultural inheritance tends to become global, cross language opinion mining becomes important. This strongly implies that besides linguistic factors, geography, semantics, and the cultural context must also be considered in order to conduct meaningful aspect mining. Another issue concerns the development of a clustered system where big data related techniques will be utilized for other countries.

REFERENCES

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *26th Annual International Conference on Machine Learning (ICML)*, pages 25–32, 2009.

[2] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pages 121–128, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34:569–603, 2009.

[6] N. Franciscus, Z. Milosevic, and B. Stantic. Influence of parallelism property of streaming engines on their performance. In *New Trends in Databases and Information Systems (ADBIS)*, pages 104–111, 2016.

[7] T. L. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Technical Report, Stanford University*, 2002.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004.

[9] J. Han, E. Haihong, G. Le, and J. Du. Survey on nosql database. In *6th International Conference on Pervasive Computing and Applications (ICPCA)*, pages 363–366, 2011.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 289–296, 1999.

[11] Y. Hu, J. L. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 248–257, 2011.

[12] D. Logothetis, C. Trezzo, K. C. Webb, and K. Yocum. In-situ mapreduce for log processing. In *USENIX Annual Technical Conference*, 2011.

[13] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *18th International Conference on World Wide Web (WWW)*, pages 131–140, 2009.

[14] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. In *50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 339–348, 2012.

[15] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012.

[16] D. Ramage, D. L. W. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 248–256, 2009.

[17] C. Sauper, A. Haghighi, and R. Barzilay. Content models with attitude. In *49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 350–358, 2011.

[18] B. Stantic and J. Pokorný. Opportunities in big data management and processing. In *11th International Baltic Conference on Databases and Information Systems*, pages 15–26, 2014.

[19] I. Titov and R. T. McDonald. Modeling online reviews with multi-grain topic models. In *17th International Conference on World Wide Web (WWW)*, pages 111–120, 2008.

[20] G. Vonitsanos, A. Kanavos, P. Mylonas, and S. Sioutas. A nosql database approach for modeling heterogeneous and semi-structured information. In *9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8, 2018.

[21] H. M. Wallach. Topic modeling: Beyond bag-of-words. In *23rd International Conference on Machine Learning (ICML)*, pages 977–984, 2006.

[22] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 783–792, 2010.

[23] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 15–28, 2012.

[24] M. Zaharia, T. Das, H. Li, S. Shenker, and I. Stoica. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. In *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2012.