Aspect-Based Community Detection of Cultural Heritage Streaming Data

Elias Dritsas*, Maria Trigka*, Gerasimos Vonitsanos*, Andreas Kanavos*,†, Phivos Mylonas‡
*Computer Engineering and Informatics Department
University of Patras, Patras, Greece
{dritsase, trigka, mvonitsanos}@ceid.upatras.gr
†Department of Regional Development
Ionian University, Lefkada, Greece
akanavos@ionio.gr
‡Department of Informatics
Ionian University, Corfu, Greece
fmylonas@ionio.gr

***Abstract*—Twitter is one of the most popular online social networks providing a huge amount of data produced by users interactions through tweets. After an appropriate analysis of this data, groups of users who share similar attributes, emotions, opinions, and preferences can be identified. Massive cultural content management are important because reviews and opinions may be analyzed in order to extract meaningful representations. In this paper, an aspect mining method of a cultural heritage scenario by taking advantage of Apache Spark streaming architecture is presented. Specifically, we propose the combination of a community detection detection algorithm, i.e. the Parallel Structural Clustering Algorithm for Networks (PSCAN), with a topic modelling methods, i.e. the Latent Dirichlet Allocation (LDA), for performing large-scale data analysis in Twitter.**

***Index Terms*—Apache Spark Streaming, Community Detection, Cultural Heritage Management, LDA, MongoDB, Stream Analysis, Topic Modelling**

I. Introduction

Social media are considered Internet-based services that allow users to create public profiles and become group member, interacting with each other for achieving either individual or team goals. Nowadays, the online social networking is very popular, since millions of people want to stay connected with friends, family, colleagues. Twitter is a social network that allows users to exchange short messages, called tweets. Except for the services it offers to its users, it is among the social networks that have helped to continuously improve data mining due to the easy-to-use API that makes the data collection very easy [7].

There is no doubt that huge amount of data is being generated daily by consumers, and/or businesses all over the world. Nonetheless, the traditional methods for data analysis are impossible to effectively process Big Data, due to their diversity, complexity, and large scale characteristics. In the era of Big Digital Data, social networks have become an integral part of humans life. The social networks are usually represented as graphs consisting of over millions of vertices and edges. The vertices represent the interactive users and the edges their relationship. A fundamental task in social networks analytics is the discovery of groups of users with common attributes. This process is called community detection and, it is addressed using clustering-based methods.

In general, the term network can refer to any interconnected group or system that interacts in a complex way to serve a purpose. Studies indicate that in social networks, the distribution of the clustering coefficient follows the power law, and it decreases when the degree of nodes increases. The clustering coefficient is an essential factor that measures the tendency of the nodes to cluster together. This characteristic implies that social networks are formed by connected communities. Discovering these communities is essential in order to understand the structure of the network where a community is defined as a group of nodes with more links between themselves and fewer external links to other nodes.

The main motivation of this study is that social media, such as Twitter, play a crucial role in cultural heritage management [16], [26]. A plethora of graph partitioning algorithms and topics automatic detection have been developed that help scientists to get an insight into the users interests/preferences about museums, monuments, and urban heritage sites through communities and topics discovery. In this paper, we address the problem of community detection and topic modelling in Apache Spark using a NoSQL database, i.e., MongoDB as data storage. The problem is addressed with the aid of PSCAN [31], a clustering algorithm for community detection using the GraphX and Streaming API of Apache Spark, and the LDA for topic modelling of users in the extracted communities [17]. The proposed framework is investigated on a Twitter dataset considering only users with followers to ensure that the corresponding graph, for communities detection, is connected.

The rest of the text is outlined as follows: In Section II, we outline some related work on the topic while Section III introduces the proposed framework for community detection and topic modeling on Twitter. Section IV details our implementation whereas Section V presents the research results. In Section VI, we summarize our contributions and future directions.

II. Related Work

Social networks analysis is strongly related to graph clustering algorithms, while text mining or analysis deals with Natural Language Processing (NLP) algorithms for topic analysis. In this section, a brief review of community detection and topic modelling methods is presented, with an emphasis on social networks and especially in Twitter.

Community is defined as a group of network nodes with dense links between the nodes [29]. Many community detection algorithms have been proposed to identify complex community structures in social networks [4], [14], [15]. Some traditional techniques for data clustering, like hierarchical, partitional and spectral clustering are sometimes adopted for graph clustering too [4], [20], [21], [22]. Authors in [23] proposed a standard methodology for community detection in Twitter using feature selection methods. Here, our focus is on the PSCAN algorithm, which was designed on the Hadoop framework and provides another parallel scheme for the MapReduce model in large networks, such as Twitter [31]. In order to identify the desired topics, the LDA model is employed [25].

An empirical study of topic modelling in Twitter using various performance metrics was implemented in [7]. Topic modelling with LDA [19], a widely used probabilistic method, has become a standard tool and as a result, various extensions have been proposed to tackle its limitations, and in particular for social networks. In [24], for addressing the LDA inadequency in tweets sparsity of short documents, several pooling techniques were suggested. The research outcomes have shown that aggregating similar tweets into individual documents significantly augments topic coherence. The concept of pooling techniques in topic modelling has been also studied in [1] with the aggregation of tweets by conversations.

Researchers have incorporated the concept of influence from the side of users to the side of networks and users' personality has been utilized as the key characteristic for identifying influential networks [8], [9], [10]. Moreover, the behavior of users on an emotional level is enhanced by introducing a new methodology that effectively aids in community detection [11], [12], [13]. Similarly, there are several ways to assess the clustering quality, namely community coherence [18].

III. Proposed Architecture

In this section, we briefly introduce the architecture of our proposed Twitter topic modeling based community detection system. A novel system that consists of two main components, which are data collection and analysis, is proposed in Figure 1 taking into account the corresponding modules. Specifically, the three modules implemented are described below.

A. Data Collection and Pre-processing

The data collection module is developed to crawl the tweets from Twitter utilizing Apache Spark Streaming and in following to store the tweets into MongoDB, a NoSQL database for scalability and scheme less data storage purpose. A similar method was introduced in [27], where a NoSQL database approach for modeling heterogeneous and semi-structured information by integrating Apache Spark with Apache Cassandra was depicted.

Twitter Streaming API has been used for the crawling of tweets as it allows high-speed access almost in real time in various subsets of public and private Twitter data. Also, it contains public tweets, filtered in various ways and is a continuous flow of data without rate limitation and with random content.

In following, the crawler traverses the Twitter and creates a social media graph where nodes are users and edges represent the "follow" connection between two users. The process creates a sample of the Twitter graph as follows: initially it retrieves the users as well as their followers, who have posted a tweet within the given time period. Subsequently, it connects users that follow each other or have a common follower through that follower.

Furthermore, we aim at creating a graph from Twitter users. Spark's GraphX library is used, where two RDDs (Resilient Distributed Datasets) [30] were used; the first include the nodes of the graph and the latter its edges.

B. Community Detection

The Parallel Structural Clustering Algorithm for Networks (PSCAN) [31] is considered for the community detection process as it is suitable for detecting community structures in big networks. Two basic input parameters are the user graph and a metric e that is employed for the edges pruning based on the value of "similarity" calculated by the algorithm itself. The similarity of the edges is calculated based on the equation:

$$sim(v,u) = \frac{|N(v) \cap N(u)|}{|N(v)||N(u)|} \tag{1}$$

where $N(u)$ are the adjacent nodes of the node u. The edges with similarity less than e are subtracted from the graph. The default value of e is 0.5 and the continuation of the process is based on label propagation.

Each community contains Twitter users with the corresponding statuses. Our goal is, for each user inside a community, to identify the "topic" of their statuses.

C. Topic Modeling

After the community detection takes place, the system performs an online aspect mining procedure. The LDA technique [2], [28] is a probabilistic model used for knowledge mining mainly on text. Topic modeling examines a document as a "bag-of-topics" representation, and its aim is to cluster each term from each post into a relevant topic. Concretely, the LDA model extracts the most widely recognized topics discussed that are represented by the most often used words, by taking as input a group of documents. The input is a term-document matrix while the output is composed of two distributions, namely document-topic θ and topic-word distribution ϕ.

Gibbs Sampling [6] algorithm was utilized to derive the distributions of θ and ϕ and the update of each topic assignments

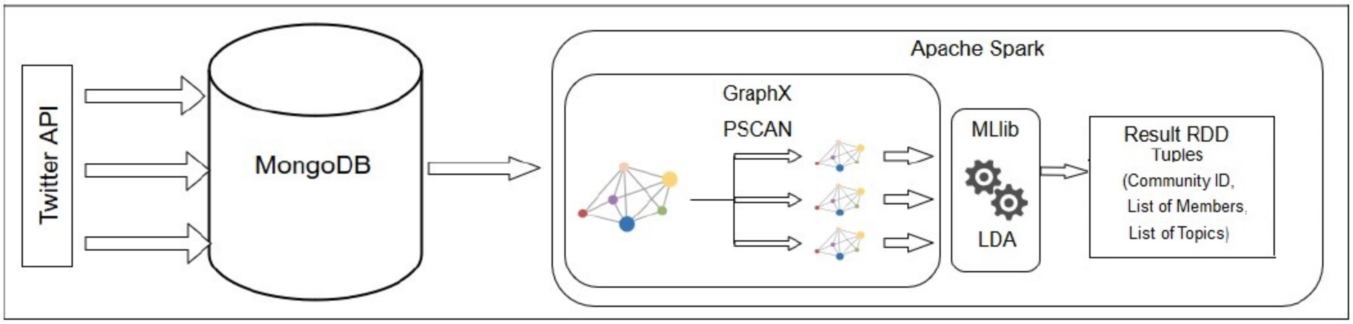

Fig. 1. Overall Architecture of the Proposed System

for each term in every document according to the probabilities is calculated using Equation 2.

$$\mathbb{P}(z_i = k|z_{\_i}, w, \alpha, \beta) \propto \frac{(n^{-i}_{(k,m,\cdot)} + \alpha)(n^{-i}_{(k,\cdot,w_i)} + \beta)}{n^{-i}_{(k,\cdot,\cdot)} + V\beta} \qquad (2)$$

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

IV. Implementation

In the context of this study, various tools were used to facilitate the implementation of community detection and topic modeling process in the Twitter dataset. The operating system was Manjaro Linux, Kernel 4.14 LTS with Apache Spark 2.2.1, Scala 2.11, MongoDB 3.6 and ML.Sparkling library.

A. Twitter Dataset

An approach based on specific topics was used for collecting tweets via a keyword search query for the generation of our test dataset. Keywords that are relevant to cultural heritage in the Greek domain were downloaded. These keywords are related to different heritages, specific tourist destinations and activities, whereas emphasis was given to Corfiot music heritage and paths in accordance to the TRUMPET project[1].

In order to facilitate the mining process of the collected data, it is necessary to apply several pre-processing steps [3], [5]. These steps include the utilization of regular expressions to remove for example unnecessary urls or the representation of emoticons with their equivalent form, e.g. lol as laugh out loud. The removal of punctuation marks and stop-words is another important step. Also, the lemmatization and tokenization processes were employed in which lexical and morphological analyses of words are taken into account in order to remove complex suffixes and to retrieve the lexical form of the term.

The filtered dataset resulted in $5,000$ tweets from $01/02/2021$ to $28/02/2021$ as we have only kept tweets posted in English language.

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. It is obvious that this model could be also used as a clustering algorithm, as it groups together words with similar meaning and assigns them to topics that were initially generated. The following Table I presents the values of β which reflects the occurrence probability or importance of a word within a topic for 10 topics with 5 maximum terms per topic.

TABLE I
LDA Matrix β for $k = 10$ topics with 5 maximum Terms per Topic

k	$\boldsymbol{\beta}_{k,1}$	$\boldsymbol{\beta}_{k,2}$	$\boldsymbol{\beta}_{k,3}$	$\boldsymbol{\beta}_{k,4}$	$\boldsymbol{\beta}_{k,5}$
1	0.01633	0.01323	0.01241	0.00895	0.00850
2	0.01973	0.01469	0.01038	0.00733	0.00662
3	0.02985	0.02652	0.01010	0.00875	0.00844
4	0.02347	0.01672	0.01529	0.01184	0.01148
5	0.02134	0.00884	0.00864	0.00701	0.00634
6	0.01480	0.01146	0.01039	0.00958	0.00926
7	0.01597	0.01067	0.01026	0.00938	0.00878
8	0.01452	0.01258	0.01170	0.00952	0.00793
9	0.01945	0.01022	0.00802	0.00771	0.00716
10	0.02125	0.01955	0.01859	0.01516	0.00921

For user evaluation of the downloaded Twitter dataset, we organized an online survey and asked students associated with the University of Patras to evaluate the results. 10 annotators were used to read the downloaded posts and specify whether the extracted aspects are correctly identified and we compared the system results to the users' responses, which were used as gold standard for the evaluation of the system's performance. The purpose of this experiment is to examine whether our approach extracts and categorizes correctly aspects from a real-time data analytics platform. Users were presented with the communities wherein each community, the corresponding user with their tweets and the identified aspects, was considered.

[1] http://trumpet.di.ionio.gr/demo/

The percentages of corrected identified tweets are presented in following Table II. We examine three options: dense community, sparse community, and in-between. It can be observed that our proposed methodology achieves notable accuracy as 70% of aspects are correctly identified, whereas 10% of aspects cannot be considered as correct.

TABLE II
PERCENTAGES OF ASPECTS EXTRACTED FROM DOWNLOADED TWEETS

Category	Percentage of Aspects
Correctly Identified	70%
In-between Identified	20%
Wrongly Identified	10%

VI. CONCLUSIONS AND FUTURE WORK

To sum up, the problem of community detection and topic modelling was investigated. The user-based experimental evaluation on a Twitter dataset verified the efficiency of PSCAN algorithm to identify communities and of LDA to extract meaningful aspects.

The proposed framework can be extended as cultural inheritance tends to become global, cross language opinion mining becomes important. This strongly implies that besides linguistic factors, semantics, and the cultural context must also be considered in order to conduct meaningful aspect mining.

ACKNOWLEDGEMENT

This research was funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014-2020) under the Regional Operational Programme Ionian Islands 2014-2020, project title: "TRaditional corfU Music PresErvation through digiTal innovation - TRUMPET", project number: 5030952.

REFERENCES

[1] D. Alvarez-Melis and M. Saveski. Topic modeling in twitter: Aggregating tweets by conversations. In *10th International Conference on Web and Social Media (ICWSM)*, pages 519–522, 2016.

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

[3] E. Dritsas, G. Vonitsanos, I. E. Livieris, A. Kanavos, A. Ilias, C. Makris, and A. K. Tsakalidis. Pre-processing framework for twitter sentiment classification. In *15th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, volume 560, pages 138–149, 2019.

[4] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[5] S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer, 2015.

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

[7] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *3rd Workshop on Social Network Mining and Analysis (SNAKDD)*, pages 80–88, 2010.

[8] E. Kafeza, A. Kanavos, C. Makris, and D. K. W. Chiu. Identifying personality-based communities in social networks. In *Advances in Conceptual Modeling*, volume 8697, pages 7–13, 2013.

[9] E. Kafeza, A. Kanavos, C. Makris, G. Pispirigos, and P. Vikatos. T-PCCE: twitter personality based communicative communities extraction system for big data. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1625–1638, 2020.

[10] E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos. T-PICE: twitter personality based influential communities extraction system. In *IEEE International Congress on Big Data*, pages 212–219, 2014.

[11] A. Kanavos and I. Perikos. Towards detecting emotional communities in twitter. In *9th IEEE International Conference on Research Challenges in Information Science (RCIS)*, pages 524–525, 2015.

[12] A. Kanavos, I. Perikos, I. Hatzilygeroudis, and A. K. Tsakalidis. Integrating user's emotional behavior for community detection in social networks. In *12th International Conference on Web Information Systems and Technologies (WEBIST)*, pages 355–362, 2016.

[13] A. Kanavos, I. Perikos, I. Hatzilygeroudis, and A. K. Tsakalidis. Emotional community detection in social networks. *Computers & Electrical Engineering*, 65:449–460, 2018.

[14] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[15] J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *19th International Conference on World Wide Web (WWW)*, pages 631–640, 2010.

[16] X. Liang, Y. Lu, and J. Martin. A review of the role of social media for the cultural heritage sustainability. *Sustainability*, 13(3):1055, 2021.

[17] X. Meng, J. K. Bradley, B. Yavuz, E. R. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17:34:1–34:7, 2016.

[18] P. Mylonas, M. Wallace, and S. D. Kollias. Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering. In *3rd Hellenic Conference on Artificial Intelligence (SETN)*, volume 3025, pages 191–200, 2004.

[19] E. S. Negara, D. Triadi, and R. Andryani. Topic modelling twitter data with latent dirichlet allocation method. In *International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 386–390, 2019.

[20] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554, 2012.

[21] M. Plantié and M. Crampes. Survey on social community detection. In *Social Media Retrieval*, Computer Communications and Networks, pages 65–85. 2013.

[22] A. Pothen, H. D. Simon, and K.-P. P. Liu. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990.

[23] W. Silva, Á. L. de Santana, F. M. F. Lobato, and M. Pinheiro. A methodology for community detection in twitter. In *International Conference on Web Intelligence (WI)*, pages 1006–1009, 2017.

[24] A. Steinskog, J. Therkelsen, and B. Gambäck. Twitter topic modeling by tweet aggregation. In *21st Nordic Conference on Computational Linguistics (NODALIDA)*, volume 131, pages 77–86, 2017.

[25] Z. Tong and H. Zhang. A text mining research based on lda topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*, pages 201–210, 2016.

[26] G. Vonitsanos, A. Kanavos, A. Mohasseb, and D. Tsolis. A nosql approach for aspect mining of cultural heritage streaming data. In *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4, 2019.

[27] 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.

[28] F. Wang, K. Orton, P. W. III, and K. Xu. Towards understanding community interests with topic modeling. *IEEE Access*, 6:24660–24668, 2018.

[29] S. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta. Large-scale high-precision topic modeling on twitter. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1907–1916, 2014.

[30] 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 Symposium on Networked Systems Design and Implementation*, pages 15–28, 2012.

[31] W. Zhao, V. S. Martha, and X. Xu. PSCAN: A parallel structural clustering algorithm for big networks in mapreduce. In *27th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 862–869, 2013.