Real Time Twitter-based Encyclopedia

Vikramaditya Agarwala
School of Information Sciences
University of Illinois at Urbana-Champaign
va22@illinois.edu

Krishnaditya Kancharla
School of Information Sciences
University of Illinois at Urbana-Champaign
kpk4@illinois.edu

Prof. Dong Wang
School of Information Sciences
University of Illinois at Urbana-Champaign
dwang24@illinois.edu

***Abstract*— The purpose of encyclopedias is to provide knowledge on a variety of subjects. Encyclopedias, on the other hand, are soon out of date in the present era and frequently gloss over the distinctive social influence and nuances that each topic possesses in their endeavor to provide objective knowledge. One attempt at a solution is to use Twitter to create a real-time encyclopedia. Hashtags are a way for microblog entries to identify particular themes or phrases. Because they are used in an ad hoc manner, these hashtags are difficult to interpret and query because their use implicitly defines their meaning. One can use hashtags as encyclopedia entries in particular because they are widely used on Twitter to link tweets to specific themes. The application will scan recent, well-liked user tweets (perhaps from a selected geographic area) and create a current and social sensing definition of the intended hashtag when the user enters the desired hashtag.**

Keywords—Twitter, Clustering, Encyclopedia, Hashtags, Jaccard, Cosine Similarity

I. INTRODUCTION

Online social media's reach has grown at an unparalleled rate in recent years. As social networking services spread throughout more regions of the world and into a wider range of market categories. The importance of information that is collectively available to the public created on these online platforms substantially rises. Social media interaction and engagements frequently reflect real-world events and dynamics; as a result, social media streams become reliable sensors of real-world events as the user bases of social networks grow larger and more active in creating content about real-world events almost in real-time. A new type of communication called microblogging allows users to share brief status updates via instant messaging, mobile phones, email, or the Internet. Since its debut in October 2006, the popular microblogging platform Twitter has experienced rapid expansion. Between Q1 2017 and Q2 2022, Twitter has shown a consistent increase in the monthly daily active users (MDAU) from 109 million to 237.8 million users [1]. This data shows tremendous potential in the application of social sensing algorithms and topic detection.

In general, the term "social sensing" refers to a group of sensing and data gathering paradigms where information is gathered from people or objects on their behalf. In this paper, we aim to leverage social sensing in order to create a real time, dynamic encyclopedia which provides current contextual description of trending topics.

II. RELATED WORK

Aiello et al. used natural language processing and topic detection methods to identify trending issues or discussion topics on Twitter[2]. They devised and proposed methods to analyze the temporal distribution of topics for managing increasingly diverse streams with several storylines developing concurrently.

Soliman et al.[3] proposed a methodology which investigated 39 million geo-located tweets and the temporal patterns in the dataset in order to categorize important locations into generic land-use types.

Although we could not find a direct study or use case of an encyclopedia or a “dynamic dictionary” based off of real time trending topics on social media, sentiment Analysis on large scale tweets was a common use-case we could find during our research.

III. PROBLEM STATEMENT

In this study, we propose a system which generates a contextual description for a trending topic to summarize the latest information which has been put up on Twitter about it. To do this, we divided the problem statement into three parts.

A. Real Time Tweet Data Ingestion

Tweets about a pre-defined collection of topics are ingested in near real time using Twitter API v2 which enables programmatic access to tweets from a certain topic or hashtag. A software module is created to collect these tweets and send it for further processing.

B. Trending Topic Detection, Analysis & Ranking

With the help of low latency distributed computing, text processing techniques are used to analyze the trending hashtags. The hashtags are then ranked based on the respective number of tweets. A vital component of information monitoring and synthesis emanating from social sources is trending topic detection. The quality of the outcomes is significantly influenced by a wide range of techniques and variables.

C. Tweet clustering using text similarity techniques

Both topic modeling and clustering have the same traits in that they are based on unsupervised learning, need a predetermined number of topics or clusters, and do not require labels. Determining the number of topics/clusters is another significant issue with topic modelling and clustering techniques. Although several algorithms have been proposed to address the issue of figuring out how many clusters there are, it doesn't seem like there is a single approach that has been demonstrated to be the most trustworthy, probably because real-world datasets are so complicated. We explore the methodologies of tweet clustering using K-Means clustering, Jaccard Similarity Index and Cosine Similarity techniques for measuring distance.

IV. SOLUTION

A. TCP Socket communication between software modules

The techniques needed for inter-process communication between applications, whether they are deployed across a local system or a TCP/IP-based network environment, are provided via a socket programming interface. In order to uniquely identify a peer-to-peer connection after it has been created, a socket descriptor is used. A server and a client process communicate with one another using TCP sockets. The code for the server is executed first, and it opens a port and waits for client connection requests. The client or server may send a message after a client connects to the same (server) port[4]. When a message is sent, either the server or the client that receives it will handle it appropriately.

B. Porter Stemming & Lemmatization

English words with frequent morphological ends are removed using the Porter stemming method[5]. Its primary function is to normalize terms as part of the process of building up information retrieval systems, which is what it does. Porter's stemmer handles context in a unified manner.
The process of lemmatization involves determining a word's normalized form. The process is analogous to seeking out a transformation to apply to a word in order to obtain its normalized version.

C. Structured Data Processing using Distributed Computing

To design a scalable system with low latency we use distributed computing for the text processing and analytics. Data is handled with the help of dataframes. Building machine learning capability on top of the several "next generation" data flow engines that extend MapReduce (Dean and Ghemawat, 2004) have been intriguing problems in recent times. Multiple operations, like ‘split’, ‘explode’ and ‘groupby’ are conducted to process the information efficiently. A fault-tolerant and all-purpose cluster computing system called Spark offers Java, Scala, Python, and R APIs as well as an engine that has been designed to handle broad execution graphs[6]. Additionally, Spark is effective at iterative calculations, making it a good choice for the creation of expansive machine learning applications.

D. Jaccard Similarity and K-Means Clustering

When comparing two texts, the Jaccard Similarity co-efficient is used to evaluate how similar they are. This refers to how similar the context of the two texts is, or how many common words there are relative to the overall number of words. According to the definition of Jaccard similarity, the intersection of two texts divided by the union of those two documents refers to the proportion of common terms over all words[7].

Unsupervised learning algorithm K-Means Clustering divides the unlabeled dataset into various clusters. Here, K specifies how many pre-defined clusters must be produced during the operation. It gives us the ability to divide the data into several groups and provides a practical method for automatically identifying the groups in the unlabeled dataset without the need for any training. Each cluster has a centroid assigned to it because the algorithm is centroid-based. This algorithm's primary goal is to reduce the total distances between each data point and its corresponding clusters. In this study, we implement the K-Means Clustering algorithm for tweets by using Jaccard Similarity as the distance measure[8].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

E. Cosine Similarity

Cosine similarity is one of the metrics used in natural language processing to compare the text in two documents, regardless of their size. A word is visualized as a vector. The text documents are visualized as vector objects in an n-dimensional space[9].

The cosine of the angle formed by two n-dimensional vectors projected in a multi-dimensional space is measured mathematically by the cosine similarity metric. The range of the cosine similarity between two documents is 0 to 1. The orientation of two vectors is the same if the Cosine similarity score is 1. The closer the value is to 0, the less similar the two documents are.

$$\text{Cosine similarity } (|x\ .\ y|) = \frac{x\,y}{||x||\,||y||}$$

V. EVALUATION

Tweet Clustering Algorithm Convergence

We documented the number of iterations it took for the algorithm to converge to understand the minimum iterations required to generate a meaningful coherent and relevant description for a particular hashtag. The smaller the number of iterations, the faster the results are displayed for the user.

VI. RESULTS

Results were documented in a tabular format where details of every experimental iteration were noted. Along with the results from the table, we also checked the word strings being returned by the algorithm manually to gauge the relevance and coherence of the output.

Experiment No.	No. of Iterations	Value of K	Sum of Squared Error (SSE)
1	2	3	11.70
2	4	4	11.00
3	3	5	10.86
4	3	6	11.06

Fig. 1. Results of the K-Means Clustering using Jaccard Distance

We could observe that the minimum SSE was when K=5 and the number of iterations were 3. After manually checking, the word phrase output, we could see that the word strings were also coherent and providing relevant information about the keyword topic.

In this case the keyword entered was “worldcup”, in order to get the latest information about the FIFA World Cup 2022. The hashtag trending for this topic was found to be “#worldcup” and the relevant description output string was “Moroccan Fans raise a big Palestinian flag during the Morocco vs Belgium Match at the Fifa World Cup 2022”.

VII. DISCUSSIONS AND FUTURE WORK

A. Scope of Reliability Improvement

As is the case with every social sensing study involving crowdsourcing, reliability of the source information is a fundamental research challenge. Due to the lack of an underlying sample frame, the validity of data obtained through crowdsourcing is frequently questioned. Reliability can be improved by identify and removing tweet from bot accounts.

The challenging and promising class of CPS applications known as "human-in-the-loop cyber-physical systems" (HiLCPSs) enhances and facilitates human interaction with the physical world to improve reliability[10]. Energy management, healthcare, automobile systems, and disaster response are a few examples of these applications.

B. Capability to cover multiple storylines for single topic

There are several real-world cases or topics where multiple concurrent events occur within a single topic. One example would be Election results, where every state has polling result news being shared at the same time and users tweet multiple subtopics within the broader topic of election. Advanced NLP and topic detection techniques could potentially be leveraged for the encyclopedia to organize and display the information.

.

REFERENCES

[1] S. D. Published by S. Dixon and N. 11, “Twitter global mdau 2022,” *Statista*, 11-Nov-2022. [Online]. Available: https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwid. [Accessed: 12-Dec-2022].

[2] L. M. Aiello *et al*., "Sensing Trending Topics in Twitter," in *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268-1282, Oct. 2013, doi: 10.1109/TMM.2013.2265080.

[3] Soliman, A., Soltani, K., Yin, J., Padmanabhan, A., & Wang, S. (2017). Social sensing of urban land use based on analysis of Twitter users’ mobility patterns. *PloS one*, *12*(7), e0181657.

[4] Kalita, Limi. "Socket programming." *International Journal of Computer Science and Information Technologies* 5.3 (2014): 4802-4807.

[5] Willett, Peter. "The Porter stemming algorithm: then and now." *Program* (2006).

[6] Meng, Xiangrui, et al. "Mllib: Machine learning in apache spark." *The Journal of Machine Learning Research* 17.1 (2016): 1235-1241.

[7] Ferdous, Raihana. "An efficient k-means algorithm integrated with Jaccard distance measure for document clustering." *2009 First Asian Himalayas International Conference on Internet*. IEEE, 2009.

[8] Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. "The k-means algorithm: A comprehensive survey and performance evaluation." *Electronics* 9.8 (2020): 1295.

[9] Gunawan, Dani, C. A. Sembiring, and Mohammad Andri Budiman. "The implementation of cosine similarity to calculate text relevance between two documents." *Journal of physics: conference series*. Vol. 978. No. 1. IOP Publishing, 2018.

[10] Schirner, Gunar, et al. "The future of human-in-the-loop cyber-physical systems." *Computer* 46.1 (2013): 36-45.