Trendspedia: An Internet Observatory for Analyzing and Visualizing the Evolving Web

Wei Kang [#1], Anthony K. H. Tung [#*2], Wei Chen [†3], Xinyu Li [*4], Qiyue Song [*5]
Chao Zhang [‡6], Feng Zhao [*7], Xiajuan Zhou [†8]

[#] *NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore*
[1] kangw@nus.edu.sg
[*] *School of Computing, National University of Singapore, Singapore*
{[2] atung, [4] lixinyu, [5] qiyue, [7] zhaofeng}@comp.nus.edu.sg
[†] *College of Computer Science, Zhejiang University, P.R. China*
[3] chenwei@cad.zju.edu.cn, [8] zhouxiajuan@zjucadcg.com
[‡] *Department of Computer Science, University of Illinois at Urbana-Champaign, USA*
[6] czhang82@illinois.edu

***Abstract*— The popularity of social media services has been innovating the way of information acquisition in modern society. Meanwhile, mass information is generated in every single day. To extract useful knowledge, much effort has been invested in analyzing social media contents, e.g., (emerging) topic discovery. With these findings, however, users may still find it hard to obtain knowledge of great interest in conformity with their preference.**

In this paper, we present a novel system which brings proper context to continuously incoming social media contents, such that mass information can be indexed, organized and analyzed around Wikipedia entities. Four data analytics tools are employed in the system. Three of them aim to enrich each Wikipedia entity by analyzing the relevant contents while the other one builds an information network among the most relevant Wikipedia entities. With our system, users can easily pinpoint valuable information and knowledge they are interested in, as well as navigate to other closely related entities through the information network for further exploration.

I. Introduction

In recent years, social media services have been proliferating at an unprecedented speed. Hundreds of millions of users are attracted to participate in these virtual communities to communicate with each other, exchange ideas, update their statuses, etc. For instance, Twitter CEO Dick Costolo revealed in June 2012 that Twitter was seeing 400 million tweets per day[1]. With such an abundance of information generated every day, however, users are often overwhelmed in the data ocean and have no idea where to derive knowledge in which they are really interested.

To help users obtain useful knowledge, some social media services like Twitter provide trending keywords by analyzing features, e.g., hashtag frequencies. However, such preliminary analysis can still lead to bias. This concern is also supported by a discovery in [1] that the majority (over 85%) of trending topics are headline news or persistent news in nature, indicating that many other potentially interesting points are dominated by the globally hottest ones and thus become concealed forever.

[1]http://www.mediabistro.com/alltwitter/tag/tweets-per-day

Existing research efforts (e.g., [2], [3]) often try to obtain knowledge from mass social media contents by summarizing the data, extracting trending topics or even making predictions. Unfortunately, many people, including ordinary and corporate users, tend to have more interest in only topics or events within certain context instead of the globally significant ones. Although some researchers alleviate the problem by introducing constraints, such as allowing keyword filtering [4] and focusing on contents published in specific geographic areas [5], users may still find the discoveries pointless in terms of their preference. Consider, for instance, the scenario where a tourist is going to visit a place that he has never been to before, or an investor plans to buy stocks of a certain company. It would be greatly beneficial if both of them have access to some well organized and continuously updated knowledge regarding how other people talk about their targets, and furthermore, if they can also obtain such knowledge of a few more closely related entities via an information network.

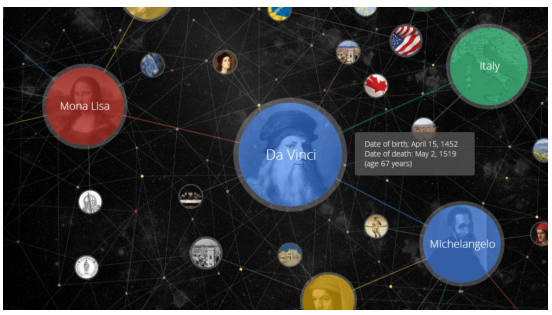

Fig. 1. Google Knowledge Graph

To solve the problem, we propose to bring proper context to social media contents which are streaming in from the Internet. We try to index these dynamic contents via Wikipedia, a well-established online encyclopedia which has entries for a large number of entities and concepts. Organizing Internet contents around Wikipedia also creates a new way to search for content on the Internet compared to conventional search

engines like Google. Emerging effort in the same direction is also adopted now by Google in the form of Google Knowledge Graph (cf. Fig. 1) that allows users to browse other related entities by linking relevant entities into a graph. However, unlike Google Knowledge Graph, we will deal with dynamic information related to each Wikipedia entity and try to extract knowledge from it with analytics and visualization tools for better exploration and understanding.

Based on the above methodology, we present a system called Trendspedia in this paper. Trendspedia aims to provide a collaborative Internet observatory platform for users to fetch and digest the information flow on the Internet with great ease. In Trendspedia, Wikipedia articles serve as a knowledge base and social media contents are continuously crawled and routed to the related Wikipedia articles for further analysis.

We introduce four data analytics tools in Trendspedia. The first three tools aim to enrich each target Wikipedia entity by extracting the hottest web contents, generating summaries and emerging events respectively through an analysis of relevant social media contents. The last one builds an information network that reflects the connectivity among relevant Wikipedia entities centered with the target. To enhance user experience, we visualize the analytics results so that users can explore them easily. For instance, summaries of the related tweets of a Wikipedia entity will be depicted as hierarchical tag clouds to allow users to view the summaries in an interactive manner.

In this manner, we effectively tackle the aforementioned challenge such that, users can not only pinpoint useful information and knowledge they really have an interest in around Wikipedia but also navigate to other closely related entities through the information network effortlessly. The contents are crawled against social media services by using the entity name of a Wikipedia article as the query string for filtering. Although we currently retrieve Twitter messages, i.e., tweets, as the major social media contents, we envision that other sources of contents can be easily incorporated into Trendspedia for more diversified analysis.

II. System Architecture

The architecture of Trendspedia is shown in Fig. 2, which consists of three components, including data storage, data processing and visualization.

The dashed box highlights the core component of Trendspedia, i.e., the data processing component, which provides ancillary services and performs different types of data analytics jobs. The ancillary services run in the background, preprocess raw data, collect statistics and assist the data analytics jobs to be done properly. The major ancillary services in Trendspedia include a Twitter Crawler, a Job Scheduler and a Tweet Analyzer. By default, the Twitter Crawler alternates to retrieve tweets containing the titles of different Wikipedia articles (i.e., entities) in a round-robin manner. Besides, the crawler also works periodically for an opened Wikipedia article, such that the more frequently an article is visited the more tweets related to the corresponding entity are crawled. The Job Scheduler maintains a job queue so as to run multiple Twitter Crawlers

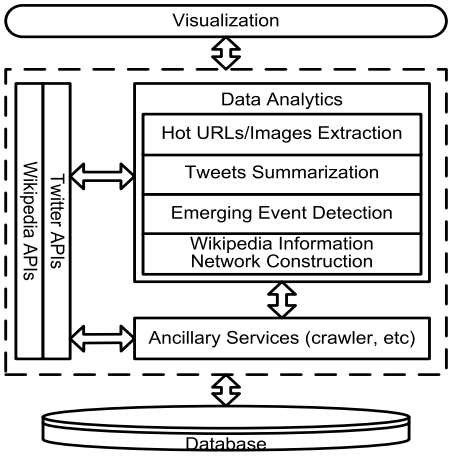

Fig. 2. System Architecture

simultaneously and to balance the query frequencies of the crawlers to avoid Twitter's API rate limit problem[2]. Having new tweets crawled for a certain entity, the Tweet Analyzer updates statistical information of the tweets continuously.

With the aid of the ancillary services, Trendspedia implements four useful data analytics tools to cater to the enrichment of Wikipedia entities. The tools are designed to provide each Wikipedia entity with (1) the most relevant and hottest URLs/images, (2) summaries of related tweets, (3) recently emerging events and (4) an information network connecting relevant Wikipedia entities together. Next, we introduce the data analytics tools in Trendspedia in more detail.

III. Data Analytics

A. Hot URLs/Images Extraction

As a growing number of tweets are crawled and attached to a certain Wikipedia entity, popular URLs that are often mentioned in those tweets can be identified such that the web contents linked to by the URLs can be retrieved and analyzed, in turn, to enrich the corresponding Wikipedia entity.

Although some URLs are different, the contents of their web pages might be similar or even exactly the same. This happens very likely especially when social media users share breaking news or emergent events, such as an Apple new product launch event and an earthquake, from popular news portals. To estimate the popularity and remove duplicates of such URLs with identical contents, the web pages are crawled and an analysis of similarity detection is conducted. Specifically, the content of a web page is first converted to a series of q-grams, which are then transformed into a vector of integers. Since the vectors of different web pages are often high-dimensional and sparse, we further compress them to have shorter length but still preserve the most important features by applying the min-wise independent permutations approach [6]. After that, the compressed vectors are compared against one another according to a similarity metric, such as the cosine similarity, to group URLs linked to similar web pages together.

[2]https://dev.twitter.com/docs/rate-limiting/1.1

We then sort the groups in terms of size in descending order, and choose one URL from each of the top ranked groups to form the hot URLs. The hot URLs are updated from time to time when more tweets with URLs are retrieved. Likewise, we extract hot images by analyzing the similarity among the color histograms of various images on the web pages.

B. Tweets Summarization

Although only relevant tweets are routed and attached to each Wikipedia entity, the continuously incoming tweets tend to discuss various aspects of it. This demands Trendspedia to be able to summarize the recently published tweets in order for users to get a multi-faceted understanding of what is going on as for a Wikipedia entity.

To this end, we propose a Formal Concept Analysis [7] (FCA) based approach for fast extraction of interesting summaries from a number of tweets. We first make use of the tweets to build a tweet-keyword matrix, where each element is either 1 or 0, indicating whether a tweet contains a keyword or not. Based on this matrix, our approach can efficiently generate a set of Formal Concepts. A Formal Concept is a sub-matrix containing a set of tweets and a set of keywords, where the keywords frequently co-occur in these tweets. The tweets in a Formal Concept are clustered together due to the common keywords they share, meaning that they are quite likely to discuss similar things. Thus, keywords in a Formal Concept naturally serve as a summary of the tweets.

Our tweets summarization approach has the following characteristics: (1) *Efficient*. Empirical experiment shows our approach runs at least an order of magnitude faster than the popular topic modeling method LDA [8]. (2) *Easy for understanding*. In a Formal Concept, the keywords and the tweets are generated simultaneously such that users can choose to view the relevant tweets if they want to explore more about the summary (i.e., keywords). (3) *Granularity customizable*. Similar Formal Concepts can be merged together to certain extent according to a user-specified density threshold, such that the resultant summaries are extracted from tweets that are of greater cohesion or diversity. (4) *Visually interactive*. We visualize the summaries as hierarchical tag clouds, which allows users to explore the summaries interactively by zooming in/out through the tag clouds.

C. Emerging Event Detection

Another useful feature of Trendspedia is its ability of analyzing tweet streams to detect emerging events for Wikipedia entities. The event detection tool enriches an entity by filtering out meaningless Twitter messages and highlighting important emerging events happening recently. For instance, when users are browsing the Wikipedia article of "Singapore", Trendspedia augments the page by listing recent events (e.g., concerts, celebration gatherings, etc) that have happened there. Such events are not readily available in the Wikipedia article, but can be mined out from the collection of relevant tweets.

Specifically, we extract the top-k emerging events by performing temporal analysis of relevant tweets in Trendspedia. The observation is that, with the outbreak of a certain event on a Wikipedia entity, the number of relevant tweets will increase sharply. As an example, when the tragic bombing hit Boston on 15 April 2013, many Twitter users were discussing this disaster online and a large number of tweets containing the keyword "Boston" were created overnight accordingly. Therefore, given a collection of relevant tweets that span over a time period $[t_s, t_e]$, we slice the time period by day and utilize the following two criteria to quantify the importance of each slice: (1) *Popularity*, the increase of tweet number compared to that of the previous slice, measuring how influential an event is; (2) *Freshness*, the time span from the slice to current time, measuring how recent an event is.

We linearly interpolates the normalizations of the above two measures to derive the score of each slice. Then, k slices with the largest scores are selected out. We consider these slices to represent k emerging events that have happened on this entity during $[t_s, t_e]$. To describe each of these k events, we further choose the top ten words that occur most frequently in the corresponding slice.

D. Wikipedia Information Network Construction

Since Wikipedia by itself is a collection of web pages linked to each other, we can perceive it as an information network, where nodes in the network represent Wikipedia entities while links indicate interconnectivity among different nodes.

In trendspedia, we extract and build a sub-network for each Wikipedia entity. Specifically, given a Wikipedia entity e, by analyzing the content of its corresponding article we construct a two-layer directed graph. Nodes in the first layer are Wikipedia entities mentioned in the content of e while those in the second layer are mentioned in the contents of the first-layer nodes. A directed edge between node a and b indicates a "contains" relationship of the two entities. Note that the resultant graph may contains loops since a node might be contained in other nodes in the same and/or a different layer. The graph we construct using Wikipedia entities is similar to the web page linkage graph, thereby enabling us to run PageRank [9] to allocate weights to different nodes.

By doing this, Trendspedia provides users a directed and weighted graph centered with an entity, the Wikipedia article of which they are reading. Although a Wikipedia API can return a list URLs of the relevant Wikipedia articles, the visualized information network in Trendspedia allows users to grasp the semantic importance and interconnectivity of relevant entities at a glance, such that it becomes much easier for them to decide where to navigate next.

IV. Demonstration

In our demonstration, we will exhibit the online real-time Trendspedia system, with all characteristics described in this paper, such that users can have an in-depth understanding of how Trendspedia integrates Wikipedia and Twitter message stream and performs data analytics as an internet observatory.

Users can log into Trendspedia via their Twitter or Weibo (the most popular microblogging service in China) accounts.

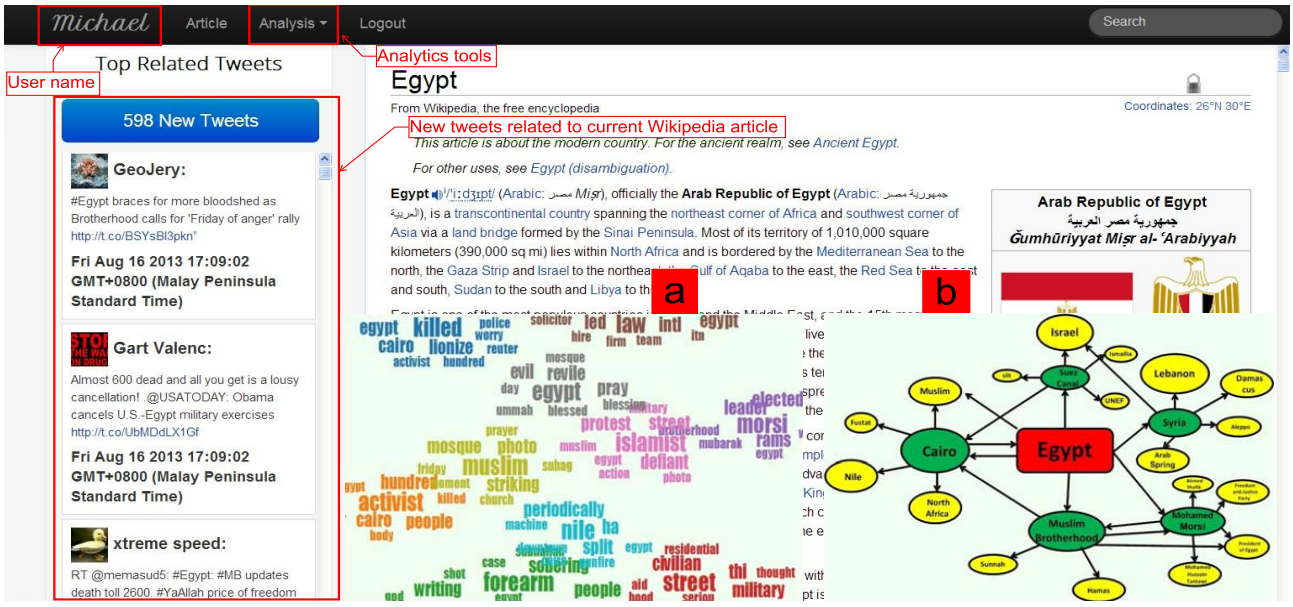

Fig. 3. Snapshot of the "Egypt" page in Trendspedia

After logging in, users can search and navigate to any page in the same way as they explore Wikipedia. Fig. 3 shows a snapshot of the "Egypt" page, where the right panel shows the corresponding Wikipedia article while the left panel presents the recently published tweets related to Egypt. As more relevant tweets are published, the number of new incoming tweets will be shown to users, such that they can read new tweets by pressing the big blue button. On the top are the user name of the account used to log into Trendspedia and three functional buttons. Users can click "Analysis" to open a dropdown list so as to see the visualized results (such as the small regions marked as (a) and (b) over the Wikipedia article in Fig. 3) of different analytics tools and click "Article" to switch back to the Wikipedia article.

Furthermore, we will introduce the system architecture and how different components are coupled with each other to support the fundamental data analytics functionalities in Trendspedia. We will explain in more detail how the data analytics tools are designed and what algorithms are adopted. In particular, we will show how we visualize the results of the analytics tools in order for users to understand the underlying discoveries in a vivid and interactive manner. Let's take "Egypt" for example. The Hot URLs/Images Extraction tool will display the hottest URLs (i.e., hyperlinks with the titles of web pages) and images related to Egypt, which are clickable for users to visit the original resources. The Tweets Summarization tool will present the tag cloud representation of extracted summaries after analyzing recently published tweets, where words in the same color stand for one extracted summary. As shown in the region marked as (a) in Fig. 3, the summaries, such as ("morsi", "brotherhood", "elected", "leader", ...) and ("killed", "cairo", "activist", "police", "reuter", ...), successfully capture and summarize the recent uprising and coup happening in Egypt in July and August 2013 from various aspects. Users can zoom in on the tag clouds interactively to view keywords with more specific meanings added in each summary. Differently yet complementarily, the Emerging Event Detection tool will provide a temporal perspective to exhibit recently bursty events along a timeline. Finally, the Information Network centered with "Egypt" visualizes the connections between "Egypt" and other entities, as shown in the region marked as (b) in Fig. 3. This allows users to explore knowledge of other closely related entities, such as "Cairo" and "Muslim Brotherhood", simply by clicking the corresponding nodes in the network.

ACKNOWLEDGEMENT

This work was partially supported by the FRC Grant R-252-000-486-112 and the SeSaMe Centre sponsored by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010, pp. 591–600.
[2] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *SIGMOD*, 2010, pp. 1155–1158.
[3] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *MDMKDD*, 2010, pp. 4:1–4:10.
[4] A. Guille, C. Favre, H. Hacid, and D. A. Zighed, "Sondy: an open source platform for social dynamics mining and analysis," in *SIGMOD*, 2013, pp. 1005–1008.
[5] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on twitter," *JASIST*, vol. 62, no. 5, pp. 902–918, 2011.
[6] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, "Min-wise independent permutations," *J. Comput. Syst. Sci.*, vol. 60, no. 3, pp. 630–659, 2000.
[7] J. Poelmans, P. Elzinga, S. Viaene, and G. Dedene, "Formal concept analysis in knowledge discovery: A survey," in *ICCS*, 2010.
[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.
[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999.