Eventera: Real-time Event Recommendation System from Massive Heterogeneous Online Media

Dongyeop Kang[1], DongGyun Han[1], NaHea Park[2], Sangtae Kim[1], U Kang[3*], Soobin Lee[1]
IT Convergence Laboratory[1], Department of Industrial Design[2], Department of Computer Science[3]
Korea Advanced Institute of Science and Technology (KAIST), Korea
[1]{dykang, handk, anne, soobinlee}@itc.kaist.ac.kr, [2]hellena1216@kaist.ac.kr, [3]ukang@cs.kaist.ac.kr

***Abstract*—Given massive heterogeneous online media, how can we summarize events, and discover causal relationships among them, in real time? Indeed we are living in a deluge of information; everyday hundreds of thousands of news articles are published, millions of postings from social media and internet forums are written, and billions of search queries are generated by Internet users. To convey user-interested news events and their big pictures for better understanding, building real-time event recommendation system is indispensable. Our proposed system, Eventera, aggregates massive online media from heterogeneous channels, summarizes them into events, discovers meaningful associations by bridging the events, and generates a sequence map of events that provides a big picture of how real life events interact with each other over time. We demonstrate how our system helps users understand events and their causal relationships effectively.**

Keywords-event recommendation; event detector; summarization; online media; sequence map

I. Introduction

We are living in an era of information overload. People read news articles online and write their opinions reacting to the articles through major news presses, portal sites, and social network services. If even more information about the topic is needed, they search related queries and find relevant web pages through search engines. We call this sequence of actions as an *active* media browsing activity that is triggered by users' interest. However, the user-triggered browsing activity sometimes requires too much effort until users obtain what they want. Moreover, the contents provided to the user from the browsing activity contain unnecessary or even biased information causing the browsing experience unpleasant and annoying. Our goal is to propose and demonstrate a *passive but personalized* media recommendation system to tackle the issues of *active* media browsing activities. In building a real-time media recommendation system, we face two critical challenges:

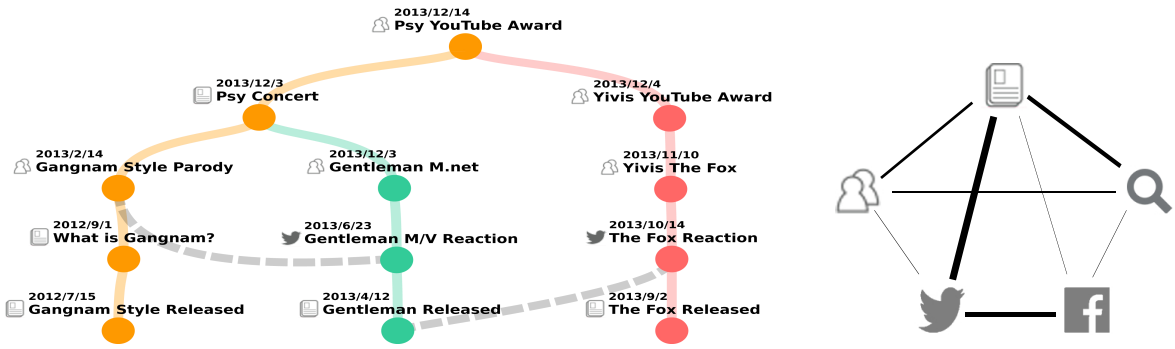
(1) Heterogeneous and massive online media. From our crawling experience, every single day hundreds of thousands of news articles (e.g., NYT), millions of online forum postings (e.g., Reddit) and social media postings (e.g., Twitter, Facebook), and billions of search queries (e.g., Google, Bing) are collected from heterogeneous media channels. To find and recommend useful information from the massive and heterogeneous online media, the challenge is to aggregate different types of media into a higher level of abstraction (called *event*) so that users can effectively understand their big picture and summarize them concisely by removing redundant information and extracting important information.

(2) Lack of association. Information overload also prevents understanding and focusing on a specific topic. Large volumes of news articles and people's reactions from social media often make us miss the big picture and the nature of the topic. Agenda setting [1], an ability of news media to influence the salience of topics of public's interests, is one of the problems caused by information overload: if a certain topic is frequently published in a short period of time, people would regard the topic as more important than it actually is. Abuse of agenda set by government or major presses happens very often in developing countries such as South Korea: an issue regarding "tax evasion by a famous politician" is often covered by other issues or gossips of sexual scandals of sports celebrities, eventually distracting people's attention from the political scandal. Here, the challenge is to associate events from diverse media so that users focus on a topic of interest, and discover hidden causal relationships for better understanding the big picture.

In this work, we propose Eventera, a real-time event recommendation system from massive heterogeneous online media. Eventera addresses the above challenges by (1) crawling massive heterogeneous online media from various channels in real time, (2) aggregating them into events, (3) mining casual relationships among the events, (4) generating their big pictures (*Sequence Map of Events* and *Interaction Map of Media* as shown in Figure 1), and (5) recommending the events to relevant users based on their profiles (e.g. in FaceBook) or past browsing histories. Using mobile and web versions of Eventera [2], we demonstrate how our system helps understand events more effectively, by tracking the origin of the events and discovering when and how they interact with other events from different media. Our major contributions are:

- **Aggregation and summarization.** From multiple sources of media channels, we crawl, aggregate and detect trending events. A few representative sentences are extracted using centroid based summarization techniques. We reduce dimensionality of hundreds of millions of heterogeneous online media into thousands of events, providing more significant and concise information to users.
- **Association.** To provide a big picture of how a certain event evolves over time and interacts with other events, we

*Corresponding author: U Kang (email: ukang@cs.kaist.ac.kr)



(a) *Sequence Map of Events* (b) *Interaction Map of Media*

Fig. 1: EVENTERA at work. (a) ***Sequence Map of Events***: for the "Psy Youtube Award" event, three different branches of events are generated over time: the orange branch for "Psy's Gangnam Style", the green branch for "Psy's Gentleman", and the red branch for "Yivis The Fox". The icon of each event shows the type of media that the event first appears: e.g., Twitter, news and communities. Solid and dotted lines mean causal relationships between consecutive events over time happened in a same branch and different branches, respectively. (b) ***Interaction Map of Media*** shows how different types of media channels (e.g., community, magazine, search engines, news, and social media) on a certain event interact with each other. The thickness of line shows the degree of interaction between the two media. For example, in (b) news and social media interact much more than others on the "Psy" event.

generate (1) *Sequence Map of Events* by bridging two events that are causally related, and (2) *Interaction Map of Media*. The associations help people understand deeper relationships between events over time from different media, and track the origin of the events.

II. DEMONSTRATING EVENTERA

We will demonstrate a user's media browsing experiences (recommendation, interaction, and exploration) with our mobile and web versions of EVENTERA system. Our scenario illustrates that when a real world event "Psy won top video of the year" happens, EVENTERA detects the event if frequency of the event rapidly increases at a certain time window, and recommends the event to a user who is interested in it. The user can further explore the event through our system. The event detection and recommendation algorithms will be explained in detail in Section III-B.

If an event is pushed or clicked, a user is directed to a summary page. The summary page first shows a representative title and a few number of important photos and sentences. The details of summarization and recommendation algorithm will be explained in Section III-B. If the user wants to explore details about the event, she can flip the summary page into the map page that contains sequence map and interaction map (Figure 1). The visualized maps help users obtain deeper insights about the event. The demonstration video is available in our project page [2].

The web version of EVENTERA system (see our project page [2]) is currently ready for use and exhibition in a demonstration. It crawls heterogeneous media in real time and recommends trending events with their summaries and sequence maps. The mobile version is ready and available, too, and we will officially release it in Google play and App Store within few months. We plan to bring a laptop running our software, a monitor, and a poster to describe the high-level ideas of our approach. We need a table, two chairs, an A0-sized (3 feet by 4 feet) poster board, and poster pins.

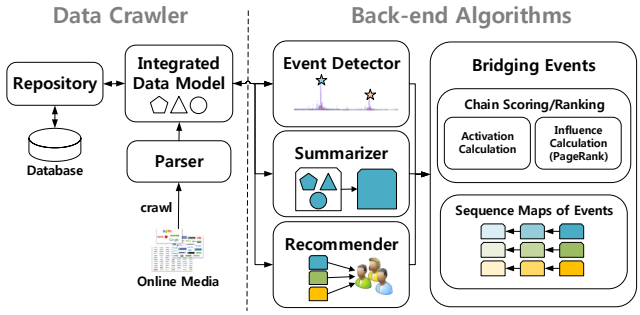


Fig. 2: Architecture of EVENTERA. The left side shows crawler: model, parser, and repository. The right side shows back-end algorithms: event detector, summarizer, recommender, and bridging events.

III. TECHNICAL DETAILS

A. Data Crawler

Due to the heterogeneity of media data to crawl, we focus on designing flexible and extendible crawler. Our crawler is implemented in Java and leverages commonly used frameworks such as Apache Maven, Spring Framework, and Hibernate. The crawler consists of three main components: *integrated data model*, *parser*, and *repository* (left side of Figure 2).

Integrated Data Model. Since there are various media services with different data models, it is hard to handle heterogeneous data models in an application. Therefore, we design an abstract but unified data model. The model contains all necessary information to recommend events, and integrates naming conventions from different services. For example, we use a variable "favorCount" to represent users' favor: "Like" in Facebook and "retweet" in Twitter. Through the integrated data model, events are retrieved by REST API calls.

Parser. It is labour-intensive to implement parsers for services with different layouts. Fortunately, a Korean portal site called Naver provides articles from various presses in the same layout. We implemented a real-time parser for Naver to crawl

meta information of news articles from 122 different presses. Since community services use table structured web pages, we also implemented a parser specialized to parse the table structured web pages. For each community we simply changed the name of table elements, and collected community postings from four major communities in Korea: HumorUniversity (HU), DcInside (DC), Ilbe, and TodayHumor (TH). To crawl Twitter and Facebook, we use open APIs of each service. To crawl search queries, we have crawled the top ten query results per ten seconds from two major search engines in South Korea, Naver and Daum, using their APIs. So far, we have collected 26 millions of news articles, 7.5 millions of social media postings, 790 thousands of forum postings and 2.5 millions of top search queries (e.g., Google Hot Trend) crawled per 10 seconds (Table I). The details and further references are in our project page [2].

TABLE I: The number of articles and postings from news media, social media, community forums, and search queries crawled by Eventera since June 2013. News media contain total 122 Korean presses, and social media contain Twitter and Facebook postings. HU, DC, Ilbe, and TH are major Korean community forums that have different political perspectives. Search queries are collected from two major search engines in Korea: Daum and Naver (k=thousand, m=million).

News	**Social**		**Communities**				**Search Queries**	
Presses	Tw	FB	HU	DC	Ilbe	TH	Daum	Naver
40m	11m	745k	259k	216k	211k	100k	3.56m	3.0m

Repository. The repository module manages database transactions. We implemented it as a singleton pattern class by leveraging Spring-framework for loose coupling and cohesion between classes. After the crawler parses data, the repository module saves data to the database. Then, the repository module is used in the back-end algorithms through the integrated data model. We ensure consistent database transaction APIs across Eventera service. In the performance perspective, we guarantee quick responsibility by leveraging 2nd level caching in Hibernate framework.

B. Back-end Algorithms

The right side of Figure 2 shows back-end algorithms used in Eventera: event detector, summarizer, recommender, and bridging-events.

Event Detector. The most important metric for recommending events is *real-timeness*: i.e., quickly detects events happening in real time. To detect trending events, we use one hour of time window so that articles, postings, tweets, and search keywords generated in the last one hour are chosen as input data. Then, the detector parses the sentences, tags Part-Of-Speech (POS), and extracts top noun phrases whose frequency rapidly increases compared to the previous time frame. To filter out commonly used noun phrases, we also calculate Inverse Window Frequency (IWF) (similar to Document Frequency) of each noun phrase and choose rarely occurring phrases whose IWF is high enough across whole time windows. The detector runs in every five minutes, and new events are added to the event repository. In the event repository, an event e consists of a keyword $w_{keyword}$ extracted by the detector, and corresponding sets of media sources: news articles m_{news}, tweets m_{tweets} and community postings $m_{communities}$ whose titles or bodies contain the keyword $w_{keyword}$:

$$e = \langle\ w_{keyword},\ \{m_{news},\ m_{tweets},\ m_{communities}\}\ \rangle, \qquad (1)$$

Another way of detecting event keywords is to directly utilize top search queries collected from search engines because they are already aggregated by the search engines as highly frequent search queries that are closely related to trending events in real world. In addition, we include some predefined keywords that are frequently used in articles and postings to represent the significance such as “breaking news”, “disaster”, etc. With these methods, we generate a list of event keywords per five minutes. Table II shows the number of detected events since June 2013. We reduce the 36 millions of activities from the massive and heterogenous media into 6,238 events.

TABLE II: The number of distinct events, and event chains with length greater than 2, detected in Eventera. The event chains are generated by the "Bridging Events" algorithm. The average length of an event chain is 3.7.

Events	Event chains (>2)
6,238	15,520

Summarizer. For each event, hundreds of articles and postings would be linked to the event. To convey concise summary of the event to users, the summarizer of Eventera needs to remove redundant information, and present only representative sentences. We adapt a simple but effective centroid based extraction method called MEAD from multiple documents [3]. MEAD scores sentences by ranking them according to a set of parameters: Centroid value (C_i), Positional value (P_i), First-sentence overlap (F_i), and Redundancy penalty (R_s). The score function for the ith sentence s_i is as follows:

$$SCORE(s_i) = w_c C_i + w_p P_i + w_f F_i - w_R R_s, \qquad (2)$$

where w_c, w_p, w_f and w_R are weight values (see [3] for details). Finally, Eventera extracts three to four top scored sentences in the summary page.

Recommender. The most famous recommendation technique is collaborative filtering (CF), which recommends items that similar groups of users are interested in. However, CF generally suffers from the “cold start” problem: users or items have not been gathered sufficiently. We propose a regularized CF algorithm on the event keyword list to recommend events to even new users without any historical data: Eventera recommends (1) globally famous events first until user logs are generated in our database or (2) relevant events for a user if the user has profile information (e.g., Facebook profile) containing keywords of the events.

Bridging Events. To associate related events, we generate a sequence map of events. Unlike previous works on associating news articles [4] or scientific papers [5] which focus on a single type of media, Eventera bridges higher level of aggregated media, *events*, that consist of heterogeneous types of media. Given a chain of n events $(e_1, ..., e_n)$, Eventera finds

TABLE III: Comparison of proposed EVENTERA with the state-of-the-art methods. EVENTERA is the only method that provides all the functionalities. (~ denotes 'unknown' or 'not directly applicable'.)

	EVENTERA	FF [6]	Kathy [7]	KeySee [8]	EventMap [5]	Presses
Data/Model						
Massiveness	✓	~	~	~	~	~
Heterogeneity	✓	news	Twitter	Twitter	~	news
Mining						
Summarization	✓	LDA	~	✓	~	~
Detection	✓	LDA	temporal	✓	~	~
Causality	✓	~	~	heuristic	✓	~
Personalization						
Recommend.	✓	~	~	~	~	~
Realtime	✓	~	✓	~	~	✓

coherent chains of the events by maximizing the objective function $Coherence(e_1, ..., e_n)$ which exploits the *influence* function. The *influence* function $Influence(w \mid e_i, e_{i+1})$ of word w from an event e_i to an event e_{i+1} is given by the PageRank value of the word in the bipartite graph between words and events. Therefore, $Influence(w \mid e_i, e_{i+1})$ is high if (1) both two events are highly connected, and (2) w is important for the connectivity. Instead of summing $Influence$ over all words, only a small set of *active* words are considered to obtain more coherent chains. Finally, our objective function is as follows:

$$Coherence(e_1, ..., e_n) = \max_{activations} \min_{i=1...n-1} \sum_{w} Influence(w \mid e_i, e_{i+1}) \mathbb{1}(w \text{ active in } e_i, e_{i+1}) \quad (3)$$

Figure 1 (a) shows an example of *Sequence Map of Events* using the seed event "Psy Youtube Award". The first orange branch shows Psy's Gangnam style syndrome on news, search engines, and communities. The second green branch is about Psy's another song "Gentleman": interestingly, active responses on social media about the music video causes Psy's next performance on M.net channel. The orange and green branches then merge into an event "Psy Concert". The most interesting part is an association between "Psy Youtube Award" and "Yivis Youtube Award" events. Both music videos are hilarious and funny, hence many reactions and parodies are made in online media. In this sense, people compare "Psy" with "Yivis", and thus the association occurs. Sometimes, the causal relationship happens between events not only from a same medium but also from different media channels. In the first branch, "What is Gangnam" event on a search engine happens after "Gangnam Style Released" event on news media (solid line). Moreover, the "Gentleman MV Reaction" event on social media in the second branch affects the "Gangnam Style Parody" event on the communities in the first branch (dotted line). Such interaction between different media help people track the origin of event more effectively. *Sequence Map of Events* provides a big picture of how certain event evolves over time, and how different media such as traditional news media and online social media interact with each other.

IV. RELATED WORKS

In this section, we compare EVENTERA with previous event recommendation systems. Table III shows the comparison of EVENTERA with the state-of-the-art methods.

Forex-foreteller (FF) [6] predicts movement of foreign currency markets on news articles, and Kathy et al. [7] developed a real-time disease surveillance system on Twitter data. However, they are limited to only a single domain such as finance or medical events, and thus miss detailed associations of heterogeneous events. KeySee [8] groups posts into event networks and tracks six types of evolution patterns. However, the heuristic method for tracking the relationship of events may not effectively discover causality between news articles, while our "bridging events" approach does even from heterogeneous media sources. To discover relationships between either articles or papers, "connecting the dot" problem [4], [5] has been studied. Again, they focus on only one type of media while our EVENTERA finds inter-relationship between *heterogenous* media channels. In the service perspective, there are many news related services: major presses (e.g., NYT), summarization (e.g., Summly), and curation (e.g., FlipBoard). However, none of the existing services provide event-level recommendation from heterogeneous channels of media, and analytical functions such as sequence map.

V. CONCLUSION

We propose and demonstrate EVENTERA, a real-time event recommendation system from massive heterogenous sources of media such as social media, news, search ranks, and community forums. Our crawler has been collecting over 36 millions of articles and postings since June 2013; currently 6.3 thousands of real world events are detected, and 15.5 thousands of their causal relationships are inferred. Through the web and mobile versions of EVENTERA, relevant events are recommended to users so that they explore big pictures of events and their association through sequence-map and interaction-map. To the best of our knowledge, EVENTERA is the first event recommendation system that aggregates massive, heterogeneous media and provides associations and summarizations for understanding the big picture of events and their relationships.

ACKNOWLEDGEMENT

This work was partly supported by the ICT R&D program of MSIP/IITP [1391104004, Development of Device Collaborative Giga-Level Smart Cloudlet Technology] and the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MSIP)(No. 2013R1A1A1064409).

REFERENCES

[1] M. McCombs and A. Reynolds, "News influence on our pictures of the world." 2002.

[2] Project Homepage: http://www.cs.cmu.edu/~dongyeok/project/Eventera.

[3] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents." Elsevier, 2004.

[4] D. Shahaf and C. Guestrin, "Connecting the dots between news articles," in *KDD*, 2010.

[5] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *KDD*, 2012.

[6] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan, "Forex-foreteller: Currency trend modeling using news articles," in *KDD*, 2013.

[7] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: Demonstration on flu and cancer," in *KDD*, 2013.

[8] P. Lee, L. V. Lakshmanan, and E. Milios, "Keysee: Supporting keyword search on evolving events in social streams," in *KDD*, 2013.