© 2015 Min Li

EKNOT: EVENT KNOWLEDGE FROM NEWS AND OPINIONS ON TWITTER

BY

MIN LI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Adviser:

Professor Jiawei Han

ABSTRACT

We present the EKNOT system that automatically discovers major events from online news articles, connects each event to its discussion on Twitter, and provides a comprehensive summary of the events from both news media and social media's point of view. EKNOT takes a time period as input and outputs a complete picture of what happened and the public's opinions. For each event, EKNOT provides multi-dimensional summaries: a) a summary from news for an objective description; b) a summary from tweets containing opinions/sentiments; c) an entity graph which illustrates the major players involved and their correlations; d) the time span of the event; and e) an opinion (sentiment) distribution.

A user-friendly interface is provided to facilitate interactive exploration of the mining results: if a user is interested in a particular event, he/she can zoom in this event to investigate its multiple aspects (sub-events). These aspects will be summarized in the same way with the above features.

Furthermore, EKNOT is built on real-time crawled news articles and tweets. The efficient data collection and processing scheme allows users to explore the dynamics of major events from the perspectives of both news and social media in near real-time.

To my parents, for their love and support.

Acknowledgments

I would like to give special thanks to my advisor, Dr. Jiawei Han, for his great support and inspiring advice on both my research and career.

I would like to thank Jingjing Wang for her great support and help during this work.

I would also like to thank other colleagues, Wenzhu Tong, Hongkun Yu, Xiuli Ma, Yucheng Chen, Haoyan Cai, for their advice and helpful discussion on my research topic.

Table of Contents

Chapter 1 Introduction 1

Chapter 2 Related Work 3
2.1 Event and Aspect Discovery 3
2.2 Text Summarization 4
2.3 Opinion Mining 5
2.4 News-Tweet Systems 6

Chapter 3 System Overview 7
3.1 Backend Architecture 7
3.2 Frontend Pages 8

Chapter 4 Major Functional Modules 11
4.1 Data Collection 11
4.2 Event and Aspect Discovery 12
4.2.1 Problem Formulation 12
4.2.2 The Generative Model 13
4.2.3 Tweets Linking 16
4.3 Joint Summarization 17
4.3.1 Co-Ranking News and Tweets to Obtain News Summaries and Tweet Highlights 18
4.3.2 Entity Graphs 19
4.3.3 Time Spans 19
4.4 Opinion Analysis 19
4.4.1 Problem Formulation 20
4.4.2 A Two-Step Classification Framework 20

Chapter 5 Case Study 22

Chapter 6 Conclusion 27

References 28

Chapter 1

Introduction

In this big data era, massive information from news media and social media is more easily accessible than ever. News agencies deliver real-time status of major events, covering every aspect with fairly formal news articles. Meanwhile, rather than passively receive the information, people post their comments about news and raise heated discussions via social media sites such as Twitter. The close relation and interaction between news and social media have long been observed. News inspires most discussions on social media (over 85% topics on Twitter are headlines or long continuing news)[1], while the social media brings public insights into news reports and may even drive the news agencies' focus. Additionally, the different natures of these two sources provide a complementary view of an event: a reasonably objective and accurate presentation of an event and a view full of opinions and sentiments from the public. Summarizing them jointly to provide a complete picture of an event can be of great interest to both policy makers and ordinary people seeking information.

To this end, we develop a system named EKNOT which effectively retrieves and discovers major events from news and connects each event to its discussions on Twitter. Given a time period, we intend to answer the following questions:

1. What are the major events?
2. Who are the key players in each event?
3. How do people talk about each event and what are their opinions?
4. When is the event and how long does the event last?
5. What are the multiple aspects (sub-events) if the event is rather big and influential? And what are the answers to the above questions for each aspect?

The contribution of the EKNOT system can be summarized as follows:

1. We build the EKNOT system that automatically retrieves real-time data from online news agencies and collects corresponding social reflections (tweets) on Twitter for each news.

2. The EKNOT system mines social events and aspects of events out of news media and social media sources per user requests. To the best of our knowledge, EKNOT is the first system that is built on event level and summarizes both sources simultaneously.

3. EKNOT provides users with comprehensive and diverse news/tweets summaries. Besides pure text, it also offers various clear and information-rich representations of the mined events to facilitate content digestion and understanding.

4. We explore the state-of-art opinion mining techniques and implement an efficient and effective opinion mining module under the proposed framework.

The rest of the paper is organized as follows. In Chapter 2, we discuss the related work of each key technical component and compare existing systems. In Chapter 3, we give an overview of the EKNOT system including the back-end architecture and front-end presentation logic. Major functional modules are introduced in Chapter 4. In Chapter 6, we provide case studies with real-world examples. Finally, Chapter 6 concludes this thesis.

Chapter 2

Related Work

Our system consists of three key technical components: event and aspect discovery, joint summarization and opinion mining. In this chapter, we review the recent literature related to these technical areas and our proposed news-tweet scenario. We will also discuss about our technical choices.

2.1 Event and Aspect Discovery

The event and aspect discovery problem can be viewed as a topic discovery process from raw documents. There has been a substantial amount of research on the topic modeling problem. Following the classic models Probabilistic Latent Semantic Indexing[2] and Latent Dirichlet Allocation[3], different topic models are developed to deal with more sophisticated modeling problems such as dynamic topic models [4, 5], entity topic models [6, 7] and hierarchical topic models [8]. We choose to take the approach proposed in our recent paper [9] which integrates multi-dimensional information for event/aspect learning with efficiency consideration. Because the unseen document is not a problem in our case, this model leaves out the priors in LDA and take a pLSI style for efficiency. We will discuss this method in detail in Chapter 4. There are also numerous works in literature focusing on event discovery on Twitter [10, 11, 12, 13]. Our work distinguishes itself from this line of work by involving the news corpus in the framework. We discover events and event aspects by investigating news articles and then retrieves reflections on Twitter.

Besides the methodology for event discovery, our proposed problem scenario is closely related to some joint studies of news media and microblogs. Zhao *et al.* [14] conduct a comparative study between traditional media and Twitter. By running separate topic models, they compare the topic categories (arts, sports, health, etc.) and topic types (event-oriented, entity-oriented, etc.) on the two sources.

They prove that, not surprisingly, Twitter and traditional news media cover a similar range of topic categories but different topic category distributions. Subasic *et al.* [15] further investigate the role of Twitter in news reporting. They claim that the major difference of the two sources lies in the sentiment aspect and conclude that Twitter users are more like commentators, not only reporting but expressing opinions and taking positions on the news. These two works clearly justify both the motivation and significance of our work. Inspired by the similarity and difference of the two sources, Gao *et al.* [16] employ cross-collection LDA (ccLDA) [17] to discover events and then construct sentence level joint summarization. Wei *et al.* [18] extract news highlights by utilizing tweets of the events. Compared to their work, our proposed method is more comprehensive, in the sense that we explicitly incorporate entities and temporal information into the model. Also, by utilizing a dynamic hierarchical topic discovery model, we provide more fine-grained perspectives by allowing users to go beyond the event level and investigate aspects of the event.

2.2 Text Summarization

The problem of automatic general text summarization has been an active research domain for decades. Mainly two categories of approaches exist in the traditional summarization techniques: extractive approaches and abstractive ones. In extractive type summarization, summary sentences are directly identified and selected from the orignal documents. On the other hand, abstractive summarization techniques will further process the document sentences or even generate new ones. Our work focuses on the extractive summarization. Various methods [19] have been proposed in this domain including feature based methods [20], cluster based methods [21], graph based methods [22] and knowledge based methods [23].

For the Twitter corpus, Inouye *et al.* [24] compare general summarization algorithms on Twitter and conclude that some sophiticated traditional algorithms give results worse than random choice and their simple frequency based summarizer Hybrid TF-IDF produces the best result. This work reveals the distinctiveness of Twitter corpus in the summarization problem. In [25], Sharifi *et al.* propose the Phrase Reinforcement algorithm to summarize the tweet posts of the same topic. Yajuan *et al.* [26] formulate tweet ranking in a unified mutual reinforcement graph to select sentences for extractive summarization. The Sumblr system

[11] proposed by Shou *et al.* utilizes a greedy algorithm to select representative tweets for summaries. Besides the importance of individual tweet, their method also takes summary diversity into consideration by adopting Maximal Marginal Relevance (MMR) [27].

For our proposed news-tweets scenario, Gao *et al.* [16] explore into the generalized Co-HITS algorithm [28] to perform co-ranking of the two different sources. However, purely relying on the co-ranking method makes the final summary sentences similar and even identical. In our EKNOT system, we combine the Co-HITS algorithm and the Maximal Marginal Relevance principle to generate summaries containing sentences with both good quality and diversity. We also employ some other scoring components such as opinion scores to accommodate the characteristics of the Twitter corpus. We will further elaborate the technique details in Chapter 4.

2.3 Opinion Mining

Opinion Mining, also known as Sentiment Analysis, has long been a heated research topic. It has been handled at many different levels of granularity including document level [29], sentence level [30] and phrase level [31]. Due to its noisy and informal nature, the Microblog data like Twitter corpus pose new challenges to this domain. Also, different from reviews, the ground truth of the sentiments on Twitter is hard to obtain without human labeling. Go *et al.* [32] propose to use distant supervision to acquire sentiment data with ground truth. Basically, positive and negative emoticons are used as noisy sentiment labels for the tweets retrieved. This method has been widely adopted in later study [33]. By formulating the problem as a classification task, popular classifiers such as Naive Bayes, MaxEnt and Support Vector Machines (SVM) with different feature sets (unigram, bigram, part-of-speech (POS) features, etc.) are experimented [32, 34]. It is worth noting that in [35], Barbosa *et al.* propose a two-step sentiment analysis classification method, which first classifies messages as subjective and objective, and then identifies the positive and negative tweets among the subjective tweets. They claim that such cascaded classifiers provide more effective and robust classification results from the biased and noisy data. More recent works on this topic employ sophisticated features such as POS-specific prior polarity features and tree kernels [36]. In the implementation of our EKNOT system, we do not incorporate

such complex features considering system efficiency. Even the simple maximum-entropy based part-of-speech tagging on a large amount of tweets requires long time computation and will eventually harm user experience of the system.

2.4 News-Tweet Systems

To the best of our knowledge, there are not many existing systems to address our proposed problem in the news-tweet scenario. In [37], Guo *et al.* propose to find the most relevant news articles to enrich a given tweet. Tsagkias *et al.* [38] focus on discovering implicitly linked social media utterances to news. Their framework derives multiple query models from a given news source and then retrieves and optimizes tweets from a target social media index. Shi *et al.* [39] follow this philosophy and develop the Insight4News system to connect news to relevant social conversations on Twitter. However, either a single tweet or a single news article has limited expressing power even after being enriched by the retrieved counterpart. Many news and tweets are in nature related, possibly complementary or redundant. Even for quite discrete news topics like car accidents, grouping similar individual news makes sense and may provide deeper insights. Thus, instead of retrieving counterpart of individual news or tweet, EKNOT is built on event level and summarizes both sources simultaneously, which clearly brings more informative and comprehensive event knowledge for users to digest.

In addition, we notice that a major event can have multiple aspects. For example, during the `Sony Pictures Entertainment Hack` [1] event happening around December 2014, several related incidents and reactions of different parties form different aspects of the event: a) the initial attack to Sony b) the film "*the Interview*" and its release; c) North Korean's Internet outage; d) U.S. accusations against North Korea. Each aspect has different focus both in the sense of key players involved and the diverse public opinions. Therefore, the mining process should be able to distinguish different aspects for each event to present a holistic view.

[1] http://en.wikipedia.org/wiki/Sony_Pictures_Entertainment_hack

Chapter 3

System Overview

3.1 Backend Architecture

Figure 3.1 illustrates the system architecture of EKNOT , which contains four major modules: data collection, event discovery, tweets linking and joint summarization. The input is a time period and the output is summaries for events and aspects.

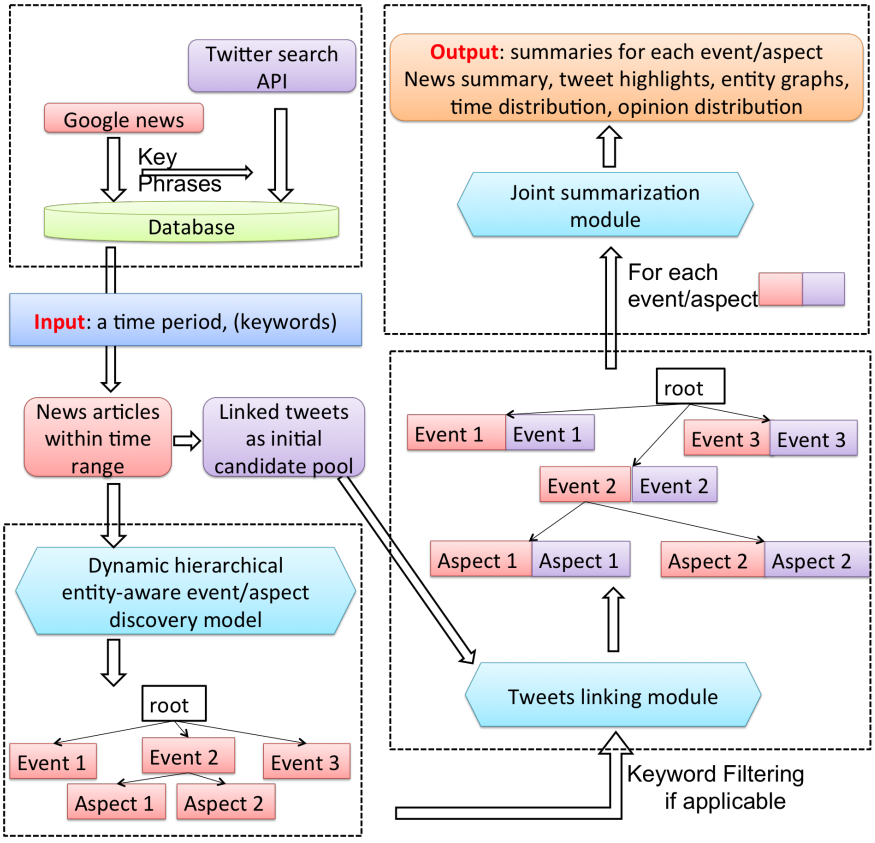


Figure 3.1: System Architecture. Throughout this paper, the red background indicates news side and the purple background indicates Twitter side.

EKNOT constantly crawls data from Google news. The key phrases extracted from each news title/snippet are used to query twitter API to obtain an initial pool of relevant tweets. Entities of the type person, organization or place are extracted from the news articles using NLP tools. Given a time period, EKNOT discovers major events/event aspects from news, in light of the high quality and broad coverage of news articles. This is achieved by employing a dynamic hierarchical entity-aware event/aspect discovery module proposed in our recent work[9]. Each event is then charaterized by an event descriptor which contains a word distribution, a time distribution and three entity distributions with respect to person, organization and place. Afterwards, the learned event descriptors are utilized to select the relevant tweets for each event and to analyze people's opinions from Twitter. At last, a joint summarization module leverages the descriptors, news articles and selected tweets to construct a news summary, tweet highlights, an entity graph, a time span and an opinion distribution for each event. Our event discovery module is instantiated by hierarchical topic modeling in a recursive manner, which allows users to zoom into a particular event interactively. Users can further investigate the event of interest and get the same style of summary for each aspect of the event.

It is worth noting that EKNOT also allows users to specify the keywords of the events they are interested in. In such case, the system searches the event clusters for these keywords after the initial topic discovery process and then only selects relevant events for further processing. This not only meet user common needs but also greatly reduces the processing time and improves the interactivity of the system.

3.2 Frontend Pages

EKNOT is designed to be a comprehensive system to provide users with event knowledge from the news-tweets sources. Currently two major page flows are provided: News/Tweets Overview and Event Discovery.

News/Tweets Overview pages allow users to check the original data corpus (the news-tweet data) before digging into event discovery.

Figure 3.2 shows the News Overview page which is available for every piece of news in the collection. Ordering and searching functions are also provided. The advanced search function enables users to search by id/keyword/title/news source.

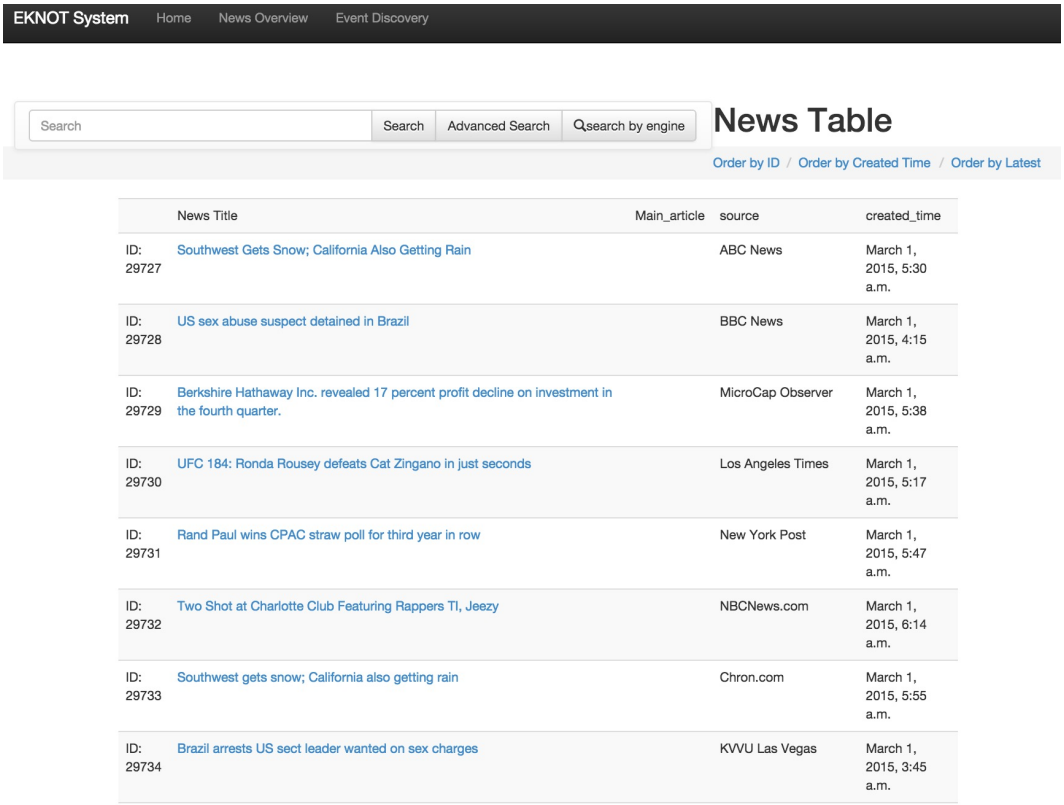


Figure 3.2: Frontend News Overview Page

By clicking on the individual news listed in the news overview page, users will be guided to the event detail page shown in Figure 3.3. This page gives users a detailed view of information related to the news including news title, time, source, orginial URL, keywords and main article. Also, tweets related to this news are provided as a raw list without ranking.

The event discovery process is shown in Figure 3.4. The entire process is designed to be simple and elegant with only necessary user controls to avoid confusion. As shown in Figure 3.5, on the input page, users can choose the time range and keywords (optional) they are interested in for event discovery. Users will then be guided to event overview page which shows the event discovery results. On this page, users can click on the individual event to further investigate its detail and aspects of the event. A detailed user case will be shown in Chapter 5.

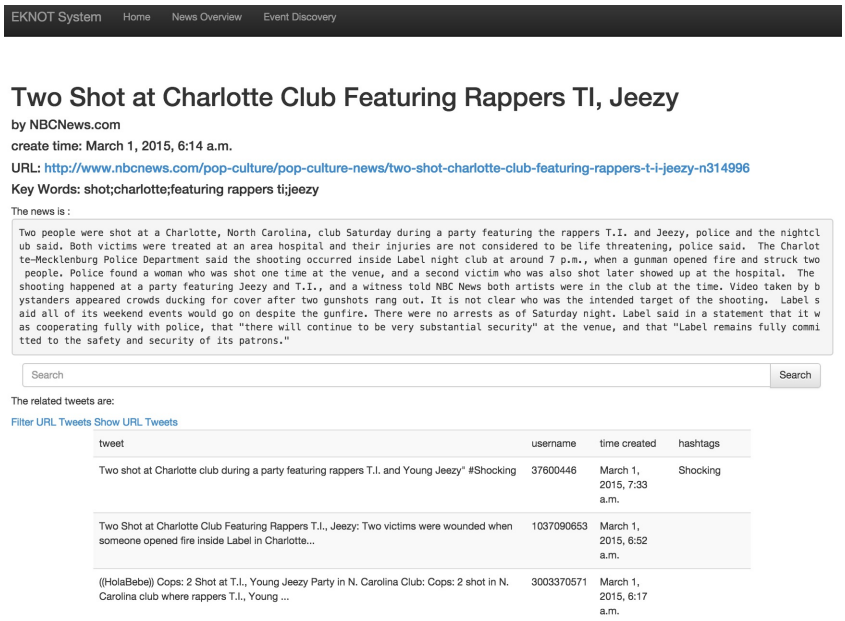


Figure 3.3: Frontend News Detail Page

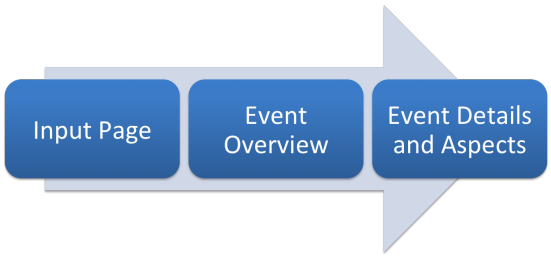


Figure 3.4: Event Discovery Page Flows

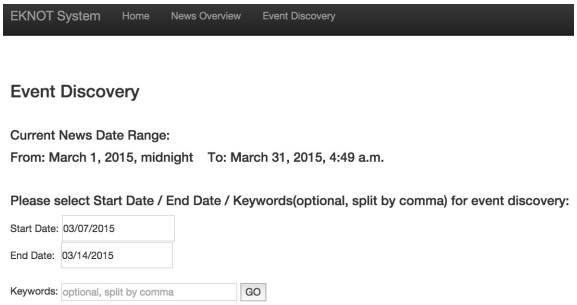


Figure 3.5: Frontend Event Discovery Input Page

Chapter 4

Major Functional Modules

4.1 Data Collection

EKNOT crawls the top stories from Google News[1] every 30 minutes. For each news crawled, we extract named entities from news title and the snippet by DBpedia Spotlight[2] with a confidence score of 0.3. If not enough named entities are identified, TextBlob[3] will be used to extract noun phrases. The system then queries the Twitter Search API[4] with extracted named entities or noun phrases. Tweets posted within one day after the news and containing at least two of the entities or noun phrases are returned. We observe that a substantial number of tweets contain URLs to news articles and their contents are the same as the news titles, which do not provide much additional information and opinions. We skip those cases and consider the tweets without URLs only.

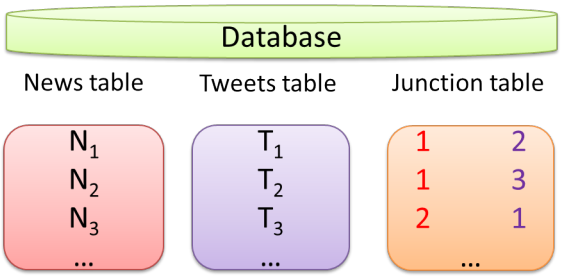


Figure 4.1: Database

Finally, news and related tweets, together with the named entity information (confidence greater than 0.5), are stored in our database which consists of a news table, a tweets table and a junction table recording the many-to-many mapping

[1] https://news.google.com/
[2] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[3] http://textblob.readthedocs.org/en/dev/
[4] https://dev.twitter.com/rest/public/search

between news articles and tweets, as illustrated in Figure 4.1.

4.2 Event and Aspect Discovery

A dynamic hierarchical entity-aware event/aspect discovery framework proposed in our recent work[9] is tailored to learn the optimal representation of news events and their multiple aspects in EKNOT . EKNOT first clusters the entire collection into events. Then if a user is interested in an event, the model decomposes the entire input collection and applies the same method to obtain the aspects of the event. In this section, we will illustrate this method in detail.

4.2.1 Problem Formulation

To formally formulate the problem, we give the necessary notations in Table 4.1 and use the following definitions.

Definition 1 (A Piece of News) *A piece of news is defined by a bag-of-words/entities model of the original news title and article with its timestamp. The entities are limited to the category of person, location or organization which are extracted from the news content.*

Definition 2 (Event/Aspect) *Events and aspects are both topics in their essence despite their different granularity. They are defined by textual, entity and temporal dimensions. Formally, an event/an aspect is defined by* $z = (\phi^{\mathtt{w}}, \{\phi^{\mathtt{e}}\}, \mathcal{N}(\mu, \sigma))$*, where 1)* $\phi^{\mathtt{w}}$ *is a multinomial distribution for words; 2)* $\{\phi^{\mathtt{e}}\}$ *is a set of multinomial distributions for entities.* $\mathtt{e}$ *can be of type person, location, or organization. We use* $\phi^{\mathtt{p}}, \phi^{\mathtt{l}}, \phi^{\mathtt{o}}$ *respectively; and 3)* $\mathcal{N}(\mu, \sigma)$ *is a Gaussian time distribution.*

Thus, our input data for this step are a collection of news within a time range which can be represented by a triple ($\mathbf{X}^{\mathtt{w}}$, $\{\mathbf{X}^{\mathtt{e}}\}$, $\mathbf{t}$), where

1. $\mathbf{X}^{\mathtt{w}}$ is a $|V^{\mathtt{w}}| \times |D|$ matrix and each entry $x^{\mathtt{w}}_{wd}$ denotes the number of times w-th word occurs in d-th news;

2. $\{\mathbf{X}^{\mathtt{e}}\}$ are three $|V^{\mathtt{e}}| \times |D|$ matrices. Each represents person, location, organization respectively. Similarly $x^{\mathtt{e}}_{ed}$ denotes number of occurrences the e-th entity occurs in the d-th news;

Name	Description
d	a piece of news
D	a news collection
$T(d)/T(D)$	the tweet collection of a piece of news / a news collection
$V^{\mathtt{w}}$	the word vocabulary of a given collection
V^{e}	the entity vocabulary of a given collection
$N^{\mathtt{w}}$	number of words in d
$N^{\mathtt{e}}$	number of entities in e
K	number of topics (events/aspects)

Table 4.1: Notations

3. $\mathtt{t}$ is a vector of timestamps where t_d denotes the timestamp of the d-th news.

With such input, our goal is to discover events if the collection is of multiple events or to mine aspects of a event if the news collection is of a same event topic.

4.2.2 The Generative Model

Given a collection of news, the generative model assumes that each news is generated by a mixture of different topics (regarded as events or aspects) governed by a multinomial topic distribution. Across three dimensions: text words, entities and time, the topic distribution should be ideally coherent. The method assumes a mixture language model by introducing a background topic B, where B is also described by multiple dimensions with the collection's word distribution $\phi_B^{\mathtt{w}}$, the collection's entity distributions $\{\phi_B^{\mathtt{e}}\}$ and the collection's temporal distribution $\mathcal{N}(\mu_B, \sigma_B)$. In this way, the common attributes of different topics will be captured by the background model, resulting in a better separation.

In Figure 4.2, we show the plate notation for the generative model. Words, entities and the timestamp are observed for each news and the parameters to be estimated are $\Theta = \{\{\theta\}, \{\phi^{\mathtt{w}}\}, \{\phi^{\mathtt{p}}\}, \{\phi^{\mathtt{l}}\}, \{\phi^{\mathtt{o}}\}, \{\mu\}, \{\sigma\}\}$.

The same as in [9], we assume the following generative process for each news in the collection D:

Let λ_B to be the proportion of the background topic. To generate each word in news d,

1. Choose a switch variable $s^{\mathtt{w}} \sim Bernoulli(\lambda_B)$.

2. If $s^{\mathtt{w}} = 1$,

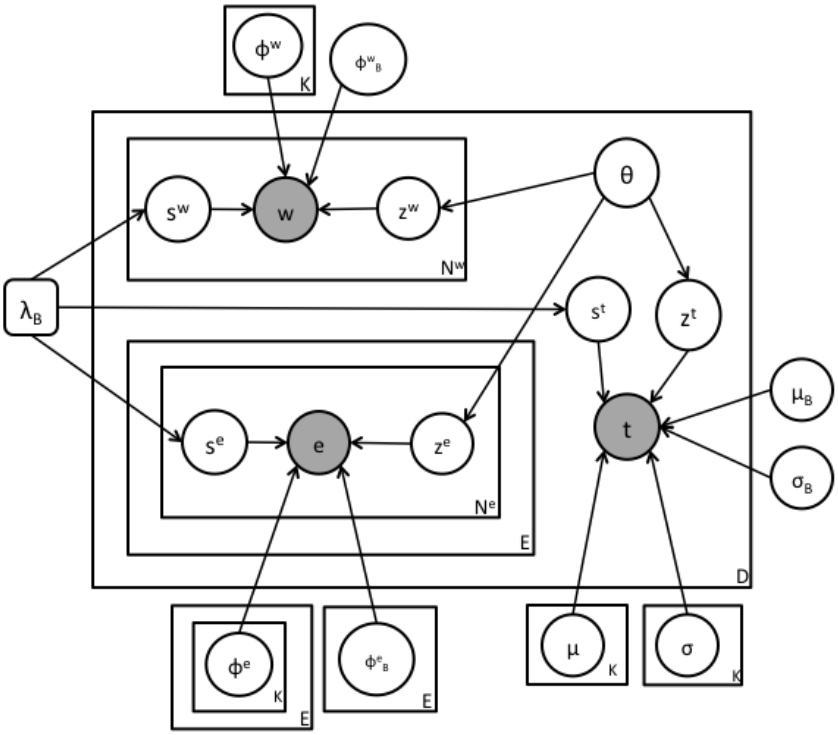


Figure 4.2: Plate Notation: News Learning Module

choose a word w from the background topic B: $w \sim \phi_B^w$;

Else,

choose a topic $z^{\mathtt{w}}$ from the topic distribution θ_d,

choose a word w from the topic $z^{\mathtt{w}}$: $w \sim \phi_{z^{\mathtt{w}}}^{\mathtt{w}}$.

To generate a timestamp t_d for news d,

1. Choose a switch variable $s^{\mathtt{t}} \sim Bernoulli(\lambda_B)$.
2. If $s^{\mathtt{t}} = 1$,

   choose a timestamp t_d from the background time distribution: $t_d \sim \mathcal{N}(\mu_B, \sigma_B)$;

   Else,

   choose a topic $z^{\mathtt{t}}$ from the topic distribution θ_d,

   choose a timestamp t_d from the topic $z^{\mathtt{t}}$: $t_d \sim \mathcal{N}(\mu_{z^{\mathtt{t}}}, \sigma_{z^{\mathtt{t}}})$.

For $\mathtt{e}$ in $\{\mathtt{p}, \mathtt{l}, \mathtt{o}\}$, to generate each entity e in news d,

1. Choose a switch variable $s^{\mathtt{e}} \sim Bernoulli(\lambda_B)$.
2. If $s^{\mathtt{e}} = 1$,

   choose an entity e from the background topic B: $e \sim \phi_B^{\mathtt{e}}$;

   Else,

   choose a topic $z^{\mathtt{e}}$ from the topic distribution θ_d,

   choose an entity e from the topic $z^{\mathtt{e}}$: $e \sim \phi_{z^{\mathtt{e}}}^{\mathtt{e}}$.

Because the above generative model in essence is a pLSI-like model without document and document-word distribution priors, we adopt the Maximum Likelihood (ML) principle for parameter inference. The final posterior distribution is given by:

$$P(z|d) \propto \alpha^{\mathtt{w}} \sum_{w} x_{wd}^{\mathtt{w}} P(z|w,d) + \sum_{\mathtt{e}} \alpha^{\mathtt{e}} \sum_{e} x_{ed}^{\mathtt{e}} P(z|e,d) + P(z|t_d)$$

,where αs are the weighting parameters for different dimensions which can be tuned. Given the news document, the topic distribution $P(z|d)$ will finally be used to determine the event/aspect the news belong to. More details can be found in [9].

Different from theoretical research, in our empirical system, how we determine the topic number K greatly affects system performance. By heuristic, we preset K to be a function of number of days $\mathcal{N}^d$ of the input collection:

$$K = \begin{cases} 25, & \text{if } \mathcal{N}^d = 1 \\ ceil(120 \cdot \log_{10} \mathcal{N}^d), & \text{otherwise} \end{cases} \tag{4.1}$$

,where the ceil function maps the float value to the smallest following integer.

It is obvious that with the log function, we are damping the increasing of K when $\mathcal{N}^d$ grows larger. The motivation of this configuration is that when users input a long time span, they are usually seeking for large and influential events. Thus, smaller and trivial events should be ignored in the result. On the other hand, during our event discovery process, if the K is too large for the news collection (some clusters are too small), EKNOT will reduce the number of events K and rerun the generative model to obtain reasonable result.

It is also worth noting that after the event discovery process, a filtering process will incur to filter out not qualified clusters. There are mainly two kinds of filtering in current EKNOT system:

1. *Keyword Filtering*. If users input the keywords of the events they are interested in, we will examine the $\phi^{\mathtt{w}}_{z^{\mathtt{w}}}(P(\mathtt{w}|z))$ for each topic resulted from previous generative model. If the keywords are not ranked top in the distribution, this irrelevant event will be discarded in the final presentation.

2. *Topic Quality Filtering*. The generative model may output some bad clusters: 1) cluster itself is too small and trivial 2) cluster news may be very diverse. Therefore, a cluster quality filtering process is adopted to further ensure the event quality. In principle, any internal cluster assessment metrics and methodologies can be applied here. Currently the distribution $\phi^{\mathtt{w}}_{z^{\mathtt{w}}}$ and the inner cluster similarity are used. Ideally, we can take all these simple metrics as features and train a classifier to distinguish the good clusters from bad ones.

4.2.3 Tweets Linking

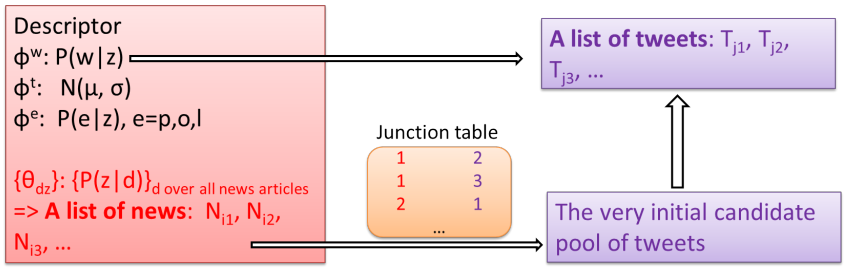


Figure 4.3: Select the candidate tweets with an event descriptor. Output: a news list and a tweet list.

Connecting a single news article to its relevant tweets is an active research area. The tweets linking module is ready to take advantage of any existing methods even though what we consider here is to connect an *event* to its relevant tweets.

The goal of this step is to maintain a high-quality news list and a high-quality tweet list for each event with an emphasis on relevance. The two lists will be used to generate summaries in the co-ranking component in Section 4.3.1 and to analyze opinions in Section 4.4. For the sake of both effectiveness and efficiency,

we design the following procedure to select tweets for each event as illustrated in Figure 4.3. EKNOT first obtains a list of news articles for each event based on $\{\theta\}$, *i.e.*, a news article d with $P(z|d)$ greater than a threshold will be selected for event z. The linked tweets (based on the junction table) of the selected news articles form the very initial pool of the candidate tweets. Then we adapt the standard information retrieval ranking function Okapi BM25[5] to rank the tweets in the initial pool. Top k words $(w_1, w_2, ..., w_k)$ in the word distribution $\phi^{\mathtt{w}}$ of the event are selected as query keywords in the BM25 method. We also leave out the IDF (inverse document frequency) in BM25. Instead, we take the $P(w|z)$ value to weight the final score. The result scoring function used is therefore given by:

$$score(D, Q) = \sum_{1}^{k} P(w_i|z) \cdot \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

where,

1. function f is the number of occurrences of w_i in the document;
2. $avgdl$ is the average document length in the document collection
3. we preset $k_1 = 1.6$, $b = 0.75$, $k = 10$ in EKNOT .

Currently in EKNOT , for each event, at most 2000 tweets will be retrieved for a single news.

Within each event, EKNOT uses the aspect descriptors to select relevant tweets from the event's list of tweets for each aspect. It obtains the news list and tweet list in the same way as it does for an event, except that the initial pool takes the event's tweet list rather than referring to the junction table.

Finally, after the pLSI-like model, topic filtering, and tweets linking, the result events / aspects (news collections and the descriptors indicating their charateristics) and their relevant tweets are passed to the next step.

4.3 Joint Summarization

To help users gain a thorough understanding of their interested event, EKNOT offers various information including a content summary from news for an objective

[5]http://en.wikipedia.org/wiki/Okapi_BM25

description, highlights from tweets for an opinion overview, an entity graph, a time span and an opinion distribution. Except for the opinion distribution, we will discuss about the techniques employed in all other summarization process in this section. The opinion mining part will be discussed in detail in Section 4.4.

4.3.1 Co-Ranking News and Tweets to Obtain News Summaries and Tweet Highlights

Our goal of this co-ranking step is to construct an objective news summary and an opinion-rich tweet summary for each event/aspect. In EKNOT , we choose news titles as candidate summary sentences for the event. The reason for this choice is that 1) news title is usually a concise and precise summary of its article given by trained editors 2) the full text article of the news title can be shown directly for further information if users click on the title sentence. The candidate of the tweet summary is given by Section 4.2.3.

EKNOT co-ranks the news and tweets considering:

1. content and temporal *consistency* with the event/aspect;
2. *coherence* between the news summary and tweet summary;
3. *coverage* and *diversity* of the news(tweet) summary;
4. whether the tweet summary contains substantial *opinions/sentiments* and represents a general trend of the public.

To instantiate the co-ranking algorithm, EKNOT combines Co-HITS[28] and the Maximal Marginal Relevance (MMR) principle[27]. At the beginning, four static score lists are computed.

- R_n^c, R_t^c: Co-HITS score for news and tweets, which captures content consistency and coherence. Co-HITS is run on the bipartite news-tweets graph, where the edge weight is the cosine similarity between the word vectors of the connected news and tweet. We impose a regularization term to enforce the score of a news/tweet to be consistent with the BM25 score with the event's word distribution $\phi^{\mathtt{w}}$.
- R_n^t, R_t^t: Temporal consistency score for news and tweets, which captures temporal consistency, is computed as the probability density of the timestamp of a news/tweet.

- R_t^s: Sentiment polarity for tweets, given by a classifier which will be explained in Section 4.4.

- R_t^p: Popularity score for tweets based on Twitter-specific features such as retweet number and favorites.

The final news ranking is given by a linear combination of R_n^c and R_n^t, while tweets ranking is given by a linear combination of $R_t^c, R_t^t, R_t^s, R_t^p$. To guarantee the content and temporal coverage as well as diversity of the summaries, each time after selecting a piece of news/tweet into the summary/highlights, we penalize the similar news/tweets in the rest of the corpus under MMR. At last, the headlines of the top ranked news are output as the news summary[6] and the contents of the top ranked tweets are output as the tweet highlights.

4.3.2 Entity Graphs

EKNOT generates an entity graph for each event/aspect with the descriptor. An entity is denoted by a node and the correlation between two entities is denoted by an edge. In our visualization, the node size is proportional to $P(e|z)$, indicating how influential an entity is. The edge width is proportional to the co-occurrence number of two entities within the event, indicating how strong the two entities are correlated. Different colors in an event's entity graph indicate different aspects and are consistent with the colors throughout the visualization.

4.3.3 Time Spans

EKNOT utilizes the Gaussian time distribution $\phi^{\mathbf{t}}$ in the descriptor to approximate the time span of each event/aspect.

4.4 Opinion Analysis

Opinion analysis provides the sentiment polarity feature for co-ranking and is used to calculate the positive/negative percentage of public opinions towards an event/aspect. Naturally, the opinions for an event are obtained from tweets.

[6] In our system, the full text of news articles are accessible by clicking the headlines.

4.4.1 Problem Formulation

According to our user scenario, we take the following definitions in the opinion mining process:

Definition 3 (Subjective Tweet) *A tweet that contains "a personal positive or negative feeling"[32] is defined as a subjective tweet. A subjective tweet is either a tweet with positive sentiment (positive tweet) or one with negative sentiment (negative tweet).*

On the other hand, we have,

Definition 4 (Neutral Tweet) *A tweet that contains pure facts such as containing only news headlines is considered as a neutral tweet.*

Such definitions emphasize that the sentiments should be personal. Even if there is a clear sentiment indication in the news headline such as "a husband kills his wife because of hatred", EKNOT will not regard a tweet like this as subjective.

Based on these definitions, we formulate the problem as a classification task, i.e. given a tweet, the opinion mining module should be able to classify the tweet as neutral, positive or negative.

4.4.2 A Two-Step Classification Framework

In order to effectively extract subjective tweets and identify their sentiments, as well as to ensure efficiency, we build a two-step classification model to determine a tweet's sentiment following [35]:

- *Step 1.* Subjectivity Classification. This classifier decides whether a tweet is subjective or neutral. Tweets classified as subjective will be passed to step 2.
- *Step 2.* Polarity Classification. This classifier determines whether a subjective tweet is positive or negative.

In both steps, EKNOT builds a binary logistic regression classifier. The features employed are

1. Unigram Features. This is a common feature adopted in sentiment analysis. In [32], Go *et al.* try unigram, bigram model in conjunction with part-of-speech (POS) features and report that unigram model outperforms all other models in the sense that bigrams and POS features do not help.

2. Linguistic Features. This category of features captures the tweet's linguistic properties such as different punctuations used.

3. Tweet Features. This category includes all the tweet features such as the presence of hashtag and the number of retweets.

4. Dictionary-based Sentiment Features. To determine sentiment of an individual word, EKNOT uses a sentiment dictionary derived from SentiWordNet[7].

As in EKNOT , thousands of tweets need sentiment classification for a single event. Efficiency should be ensured during the feature computation process. Thus, we do not incorporate complex features such as Maximum-Entropy based part-of speech tagging and syntax tree.

In order to generate large and labeled tweet data for the training process, we follow the paper [32] by employing distant supervision. The neutral set is formed by news titles, which is in accordance with our definition for neutral tweets. The positive and negative sets are obtained by inspecting the emoticons of the tweets we retrieved in data collection. Tweets with positive emoticons such as :) are automatically labeled as positive, while those with negative emoticons such as :(are labeled as negative. Step 1 is trained on all the three sets considering both positive and negative sets as subjective. Step 2 is trained only on the positive set and the negative set.

Finally, The sentiment polarity score weighted by its relevance to an event/aspect is aggregated to compute the positive and negative percentage of public opinions.

[7]http://sentiwordnet.isti.cnr.it/

Chapter 5

Case Study

In this section, we will study a real-world example to show the effectiveness of the EKNOT system. Though we do not have the ground truth data, we can assess the popular event from our knowledge.

Here let us use `03/07/2015-03/11/2015` totally five days as an example. EKNOT returns the event display page which displays all the major events within this period, as illustrated in Figure 5.1 (we only show the screenshot for two events here because of space limit). In addition to the news summary and tweet highlights, a word cloud is also presented. Entity graphs and opinion distributions are not displayed on this page to guarantee page loading efficiency as well as readability. For this time period, totally 52 events are shown. In Figure 5.1, we can see that the result events are quite reasonable:

1. Event 2 is about the apple event on March 9th which releases the apple watch and the new macbook;

2. Event 3 is about the missing plane Malaysia Airlines Flight MH370, because it is around one year anniversary of its disappearance.

3. Event 4 is mainly about the White House lockdown on March 7th.

Users can choose any event to investigate the event details and inspect its aspects. By clicking the button below the summaries, users will be navigated to an "event details" page which displays the entire summary of the event including the news summary, tweet highlights, the entity graph and the opinion analysis. Here we take event 2, the apple event happened on March 9th as an example. Figure 5.2 shows the event detail page for this event.

To inspect the aspects of an event, users can click the "Event Aspects" tab on the upper-left of the event details page. As illustrated in Figure 5.3, the event aspects page will first present users with the time span of different aspects. Then detailed information for each aspect including news summary, tweet highlights,



Figure 5.1: Event Display Page

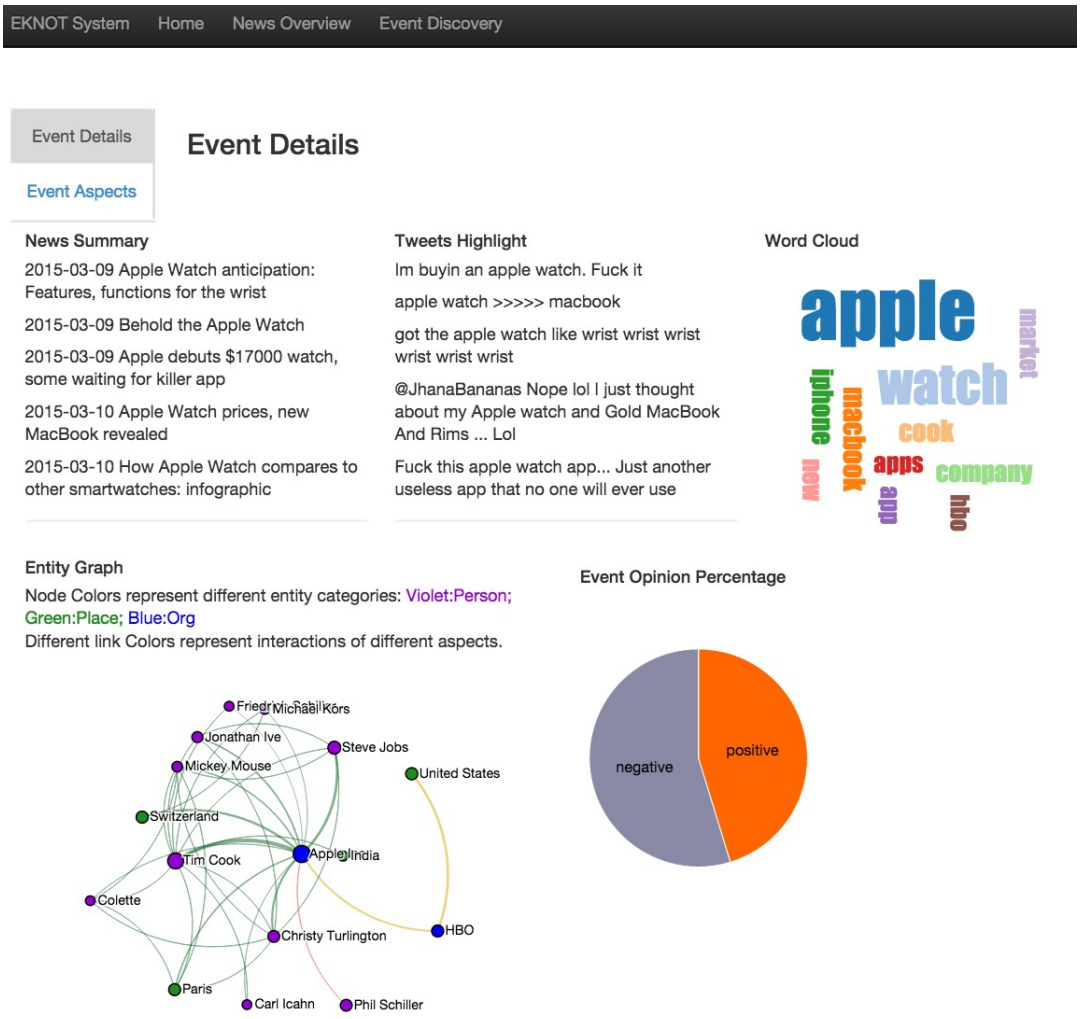


Figure 5.2: Event Details Page

word cloud, entity graph and opinion distribution is showed separately.

Because of the limit of page space, the entire aspect discovery result cannot be covered in Figure5.3. Here we directly give a brief summary of all the result aspects:

- Aspect 1 is all about apple watch. From the time span graph we can see that this aspect lasts a quite long time because apple watch raises heated discussions both before and after the Apple Event.
- Aspect 2 is about the new macbook, which has its peak during the Apple

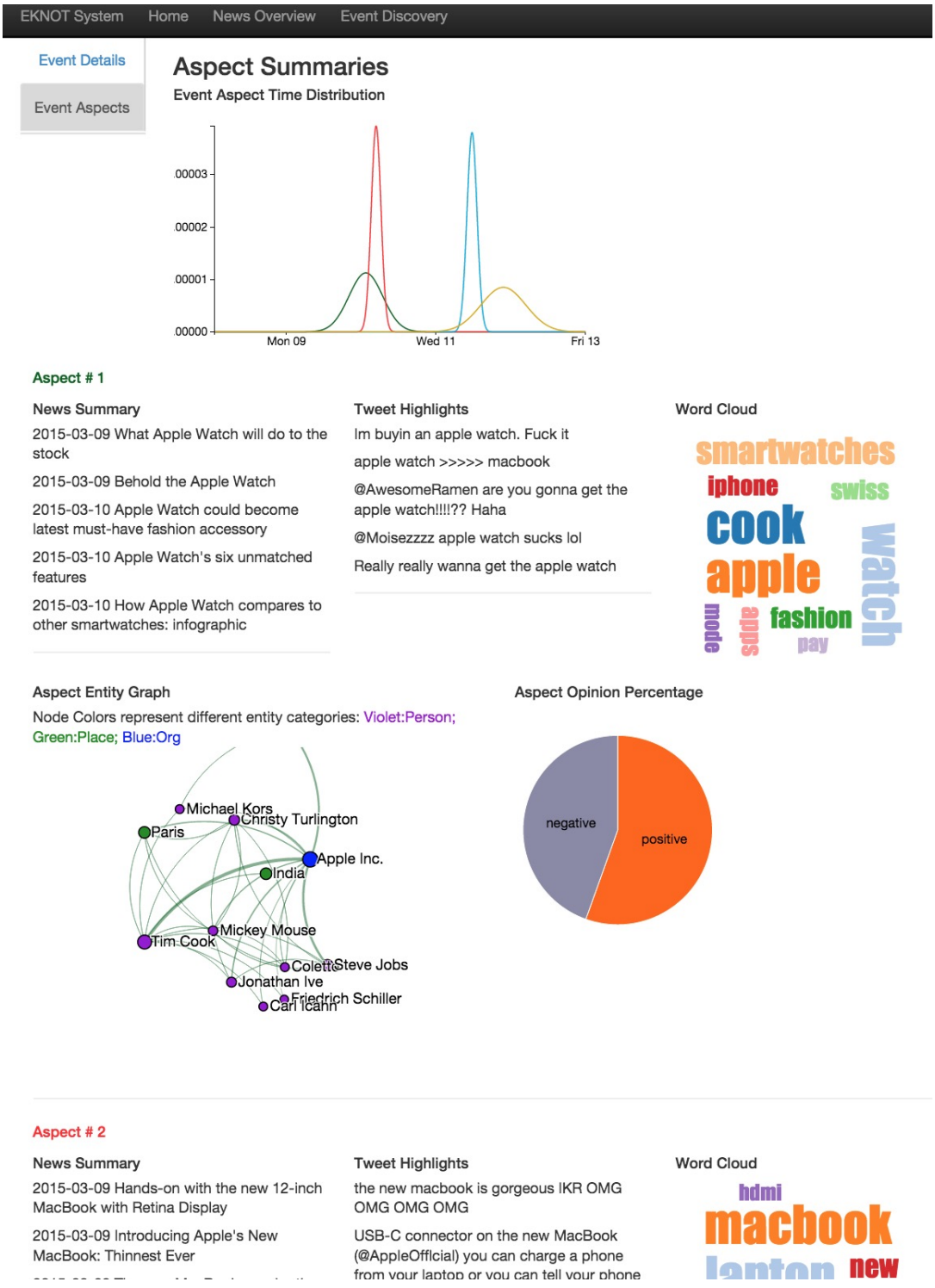


Figure 5.3: Aspect Details

Event.

- Aspect 3, the blue line in the time distribution graph, talks about the 11-hour

apple service outage on March 11th.

- Aspect 4 is mainly about the launching of HBO streaming service on Apple device. It is represented as the yellow line on the time distribution graph.

Besides the news summary, opinions from the tweets and other informative graphs are presented. It is quite obvious that these aspect results perfectly match what happened around the Apple Event on March 9th.

Chapter 6

Conclusion

In this thesis, we propose a general framework to perform event discovery, summarization and analysis based on the scenario that combines traditional news media with social media. We develop the EKNOT system which provides users with informative and well-organized knowledge for automatically discovered events including news summary, tweet highlights, entity graph, opinion analysis and event aspects information.

Many technical components are involved in the development of EKNOT including topic modeling, text summarization and opinion mining. We leverage the related state-of-art techniques, design practical methods and implement a compete system to handle the real-world task in an effective and efficient manner.

Besides the functional values of the system, EKNOT also offers a valuable platform for following explorations and experiments of further related researches because of the comprehensive and high-quality event data it generates. Much work can be done following this line. For example, how to perform online streaming event discovery may be a good research topic and the result solution may be a great improvement to the current system.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 591–600, ACM, 2010. 1

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, (New York, NY, USA), pp. 50–57, ACM, 1999. 3

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. 3

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), pp. 113–120, ACM, 2006. 3

[5] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, (New York, NY, USA), pp. 424–433, ACM, 2006. 3

[6] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, (New York, NY, USA), pp. 680–686, ACM, 2006. 3

[7] H. Kim, Y. Sun, J. Hockenmaier, and J. Han, "Etm: Entity topic models for mining documents associated with entities," in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, (Washington, DC, USA), pp. 349–358, IEEE Computer Society, 2012. 3

[8] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, pp. 7:1–7:30, Feb. 2010. 3

[9] S. for double-blind review, "Mining multi-aspect reflection of news events in twitter: Discovery, linking and presentation," in *Submission*. 3, 8, 12, 13, 15

[10] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, (New York, NY, USA), pp. 42–51, ACM, 2009. 3

[11] L. Shou, Z. Wang, K. Chen, and G. Chen, "Sumblr: Continuous summarization of evolving tweet streams," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, (New York, NY, USA), pp. 533–542, ACM, 2013. 3, 5

[12] J. Vosecky, D. Jiang, K. W.-T. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in twitter," in *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, (New York, NY, USA), pp. 879–884, ACM, 2013. 3

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 851–860, ACM, 2010. 3

[14] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, (Berlin, Heidelberg), pp. 338–349, Springer-Verlag, 2011. 3

[15] I. Subavsic and B. Berendt, "Peddling or creating? investigating the role of twitter in news reporting," in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, (Berlin, Heidelberg), pp. 207–213, Springer-Verlag, 2011. 4

[16] W. Gao, P. Li, and K. Darwish, "Joint topic modeling for event summarization across news and social media streams," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (New York, NY, USA), pp. 1173–1182, ACM, 2012. 4, 5

[17] M. Paul and R. Girju, "Cross-cultural analysis of blogs and forums with mixed-collection topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, (Stroudsburg, PA, USA), pp. 1408–1417, Association for Computational Linguistics, 2009. 4

[18] Z. Wei and W. Gao, "Utilizing microblogs for automatic news highlights extraction," in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 872–883, 2014. 4

[19] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," *Journal of Computer Science*, vol. 8, no. 1, pp. 133–140, 2011. 4

[20] M. S. Binwahlan, N. Salim, and L. Suanmali, "Swarm diversity based text summarization," in *Proceedings of the 16th International Conference on Neural Information Processing: Part II*, ICONIP '09, (Berlin, Heidelberg), pp. 216–225, Springer-Verlag, 2009. 4

[21] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, pp. 919–938, Nov. 2004. 4

[22] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, pp. 457–479, Dec. 2004. 4

[23] L. Li, D. Wang, C. Shen, and T. Li, "Ontology-enriched multi-document summarization in disaster management," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, (New York, NY, USA), pp. 819–820, ACM, 2010. 4

[24] D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 298–306, IEEE, 2011. 4

[25] B. Sharifi, M.-A. Hutton, and J. Kalita, "Summarizing microblogs automatically," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, (Stroudsburg, PA, USA), pp. 685–688, Association for Computational Linguistics, 2010. 4

[26] D. Yajuan, C. Z. WEIF uRu, Z. M. Heung, and Y. SHUM, "Twitter topic summarization by ranking tweets using social influence and content quality," in *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 763–780, 2012. 4

[27] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *ACM SIGIR*, SIGIR '98, (New York, NY, USA), pp. 335–336, ACM, 1998. 5, 18

[28] H. Deng, M. R. Lyu, and I. King, "A generalized co-hits algorithm and its application to bipartite graphs," in *ACM SIGKDD*, pp. 239–248, ACM, 2009. 5, 18

[29] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (Stroudsburg, PA, USA), pp. 417–424, Association for Computational Linguistics, 2002. 5

[30] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (New York, NY, USA), pp. 168–177, ACM, 2004. 5

[31] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, (Stroudsburg, PA, USA), pp. 347–354, Association for Computational Linguistics, 2005. 5

[32] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," tech. rep., Stanford University. 5, 20, 21

[33] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREC*, vol. 10, pp. 1320–1326, 2010. 5

[34] A. Bermingham and A. F. Smeaton, "Classifying sentiment in microblogs: Is brevity an advantage?," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, (New York, NY, USA), pp. 1833–1836, ACM, 2010. 5

[35] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *COLING: Posters*, COLING '10, (Stroudsburg, PA, USA), pp. 36–44, Association for Computational Linguistics, 2010. 5, 20

[36] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, LSM '11, (Stroudsburg, PA, USA), pp. 30–38, Association for Computational Linguistics, 2011. 5

[37] W. Guo, H. Li, H. Ji, and M. T. Diab, "Linking tweets to news: A framework to enrich short text data in social media," in *ACL*, pp. 239–249, 2013. 6

[38] M. Tsagkias, M. de Rijke, and W. Weerkamp, "Linking online news and social media," in *WSDM*, WSDM '11, (New York, NY, USA), pp. 565–574, ACM, 2011. 6

[39] B. Shi, G. Ifrim, and N. Hurley, "Insight4news: Connecting news to relevant social conversations," in *ECML PKDD*, pp. 473–476, 2014. 6

[40] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *ACL Student Research Workshop*, ACLstudent '05, (Stroudsburg, PA, USA), pp. 43–48, 2005.