Ranking Events Based on User Relevant Query

Xiangfei Kong[1,2], Wenji Mao[1,2]
[1]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Computer and Control Engineering, University of Chinese Academy of Sciences
Beijing,China
{kongxiangfei2015, wenji.mao}@ia.ac.cn

***Abstract*—Given a collection of event-related documents, event ranking generates a list of ranked events based on the input query. Ranking news events, which takes event related news documents for the generation of ranked events, is both an essential research issue and important component for many security oriented applications, such as public event monitoring, retrieval, detection and mining. Previous related work solely relies on queries of event relevant aspects, and user relevant aspects of queries that are critical for security applications are totally ignored. In this paper, we deal with the problem of news ranking by incorporating user relevant information into the input query, from the cluster of relevant new documents and comments. Given an input query, which contains event related objective aspects(e.g. actors, locations, date) and user related subjective aspects(e.g. public attention and opinion polarity), we develop a Learning-to-Rank framework to integrate aspect-level correlation between query and event. Experiments on a crawled large news corpus show the effectiveness of our proposed approach compared to several baseline models.**

Keywords–news event ranking*;*learning to rank*;*user relevant information

I. INTRODUCTION

Reading news online has become one of the most important means for acquiring information. When a public social event occurs, various media from different platforms will flock to report it. Meanwhile, when reading news to learn about the latest events, people incline to express their views and opinions, which are equally important for characterizing events. In security-related applications, these user relevant event aspects reflect public demand and individualized needs. It can provide valuable information to facilitate government decision maker, security official and information seeker. Moreover, fulfilling a event query such as “Event happened in Beijing which has a positive social influence or good public opinion”, which includes both event related objective aspects and user related objective aspects, will extend the functionality of existing event-related query systems greatly, and have promising application prospects in many security oriented applications.

To address the aforementioned issues, we develop a Learning-to-Rank framework that aims to address the problem of ranking events from a large news corpus based on the aspect-level correlation between extracted events and given user query containing event relevant aspects and user relevant information.

The remainder of the paper is structured as follows: Section II presents related works. Section III defines the problem formally. Section IV describes our approach in detail. Experimental evaluation are presented in Section V and Section VI concludes the paper.

II. RELATED WORK

Previous works related to news event ranking involve news ranking and event-based question answering. Now we discuss these works in order and explain the difference between our work and the related research.

News Ranking: News ranking is an important research issue and has drawn much attention in previous researches[1–4]. Gwadera et al[1] mine sequential patterns in a time window of news streaming to rank them based on story timeliness and content authority. In [2], the authors first cluster the entire collection into a fixed number of topics and then proposed a language model approach to combine both the cluster-based importance of an article and its temporal profile contribution to obtain the final score of the article on which it is ranked.

Event-based Question Answering: Liu et al.[5] represent each event as a tuple of three elements: time, location and topic. By query semantic parsing, the corresponding time, location and topic keywords are extracted. To get the most relevant results, they retrieve these keywords on *Google News* and then cluster the results using *LDA* and *DBSCAN*. The event importance is evaluated by the entropies of event keywords. Bechberger et al.[6] introduce the *NewsTeller* system which retrieves a news event based on a user query and the user's general interests. Authors label each event manually as ‘bad’(0), ‘fair’(1), ‘good’(2), ‘excellent’(3), ‘perfect’(4) with user query and user model, then train the regression model with Random Forest. The output is the most related event displayed by an abstract of the event.

Although *News Ranking* can rank news documents to a certain extent, they can only deal with simple queries without considering abundant user relevant information. Moreover, related works can only provide coarse ranking of news documents, we need more fine-grained treatment of event-level ranking in many security oriented applications. In addition, news event ranking should integrate aspect-level correlation and other event/user features comprehensively, it is important to characterize the relevance between input query and event.

978-1-5090-6727-5/17/$31.00 ©2017 IEEE

III. PROBLEM DEFINITION

In this study, we aim to rank a list of events extracted from a large news corpus based on a given user query which contains event related objective aspects and user related subjective aspects. We formalize the problem of news event ranking based on event aspects and user aspects information and introduce the mathematical symbols used in this paper.

Given a collection of documents $\mathcal{D} = \{d\}$ containing news articles published by several news agencies and corresponding comment documents $\mathcal{R} = \{r\}$, let $E = \{e_1, e_2, ..., e_n\}$ be a set of events extracted from $\mathcal{D}$, a news event e_i is reported through a cluster of news documents $D_i = \{d_{i1}, ..., d_{in_i}\} \subset \mathcal{D}$, accompanied by a set of comment documents $R_i = \{r_{i1}, ..., r_{in_i}\} \subset \mathcal{R}$ that contain the opinions expressed by user about the news articles, where r_{ij} is the concatenation of all the user review documents of the news article d_{ij}. Each news event e_i is characterized by (D_i, R_i).

We cast the problem as a text ranking problem where the text collection $\mathcal{C}$ consists of news articles and comment documents for all the events, that is $\mathcal{C} = \{(D_1, R_1), ..., (D_n, R_n)\}$, and the query Q contains event keywords describing various aspects.

IV. PROPOSED APPROACH

In all, our approach pipeline contains three main modules: Query Parsing and Expansion, Event Extraction and Event Ranking. We will illustrate explicitly in the following.

A. *Query Parsing and Expansion*

Given a query Q, we aim to get all of the query aspects which contain additional user information than general event query. In detail, we firstly use the Stanford Named Entity Tagger[1] to identify the basic event aspects: location, organization, date, person and time[7]. Next, we write a script to get the user information through syntactic parsing and self-defined extract patterns(see in Table I). The simple patterns are then combined in a tree-like manner to obtain more valuable opinion phrases.(*N* indicates a *noun*, *A* an *adjective*, *V* a *verb*, *h* a head term, *m* a *modifier*)[8].

TABLE I
SELF-DEFINED EXTRACTION PATTERNS

Methods
$amod(N, A) \rightarrow < N, A >$
$acomp(V, A) + nsubj\ (V, N) \rightarrow < N, A >$
$cop(A, V) + nsubj(A, N) \rightarrow < N, A >$
$dobj\ (V, N) + nsubj\ (V, N) \rightarrow < N, V >$
$< h1, m > +conj\ and(h1, h2) \rightarrow < h2, m >$
$< h, m1 > +conj\ and(m1, m2) \rightarrow < h, m2 >$
$< h, m > +neg(m, not) \rightarrow < h, not + m >$
$< h, m > +nn(h, N) \rightarrow < N + h, m >$
$< h, m > +nn(N, h) \rightarrow < h + N, m >$

B. *Event Extraction*

To determine the number of events accurately, we resort to agglomerative clustering method and self-defined optimization function to find the optimal cluster threshold. v_i and v_j are the vector representations of news document d_i and d_j. $dist(vec_i, vec_j)$ represents the distance of d_i and d_j. In Equation 1, the solution of the function depends on the value of the independent variable: $margin$. If I_{ij} equals 1, it indicates d_i and d_j belongs to the same cluster.

$$\begin{gathered}\arg\max f\left(margin\right) = \\ \arg\max \sum_{j \in D} \sum_{i \in D} sign\left[I_{ij} dist_k\left(v_i, v_j\right) > \varepsilon\right] \qquad (1) \\ I_{ij} = \begin{cases} 1 & dist_k\left(v_i, v_j\right) \leq margin \\ 0 & else \end{cases}\end{gathered}$$

C. *Event Ranking*

Given a user query, we obtain Q_{parsed} containing query aspects and user information through *Query Parsing and Expansion*. When we have extracted a list of events $E = \{e_1, e_2, ..., e_n\}$ through *Event Extraction*, we then deal with $(D_i, R_i) = (\{d_{i1}, ..., d_{in_i}\}, \{r_{i1}, ..., r_{in_i}\})$ of each e_i to get event aspects and user information. We adopt the same approach as Query Parsing. In this way, we can construct ranking features fully considering the correlation of event and query on aspect level.

To consider event's sentiment tendency on specific aspect separately. Here we adopt SentiWordNet 3.0 to determine the polarity of terms. In SentiWordNet, each synset(t) is associated with three numerical scores: $Pos\left(t\right)$, $Neg\left(t\right)$ and $Obj\left(t\right)$, measuring the degree of how positive, negative and objective the word is. Thus each term's polarity score is calculated by

$$polarityScore\left(t\right) = \frac{\sum\limits_{w \in senset_t} \left(Pos(w) - Neg(w)\right)}{\left|senset_t\right|} \qquad (2)$$

To evaluate the polarity of user information in R_i, we use terms in the neighborhood of the user information keywords $ut(O\left(ut, k_n\right)$ in Equation 3). The relevance value of query and event in $info_k$ is calculated by Equation 4. σ is the sigmoid function to scale the relevant score.

$$p_{R_i}\left(info_k\right) = \sum_{ut \in info_k} \sum_{t \in O(ut, k_n)} polarityScore(t) \qquad (3)$$

$$rel_{info_k} = \sigma\left(p_{R_i}\left(info_k\right), p_{D_i}\left(info_k\right)\right) \qquad (4)$$

The event ranking problem aims to find a ranking of k most relevant news events on a given query Q. Here we resort to *SVMRank*[9] to construct the ranking model integrating relevant features. The relevant features we construct contain *BM25*, *TFIDF*, aspects that comments involve, event polarity. Since *SVMRank* is a pairwise *Learning to Rank* method, we refer to the partial ordering relation in Gdelt datasets[2] to train the model.

[1] http://nlp.stanford.edu:8080/ner/

[2] http://www.gdeltproject.org

V. EXPERIMENTAL EVALUATION

A. Datasets

We collects news from GDELT Project, which monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages. We crawled a dataset of only English news documents and corresponding comments amounting to 268,215 news documents on 02/01/2017.

B. Relevance Measure

We take the event's centroid news article as the representation of the event. Then the standard result of the generated event query is just the news document. The experimental results are evaluated by *nDCG@k* values because users care about only the *top-k* events that closely match the user's preferences.

C. Results Analysis

TABLE II
nDCG@k RESULTS WITH DIFFERENT METHODS

Methods	$nDCG$@10	$nDCG$@20	$nDCG$@30	$nDCG$@50
1	0.598	0.765	0.815	0.852
2	0.195	0.367	0.456	0.536
3	0.667	0.823	0.868	0.898

[1]Event Ranking Baseline Using the date, location, actor and event type as relevant aspects.

[2]Sentiment Information Only Using sentiment information only.

[3]Event Ranking Fused With Sentiment Information Ranking events based on event related subjective aspects and user related subjective aspects(Sentiment Information).

To evaluate the effectiveness of ranking events based on user relevant query, we conduct experiments with and without sentiment information in [1] and [3]. We observe that the naive baseline using only the basic event aspects is inadequate, our additional aspects which incorporate polarity information can improve event ranking performance on *nDCG@k* metric. We also found that different numbers of events used for relevance judgments *nDCG@k* cause inconsistent results. When only top k relevant events are used for relevant judgments, the performance divergence between two methods is more significant. The comparable performance also demonstrates that comments about news event can characterize event better. We also validate the effectiveness of event sentiment polarity by only using event sentiment information in [2]. Since event polarity usually does not involve the concrete aspects about event, the ranking results are unfavorable. In the end, we compare the performance of our approach that considers all the aforementioned user information in event and query.

VI. CONCLUSION

Ranking news events based on user relevant query is beneficial and critical for many security-oriented tasks. We identified several limitations of related works. Consequently, in this paper we firstly use a cluster of news documents to represent event. We demonstrate the effectiveness of clustering by experiments in comparison with news ranking. Then given user's event query containing event aspects fused with user information, we rank events based on the aspect-level correlation between query and event and event importance. By conducting experiments on a large news corpus, we demonstrate the effectiveness of our approach. We observe that by fusing more abundant and valuable event information, event ranking can have more broad application scenarios. Additionally, it also refines the performance of current event ranking problem.

ACKNOWLEDGEMENT

This work is supported in part by NSFC Grant #71621002, the Ministry of Science and Technology of China Major Grant #2016QY02D0205, and CAS Key Grant #ZDRW-XH-2017-3.

REFERENCES

[1] Robert Gwadera and Fabio Crestani. Mining news streams using cross-stream sequential patterns. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 106–113, 2010.

[2] Yeha Lee and Jong-Hyeok Lee. Identifying top news stories based on their popularity in the blogosphere.

[3] Koichiro Yoshino and Tatsuya Kawahara. Information navigation system based on pomdp that tracks user focus. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 32, 2014.

[4] Koichiro Yoshino and Tatsuya Kawahara. News navigation system based on proactive dialogue strategy. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 15–25. Springer, 2015.

[5] Xueliang Liu and Benoit Huet. Event-based cross media question answering. *Multimedia Tools and Applications*, 75(3):1495–1508, 2016.

[6] Lucas Bechberger, Maria Schmidt, Alex Waibel, and Marcello Federico. Personalized news event retrieval for small talk in social dialog systems. In *Speech Communication; 12. ITG Symposium; Proceedings of*, pages 1–5. VDE, 2016.

[7] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

[8] Samaneh Abbasi Moghaddam. *Aspect-based opinion mining in online reviews*. PhD thesis, Applied Sciences: School of Computing Science, 2013.

[9] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.