Knowledge-Based Systems 76 (2015) 17–29

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

NewsMiner: Multifaceted news analysis for event search

Lei Hou [a,*], Juanzi Li [a], Zhichun Wang [b], Jie Tang [a], Peng Zhang [a], Ruibing Yang [a], Qian Zheng [a]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[b] College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 13 November 2014
Accepted 16 November 2014
Available online 3 December 2014

Keywords:
Event
Link
News mining
Social content
Knowledge linking
News search

ABSTRACT

Online news has become increasingly prevalent due to its convenience for information acquisition, and meanwhile the rapid development of social applications enables news generate and spread through various ways at an unprecedented rate. How to organize and integrate news from multiple sources, and how to analyze and present news to users are two challenging problems. In this article, we represent news as a link-centric heterogeneous network and formalize news analysis and mining task as link discovery problem. More specifically, we propose a co-mention and context based knowledge linking method and a topic-level social content alignment method to establish the links between news and external sources (i.e. knowledge base and social content), and introduce a unified probabilistic model for topic extraction and inner relationship discovery within events. We further present a multifaceted ranking strategy to rank the linked events, topics and entities simultaneously. Extensive experiments demonstrate the advantage of the proposed approaches over baseline methods and the online system we developed (i.e. NewsMiner) has been running for more than three years.

1. Introduction

Online news has become increasingly prevalent in our daily life for its convenience to acquire information. Report [1] from China Internet Network Information Center shows that there are about 618 million Internet users in China by the end of 2013 and 79.6% of them access news information online. Meanwhile, with the rapid development of social networks and applications, news finds various spreading ways, such as Wikipedia, blogs and Twitter. Taking the *Malaysia MH370* as example, the corresponding page in Wikipedia[1] was created within minutes, 3,434 news articles were published on sina,[2] accompanying with which millions of comments were posted by social users. It is obviously infeasible for users to go through even part of them to understand this catastrophe event. Therefore, we need to find an effective way to help users understand and explore such huge amount of news information from multiple sources.

News search has been attracting much attention from research and industry, and a number of systems have been developed, such as Google News, Yahoo! News and Digg. There are two main directions that enhance news search, one is topic detection and tracking and the other is incorporating with public knowledge base. Topic Detection and Tracking (TDT) addresses many problems that relate to news organization, event modeling and event detection etc, e.g. Google organizes news in collections of similar articles. The limitation is that TDT cannot provide extra information beyond news collections. A knowledge base is a database of entities as well as their attributes and relations, which can provide extra information about news events and related entities. A famous application is the knowledge graph released by Google which embeds structural knowledge into search results. Another related and well-known system is the Microsoft EntityCube.[3] However, knowledge graph aims at general search and EntityCube pays more attention on the relationship between entities while both of them cast little glances at news content.

Facing large amount of various types of news from multiple sources, people may want to have a better understanding about how news media report, what the users' comments focus on, and how it is related to the previous news. If necessary, they probably want the background knowledge and more detailed information about the related named entities. However, most of existing news services always present the latest news without in-depth analysis, which makes valuable information among similar things be lost.

* Corresponding author. Tel.: +86 (010) 62789831; fax: +86 (010) 62781461.
E-mail addresses: houlei@keg.cs.tsinghua.edu.cn (L. Hou), ljz@keg.cs.tsinghua.edu.cn (J. Li), zcwang@bnu.edu.cn (Z. Wang), tangjie@keg.cs.tsinghua.edu.cn (J. Tang), zp@keg.cs.tsinghua.edu.cn (P. Zhang), yangruibing68@163.com (R. Yang), zy@keg.cs.tsinghua.edu.cn (Q. Zheng).

[1] http://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370.
[2] http://roll.news.sina.com.cn/s_mlxyhkhbsl_all/index.shtml.
[3] http://entitycube.research.microsoft.com/index.aspx.

http://dx.doi.org/10.1016/j.knosys.2014.11.017
0950-7051/

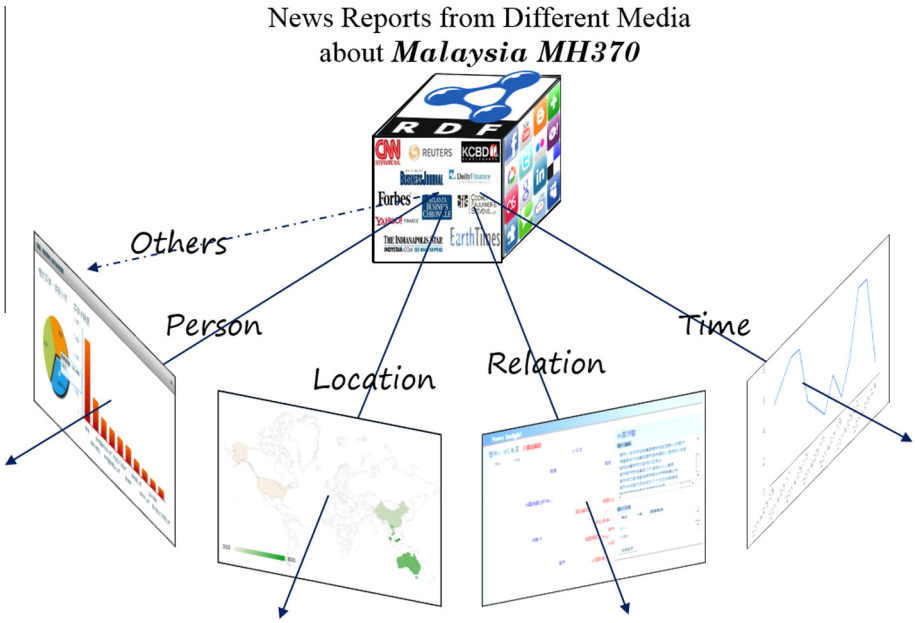


Fig. 1. Multiple Facets of *Malaysia MH370*: *Person* lists the key people in the event, *Location* presents the related locations with colors expressing their importance, *Time* plots the topic trends, and *Relation* computes the relationships among all the related topics and named entities.

Therefore, it is urgent to analyze and explore different aspects of news to satisfy people's information need.

As we know, event is the way that information providers gather, edit, deliver and achieve news information [2], and it is also the natural way and focus which information consumers capture the thing happened in the world. Employing event to organize and manage news can help integrate news from multiple sources and different facets, so that we can turn news collection into multifaceted information (person, location, time, etc.) as shown in Fig. 1. Furthermore, the analysis results accumulates to serve as the historical event knowledge base which can be used as the semantic description of happened events and to predict the trends of similar events in the future. Therefore, we aim at investigating news analysis from the perspective of event. It is a non-trivial problem and raises some challenging problems:

1. How to model the event which consists of different types of news in a unified model?
2. How to provide background knowledge for news incorporating with public knowledge base?
3. How to integrate official news reports with corresponding social media?
4. How to analyze event news and present them in a user-easy-understanding form?

To tackle these issues, we view them as link discovery processes. Linking named entities to public knowledge base brings background knowledge for news, aligning comments to news articles establishes links between news and social content, and event analysis discovers the links among topics and named entities within event. Therefore, we need a unified model which can express both internal links within events and external links among various sources.

In this article, we first represent the news as a link-centric heterogeneous network in three semantic levels, namely events, topics and entities. Different types of links are formalized as the edges within levels, the edges across different levels, and edges across different sources. Then we propose a co-mention and context based entity linking method and a topic-based social content alignment for link establishment. We further employ event based entity topic model to extract latent topics from news and calculate the relationships among events, topics and entities. Finally, we apply a multifaceted ranking strategy to rank news related information simultaneously incorporating with the content and structural information. Based on these techniques, an event based news search system named NewsMiner[4] is developed. Currently, we have collected 5865 events from *sina* special news, involving 549,607 news articles, 478,038 (22,983 distinct) persons, 913,982 (15,031) locations and 165,604 (7086) organizations. Furthermore, we build an event knowledge base to provide high quality news search experience, predict the events, topics and trends for new events.

The contributions of this article include:

- We analyze the necessity for event oriented news analysis and present an event based news mining and search framework, where news is formalized as a link-centric heterogeneous network to facilitate multi-source news integration and analysis.
- We propose a co-mention and context based entity linking method, building knowledge links between named entities and public knowledge bases.
- We propose a topic level alignment method for linking social content to news, in which we take the news as background knowledge of users' comments.
- We implement a probabilistic model and multifaceted ranking strategy for news analysis and search.

The rest of the article is organized as follows. Section 2 reviews the related literatures and analyzes the main differences from our methods. We present the system framework and the link centric news representation in Section 3. The detailed algorithms of four main components are described from Sections 4–7. Extensive experimental results are reported in Section 8, and finally Section 9 concludes our work.

2. Related work

There are several lines of researches that are related to our work, and we present some of the related literatures as follows:

[4] http://www.newsminer.net/.

2.1. News representation

News Representation The IPTC (*International Press Telecommunications Council*) specified standards for exchanging news since 1979 and had drawn up *NewsML* and *EventsML*, which were standards for conveying news and event information in news industry environment [2]. With the growing of the HTML pages, *rNews* was developed to specific the terminology and data model required to embed news specific meta data into HTML documents in October 2011 in [3]. Google, Microsoft and Yahoo! proposed a common markup vocabulary covering all the items in event [4].

In this article, we present a link-centric news representation which organizes news in three semantic level and links news to different types of information including social content and public knowledge base. Guided by the model, we develop a ranking strategy on the heterogenous network and visualize the search results.

2.2. Topic detection and tracking

Topic Detection and Tracking (TDT) is a multi-site research project, which discovers the topical structure in unsegmented streams of news reporting as it appears across multiple media and in different languages.[5] It ran from 1998 to 2004 and comprehensively addressed many important issues for better news understanding. There are many publications, like [5–7], and some of them employed named entities, e.g. [8–10].

Leskovec et al. [11] developed a framework for tracking short, distinctive phrases that travel relatively intact through on-line text. Li et al. [12] designed a flexible topic-driven framework for news exploration and Shahaf and Guestrin [13] investigated the method to construct connections between different pieces of information to form a chain rounding a specific topic. Lin et al. [14], Mei and Zhai [15] proposed ETP to discovering and summarizing the evolutionary patterns of themes in a text stream. Pouliquen et al. [16] developed EMM system for cross-lingual story tracking, gathering, grouping and linking news over time. Shan et al. [17] developed Tianwang for event extraction and retrieval on news-related historical data.

Actually, we do not work on event detection and tracking while we pay more attention on in-depth event analysis and try to figure out the common patterns they share in the future work.

2.3. Knowledge linking

Knowledge linking has been extensively studied, and existing methods can be classified into two broad categories as described in [18]:

Independent linking where each entity was mapped to a knowledge base entity independently. It compared the context of the entity with the text meta data associated with the entity in the knowledge base [19–22]. They differed in the features used (e.g. BOW, Wikipedia categories) and comparison techniques(cosine similarity or classifier). The main drawback was that they ignored the interdependence among entities.

Collective linking which works based on the hypothesis that a document usually referred to topically coherent entities. Besides context similarity, it employed the interdependence among entities in disambiguation. Typical usage were: the number of common Wikipedia pages that refer Wiki pages of these two entities in [23], the sum of pair-wise dependencies in [22,23] and the global interdependence in [24,25], etc. Moreover, Shen et al. [26] linked named entities in text with a knowledge base by leveraging the rich semantic knowledge embedded in the Wikipedia and the taxonomy of the knowledge base.

They mostly focused on linking entities to knowledge bases like DBpedia [27,28], while did not evaluate the effectiveness on Chinese knowledge base (e.g., *Hudong*, *Baidu*). It should be noted that the taxonomy tree in *Hudong* is inconsistency [29], which might have side effect on the existing methods.

2.4. News and social content

Social content analysis has attracted much attention of many researchers. Mei et al. [30] defined the general problem of spatiotemporal theme patterns discovery and proposed a probabilistic mixture model which explains the generation of themes and spatiotemporal theme patterns simultaneously in web blogs. Yang et al. [31] studied to leverage social information for web document summarization and proposed dual wing factor graph for summarizing news by incorporating with Twitter. Phan et al. [32] presented a general framework for building classifiers that deal with short and sparse text & Web segments by making the most of hidden topics discovered from large scale data collections.

Lu and Zhai [33] studied how to automatically integrate opinions expressed in a well-written expert review with lots of opinions scattering in various sources such as blogs, spaces and forums, and a semi-supervised model was raised for tackling it. Hong et al. [34] studied the problem of modeling text streams from two different sources – Twitter and Yahoo! News, addressed both their individual properties and their inter-relationships. Recently, Sil et al. [35,36] proposed to read news along with comments and gave a supervised matching method.

In this article, we propose news-comment topic model [37] to extract topic features from news and social content by taking news as the prior information, and then use them for linking news segments and social content in topic dimension.

2.5. Topic model

Considerable work has been conducted for learning topics from text. Two of the most popular models are the probabilistic latent semantic indexing (pLSI) proposed by Hofmann [38] and Latent Dirichlet Allocation (LDA) introduced by Blei in [39]. For modeling annotated documents, some variants were proposed, like in [40–42].

In order to adapt to different application scenarios, researchers have extended basic topic models to include other information contained in text documents. Rosen-Zvi et al.'s [43], Steyvers et al.'s [44] author-topic model used the authorship information together with the words to learn models. McCallum et al. [45] studied topic model in social network analysis and the proposed ART model learned topic based on emails sent between people. Tang et al. [46] proposed ACT model to simultaneously model the topical aspects of different types of information in the academic network, and [47] presented (D) ICT model to characterize the topical evolution of different types of objects (inventors, companies) within the patent network.

Compared with these prior work, our news-comment model pays more attention on content interaction between news articles and social content, and event based entity topic model [48] focuses on modeling news and inside named entities simultaneously within events.

3. NewsMiner framework and news representation

3.1. Framework

Fig. 2 shows the proposed framework which consists of six major components:

[5] http://projects.ldc.upenn.edu/TDT/.

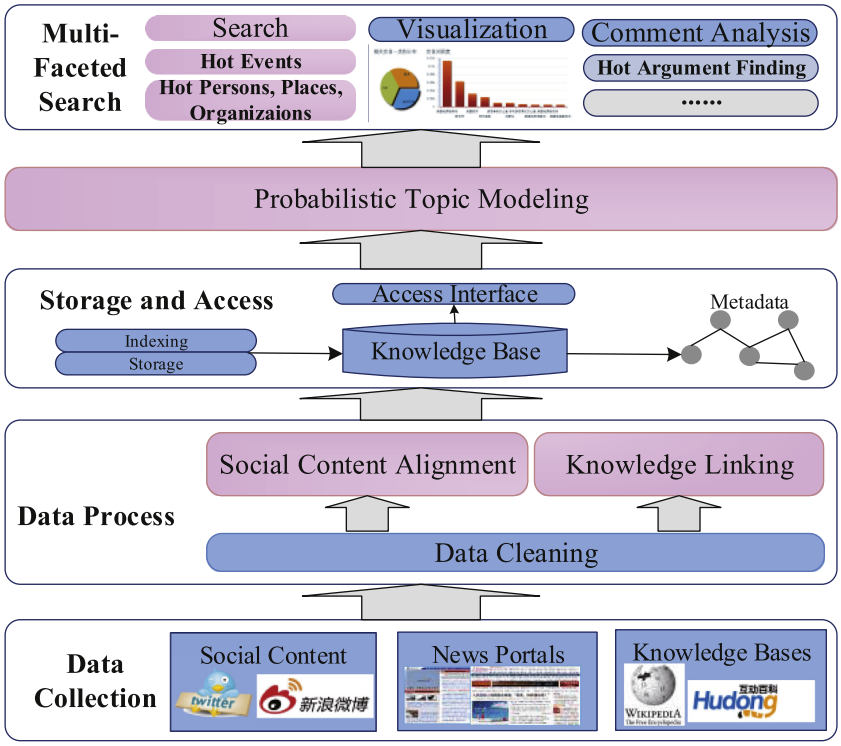


Fig. 2. Architecture of NewsMiner.

News collection and extraction. We collect multi-source news automatically, and perform preprocessing after crawling to extract the named entities and construct a heterogeneous information network containing event, news, comments, named entities and knowledge base entries.
Storage and access. It stores and indexes the extracted data in the news database.
Knowledge linking. It provides background knowledge for the extracted entities from news by linking them to the entries in *Hudong*[6] through our co-mention and context based method.
Social content alignment. We propose news comment topic model for news and social content integration, which helps provide a full view of the event.
Probabilistic topic modeling. We introduce event based entity topic model to simultaneously model words and named entities, and calculate their relationships based on the results.
Multifaceted event search. It provides several services, such as event search, hot entity search, argument finding, as well as a flash plug-into visualize the search result.

At present, we maintain a web crawler for news collection. It takes special news in news portals as event news, the comments issued by news readers after each news and the posts published on *Weibo* as the social content, and news-related entries in *Hudong* as background knowledge. For storage and access, we utilize the classical methods based on *MySQL* and *Apache Lucene*.[7] Therefore, knowledge linking, social content alignment, probabilistic topic modeling, and multifaceted search, are four key components in the framework.

3.2. News representation

First of all, it is important to find an appropriate way for news representation. It is required to express various news-related information and links among them. Therefore, we have the following definition:

News and related information can be formalized as a heterogeneous network $G = \{V, E\}$. *The vertices set* V *includes different objcts: news articles* $D = \{d_1, d_2, \ldots, d_{N_D}\}$, *events* $S = \{s_1, s_2, \ldots, s_{N_S}\}$, *topics* $T = \{t_1, t_2, \ldots, t_{N_T}\}$, *social content* $C = \{c_1, c_2, \ldots, c_{N_C}\}$, *named entities* $NE = \{ne_1, ne_2, \ldots, ne_{N_{NE}}\}$ *and knowledge base entries* $KB = \{kb_1, kb_2, \ldots, kb_{N_{KB}}\}$. E *is the set of the edges among them, with each weighted edge* $(v_i, v_j, p) \in E$ *suggests that there is a relationship between objects* v_i *and* v_j *with a probability* p.

These objects are correlated as follows: a news document $d \in D$ contains a vector w_d of N_d words, in which each word w_{di} is chosen from a vocabulary. It is usually published under an event $s_d \in S$ about a specific topic $t_d \in T$, associated with a collection of social content $c_d \in C$, and it contains a set of named entities $\{ne_d\} \subset NE$ which can be linked to knowledge base entries $\{kb_d\} \subset KB$.

Taking the types of the vertices into account, we can draw a three-level representation for the news network, namely event level, topic level and entity level as shown in the middle part of Fig. 3. The original news articles and corresponding social content are on the left, and the knowledge base is on the right part. Such representation provides a flexible way to explore news from different dimensions so that users can not only capture their interested events, topics or entities, but also can acquire the background knowledge or maybe historical information about the news.

Event. An event s is defined as the things happened at a specific time and location and it is the natural way to describe the news (e.g. in 5W1H[8]). For example, *Japan earthquake, American president election, world economic crises* are events and Fig. 3 presents three similar events about *earthquake*. Actually, event and its related information (i.e. latent topics, named entities, social content and knowledge base) form a subgraph of our defined network.

[6] http://www.hudong.com/, the biggest Chinese knowledge base.
[7] http://lucene.apache.org/.
[8] Who, When, Where, What, Why, How.

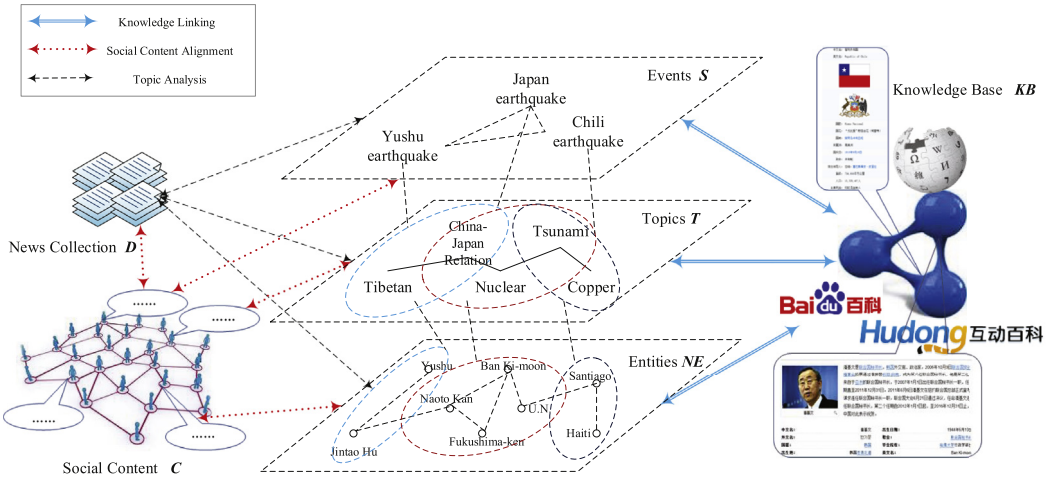


Fig. 3. News representation.

Topic. News often reports a specific event from various aspects, and topic is introduced to express a subject or theme related to the event. Conceptually, it can be described by a cluster of words that frequently occur together. Mathematically, a topic t is defined as a probability distribution over words $p(w|t)$.

Different people may care about different topics, for example, in *Job's Death*, economists may concern its influence on economic while Apple fans may pay more attention on how Apple products evolve. Different events may share common topics, e.g. we can find the topic of *international aid* in both *Japan earthquake* and *Yushu earthquake* in Fig. 3. Topics themselves are related to each other within or across events, for example, *casualty and loss* is related, to some extent, to *earthquake influence* in *Chili earthquake*.

Named entity. Named entities give the detailed answer of 5W1H, which consists of persons, locations and organizations. In our definition, a named entity *ne* is associated with a probability distribution over all related topics (same with normal words), and may be linked to public knowledge bases.

Different events have different important entities, *Steven Jobs* is the most important person in *Job's Death* while *Barack Obama* and *Mitt Romney* are more important in *American president election*. Even in the same event, the important entities in different topics may be different, *Red Cross Society* is important to topic *international aid* while *PLA* appears more in topic *rescue* in *Yushu earthquake*.

Social content. Social content is usually generated by information consumers, and conveys their understanding or opinions over events, topics or specific entities. It can be expressed through various ways while we only include the textual ones in our system, like comments following the news, blogs and microblogs. Each piece of social content c is associated with a news article, and it is also characterized as a vector with words chosen from a vocabulary.

Knowledge base. A knowledge base is a collection of entries, with each entry $kb_i \in KB$ containing *title*, *content* and probably *infobox*, where *title* is a string label, *content* can be represented as a vector with words chosen from a vocabulary, and *infobox* is structured information expressed by a set of attribute–value pairs.

The definitions show us a heterogenous news network, and now we are ready to present our work on dealing with the key issues mentioned above.

4. Co-mention and context based knowledge linking

Named entities are the key components to identify 5W1H in news and here we focus on three kinds of entities, namely *person*, *location* and *organization*. Recently, many large-scale knowledge bases have been built, such as *Wikipedia* in many languages, *Hudong* and *Baidu Baike* in Chinese, which contain rich information about the entities. Linking entities in news to these knowledge bases can provide background knowledge about entities as well as help update knowledge bases. The task of knowledge linking is described as follows:

Given a knowledge base $KB = \{kb_1, kb_2, \ldots, kb_m\}$, and a set of named entities $ne_{d_j} = \{ne_{d_j1}, ne_{d_j2}, \ldots, ne_{d_jn}\}$ in news document d_j, entity linking is, for each $ne_{d_ji} \in ne_{d_j}$, finding an equivalent entry $kb_k \in KB$.

A simple method for knowledge linking is to match entity labels with the entry titles in knowledge bases. Through observations, we find there are two types of features that may benefit the linking task:

Context. The probability that a named entity in news is linked to the knowledge entry is proportional to their context similarity.

Co-mention. If the target entity ne_i is mentioned along with entity ne_j in news, and ne_j also appears in the candidate knowledge entry kb_k, then the chance that ne_i and kb_k refer to the same entity is higher. For example, *Jordan* and *Bull* appear in same news, then *Jordan* is very likely to be the basketball player instead of the professor since they are also co-mentioned in some knowledge entries.

Therefore, we propose a simple and efficient co-mention and context based approach to build links between entities in news and public knowledge base, which can be divided into the following three steps:

Candidate selection. For each entity ne_{d_ji} from news article d_j, a list of entities with similar names $\{kb_1, kb_2, .., kb_m\}$ in KB are selected as the linking candidates. Name similarity is computed based on edit-distance of entity names between ne_{d_ji} and each entity kb_k in KB. Entity pairs $< ne_{d_ji}, kb_k >$ with similarities larger than a threshold σ are chosen as the candidates. Given the source string l_s and target string l_t, the edit distance is defined as:

$$Ed_{sim}(l_s, l_t) = 1 - \frac{|\{ops\}|}{max(length(l_s), length(l_t))} \quad (1)$$

where $|\{ops\}|$ indicates the minimum number of operations to transform l_s to l_t, and $length(\cdot)$ returns length of input string.

Entity disambiguation. Each named entity is associated with a vector which contains two types of features: context features are words extracted from news articles or knowledge base entries and weighted by word frequency; co-mention features are named entities which co-occurred with the entity in news articles or knowledge base entries and weighted by predefined values according to their types (e.g. an organization often contributes more than a location when setting up a person's feature vector). The entity pair with the largest cosine similarity is output as the linking result.

Infobox-based pruning. As defined in Section 3.2, the infobox of a knowledge entry *kb* consists of several attribute–value pairs that describe the entity-related information. Through observation, we find an organization often appears in knowledge base with different titles but similar infoboxes. Thus we conduct an extra step which tries to merge them by using the infobox information. Given two linking result entries kb_i and kb_j in *Hudong* with the infoboxes $I(kb_i)$ and $I(kb_j)$, we use the Jaccard similarity to measure the infobox similarity between them, and merge them if it is larger than a predefined threshold δ.

5. Link-based topic model for social content alignment

Social network applications encourage people to share their understanding and opinions about news which form social content, a light-weight mechanism for users' participation. But popular topics attract lots of social content, making it difficult for users to keep track of and assimilate the information. Social content alignment is trying to find the links between social content and news articles, which can not only provide additional knowledge about the news, but also can help user know what readers care about and bring more interactions. And our problem is defined as follows:

Given news segment collection $SG = \{sg_1, sg_2, sg_3, \ldots, sg_N\}$ and the corresponding social content $C = \{c_1, c_2, c_3, \ldots, c_M\}$, social content alignment aims at generating a set of matching triples $\{(c_i, sg_j, p_{ij}) | c \in C, sg \in SG, \text{ and } p_{ij} \in [0,1]\}$ which means c_i comments on sg_j with a probability p_{ij}.

It is difficult to perform matching just using the words because comments are usually short texts and they may use different words from news articles to discuss the same topic. As mentioned in Section 3.2, news reports different topics of one event and users often issue comments on their interested ones. Therefore we could employ topic information for the alignment task. We extract topics from news and comments, then represent news segments and comments by the obtained topic distribution and original word-level features, and finally calculate the relatedness between comments and news segments. For each comment, sort the related news segments and take the most-related one as its aligned result. Obviously, topic extraction is the most important component in the whole alignment process.

Through observation, we find that posts issued by users often leverage news articles. Therefore, we view the news as the prior information on which the users issue their comments, and propose the news comment topic model (as shown in Fig. 4) whose joint distribution is:

$$p(w,z,\theta_c,\theta_n,\phi|\alpha,\beta) = \prod_{m=1}^{M}\left[\lambda\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{c_{mk}}^{\alpha-1+n_{mk}} + (1-\lambda)\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{n_{mk}}^{\alpha-1+n_{mk}}\right]\prod_{i=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{j=1}^{W}\phi_{ij}^{\beta-1+n_{ij}} \quad (2)$$

where $\Gamma()$ is the *Gamma* function, α, β stand for prior distributions, θ_c, θ_n, ϕ are the matrixes representing the topic distribution over comments (and news) and word distribution over topics, λ is the impact factor that measures how news affects the generation of social content, w, z are words and topics and W, K are the number of them.

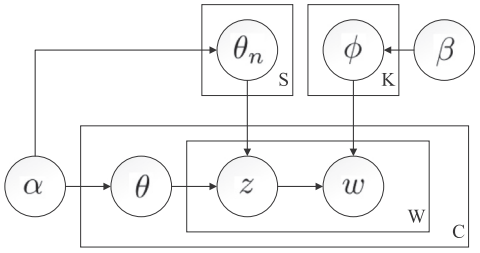


Fig. 4. Graphical representation of news comment topic model.

Following [43]'s guide, we choose Gibbs sampling for inference. As for the hyper parameters α and β, we take fixed values for simplicity (i.e., $\alpha = 50/K$, $\beta = 0.01$). In the sampling procedure, we maintain two count matrixes and assign a topic for every word at each iteration according to the posterior distribution in Eq. (3).

$$p(z_i = t | w_i, \mathbf{z}_{\neg i}, \mathbf{w}_{\neg i}, \alpha, \beta, \lambda) \propto \frac{(1-\lambda)\cdot n_{tc_{\neg i}} + \lambda \cdot n_{td} + \alpha}{\sum_{t'}[(1-\lambda)\cdot n_{t'c_{\neg i}} + \lambda \cdot n_{t'd}] + K\alpha} \frac{n_{wt_{\neg i}} + \beta}{\sum_{w'} n_{w't_{\neg i}} + W\beta} \quad (3)$$

where n_{tc}, n_{td} and n_{wt} represent the times that topic t has been sampled in comment c or news article d, the times that word w has been sampled in topic t respectively and '$\neg$' means exclusion.

6. Event-based entity topic model

Knowledge linking and social content alignment build links across different information sources, but how can we know the relationships between objects within events is another important problem in this article.

Statistical topic models, like pLSI and LDA have been highly successful at extracting the latent topics. However, news is not plain text, but a heterogeneous network which conveys event information about *who*, *what*, *when* and *where*,etc. Traditional statistical topic models do not distinguish between normal words and named entities so that they cannot explicitly address the textual interactions between named entities and topics. Thus we adapt entity topic model which is firstly proposed in [48] for topic extraction within event, and then infer the relationships among the entities and topics. It's important to note that we take the news articles in a specific event as the model input so that we named it event-based entity topic model.

Specifically, we employ CorrLDA2 [48] to generate both word topics and entity topics through a unified probabilistic model where word topics are the primary classes and may relate to different entity topics. The generation for normal words is the same with LDA, and when generating named entities, it first selects a super word topic, then samples an entity topic based on a multivariate distribution, finally samples an entity based on another multivariate distribution (readers who are interested in this part can refer [48] for details).

Let us briefly introduce notations, d, w, ne denote news article, word and named entity while the corresponding capital letters stand for their numbers. K and $\tilde{K}$ represent the number of normal word topics z and entity topics $\tilde{z}$. $\theta, \tilde{\theta}$ are topic models of news and $\phi, \tilde{\phi}$ represent the word distributions of topics. $\alpha, \beta, \tilde{\beta}, \gamma$ are *Dirichlet* priors for topic, word, entity and the relationships between two types of topics.

For parameter estimation, we also apply Gibbs sampling algorithm for its ease of implementation. The updating rules used for sampling are listed in Eq. (4) and (5):

$$p(z_i = t|w_i, \mathbf{z}_{\neg i}, \mathbf{w}_{\neg i}, \alpha, \beta) \propto \frac{n_{td_{\neg i}} + \alpha}{\sum_{t'} n_{t'd_{\neg i}} + T\alpha} \frac{n_{wt_{\neg i}} + \beta}{\sum_{w'} n_{w't_{\neg i}} + W\beta} \quad (4)$$

$$p(\tilde{z}_i = \tilde{z}|\tilde{w}_i = ne, \mathbf{z}, \tilde{\mathbf{z}}_{\neg i}, \tilde{\mathbf{w}}_{\neg i}, \tilde{\beta}) \propto \frac{n_{td}}{N_{w_d}} \frac{n_{\tilde{z}\tilde{d}_{\neg i}} + \gamma}{\sum_{\tilde{z}'} n_{\tilde{z}'t_{\neg i}} + \widetilde{T}\gamma} \frac{n_{ne\tilde{z}_{\neg i}} + \tilde{\beta}}{\sum_{ne'} n_{ne'\tilde{z}_{\neg i}} + E\tilde{\beta}} \quad (5)$$

Based on the modeling results, we can calculate the correlations between topics and entities. The topic entity relation can be directly derived from the estimated results, and a cosine similarity based method is chosen to define the correlation between two topics z_i and z_j [12]:

$$sim(z_i, z_j) = \frac{\sum_{k=1}^{W} p(w_k|z_i)p(w_k|z_j)}{\sqrt{\sum_{k=1}^{W} p(w_k|z_i)^2 \sum_{k=1}^{W} p(w_k|z_j)^2}} \quad (6)$$

where $p(w_k|z_i)$ can be found in the result matrix ϕ. Similarly, the weight of relations between two entities ne_i and ne_j can be calculated as follows:

$$sim(ne_i, ne_j) = \frac{\sum_{k=1}^{NE} p(z_k|ne_i)p(z_k|ne_j)}{\sqrt{\sum_{k=1}^{E} p(z_k|ne_i)^2 \sum_{k=1}^{E} p(z_k|ne_j)^2}} \quad with$$

$$p(z_k|ne) = \frac{p(ne|z_k)p(z_k)}{p(ne)} \quad (7)$$

where $p(z_k), p(ne)$ are the probabilities of the topic and entity within an event.

7. Multifaceted event ranking

The previous three sections build the links in the news network, while in this section, we turn to news search over the obtain network. The search requirements over heterogeneous news network give rises to several challenging issues and make them different from general keyword based search engines. Firstly, the information seeking practice [49] is not only about news itself, but also about other characteristics of news, like 5W1H. Secondly, a query from a user does not mean he/she wants to search news merely containing these words, perhaps his/her intention is to find news on the corresponding topics. So we formalize our task as follows:

Given a heterogeneous news network $G = \{V, E\}$, and a query $q = \{w_1, w_2, \ldots, w_n\}$, multifaceted ranking is to leverage the information from both *textual content* and *network structure* to obtain accurate ranking results for news, events, topics and named entities.

By this means, what we present to users is not a news list but the articles organized by events and the related named entities so that users can easily access the interested parts, making news browsing more understandable. Particularly, we propose a multifaceted ranking strategy, namely, ranking the news-related objects in different levels (i.e. corpus, event and topic): at corpus level, we find the most related events and named entities to the query; while at event level, we present the representative topics, entities as well as their relations; and readers can access the selected typical news and hot entities for their interested topics at topic level.

News article is the basic item in the ranking process, so we assign a score for each news by incorporating with the relevance score in *Lucene* [50] as well as the topic model information:

$$score(d|q) = \lambda score_L(d|q) + (1 - \lambda) \sum_{w \in q} p(d|w) \quad (8)$$

where $score_L$ returns the relevance score in Lucene and $p(d|w)$ can be derived from the topic modeling results. In practice, the topics are usually general to a given query while the relevance score is relatively specific. Combining these two scores achieves a balance between generality and specificity, thus could improve the ranking performance.

Table 1
The overall performance for entity linking.

Dataset	Entity	Precision (%)	Recall (%)	F1-Score (%)
KBP2012	*Per*	75.2	61.1	67.4
	Loc	43.9	25.5	32.3
	Org	61.3	43.8	51.1
CNnews	*Per*	94.9	76.7	84.9
	Loc	95.2	84.9	89.7
	Org	98.6	72.9	83.8

For other types of objects, we can combine all news associated with each object to create a virtual document, and then use a similar formula to calculate the score of them. But considering both event and topic are related to several news articles in our representation, their scores can be obtained via summing overall related news articles, namely:

$$score(s|q) = \sum_{d \in E} score(d|q)$$

$$score(t|s, q) = \frac{\sum_{d \sim t} score(d|q)p(t|d)}{p(t|s)} \quad (9)$$

where $d \sim t$ returns all the news articles related to topic t, and $p(t|s)$ is used for normalization and it gives the probability that a topic appears in the given event. The difference between event and topic lies in $p(t|d) = \theta_{dt}$ since news belongs to an event completely while it is related a topic over a probability distribution.

As mentioned in Section 3.2, the entity importance is different in different levels, for example, the important entity in an event may not be so significant in a specific topic and vice versa. Therefore, when ranking named entities, we calculate three different scores using Eq. (10) for the same named entity *ne*, indicating its importance in the search results, within an event and within a topic respectively.

$$score(ne|q) = \sum_{s} score(ne|s, q) \times score(s|q)$$

$$score(ne|s, q) = \sum_{t \in s} score(ne|t) \times score(t|s, q)$$

$$score(ne|t, q) = \eta \times p(ne|t) + (1 - \eta) \times \sum_{ne' \in N(ne)} p(ne'|t) \times sim(ne, ne') \quad (10)$$

where $N(ne)$ includes all the entities related to ne, $sim(ne, ne')$ is calculated using Eq. (7), and η controls a trade-off between its own score and neighbors' influence.

For each query, our system returns top $N(= 2000)$ news articles, ranks the related events, topics as well as named entities, and then presents the results in three levels as defined in Section 3.2. As for the parameters, we take fixed values experimentally (e.g. λ = 0.6, η = 0.5).

8. Experiment

In this section, we provide experiments to evaluate the effectiveness of major components in our proposed news search framework.

8.1. Knowledge linking

Datasets. We use two datasets, one is the most popular KBP2012[9] and the other comes from Tsinghua news and sina news (denoted as CNnews). The evaluation for KBP2012 is based on 2226 queries (918 persons, 602 locations and 706 organizations),

[9] http://www.nist.gov/tac/2012/KBP/index.html.

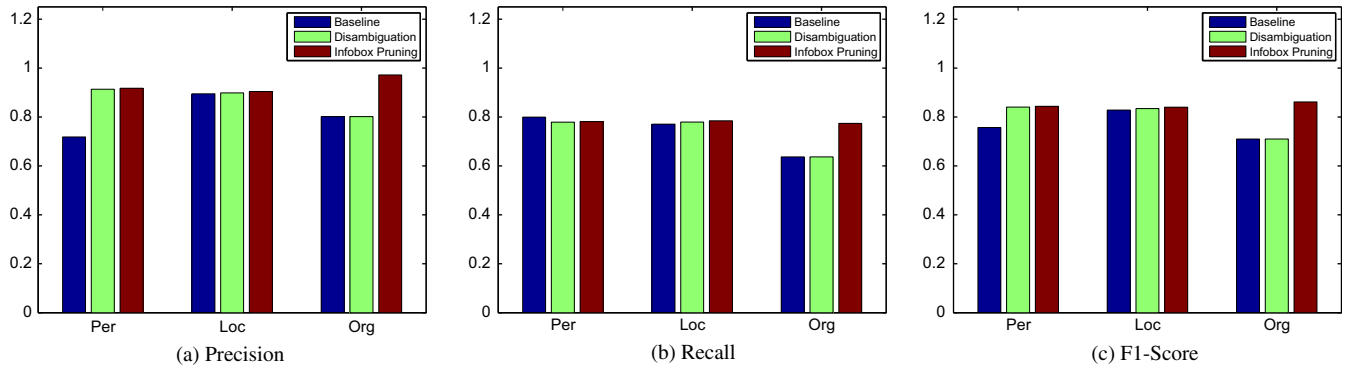


Fig. 5. Result on stepped experiment.

while for CNnews, we randomly selected 100 news articles and manually established the entity links between news articles and *Hudong* knowledge base as golden standard.

Result and analysis. First we present the overall performance of our algorithm on linking three types of entities, and Table 1 shows the results. We can see that our method achieves pretty good performance on CNnews, while performs not so well on KBP2012. Specifically, it could achieve comparable results with other methods presented in the reports[10] in person and organization linking (even better than some of them), but relatively poor results in location linking (about −10% than the best performance). One reason for that is our method is designed for Chinese entity linking without dealing with some important issues in English (e.g. abbreviation). Then we investigate the null and wrong links and draw the following conclusions:

- In CNnews, most null and wrong links in person entity linking come from the translation of foreign names. One foreign name can be expressed by several choices in Chinese and there are usually some special symbols inside(e.g. "–") which are beyond the ability of the tagging tools.
- In KBP2012, location (also named GPE in the track) is the most difficulty entity type to be linked. There are two reasons for that: it often has misleading document context which makes our method fail at the very beginning; there is not enough information in knowledge base for location disambiguation.

Then we show the experimental results by steps on *CNnews* (in Fig. 5) to validate the effect of each step, from which we can see:

- Compared with location and organization, person entities get a great improvement (19.5% in precision and 8.37% in F1-Score) during the disambiguation step because most of the ambiguous problems come from person entities.
- For organization, infobox pruning get a significant improvement (15% in average). We find some entities are not extracted or extracted with wrong entity types due to the lack of knowledge, which leads to the missing or wrong results. Infobox pruning can improve the performance by utilizing structural information in knowledge base.
- The result for location entities has a slight change (within 1%) which tells us in most cases the name-based matching method is somehow enough for location.

Using our proposed entity linking algorithm, we processed 549,607 news articles from *sina*, discovered 1,557,642 (45,100 distinct) entity links to *Hudong*. All these links are provided in NewsMiner system, which facilitates users to acquire background knowledge.

[10] http://www.nist.gov/tac/publications/2012/papers.html.

Table 2
Description on social content.

Sources	Descriptions	News	Comments
sina	A very popular Chinese news portal	61	29,544
cnn	One of the most popular news website	1303	62,225
bbc	The most famous news website in the UK	336	10,264
mtv	One of the most popular MTV site	176	9848
espn	One of the world's leading sports media	171	4320
mashable	The world's largest tech. blog	2940	114,441

8.2. Social content alignment

Datasets, baselines. To the best of our knowledge, no existing benchmark dataset is available for social content alignment evaluation. Thus we gather news articles along with social content from popular news websites and social media as shown in Table 2. *sina* consists of the sina daily hot news[11] following by users' comments from November 1 to December 31, 2011, and the others come from Twitter which is divided into five sub classes by news sources [31]. We use *SA-NCT* to denote our method, and compare it with the following three methods:

- *SA-VSM*: a similarity based method which employs TF-IDF and cosine similarity for measuring the relatedness between news segment and comment. It's the simplest way for social content alignment which we chose as our baseline method.
- *SA-SVM*: a classification based method which builds classifiers on news segments through which comments are classified. It is a supervised method, and we use five-folder cross-validation using libsvm[12] in our evaluation.
- *SA-LDA*: a topic based method which only replace our news-comment topic model by standard LDA to validate the effectiveness of our proposed model.

Result and analysis. The results are presented in Table 3 and the best performances in the comparisons are highlighted in bold. From the results, we can see that:

- SA-NCT outperforms other three methods in most cases and even though in *mtv* and *espn*, its performance is very close to the best approach.

[11] http://news.sina.com.cn/hotnews/.
[12] http://www.csie.ntu.edu.tw/cjlin/libsvm/.

Table 3
Results for news comment alignment.

	Metrics	sina	cnn	bbc	mtv	espn	mash	avg.
SA-VSM	Precision (%)	**73.2**	**66.3**	**71.2**	70.8	**72.6**	**69.4**	**70.6**
	Recall (%)	36.6	35.3	38.0	42.6	43.5	33.6	38.3
	F1-Score (%)	48.8	46.1	49.6	53.2	54.4	45.3	49.6
SA-SVM	Precision (%)	63.4	47.3	59.4	69.2	62.3	50.4	58.7
	Recall (%)	50.6	24.5	**49.8**	**59.5**	**59.7**	30.4	45.8
	F1-Score (%)	56.3	32.3	54.2	**64.0**	**61.0**	37.9	50.9
SA-LDA	Precision (%)	53.4	47.2	47.23	51.3	49.7	48.8	49.6
	Recall (%)	21.5	19.0	17.4	22.5	19.8	20.1	20.0
	F1-Score (%)	30.7	27.1	25.4	31.3	28.3	28.5	28.6
SA-NCT	Precision (%)	65.3	62.4	68.9	**70.9**	69.9	63.2	66.7
	Recall (%)	**51.8**	**39.5**	45.9	57.0	49.7	**35.8**	**46.7**
	F1-Score (%)	**57.8**	**48.4**	**55.1**	63.2	58.1	**45.7**	**54.7**

- SA-VSM obtain high precision, while our SA-NCT achieves relatively balance precision and recall and thus we outperforms it in the overall F1-score. Through investigation, we can find that SA-VSM always retrieves comments that use same words with news and the strict restriction make it miss many results, but SA-NCT unitizes the topical information to enhance the comment representation which benefits the alignment.
- The reason why SA-NCT outperforms SA-LDA where topical information is also employed is that: the comments are usually very short while LDA is based on word co-occurrence so that the modeling result is not satisfactory; but our NCT model employs the news as background knowledge which alleviates the sparseness problem to some extend.
- SA-NCT achieves relatively better on dataset with less news articles than dataset with more news (i.e. *cnn* and *mashable*), possibly because news conveys useful information as well as noise and more news will make the topics more diverse, which eventually leads to the performance drop.

8.3. Event based entity topic modeling

Datasets. We prepare the following three labeled datasets for event-based topic modeling evaluation (statistics are shown in Table 4):

- *TDT2*[13] is designed to include six months (January 4 to June 30, 1998) of material drawn on a daily basis from six English news sources.
- *Xinhua* is a Chinese news corpus from *Xinhua News Agency* for clustering evaluation.
- *Snow* comes from sina special news reported on Snow Disaster in South China from January 20 to April 25, 2008.

The first two have ground truth while for *Snow*, we randomly select 1000 reports, ask five annotators to provide human classification on each document and leave 568 documents with high agreement (at least three annotators give the same answer).

Another difficult issue for topic modeling is to determine the number of topics. We have the following rules: an event will not be processed until the number of news reaches 50; K (the number of topics) is set to be 5 if the number of news is between 50 and 200; K increases at the speed of 5 topics per 200 news; we set a maximal value (i.e. 50) based on the hypothesis that there cannot be too many topics for a specific event.

Result and analysis. Table 4 shows the performance of our algorithm on three labeled datasets, and we can observe that the performance on *TDT2* is not so satisfactory since it covers so many topics that makes it too sparse (a topic in *TDT2* consists of 18 articles at most), and that's why we limit the number of topics when analyzing event. We have the best performance on *Snow* since it talks about only one thing, which proves that our model work better on news within a specific event, in other words, event-based modeling makes sense.

[13] http://projects.ldc.upenn.edu/TDT2/.

Table 4
Results for topic modeling.

Dataset	# of news	# of topics	Precision (%)	Recall (%)	F1-Score (%)
TDT2	777	115	43.9	57.2	49.7
Xinhua	6000	6	89.2	64.9	67.6
Snow	3981	10	96.7	93.8	95.2

The result above better explains why we employ event-based model, and now let's turn around to some examples on real news. Table 5 lists three similar topics from different earthquake events, and we have the following observation:

- Entity topic model can distinguish normal words and named entities well as we can find few named entities in the top words.
- All the three topics are talking about the cause of the earthquakes (the word *geology* in *Yunnan*; *fault*, *plate* in *Haiti* and *Richter* in *Myanmar*) and their potential subsequent disasters, for example, *Yunnan* cares more about *weather* because it may lead debris flow while *tsunami* happened in *Haiti* due to it is coastal country.
- *Haiti* earthquake attracts more attention all over the world. One reason is that we can find *global* in its keywords, the other is that *Ban Ki-moon*, *Bill Clinton* are in its entity list and many famous news agencies reported it while in other two we can only some local persons and location names.

8.4. System overview and case study

System overview. NewsMiner is an event-based news search system by utilizing the methods mentioned in previous sections. Currently, the beta version has been already successfully running online for more than three years. Given a query, it returns the most-related events and named entities. If user is interested in some events, he/she can move forward to the analysis results as well as the event report. Fig. 6 presents its homepage and some major functions.

Up to June 2014, we have collected 5865 events, 549,607 pieces of news from 16 different categories, and after analysis 16,784 topics and 45,100 distinct named entities are extracted from the news collection. Table 6 shows statistics over the news collection and analysis results.

Table 5
Example for topic modeling.

Event	China (Yunnan)		Haiti		Myanmar	
Word	disaster	0.0253630	Haiti	0.05098583	seisesthesia	0.0350457
	geology	0.0150538	expert	0.01929279	time	0.0312262
	zone	0.0119751	fault	0.01143411	origin	0.0244014
	epicenter	0.0093774	tsunami	0.01133922	depth	0.0202377
	Yiliang	0.0090212	plate	0.00798098	Richter	0.0198068
	casualty	0.0077721	global	0.00683332	resident	0.0155545
	expert	0.0075416	outburst	0.00499315	Bangkok	0.0146670
	Yunnan	0.0074382	movement	0.00497235	capital	0.0120510
	weather	0.0072492	cause	0.00476049	measure	0.0111912
Entity	*Per*	Jiheng Li	*Per*	Ban Ki-moon	*Per*	Thein Sein
	Per	Binghui Luo	*Per*	Bill Clinton	*Loc*	Myanmar
	Per	Huiyan Liu	*Per*	Guosheng Qu	*Loc*	Thailand
	Loc	Shaotong	*Loc*	Haiti	*Loc*	Lao
	Loc	Yiliang	*Loc*	Port Prince	*Org*	AP
	Loc	Bijie	*Loc*	Dominicana	*Org*	Reuters
	Org	Xinhua	*Org*	AFP	*Org*	AFP

Fig. 6. Snapshot for NewsMiner.

Table 6
Statistics on some categories (QTY: quantity; PCT: percent).

Category	News		Event		Topic	Named entity		
	QTY	PCT (%)	QTY	PCT (%)		Per	Loc	Org
Sport	155,498	28.29	705	12.02	5452	6987	4562	1368
Inland	121,279	22.07	383	6.53	3170	11,378	10,523	2893
International	79,201	14.41	274	4.67	2431	2282	2944	806
Entertainment	5707	1.04	73	1.24	361	1014	563	89
Education	12,183	2.22	56	0.95	330	1794	1605	1381
Sci. & Tech.	13,820	2.51	137	2.34	755	1646	1025	463
Economy	82,010	14.92	661	11.27	3811	6144	3931	2169

Currently, we have some manual edits on the analysis result to ensure the correctness and make it more understandable. We are sure of that the accumulation of event knowledge will release us from that kind of manual work.

Case study. Now we present a case study to demonstrate the advantages of our system. We assume that, when users perform news search, they probably want to know when and where the event happens, who is involved and what other events the person is involved in, etc. When query is specified, we try to answer these questions and Table 7 shows parts of the results when different queries are given.

If the query is a normal word, taking *earthquake* as example, we list the recent earthquake events and the related entities like *Ban Ki-moon*. Starting with a person, we can explore his recent activities and behaviors. For *Steve Jobs*, the related events include *Apple WWDC (Worldwide Developers Conference) 2010*, *Apple WWDC 2011*, *Steve Jobs' Death*, *and the Release of iPad 2* which are all activities *Jobs* took part in, so that users can find their interests quickly. Moreover, we can track the historical information, for instance, we can get how the event evolves from *Western military intervention* to *Gaddafi's Death* when searching *Libya*.

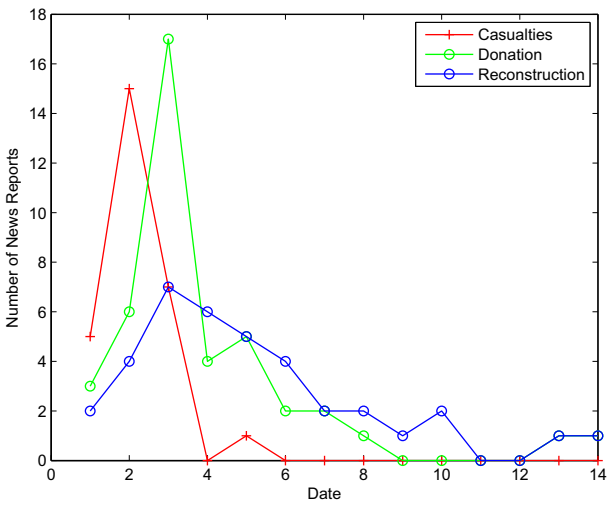
Table 7
Search results in NewsMiner.

Query	Event list	Hot entities
Jobs	Apple WWDC 2010	Steve Job
	Apple WWDC 2011	Kai-Fu Lee
	Steve Jobs' Death	Google
	the Release of iPad 2	Microsoft
	Jobs' on leave	
Libya	Western military intervention	Gaddafi
	Multinational military intervention	UNSC
	Libyan opposition scored capital	AU
	Gaddafi's death	LAS
Earthquake	Japan Earthquake	Ban Ki-moon
	Yushu Earthquake	Jintao Hu
	Chili Earthquake	Santiago
	Haiti Earthquake	U.N.

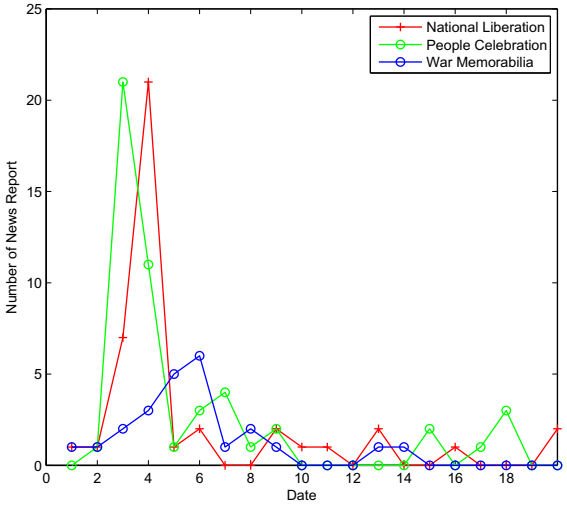
What's more, we can go deep into the interested events for better understanding. Taking *Gaddafi's Death* as an example, the related topics include *Wounded and Died*, *War Memorabilia*, *National Liberation*, *People's Celebration* and we further explore how the hotness of the topic changes over time. Fig. 7 shows some examples on the trends of the topics (we pick three topics from each event).

(1) As shown Fig. 7(a), we can see the *Casualties Statistics* comes right after the earthquake happens, then people *donate* to help the disaster area, and the *Reconstruction* of home is carried all the time.
(2) Fig. 7(b) shows two topics in *Gaddafi's Death*: *National Liberation* and *People's Celebration*. We can see that there is some delay between them, and it makes sense because *People's Celebration* goes before *National Liberation* naturally. And the *War Memorabilia* follows them to report some major points in the war.
(3) *Jobs* died on October 5th and lots of reports were published in the following two days. His *Biography* was scheduled to publish on 24th, thus the report on that lasted from his death to the release. Moreover, his death is a lost of world and it has a far-reaching influence on electronic industry which can be found in Fig. 7(c).

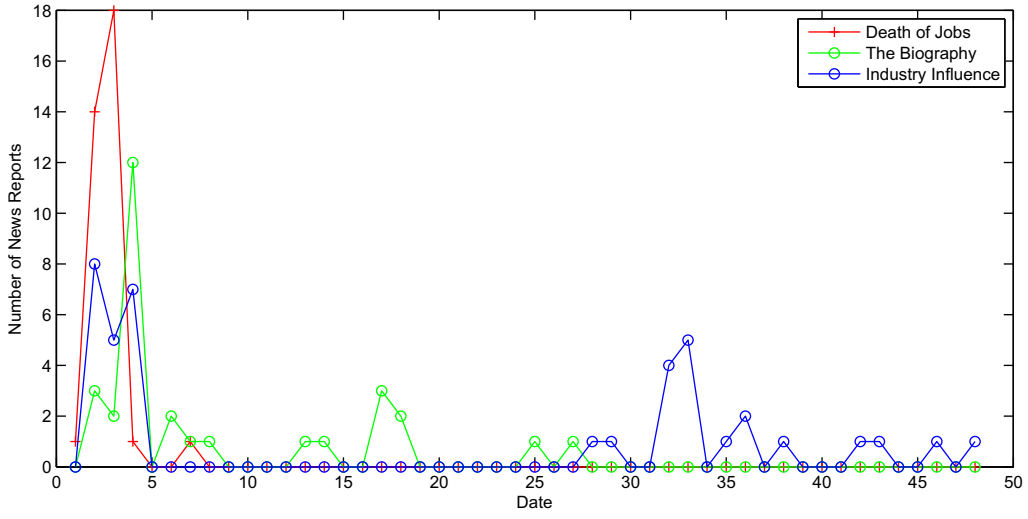
In conclusion, NewsMiner presents results in event-topic-entity with historical information instead of listing huge amount of latest



(a) Earthquake



(b) Gaddafi



(c) Steven Jobs

Fig. 7. Examples on topic trends.

news articles. Moreover, we extend the existing work (NewsInsight [12], EMM [16], etc.) by adding more functions, like in-depth event analysis, knowledge linking and social content alignment.

9. Conclusion

In this article, we propose a link centric news mining and search framework to tackle some important issues in news analysis. By considering the semantic nature of news, we propose a three-level representation for news event to integrate different types of news. Then we implement entity linking, social content alignment, and event based entity topic modeling to complete and enrich this representation. For result presentation, we develop an event based multifaceted ranking strategy to rank related objects simultaneously. Besides extensive evaluation to demonstrate the advantages of our proposed methods, we also give a brief introduction of the system implementation along with a case study.

News mining is an enduring and interesting problem. Thus there are many potential future directions of this work. First, we can explore the pattern that similar events may follow in topic level and make use of them to guide the analysis of new events. We can also perform more in-depth research on social content after aligning them to news articles, e.g. sentiment analysis. Finally, improving our search and browsing experience with social information added in is an intriguing direction for future research.

Acknowledgment

The work is supported by the NSFC (No. 61035004), NSFC-ANR (No. 61261130588), 863 High Technology Program (2011AA01A207), European Union 7th Framework Project FP7-288342, and THU-NUS NExT Co-Lab.

References

[1] C. I. N. I. Center, Statistical Reports on the Internet Development in China, Technical Report, CNNIC, 2013.
[2] I. P. T. Council, Eventsml-g2: a solution for collecting and distributing structured event information, 2014, <http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/>.
[3] I. P. T. Council, rnews, 2011, <http://dev.iptc.org/rNews>.
[4] Google, Microsoft, Yahoo!, Schemas – schema.org, 2012, <http://www.schema.org/docs/schemas.html>.
[5] J. Allan, Topic detection and tracking: event-based information organization, Kluwer Academic Publishers, Norwell, MA, USA, 2002.
[6] J. Allan, R. Papka, V. Lavrenko, On-line new event detection and tracking, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 37–45.
[7] Y. Yang, T. Pierce, J.G. Carbonell, A study of retrospective and on-line event detection, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 28–36.
[8] G. Kumaran, J. Allan, Text classification and named entities for new event detection, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 297–304.
[9] C. Shah, W.B. Croft, D. Jensen, Representing documents with named entities for story link detection, in: Proceedings of the 15th ACM Conference on Information and Knowledge Management, 2006, pp. 868–869.
[10] K. Zhang, J. Li, G. Wu, New event detection based on indexing-tree and named entity, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 215–222.
[11] J. Leskovec, L. Backstrom, J.M. Kleinberg, Meme-tracking and the dynamics of the news cycle, in: Proceedings of the 15th ACM International Conference on Knowledge Discovery in Data Mining, 2009, pp. 497–506.
[12] J. Li, J. Li, J. Tang, A flexible topic-driven framework for news exploration, in: Proceedings of the 15th ACM International Conference on Knowledge Discovery in Data Mining (Demo), 2009, pp. 338–349.
[13] D. Shahaf, C. Guestrin, Connecting the dots between news articles, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011, pp. 2734–2739.
[14] C.X. Lin, B. Zhao, Q. Mei, J. Han, Pet: a statistical model for popular events tracking in social communities, in: Proceedings of the 16th ACM International Conference on Knowledge Discovery in Data Mining, 2010, pp. 929–938.
[15] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining, 2005, pp. 198–207.
[16] B. Pouliquen, R. Steinberger, O. Deguernel, Story tracking: linking similar news over time and across languages, in: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization at 22nd International Conference on Computational Linguistics, 2008, pp. 49–56.
[17] D. Shan, W.X. Zhao, R. Chen, B. Shu, Z. Wang, J. Yao, H. Yan, X. Li, Eventsearch: a system for event discovery and retrieval on multi-type historical data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 1564–1567.
[18] C. Wang, K. Chakrabarti, T. Cheng, S. Chaudhuri, Targeted disambiguation of ad-hoc, homogeneous sets of named entities, in: Proceedings of the 21st International World Wide Web Conference, 2012, pp. 719–728.
[19] R.C. Bunescu, M. Pasca, Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of 11st Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 9–16.
[20] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R.V. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, J.Y. Zien, Semtag and seeker: bootstrapping the semantic web via automated semantic annotation, in: Proceedings of the 12th International World Wide Web Conference, 2003, pp. 178–186.
[21] R. Mihalcea, A. Csomai, Wikify!: linking documents to encyclopedic knowledge, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, 2007, pp. 233–242.
[22] D.N. Milne, I.H. Witten, Learning to link with wikipedia, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008, pp. 509–518.
[23] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of wikipedia entities in web text, in: Proceedings of the 15th ACM International Conference on Knowledge Discovery in Data Mining, 2009, pp. 457–466.
[24] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 782–792.
[25] X. Han, L. Sun, J. Zhao, Collective entity linking in web text: a graph-based method, in: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 765–774.
[26] W. Shen, J. Wang, P. Luo, M. Wang, Linden: linking named entities with knowledge base via semantic knowledge, in: Proceedings of the 21st International World Wide Web Conference, 2012, pp. 449–458.
[27] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives, Dbpedia: a nucleus for a web of open data, in: Proceedings of the Joint Conference of 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, 2007, pp. 722–735.
[28] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia – a crystallization point for the web of data, J. Web Semantics (2009) 154–165.
[29] Z. Wang, Z. Wang, J. Li, J.Z. Pan, Building a large scale knowledge base from chinese wiki encyclopedia, in: Proceedings of Joint International Semantic Technology Conference 2011, 2011, pp. 80–95.
[30] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: Proceedings of the 16th International World Wide Web Conference, 2007, pp. 171–180.
[31] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, J. Li, Social context summarization, in: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 255–264.
[32] X.H. Phan, M.L. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th International World Wide Web Conference, 2008, pp. 91–100.
[33] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, in: Proceedings of the 17th International World Wide Web Conference, 2008, pp. 121–130.
[34] L. Hong, B. Dom, S. Gurumurthy, K. Tsioutsiouliklis, A time-dependent topic model for multiple text streams, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 832–840.
[35] D.K. Sil, S.H. Sengamedu, C. Bhattacharyya, Readalong: reading articles and comments together, in: Proceedings of the 20th International World Wide Web Conference (Poster), 2011a, pp. 125–126.
[36] D.K. Sil, S.H. Sengamedu, C. Bhattacharyya, Supervised matching of comments with news article segments, in: Proceedings of the 20th ACM Conference on Information and Knowledge Management, 2011b, pp. 2125–2128.
[37] H. Xia, J. Li, Plink-lda: using link as prior information in topic modeling, in: Proceedings of 17th International Conference on Database Systems for Advanced Applications, 2012, pp. 213–227.
[38] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 50–57.
[39] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Machine Learning Res. (2003) 993–1022.
[40] D.M. Blei, M.I. Jordan, Modeling annotated data, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 127–134.

[41] D.M. Blei, J.D. McAuliffe, Supervised topic models, in: Proceedings of the 21st Annual Conference on Neural Information Processing Systems, 2007, pp. 121–128.
[42] C. Wang, D.M. Blei, F.-F. Li, Simultaneous image classification and annotation, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 1903–1910.
[43] M. Rosen-Zvi, T.L. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 487–494.
[44] M. Steyvers, P. Smyth, M. Rosen-Zvi, T.L. Griffiths, Probabilistic author-topic models for information discovery, in: Proceedings of the 10th ACM International Conference on Knowledge Discovery in Data Mining, 2004, pp. 306–315.
[45] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on enron and academic email, J. Artif. Intell. Res. (2007) 249–272.
[46] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: extraction and mining of academic social networks, in: Proceedings of the 14th ACM International Conference on Knowledge Discovery in Data Mining, 2008, pp. 990–998.
[47] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, A.K. Usadi, Patentminer: topic-driven patent analysis and mining, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 1366–1374.
[48] D. Newman, C. Chemudugunta, P. Smyth, Statistical entity-topic models, in: Proceedings of the 12th ACM International Conference on Knowledge Discovery in Data Mining, 2006, pp. 680–686.
[49] M. Hertzum, A.M. Pejtersen, The information-seeking practices of engineers: searching for documents as well as for people, Inform. Process. Management (2000) 761–778.
[50] M. McCandless, E. Hatcher, O. Gospodnetic, Lucene in Action, second ed., Covers Apache Lucene 3.0, Manning Publications Co., Greenwich, CT, USA, 2010.