A Dynamic Reconstruction Approach to Topic Summarization of User-Generated-Content

Zhao-Yan Ming
School of Computing
National University of Singapore
mingzhaoyan@gmail.com

Jintao Ye
MOZAT PTE.LTD.
77 Science Park Drive
05rjgcyjt@gmail.com

Tat-Seng Chua
School of Computing
National University of Singapore
chuats@comp.nus.edu.sg

ABSTRACT

User generated contents (UGCs) from various social media sites give analysts the opportunity to obtain a comprehensive and dynamic view of any topic from multiple heterogeneous information sources. Summarization provides a promising means of distilling the overview of the targeted topic by aggregating and condensing the related UGCs. However, the mass volume, uneven quality, and dynamics of UGCs, pose new challenges that are not addressed by existing multi-document summarization techniques. In this paper, we introduce a timely task of dynamic structural and textual summarization. We generate topic hierarchy from the UGCs as a high level overview and structural guide for exploring and organizing the content. To capture the evolution of events in the content, we propose a unified dynamic reconstruction approach to detect the update points and generate the time-sequence textual summary. To enhance the expressiveness of the reconstruction space, we further use the topic hierarchy to organize the UGCs and the hierarchical subtopics to augment the sentence representation. Experimental comparison with the state-of-the-art summarization models on a multi-source UGC dataset shows the superiority of our proposed methods. Moreover, we conducted a user study on our usability enhancement measures. It suggests that by disclosing some meta information of the summary generation process in the proposed framework, the time-sequence textual summaries can pair with the structural overview of the topic hierarchy to achieve interpretable and verifiable summarization.

Categories and Subject Descriptors

H.3.1 [**[Information Storage and Retrieval**]: Content Analysis and Indexing abstracting methods; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Experimentation

CIKM'14, November 03–07, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661936.

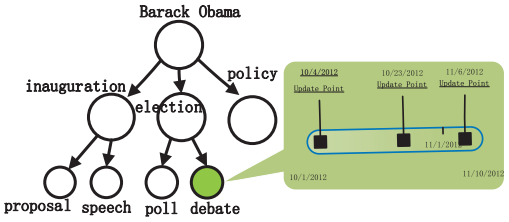

Figure 1: An illustration of the time-sequence topic summary for dynamic social media contents. On the left is an automatically generated high-level topic overview. When a subtopic is selected, the time-sequence summaries are shown at the update points along a timeline on the right (enlarged in Figure 2).

Keywords

Topic Hierarchy, User Generated Content, Update Summarization

1. INTRODUCTION

User generated contents (UGCs) from social media sites such as Twitter, Facebook, Yahoo! Answers, and Bloggers, give analysts the opportunity to obtain a comprehensive set of information on any topic. However, it is also very challenging for the analysts to get an overview, track the evolution, and gain insight on the topic of interest. There is a great need for an automatic summarization system that can distill an overview from the multi-source heterogeneous UGCs to support inquiries in temporal and topical dimensions.

Towards this goal, we propose a novel time-sequence topic summarization (TTS) task. An envisioned summary result can be illustrated by Figure 1. In particular, a high level topic summary generated from the target UGCs is presented on the left. With the high-level overview as a navigation aid to detailed topics, the time-sequence prose summary is presented along a timeline beside the subtopics of focus. Such summaries can serve as the starting point from which the analysts can perform subsequent actions to explore the data and validate the hypothesis that they may have in certain time periods and on subtopics of different granularities.

Given the huge volume, the heterogeneous sources, and the dynamics of the UGCs, this novel summarization task is extremely challenging for existing methods. To tackle the scale issue of the UGCs, high-level term-based summaries, such as key word cloud and topic model [4], achieve scalable abstraction by presenting the term or term-co-occurrence statistics. However, these approach-

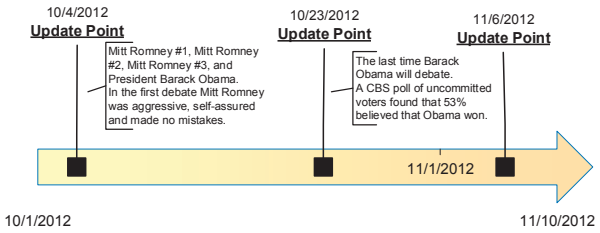

Figure 2: Time-Sequence summarization: given the continuously generated multi-source UGCs, the update points are detected and the associated prose summaries are emitted along the timeline.

es are either too shallow and context-free for in-depth analysis, or not easily interpretable for specific tasks [5]. Given the high level overview, there is still a strong need for a human-readable sentence based summary (we call it prose summary in this work). Multi-document summarization (MDS) that usually extracts salient sentences from the target document set provides a plausible solution. However, MDS is optimized for standard evaluations on static and relatively small content of trustful quality, and may not be able to generate meaningful summaries for large scale heterogeneous UGCs.

An appealing solution here is the combination of high-level overview and a human-readable sentence based summary. The existing visual analytic systems such as Tiara [31] and HierarchicalTopic [7] that present both the topical overview and the associated data exhibit similar idea. However, the natural language summary is usually missing, while only the raw content for a component subtopic is displayed directly.

The dynamics of UGCs provide timely information. However, it also raises issues about when to produce a new summary and how to update upon the old one, on a timeline such as the example presented in Figure 2. Existing approaches mostly take a subtractive approach where the redundant content from the reference summary is removed from the new summary. Despite its simplicity, this approach loses consistency in the summarization process, because the reference summary does not influence the generation of the new summary. It is also different from a human approach where the previous summary is considered during the generation of new summarization, rather than as a post process on the new summary. Moreover, the study on the update point detection is still rare in literature.

To address the above issues, we propose a temporal and topical summarization scheme that accounts for the dynamicity and the massive amount of UGCs from a novel data reconstruction perspective. In our proposed scheme, the summary sentences are selected to reconstruct the original documents. The proposed scheme has two key features:

First, TTS tackles the scale issue by employing a topic hierarchy as the high level overview, with which users can specify the desired level of generalisation by browsing through the hierarchy. The topic hierarchy helps MDS scale up by naturally dividing the content into subtopics. The subtopic relation is also naturally added into the representation of the UGCs to be summarized.

Second, to adapt to the need of time-sequence summarization, a dynamic data reconstruction model is proposed to smoothly generate the update summaries by minimizing the reconstruction errors. The update point is also naturally determined by monitoring the changes in reconstruction errors.

We conduct extensive experiments on a real world dataset of UGCs from some popular social media services. The results show the effectiveness of the proposed framework and its components. Moreover, a user study suggests that by disclosing some meta information about the summary generation process in the proposed framework, the generated summaries can pair with the visual overview of the topic hierarchy to achieve interpretable and verifiable text summarization.

The remainder of this paper is organized as follows. We introduce related work in Section 2. Our problem formulation is detailed in Section 3. Our strategy for dynamic multi-source UGC summarization, is described in Section 4. Section 5 presents and discusses the experimental results and Section 6 concludes the paper.

2. RELATED WORK

2.1 Traditional Document Summarization

Multi-document summarization techniques have long been studied in the area of information retrieval. Previous work mostly uses document word frequency to evaluate the importance of the candidate sentences. To take the sentence relation into consideration, LexRank [8] achieved good summarization performance by modeling graph-based lexical centrality as the sentence salience measure. [28] further proposed a cluster-based Markov Random Walk model on the sentence graph similar to that of LexRank. To represent the documents in a higher level of representation than Bag-of-Words, statistical topic models are well recognized and widely adopted for revealing the underlying topics of a collection and thus guiding the selection of sentences that cover the topics [2, 29, 10]. More recently, [12] proposed to build a summary from a data reconstruction perspective where sentences that best reconstruct the original documents are selected.

To incorporate a deeper understanding on the topics to be covered, TREC introduced the guided summarization based on some pre-specified topics or automatically extracted aspects using NLP techniques [35, 34]. Lin et al. [37] explored the category-specific information in multi-document summarization as the common ground for sentence selection. Despite the success of these summarization models, they are designed for static and small scale document set; they are not directly applicable to multi-source UGCs.

Besides topic coverage, the temporal dimension is attracting attention as shown in the TREC update summarization track [6] and temporal summarization track.[1] The update summary aims to inform users of new information about a topic given the arrival of some new documents, assuming that the user has already read the previous documents. Towards a unified framework for updating multi-document summarization, Wang and Zhou [30] employed a topic modeling approach for salience determination and a dynamic modeling approach for redundancy control. Temporal summarization is formally introduced by Allen et al. [1] in 2001. It includes the update function in a sequential way on a timeline. While one of our tasks is to generate time-sequence summaries, the user contributed nature of UGCs entails new challenges such as the large volume and unknown reliability. Therefore, in this work we propose a novel data reconstruction approach to meet the new requirements.

2.2 UGC Summarization

UGC summarization has gained popularity in recent years following the booming of social media services. Great efforts have been devoted to the summarization of microblogs. Earlier work by

[1] http://www.trec-ts.org/

Harabagiu et al. [11] introduced a microblog summarization framework combining two relevance models: an event structure model and a user behavior model. Sharifi et al. [25] proposed the Phrase Reinforcement algorithm to pick a single tweet to summarize multiple tweet posts on the same topic. Inouye et al. [15] proposed a hybrid TFIDF algorithm and a cluster-based redundancy removal scheme to generate multiple tweet summaries. Considering the posting time of microblog, Takamura et al. [27] proposed a stream summarization model based on the p-median problem for tweets on the timeline. More recently, Shou et al. [26] proposed novel data structures for efficient continuous summarization in a system named Sumblr, that explores the scalability and efficiency issues facing the large volume dynamic tweet stream. To account for user interest, [23] utilized the users' historical tweets and social circles to generate personalized time-aware tweet summaries. Besides microblogs, event-related updates in Wikipedia editing history are also targets of temporal summarization in recent works [9].

Customer reviews are another popular subject of the UGC summarization problem. Hu and Liu [13, 14] presented the early work on summarizing opinions in customer reviews. Zhai et al. [33] explored the use of structured ontologies and social networks for the generation of enhancing the opinion summaries. As in guided summarization, aspects are recognized in reviews for accurate opinion summarization [17]. To generate a more informative summary for understanding the opinions, Kim et al. [16] proposed to select sentences with high explanatoriness.

However, to the best of our knowledge, there is no prior work on the dynamic summarization of multi-source UGCs for a given topic. The most related work we found is Ren et al.'s [23] on tweet summarization enriched by the corresponding Wikipedia articles. It is shown to effectively add semantics [19] and thereby enhance the quality of the tweet summary. In this paper, we use UGCs from multiple sources to capture a complete picture of a topic.

3. PROBLEM FORMULATION

We consider the problem of generating summaries for dynamic multi-source UGCs. The task is modeled in two orthogonal dimensions: temporal and topical. For the topic summary, we assume that the UGCs to be summarized are on a given topic. The topic is outlined by a topic hierarchy with subtopics and relations. The topic summary can thus be generated for any subtopic in the hierarchy. For time-sequence summary, we assume that the user is continuously reading and expecting updates in a time period. At any given time, a new summary is to be generated for the UGCs since the last summary, that will include updates of any new information that is not covered by the previous summary. In the following, we will first give our definitions for topic summary and the time-sequence summary, followed by the formal task definition.

Preliminary 1. A *Topic Hierarchy* (TH) is defined as a tree that consists of a set of unique nodes of topic set $T : \{\tau_1, ...\tau_{|T|}\}$ and a set of parent-child relations R between T. The root of TH is the most general topic and the leaves the most detailed topics.

Given a document collection U of a focused topic in the time period $[t_0, t_p]$, TH that is extracted from U can serve as its topical outline, or a high-level summary. U can be organized into a hierarchy of UGCs that shares the same structure as TH, with each node contains a cluster of objects under the corresponding topic in TH.

Definition 1. A **Topic Summary** TS for a collection of UGCs U on a topic T is defined as a high level topical outline $TO : \{to_i, ..., to_{|T|}\}$ containing $|T|$ subtopics and a set of sentence-based summaries $SS : \{ss_1, ..., ss_{|T|}\}$ for each subtopic to_i in TO. Without loss of generality, we define the topical outline as a topic hierarchy TH_T on T that organizes U; and ss_i as a set of natural language sentences selected from UGCs depicting a node topic τ_i in TH_T.

With sentence-based summaries for subtopics at all levels of the TH, the topic summary of a collection provides a flexible way of exploring a large text collection at various levels of granularity.

The second focus of our UGC summarization task is on the temporal dimension of the collection, namely, the generation of a sequence of summaries with respect to time.

Definition 2. A **Time-Sequence Summary** TSS for a collection of time sequenced UGCs U is defined as a sequence of summaries $\{tss_1, ..., tss_{N_t}\}$ dynamically generated based on U along the time period $[t_0, t_p]$. An tss_i corresponds to a detected update point t_i in the time period $(t_{i-1}, t_i]$, in which there is significant amount of new content as compared to the previous summary tss_{i-1}.

Formally, the time-sequence and topic summarization (TTS) task is defined as follows:
Input: a topic T and its associated UGCs from multiple sources $U : \{u_1, u_2, ..., u_j, ..., u_n\}$ where u_j is the content from source j, from time t_0 till the present or an end of interest time t_p.
Output: the output at time t is a topic summary TS consisting of a topic outline in the form of a topic hierarchy TH and a set of prose summary SS for each subtopic node τ. During the whole period, the sequence of prose summaries under a subtopic τ adheres to the update points during $[t_0, t]$, forming a time-sequence summary SP_τ.

Specifically, we envision a summary that combines the strength of the term/topic based summary and the extractive summary as shown in Figure 1. For the topic summary, we choose the topic relation graph for its better expressiveness; and for extractive summary, we choose a data reconstruction based summarization model that naturally integrates with the topic summary.

4. TIME-SEQUENCE TOPIC SUMMARIZATION OF UGCS

In this section, we detail our time-sequence topic summarization method. First, at a given time t, we extract the high level topic summary in the form of topic hierarchy. The topic hierarchy organizes the UGCs into levels of subtopics, which scale down the raw UGCs into manageable size for the subsequent textual summarization. Second, we consider the time period of interests and generate the time sequence textual summary based on the organized UGCs with a novel dynamic reconstruction perspective.

4.1 Topic Summary Generation

In the following, we first introduce our method for topic hierarchy construction from UGCs. The topic hierarchy is used as the high level outline in our framework. Second, we organize the UGCs according to the topic hierarchy. We then propose a topic hierarchy based sentence representation scheme that enhance the usual term-based representation for the textual summary generation. Finally, we discuss other options of high level summary choices.

4.1.1 Topic Hierarchy Construction

A topic hierarchy summarizes a set of documents on a specific topic using a set of nodes (subtopics) and edges (relations). Given the multiple sources of UGCs contributed by different users, we expect the topic hierarchy constructed to reflect the interests of multiple users, thus represent a general outline for the main topic. Here,

a three-step approach is adopted: first we extract the key subtopics from the document set to be summarized; next we establish the relations between these subtopics using linguistic and statistic evidences [32, 36]; and finally we produce the topic hierarchy from the subtopic relation graph. While the general flow is similar to the state-of-art approaches, our approach is customized to the summarization task with distinct features for handling the UGCs.

First, we extract the salient noun phrases (NP) from the focused document set using tf-idf based keyword extraction. NP length heuristic (1-3 words) and frequency heuristic (document frequency higher than three) are applied to eliminate possible noise. Instead of choosing the keywords that appear in more than one UGC sources to ensure quality, we impose a text quality threshold to discard NPs from UGC pieces that are not well written (cf. Section 4.1.2). This is to ensure that low quality UGCs are not contributing to the key subtopics to be covered in the summaries.

Second, mid and high level subtopic terms in the hierarchy are enriched from external resources such as Wikipedia and WordNet. According to [36], UGCs usually do not contain relatively abstract terms and thus the higher level ones need to be harvested elsewhere. However, introducing too many subtopics from outside the document set may cause the summaries not to be faithful to the UGC sources. We thus further restrict that the lowest two levels of the hierarchy to contain subtopics from the original documents.

Third, the pairwise relatedness between two extracted keywords are calculated in order to connect the subtopics. We mainly adopt the cosine similarity of the contextual distribution of the two terms, while relation evidences from external resources, such as pointwise mutual information based on the Wikipedia articles and path distance on WordNet are also used [32, 36]. In this work, we further impose that if two terms are both from the original documents, the relatedness measures should come from the original documents as well, without using the external resources.

Finally, the pairwise relation based sub-topic graph is pruned in order to generate a valid topic hierarchy. In particular, an iterative approach is taken where at each step, one sub-topic is added into the hierarchy. This approach is readily extendable for our requirement of incremental update of the hierarchy along the time. Different from [36], we keep all the historical nodes in the hierarchy to show the complete overview of the ever accumulating UGCs. When the topic hierarchy is presented, the active/inactive nodes (subtopics) can be highlighted using visual clues.

4.1.2 UGC Organization Using Topic Hierarchy

Besides the role of the high level overview, the topic hierarchy is used as a content organizer. Given a collection of documents on a topic τ, we now partition and map them into the categories that are defined by a topic hierarchy on τ, such that the formed document clusters CO_1, CO_2,..., CO_k are organized in a similar hierarchy. During the topic hierarchy extraction process, a sentence is naturally assigned to the topic(s) that it contributes to. For those sentences that do not contribute to any topic directly, we use a topic assisted clustering approach [21] to assign them to the hierarchy.

Given the organized content in the corresponding DH, the topic summarization can be done in a bottom-up approach. For a leaf node with data d, the summary is generated on the data directly, with $X^* = \arg\min \mathcal{L}(d, X, \mathbf{A})$. For a non-leaf node, the summary $X^* = \arg\min \mathcal{L}(X_c, X, \mathbf{A})$ is generated on the aggregated summary sentences X_c from its child nodes.

Now only the leaf level subtopics that contain content at a much smaller scale are summarized on raw content. The intermediate level nodes, which have more data to be summarized and thus may consume more computational resources, are now handling fewer sentences that constitute the summaries of the child nodes. As a result, the topic hierarchy based content organization and the progressive summarization greatly reduce the computation costs, and helps in managing the scale issue of the UGC summarization task.

4.1.3 Topic Hierarchy Based Sentence Representation

The topic structure provides the key subtopics to be covered in the summary, thus can be emphasized in the sentence representation. In the original data reconstruction framework, the sentences are represented by the individual terms and the reconstruction is happening in the term space. A critical drawback here is that term-based representation is suboptimal for reconstructing the meaning of a document collection. Intuitively, a subspace that approximates the original one in terms of topics is closer to the actual needs of the summarization task. Therefore, in this work, we propose to use both the terms and the topics to represent the sentences $u :< \boldsymbol{\omega}, \boldsymbol{\tau} >$ for reconstruction. Here $\boldsymbol{\omega} : \{\omega(t)\}$ is the tf-idf weighted term vector, and $\boldsymbol{\tau}$ indicates the presence of the subtopics that the sentence covers.

We further propose to calculate the *Subtopic Specific Importance* (SSI) of terms based on their subtopic memberships and interpolate it with the tf-idf based term weight. Similar ideas are explored in domain-specific term weighting methods such as [20], where terms in a vocabulary is weighted to reflect its specificity [22]. As the UGCs are organized under subtopics, we can thus calculate SSI using the frequency statistics of the terms in constituent subtopics and the UGC collection. For each term, we consider a) its relevance to a subtopic by counting its appearances in the sentences within the subtopic and b) its specificity in the subtopic as compared to the whole collection. This results in the proposed term weighting function as follows:

$$\begin{aligned} \omega(t) &= \pi\omega_{TFIDF}(t) + (1-\pi)\omega_{SSI}(t) \\ \omega_{SSI}(t) &= \lambda\omega_{TLF}(t,\tau) + (1-\lambda)p_\tau(t)log\frac{p_\tau(t)}{p_D(t)} \end{aligned} \quad (1)$$

where $\omega_{TLF}(t,\tau)$ is the normalized subtopic level frequency of term t which represents its relevance to the subtopic τ. $p_\tau(t)log\frac{p_\tau(t)}{p_D(t)}$ is the KL divergence that represents the term's specificity to τ. $p_\tau(t)$ and $p_D(t)$ are the subtopic level probability and collection level probability of t. Here λ is empirically set as 0.5.

4.1.4 Discussion on High-level Summary Choices

We choose the topic hierarchy as the high-level summaries for the topic summary. But our framework is not limited to this choice. The topic summary may be based on other types of high-level summaries such as keyword cloud, topic models, and metro map [24]. These methods all provide an overview of the collection and organize the collection at term/topic/document level. Still, the topic hierarchy possesses some quality that is better than the other choices as discussed below.

Keyword cloud is a popular visualization method to generate the high-level summary for unstructured texts. However, the shallow bag-of-words representation and the missing context make the term-based summary not adequate for in-depth analysis. Compared to the keyword cloud where the keywords and their frequencies can be ambiguous without context and background knowledge, the topic hierarchy are more constrained and contain more information in terms of the topic relations and the subtopic word selection.

Statistical topic models present the underlying topics as a list of probable terms [4] to the analysts. However, the latent topics impose some interpretation overhead for the analysts and are of limited value when the task become more specific [5]. Compared to topic model, topic hierarchy is more interpretable and presentable.

Moreover, topic hierarchy presents the explicit topic terms arranged from the general to the specific are more friendly for understanding and interface design.

Metro maps [24] that connect relevant articles like metro lines are another choices of structured overview. However, the basic units, or the vertices in the maps, are the individual articles, which is coarser than we need in the summarization task. Here we need more detailed focus on the micro level structure that captures the relations between ideas and concepts. Still, it is possible to use metro maps as the high level summaries and use ordinary MDS methods separately to generate textual summaries for the lines. This however is not in line with our target to find a natural union of high level and textual summaries.

Therefore, topic hierarchy is a better alternative in presenting the explicit topic and their relations, and as the aids for generating textual summaries. It also solves the scale issue with high-level abstraction and easily perceivable presentation.

4.2 Time-Sequence Summarization

In the previous subsection, the UGCs are organized in a topic hierarchy that also provides as the topic summary of the contents. In this subsection, we introduce our method for generating the time-sequence textual summary for the time-stamped UGCs on the topic hierarchy. As the summary is incrementally updated as time goes by and new contents flow in, to generate the time-sequence summary, there are two major tasks: determine the time point where a summary should be emitted to reflect the significant changes in the new contents, and generate the summary by updating from its previous summary.

4.2.1 The Data Reconstruction Framework

We start by introducing a novel data reconstruction perspective [12] as our basic framework. From the original sentence space, data reconstruction tries to find a subspace of sentences that can best reconstruct the original sentences. In particular, a sentence u_i from the target collection U can be approximated by the linear combination of the selected summary sentences $X : \{x_1, x_2, ..., x_m\}$,

$$u_i \approx \varphi(X, \mathbf{a_i}) = \sum_{j=1}^{m} x_j a_{ij} \tag{2}$$

where $\mathbf{a_i} : \{a_{ij}\}$ is the combination parameter, and m is the number of summary sentences. Summing up the reconstruction errors of the whole collection of n sentences, the overall loss can be written as:

$$\mathcal{L}(U, X, \mathbf{A}) = \sum_{u_i \in U} ||u_i - \varphi(X, \mathbf{a}_i)|| \tag{3}$$

where $\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_n}]^T$, U is the original sentence set, X is the selected summary sentences, and $\varphi(X, \mathbf{a_i}) = \sum_{j=1}^{m} x_j a_{ij}$ is a linear reconstruction function. The objective is to find the set of sentences X^* that minimize the overall reconstruction error:

$$\begin{aligned} X^* &= \arg\min_{X, \mathbf{A}} \sum_{u_i \in U} ||u_i - \varphi(X, \mathbf{a}_i)|| \\ s.t.\ & X \subset U, \mathbf{A} \in \mathbb{R}^{n \times m} \end{aligned} \tag{4}$$

4.2.2 Dynamic Reconstruction for Summary Update

Our task of generating the time-sequence summary is to emit updates along a timeline as time goes by and new contents flow in. However, the basic data reconstruction framework assumes that the content is static. We propose a dynamic reconstruction model that continuously minimizes the reconstruction errors for generating summary updates.

Given a time t_{k+1}, and the existing summary X^{t_k} at a previous time t_k, the update summary $X^{t_{k+1}}$ is generated for the UGCs produced during t_k to t_{k+1}. As we assume that the users have already read X^{t_k}, the update summary should only report the new content regarding what happened after t_k. In other words, an update summary $X^{t_{k+1}}$ is thus a function of X^{t_k} and $U_{t_k \rightsquigarrow t_{k+1}}$.

To generate the update summary, the **subtraction** approach is well adopted in the literature [30, 26] to deal with the redundant information in the previous summary. Generally, an auxiliary direct summary $(X^{t_{k+1}})'$ is first generated based on the new content using : $U_{t_k \rightsquigarrow t_{k+1}} :\longrightarrow (X^{t_{k+1}})'$. Then the previous summary is subtracted from the auxiliary summary so that the known information is removed:

$$X^{t_{k+1}} = (X^{t_{k+1}})' \ominus X^{t_k} \tag{5}$$

The subtraction operation $\ominus$ can be done by identifying the similar sentences between the current summary and the previous summary, measuring by the vector space model or language model. However, the subtraction approach loses the consistency as the reference summary does not influence the generation of the new summary. It is also different from a human approach where the previous summary is considered at the update summarization time, rather than a post process on the new summary.

We thus propose a more intuitive approach called **Dynamic Reconstruction** for summary update. It models the update summarization as a continuing process. Within the data reconstruction framework, we assume that the summary space is acquiring more dimensions to accommodate the growing document space, in order to maintain a relatively low reconstruction errors. In practice, each update summary is incrementally constructed based on the previous summary.

$$\tilde{X}^{t_{k+1}} = \arg\min \sum_{u_i \in U_{t_k \rightsquigarrow t_{k+1}}} ||u_i - \varphi(X^{t_k} \cup X^{t_{k+1}}, \mathbf{a}_i)|| \tag{6}$$

where $U_{t_k \rightsquigarrow t_{k+1}}$ is the new UGCs produced since the last summarization time t_k. The sentences in $X^{t_{k+1}}$ are selected from $U_{t_k \rightsquigarrow t_{k+1}}$. Instead of subtracting the previous summary X^{t_k}, we force it to remain in the summary space to account for any known information from the UGCs produced before t_k. Note that $X^{t_k} + X^{t_{k+1}}$ may not be the best to reconstruct $U_{t_k \rightsquigarrow t_{k+1}}$ directly according to Eq. 4, as the objective function in Eq. 6 is optimized for the best update summary $X^{t_{k+1}}$. For this reason, the update summary generated by a subtractive approach is not optimized for the updating purpose.

4.2.3 Update Point Detection

Now a natural question is, when should we generate an update summary in a continuous process? Considering an extreme case that new UGCs coming in but no new information is given, the previous summary X_{t_k} should well cover the new data, thus $X^{t_{k+1}}$ is empty, indicating that no update is needed. Therefore, an update point should be the one by which enough new information is presented in the coming UGCs. In the following, we will detail our method on update point detection within the data reconstruction framework.

The task of update point detection is to evaluate whether it is worth emitting an update summary for the new UGCs accumulated till a current time t_{k+1} since a previous summary time t_k. Specifically, we monitor the fitness of the old summary for the new content. Under the data reconstruction framework, the fact that the previous summary does not fit the new content signals the need to

Category	Topic	blog	cQA	tweet
Products	IPhone 5	218	1,854	405,036
	IPad mini	230	972	2,287
	Xbox kinect	233	1,746	28,632
Politicians	Barack Obama	265	1,447	416,354
	Mitt Romney	229	1,252	2,043
	Hillary Clinton	227	1,136	756
Fashion Brands	Chanel	253	1,087	2,278
	Estee Lauder	285	1,375	15,823
Corporations	Facebook Inc.	274	1,135	398,035
	Microsoft Corp.	252	1,253	403,983
	Blizzard Inc.	247	1,046	1,722

Table 1: Statistics on the sizes of UGCs from each source.

produce a new summary. In particular, we use the per sentence average reconstruction error to denote the fitness level of the summary X for the content U:

$$RE^{ref}_{avg} = \mathcal{L}(U, X^{ref}, \mathbf{A})/|U|$$
$$RE_{avg} = \mathcal{L}(U', X^{ref}, \mathbf{A}')/|U'| \quad (7)$$

where RE^{ref}_{avg} is the reference fitness based on the previous summary and the content. RE_{avg} denotes the fitness of the reference summary X^{ref} for the new content U'.

Denoting the fitness change as $\frac{RE_{avg}}{RE^{ref}_{avg}}$, an update point t_{k+1} is detected based its previous neighbouring point t_k when the change exceeds a update threshold δ:

$$\frac{\mathcal{L}(U^{t_{k+1}}, X^{t_k}, \mathbf{A})/|U^{t_{k+1}}|}{\mathcal{L}(U^{t_k}, X^{t_k}, \mathbf{A}')/|U^{t_k}|} > \delta \quad (8)$$

where $\delta > 1.0$, which indicates that the per sentence reconstruction error on the new content increases. According to the principle of reconstruction error minimization, an update summary is needed in such cases for better approximation of the new content.

5. EXPERIMENTS

5.1 Datasets

We collected the UGCs from blogs, a community question answering (cQA) service, and twitter, to form a corpus of four categories: Products, Politicians, Fashion Brands and Corporations. We crawled blogs for a topic by submitting the topic name (and variations of the names) as the keyword query into the Google Blog search engine[2] and collecting the top 300 returned blogs. For cQAs and tweets, we issued the keyword queries into the Yahoo! Answer API and the Twitter API to obtain the relevant data on the target topic. We restricted the time stamps in our dataset to within June 1, 2012 and December 1, 2012.

After sentence segamentation and non-English UGCs filtering through automatic language identification [3], a brief statistics of our corpus is presented in Table 1. Note that the UGCs from different sources are quite imbalanced, which is typical in the multi-source summarization problem.

5.2 Evaluation Methods

Evaluation is a challenging part of summarization research, which is more evident given that the data in our experiment is very big and diverse for manually reading and summarizing. Previous work such as [26] aggregates results from all the baseline methods. While it produces objective references for evaluation, it may suffer from the problems that all the baselines produce low quality summaries in some cases. In this work, two computer science graduate students who are not directly related to this research manually build the model summaries for the system summaries to evaluate against.

We take two steps in labeling the results. First, we evaluate the quality of the topic hierarchy according to the method used in [36] where we first check the quality of the subtopics extracted and then the correctness of the relation between the subtopics. After that, we assume that the topic hierarchy is fixed.

Then, based on the topic hierarchy organized data, the summary sentences are first selected for each leaf level group and put together as the candidates for the higher levels. Thus the sentence selection at higher levels deals with a manageable set of the "preselected" sentences. In particular, two evaluators generate a sequence of 250-word summaries for each topic in a three-pass manner: relevance sentence labeling, long summary labeling, and limited length summary labeling. First, they read the UGC sentences in the time order and label the content that are closely relevant to the topic. Sentences that cannot express a standalone message or simply irrelevant, are discarded. After the first round, the evaluators have formed some basic impression of the subtopic. In the second pass, they start to pick up the summary sentences for the subtopic without considering the summary length. The updating points are also determined at this round. Besides a time mark, the evaluators have to give a reason for setting it as a point for update. In the final round, 250-word update summaries are generated for each period between two update points. The evaluators are instructed to select sentences that cover novel content as compared to the summary in the previous time period. Note that two evaluators are involved by going through the three sequential steps individually. After each step, they compare the results and resolve the differences by discussion. The manual labeling work takes about three weeks to finish, which also indicate that manual summarization is not feasible for large heterogeneous UGCs.

Note that we do not report inter-annotator agreement. In our initial study, the inter-annotator agreement is lower than 0.5 (Cohen's kappa coefficient). Because of this, we find that asking the two annotators to discuss about their difference and make a final decision is better choice for this research.

5.3 Overall Summarization Performance

We evaluate the overall summary quality with the well-tuned update points and the topic hierarchy (parameter tuning will be presented in later sections). The question to be answered here is, given the update points and the organized UGCs within the periods, how well our system can generate sentence based propose summaries as compared to the baseline methods. In particular, the following state-of-the-arts summarization methods are compared.

(1) Sum_LDA [10]: This is a topic modeling based summarization method, where the sentences are selected to cover the latent topics discovered by LDA.
(2) LexRank [8]: It is a graph-based sentence salience model where sentences get "vote" from connecting ones in a similarity graph.
(3) DSDR [12]: A data reconstruction based summarization model where sentences are selected to approximate the original set by minimizing the overall reconstruction error.

While our methods are designed for naturally update summary generation, the baseline updates are achieved by subtraction: removing the similar sentences that are present in the previous summary (cf. Section 4.2.2). In particular, at each update point, we generate a 500-word base summary using the UGCs from the last update point, and then remove the part that covered by the previous

[2]http://www.google.com/blogsearch

Table 2: Comparison of the proposed temporal topic summarization method and its variations with the four baselines, in terms of ROUGE-1. † and ‡ denote significant differences (t-test, p-value<0.05) over Sum_LDA and DSDR respectively.

Methods	Product	Fashion	Politician	Corporation	Overall
Sum_LDA	0.396	0.416	0.235	0.249	0.314
LexRank	0.276	0.283	0.239	0.259	0.265
DSDR	0.312	0.326	0.244	0.323	0.302
TTS-TH	$0.361^{\ddagger}$	$0.411^{\ddagger}$	0.259	$0.338^{\dagger}$	$0.328^{\ddagger}$
TTS	$0.397^{\dagger\ddagger}$	$0.421^{\dagger\ddagger}$	$0.345^{\dagger\ddagger}$	$0.383^{\dagger\ddagger}$	$0.381^{\dagger\ddagger}$

summary. If the remainder is longer than 250 words, we select the top 250 according to the sentence rank in the original summary.

For our methods, besides the full TTS model, we also add a variant by removing the term weight brought in by topic hierarchy organization (TTS-TH), in order to study the effects of the component on the overall summarization performance.

We test the summarization performance using the standard ROUGE score [18]. The ROUGE-1 F1 measure based on the unigram overlap between the reference summary and the system summary is reported for its proven effectiveness.

5.3.1 Results and Discussion

The averaged ROUGE-1 results are summarized in Table 2. By comparing the results of different methods, we can draw the following observations: (1) Among the baselines, Sum_LDA achieve the best performance, followed by DSDR and LexRank. Though Sum_LDA has slightly higher performance than DSDR, the difference is not statistically significant. This shows that data reconstruction is a promising summarization framework, though the computational cost is high.

(2) The proposed TTS and its variants outperform the two best baselines with statistic differences. Removing the topic hierarchy related weights from term representation, TTS-TH , leads to about 30% degradation from TTS. This indicates that the new assumptions we made on the base data reconstruction summarization and sentence topical representation, are effective in contributing to the overall improvement. The clear contribution of topic hierarchy and the good performance of Sum_LDA among baselines indicate that capturing topical level semantics can play an important role in summary generation methods.

(3) Some categories are inherently more difficult to summarize than others. From the horizontal entries, we can see that Product and Fashion usually obtain the best results and Politician the worst. For one reason, the easy categories usually have clearly defined topic hierarchy, while the difficult categories usually have fuzzy or relatively low quality hierarchies. Another reason is that the easy categories tend to have UGCs from multiple sources, while the difficult categories do not trigger as much interest from users of different platforms. Together with the study on the contribution of UGC sources, we may conclude that aggregating more UGC sources is helpful for TTS to generate high quality summaries.

5.3.2 Efficiency Analysis

Theoretically, the efficiency bottleneck of our framework is at the computation of the solution of reconstruction objective functions, which have a complexity of $O(n^3)$. However, as the UGCs are divided into the subtopics according to topic hierarchy, n is usually of manageable size. The exceptions are that some popular subtopics may attract far more content than the rest. These big sizes usually indicate that these subtopics can be further divided into finer subtopics in the upstream topic hierarchy construction algorithm.

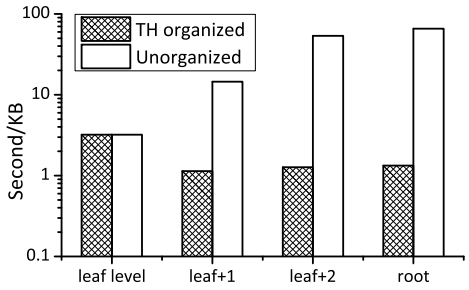

Figure 3: The influence of topic hierarchy organization on summarization efficiency.

The topic hierarchy is one of the major instruments we employ to tackle the scalability issue of multi-source UGC summarization. Now we compare the efficiency and quality difference when the UGCs are organized by the topic hierarchy, or not. Figure 3 evaluates the efficiency aspect of using topic hierarchies. Here organization refers the method we have discussed in Section 4.1.2. We can see that, with topic hierarchy organization, the runtime per KB data is almost constant for topic summarization at different levels. Except for leaf level subtopics, the higher levels are all working on the "pre-selected" sentence sets from its child nodes, and thus the computational load is relatively stable. For unorganized data, the computational costs of the data reconstruction framework increases greatly with the size of the input.

Besides dividing the contents using topic hierarchy, the sentence quality score can also be used to reduce the number of sentences for reconstruction. Instead of using the quality score directly as a sentence prior, a threshold can be imposed to remove the relatively low quality ones. We use a conservative threshold of 0.5 in our experiments.

5.4 On High Level Topical Summarization

In the subsection, we design experiments to answer the following two questions: (1) how well does the topic summary in the form of topic hierarchy present the high level overview of the UGCs around a root topic? (2) given the topic hierarchy organized UGCs, how does the topic hierarchy influence the quality of the textual summary generated on the organized UGCs?

5.4.1 Key Subtopic Discovered

To evaluate the quality of the topic hierarchy as the high level summary, we assume that the methods that discover more key subtopics are better. On average, each root topic in our dataset obtains a topic hierarchy of 54.7 nodes, 53.3 edges, and with a depth of 3.8.

To measure the quality of the discovered topics quantitatively, we compare the our proposed method (TH) with three methods: (1) DFreq method, in which all the topics are ranked according to their document- level frequency in the social media contents; (2) IFreq method, in which the topics are ranked by the number of items that contain this topic in its topic hierarchy; (3) TM method, a topic model based method where the top ten words of each topic are kept. For all the compared methods, their top 100 topics for the three root topics are manually annotated by three graduate annotators. Finally, the precision of each method is reported in Table 3.

From the result we can see the effectiveness of the TH framework on key subtopic discovering. Specifically, TH improves over

Table 3: Precision on key subtopics discovered.

Method	IPhone 5	Barack Obama	Chanel
DFreq	0.35	0.38	0.34
IFreq	0.34	0.30	0.29
TM	0.36	0.28	0.33
TH	***0.48***	**0.51**	**0.53**

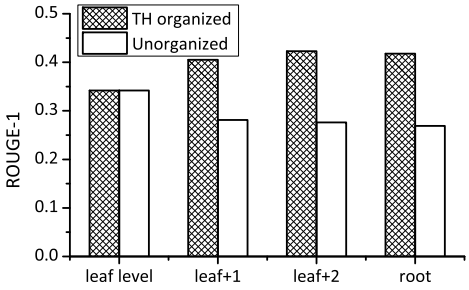

Figure 4: The influence of topic hierarchy organization on summarization quality.

the average of the three baselines by 37.1%, 59.3%, and 65.6% on the three topics respectively. Though an indirect measure on the topic hierarchy's effect on summarization, the correctly captured key subtopics may lead to better topic coverage in sentence selection.

5.4.2 *Effects on Textual Summary Generation*

In Figure 4, we compare summarization performance with and without the topic hierarchy in terms of ROUGE-1 measure. For without setting, we represent sentences with only term vectors. For each subtopic, the raw contents are fed as the input to the DSDR based summarization method. From the results, we can see that at leaf level the performance is the same. This is because at leaf level, the topic information is not used even in the with-topic-hierarchy setting. As the levels increase, the performance gaps start to show. The without setting performs even worse than at leaf level, as it has to handle much bigger input size. On the other hand, the with setting performs better than at leaf level, showing the effects of topic hierarchy assisted sentence representation.

To see whether the size of the topic hierarchy has an effect on the summarization performance, we deliberately remove the leaf levels of each hierarchy and rerun the system to generate summaries for the root topic. As a general trend, all the root topics get a worse summary, with an average 13.1% drop of ROUGE-1 score. This may indicate that a comprehensive modeling of the subtopics is helpful in generating high quality summaries, which however means higher computation costs too.

Meanwhile, by checking Figure 3 and Figure 4 together, we find that both the summarization quality and efficiency are enhanced with *less* but more relevant data feeding into the higher levels. We thus may conclude that organizing UGCs by topic hierarchy is indeed a practical solution to the scale issue.

5.5 On Update Point Detection

5.5.1 *Update Threshold Tuning*

We use the per sentence average reconstruction error change rate to determine the update point for summarization. In this subsection, we are trying to find out the optimal update threshold so as to capture as many correct update points as possible. Here, we care about the accuracy (measured by precision: the ratio of correctly identified points versus the total points by the method), as well as the number of points detected (measured by recall: the ratio of the correctly identified points versus the gold standard points).

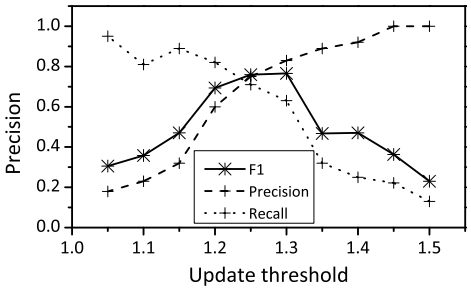

Figure 5: Effect of varying update thresholds δ on the proposed update point detection.

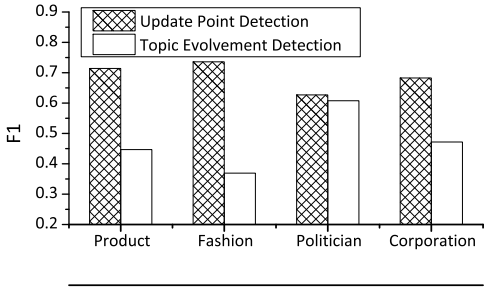

Figure 6: Comparison with the topic evolvement detection method in [26].

As it is not practical to match an exact time point t_n of gold standard, we relax each gold point into a "grace" period with a left and right boundaries of the timeline. The "grace" period is not a fixed length as the gold points can be very close to each other or far apart. We thus relax by a fraction (10%) of the period before and after the gold point, namely, $[t_n - \frac{t_n - t_{n-1}}{10}, t_n + \frac{t_{n+1} - t_n}{10}]$.

To set the appropriate update threshold, we use one topic from each of the four categories as the development data. In Figure 5, we plot the precision, recall and F1 for the threshold in $(1.0, 1.5]$ with a step size 0.05. We can see that both the precision and recall reach the highest at $1.24 - 1.30$. From 1.0 to 1.5, the precision goes up monotonically, the recall goes down monotonically, while the F1 peaks in the middle. By examining the actual points detected, we find that this outcome is mainly caused by the number of points detected. When the threshold is low, a total of more than 100 points are returned, which well cover the desired points but with low accuracy. When the threshold is at around 1.4, the three to six returned points are almost all correct but the recall is too low. Hence we set the update threshold at 1.25 and use it as the optimal update setting for the rest of our experiments.

5.5.2 *Comparison with the Prior Art*

We compare our update point detection method with the "topic evolvement detection" (TED) algorithm, a state-of-the-art method in the Sumblr system [26]. TED is designed to detect sub-topic changes in the tweets stream in order to determine when to put a time node on the timeline, where the time nodes serve the similar purpose as the update points in our task. For easy comparison, we use a modified TED ($modTED$): we use the base data reconstruc-

tion summary to represent the content cluster of a fixed number of sentences, and adopt the TED formula to determine the points.

From Figure 6 , we can see that our method is able to capture the update points more accurately than the topic evolvement detection in terms of F1 in all four categories. While the other three categories show a more significant difference, the Politician category obtains similar performance.

These can be explained by the two major differences in the design of the two approaches. While both approaches are concerned about the new content, our approach also takes the previous neighboring update summary into consideration. However, $modTED$ only clusters the new content and monitors the divergence of the successive content clusters, without considering the previous update. The other major difference is in the definitions of topic change in the two methods. Our method assumes that if the new content is well covered by the previous summary, there is no need to emit an update. Hence, in our approach, even if the changes are gradual and accumulative, we set up a new update point when there is sufficient information comes in. However, the $modTED$ method is more concerned about whether there is a burst between successive incoming new content. This may also explain the performance difference between categories, where in Politician the presidential campaign entails more burst subevents, whereas the other categories contain mainly slow evolving subevents.

When the size of the topic hierarchy and the number of subtopics are concerned, our method also demonstrates higher flexibility. By examining the results at different levels, we find that our method produces better update points than $modTED$ at lower levels, especially leaf levels. As leaf levels are specific subtopics thus many do not have sudden changes, $modTED$ has difficulty to capture those 'slow' events. The above results and analysis suggest that our definition and implementation of topic evolvement is more general in that we can capture both gradual and sudden changes.

5.6 Usability and User Study

Besides the quantitative evaluation presented above, measures are to be taken to make the summary results useful in real world applications. In this subsection, we focus on improving and evaluating the usability of the proposed summarization method. In particular, we display meta information about the summary generation to enhance interpretability and refine the resultant summaries to enhance readability. We then conduct a user study to verify the effectiveness of the refinements as well as the structured outline as compared to other high level summaries.

5.6.1 Usability Enhancement

Besides the performance comparison in terms of the standard ROUGE score, we further refine the resultant summaries to make it better suitable for practical usage. In particular, the following post-processes are performed. First, we add meta information on the topic hierarchy. (a) For each node, we add the numbers of raw documents for each source and a link to all the raw contents. This aims to provide the means for the users to verify the importance of a subtopic. (b) For active nodes (those subtopics with a significant amount of new contents), we mark them with green, which shows that they are "increasing". (c) For each summary sentence, the source type and the link to the original documents are provided. This provides means for the users to verify the credibility of a selected sentence. (d) The summary sentences that greatly reduce the reconstruction error (a threshold is set) is highlighted in bold font. By doing this, we reveal the strong candidacy of a sentence selected, in order to give some sense about our selection criteria.

In addition, we modify the selected sentences to enhance the readability of the summaries. For sentences in non-leaf node, we arrange them in order that cover subtopics from general to specific, from popular to less talked subtopics. The sentence-subtopic memberships are determined in the summary generation process, while the popular subtopics are simply determined by the size of UGCs they cover. We also remove irrelevant clause of long sentences to make the summaries more compact.

Table 4: User study: effectiveness of the four usability enhancement measures (a-d), compared to without the measures.

	a	b	c	d
Interpretability	1.2	1.5	0.8	1.3
Ease of Verification	1.7	0.9	1.6	0.7

5.6.2 User Study

Besides the standard evaluation, we conduct further user studies to verify the usability of our methods. In particular, we verify the effectiveness of the usability enhancement measures. Six computer science graduate students who are not involved in this research give feedbacks on pairs of summaries that are on the same data but generated by different methods. They may first get an overview of the topic by reading the outline summaries and then drill down to any specific subtopic for the corresponding prose summaries. Specifically, each of them are presented with 15 pairs of summaries that are randomly selected. They are asked to rate the comparison in four scales: worse (-1), no improvement or the same (0), marginal improvement (1), and significant improvement (2). We report the averaged results below.

The users give feedbacks on two usability measures: the interpretability and the ease of verification (checking the raw contents to see whether the summary sentence is representative), comparing our proposed method with and without the usability enhancements. The ratings of are given on each measure (a-d) one by one. The averaged scores are summarized in Table 4. We can see that measures (a) and (b) greatly enhance the ease of verification; while (b) and (d) improve the interpretability the most. Besides quantitative feedback, the users also comment that the interpretability is a critical factor that affects the overall subjective impression on the systems.

6. CONCLUSION

This paper has proposed a social media content summarization framework that generates topical time-sequence summaries for analysts to gain a dynamic overview on any topic of interest. Such summaries can serve as the starting point, from which the analysts perform subsequent actions to explore the data and validate the hypothesis that he/she may have. We have proposed to dynamically generate the topic hierarchy as the structural summaries, and also used them to organize the UGCs and to augment the sentence term vectors. To capture the evolvements of events in the contents, we have proposed a unified dynamic reconstruction approach to detect the update points and generate the time-sequence textual summary. Empirical results on four categories of diverse topics demonstrated the effectiveness and efficiency of our method. A user study also has showed that the structural and textual summaries are easier to interpret and verify with our usability enhancement measures. For future work, we plan to explore the topic hierarchy for automatically adjusting the level of details presented in the summaries.

7. ACKNOWLEDGEMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore

Funding Initiative and administered by the IDM Programme Office, and the Natural Science Foundation of China under Grant No. 61472059.

8. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR*, pages 10–18. ACM, 2001.

[2] R. Arora and B. Ravindran. Latent dirichlet allocation based multi-document summarization. In *AND*, pages 91–97, New York, NY, USA, 2008. ACM.

[3] S. Carter, W. Weerkamp, and M. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, pages 1–21, 2013.

[4] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

[5] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: designing model-driven visualizations for text analysis. In *SIGCHI*, pages 443–452. ACM, 2012.

[6] H. T. Dang and K. Owczarzak. Overview of the tac 2008 update summarization task. In *TAC*, pages 1–16, 2008.

[7] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2002–2011, 2013.

[8] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479, 2004.

[9] M. Georgescu, D. D. Pham, N. Kanhabua, S. Zerr, S. Siersdorfer, and W. Nejdl. Temporal summarization of event-related updates in wikipedia. In *WWW Companion*, pages 281–284, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[10] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *HLT:NAACL*, pages 362–370. Association for Computational Linguistics, 2009.

[11] S. M. Harabagiu and A. Hickl. Relevance modeling for microblog summarization. In *ICWSM*, 2011.

[12] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. Document summarization based on data reconstruction. In *AAAI*, 2012.

[13] M. Hu and B. Liu. Mining and summarizing customer reviews. In *SIGKDD*, pages 168–177. ACM, 2004.

[14] M. Hu and B. Liu. Opinion extraction and summarization on the web. In *AAAI*, volume 7, pages 1621–1624, 2006.

[15] D. Inouye and J. K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *SocialCom*, pages 298–306. IEEE, 2011.

[16] H. D. Kim, M. G. Castellanos, M. Hsu, C. Zhai, U. Dayal, and R. Ghosh. Ranking explanatory sentences for opinion summarization. In *SIGIR*, pages 1069–1072. ACM, 2013.

[17] S. Kim, J. Zhang, Z. Chen, A. Oh, and S. Liu. A hierarchical aspect-sentiment model for online reviews. In *AAAI*, 2013.

[18] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[19] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM*, pages 563–572. ACM, 2012.

[20] Z.-Y. Ming, T.-S. Chua, and G. Cong. Exploring domain-specific term weight in archived question search. In *CIKM*, pages 1605–1608. ACM, 2010.

[21] Z.-Y. Ming, K. Wang, and T.-S. Chua. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *SIGIR*, pages 2–9, New York, NY, USA, 2010. ACM.

[22] Z.-Y. Ming, K. Wang, and T.-S. Chua. Vocabulary filtering for term weighting in archived question search. In *Advances in Knowledge Discovery and Data Mining*, pages 383–390. Springer, 2010.

[23] Z. Ren, S. Liang, E. Meij, and M. de Rijke. Personalized time-aware tweets summarization. In *SIGIR*, pages 513–522, New York, NY, USA, 2013. ACM.

[24] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec. Information cartography: creating zoomable, large-scale maps of information. In *SIGKDD*, pages 1097–1105. ACM, 2013.

[25] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing microblogs automatically. In *HLT:NAACL*, pages 685–688. Association for Computational Linguistics, 2010.

[26] L. Shou, Z. Wang, K. Chen, and G. Chen. Sumblr: continuous summarization of evolving tweet streams. In *SIGIR*, pages 533–542. ACM, 2013.

[27] H. Takamura, H. Yokono, and M. Okumura. Summarizing a document stream. In *Advances in Information Retrieval*, pages 177–188. Springer, 2011.

[28] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *SIGIR*, pages 299–306, New York, NY, USA, 2008. ACM.

[29] D. Wang, S. Zhu, T. Li, and Y. Gong. Multi-document summarization using sentence-based topic models. In *ACL-IJCNLP*, pages 297–300. Association for Computational Linguistics, 2009.

[30] H. Wang and G. Zhou. Toward a unified framework for standard and update multi-document summarization. *TALIP*, 11(2):5:1–5:18, June 2012.

[31] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *SIGKDD*, pages 153–162. ACM, 2010.

[32] H. Yang and J. Callan. A metric-based framework for automatic taxonomy induction. In *ACL*, pages 271–279. Association for Computational Linguistics, 2009.

[33] C. Zhai, J. Han, D. Roth, and P. Tsaparas. Opinion integration and summarization. *Thesis*, 2012.

[34] R. Zhang, W. Li, and D. Gao. Generating coherent summaries with textual aspects. In *AAAI*, 2012.

[35] R. Zhang, Y. Ouyang, and W. Li. Guided summarization with aspect recognition. In *TAC*, 2011.

[36] X. Zhu, Z.-Y. Ming, X. Zhu, and T.-S. Chua. Topic hierarchy construction for the organization of multi-source user generated contents. In *SIGIR*, pages 233–242, New York, NY, USA, 2013. ACM.

[37] L. Ziheng, K. Min Yen, and L. Tan Chew. Exploiting category-specific information for multi-document summarization. In *COLING*, 2012.