Topic Chains for Understanding a News Corpus

Dongwoo Kim and Alice Oh

KAIST
Computer Science Department
Daejeon, Korea
dw.kim@kaist.ac.kr, alice.oh@kaist.edu

Abstract. The Web is a great resource and archive of news articles for the world. We present a framework, based on probabilistic topic modeling, for uncovering the meaningful structure and trends of important topics and issues hidden within the news archives on the Web. Central in the framework is a *topic chain*, a temporal organization of similar topics. We experimented with various topic similarity metrics and present our insights on how best to construct topic chains. We discuss how to interpret the topic chains to understand the news corpus by looking at long-term topics, temporary issues, and shifts of focus in the topic chains. We applied our framework to nine months of Korean Web news corpus and present our findings.

1 Introduction

The Web is a convenient and enormous source for learning about what is happening in the world. One can go to the Web site of any major news outlet or a portal site to get a quick overview of the important issues of the moment. However, it is difficult to use the Web to understand what has been happening over an extended period of time. We propose a computational framework based on probabilistic topic modeling to analyze a corpus of online news articles to produce results that show how the topics and issues emerge, evolve, and disappear within the corpus.

The problem of understanding a corpus of news articles over an extended period of time is challenging because one has to discover an unknown set of topics and issues from a large corpus of disparate sources, find and cluster similar topics, discover any short-term issues, and identify and display how the topics change over time. A narrower but similar problem has been studied in the TDT (topic detection and tracking) field [1] where the goal is to identify new events and track how they change over time. The events, however, are defined as happenings at certain places at certain times, and so they compose a small subset of general news topics and issues. For example, an earthquake in Haiti is an event, but the prolonged decline of real estate sales is not. The latter makes up a large portion of news, but the TDT community would only cover the former, whereas our research covers both. The probabilistic topic modeling community offers solutions such as Dynamic Topic Models [2] and Topics Over Time [3] for discovering topics

and looking at how they change over time, but those models do not capture how topics newly emerge and disappear because they assume the same set of topics exist from the beginning through the end of the time-series data.

We propose *topic chains*, a framework for analyzing a sequential corpus, composed of similar topics appearing within a specified sliding window. Topic chains present a temporal and similarity-based organization of topics found by latent dirichlet allocation (LDA) [4]. Topic chains can be used to identify general topics, such as *labor unions* or the *stock market*, which occur in long topic chains. Short-term issues, such as the *death of Michael Jackson*, can be seen in short topic chains. Some short-term issues can be embedded within a long topic chain because they are related to general topics. One example of such issue is the recall of Toyota cars which is related to the general topic of the automobile industry. Those issues embedded within general topics can be identified by looking at *focus shifts* shown by words that change significantly within the topic chain.

Our contributions can be summarized as follows:

- We compare six frequently used similarity metrics using log likelihood of data for finding similar topics. We show that the two most frequently used metrics, cosine similarity and KL divergence, do not give the best results.
- We define and construct *topic chains* using the best similarity metric we found. We then illustrate how to further analyze topic chains to identify general topics and short-term issues.
- Overall, we propose a framework for understanding how topics and issues emerge, evolve, and disappear through time in a corpus of online news articles. This framework includes a set of analyses for a sequential corpus that other similar tools do not provide.

2 Related Work

This work can be positioned with respect to three related research areas: topic and event detection and tracking, probabilistic topic modeling, and temporal news mining.

Topic detection and tracking (TDT) is a well-studied task, summarized in Allan's book [1], and followed up by a line of research around *event threading* [5–7]. Both TDT and event threading solve a narrowly defined problem of looking for articles related to one or more *events*, where an *event* is defined as something that happens at a certain place at a certain time in the real world. We solve a much broader problem of discovering all topics and issues that occur in the corpus, whether or not they are directly related to concrete events in the world. Also, our definition of *issues* is more general than the definition of events by the TDT task. For example, the H1N1 influenza *issue* of 2009 is a series of related events such as deaths, vaccinations, and travel warnings, as well as non-events such as the safety of the vaccine, spatiotemporal course of the pandemic, and susceptibility of populations. We borrow two central aspects of the TDT task which are the discovery of new events and the evolution of events over time.

We substitute our general definition of topics for their events such that our framework discovers new topics and how they evolve over time.

Probabilistic topic models [8] such as the frequently used latent dirichlet allocation (LDA) [4] discover all topics, regardless of event-like characteristics, that are highly represented in a corpus, and extensions to LDA, [9, 2, 10, 3] consider the temporal aspect of the corpus as well. In [9], Wang et al. worked with asynchronous text streams to find common topics from documents with different timestamps. They found highly discriminative topics from asynchronous data and synchronized the documents according to topics. With dynamic topic models (DTM) [2], Blei and Lafferty analyzed how topics evolve over time in a sequential corpus, and they demonstrated how topics in the journal *Science* changed from 1881 to 1999. One limitation with DTM is that it only models the changes of word distributions within the topics and assumes the set of topics stays constant throughout the corpus, so it does not model how topics appear and disappear over time. The same limitation exists for the topic trend detection in [10]. With Topics over Time (TOT) [3], Wang and McCallum jointly model topics and timestamps to analyze when in the sequential data the topics occur. This model can discover when new topics appear and then disappear, but in this model, the topics stay the same over time. In our framework, different but similar topics form a topic chain so we can observe how the topics evolve over time.

Previous work on temporal news mining include [11–13]. Leskovec et al. [11] look at the news cycle by tracking how *memes* travel widely through the media sites and blogs. While this approach is very interesting, it does not capture the broad and overall picture of what topics and issues emerge and spread through the media sites. Shahat and Guestrin's work [12] looks at how two news articles can be connected through a series of articles in between them to form a coherent chain of articles. This is an effective solution to get a big picture of the story that connects two news articles. Mei and Zhai's work [13] is probably the closest to our work, but they work with data that is filtered for specific topics, such as the Asia Tsunami. They extend this work in [14] to include the spatial dimension. Our work aims to present an overall picture of topics and issues including how to identify general topics as well as temporal issues.

3 Overall Framework

Suppose there is a corpus of twelve months of news articles from major online newspapers that a user wishes to understand. A good way to do that is to break down the problem into finding the following details about the corpus:

- *Topic*: a *topic* is a major subject discussed in the corpus. Examples are "winter olympics", "healthcare reform", "the stock market".
- *Long-Term Topic*: if a *topic* lasts for a long time, we say it is a *long-term topic*. Examples are "the stock market", "Afghanistan war", "education".

– *Temporary issue*: if a *topic* lasts for a short time, we say it is a *temporary issue*. Examples are "the winter olympics", "earthquake in Haiti", "death of Michael Jackson".
– *Focus Shift*: a topic chain exhibits different focuses for each individual topic in the chain. An example of a focus shift is "Greece, moratorium" to "Europe, recession" in the "economy" *long-term topic*.

We propose a framework to analyze the corpus to find the *topics, long-term topics*, *temporary issues*, and *focus shifts*. In this section, we explain the parts that compose the overall framework.

1. **Discovering Topics**: We discover the topics in the corpus with latent dirichlet allocation (LDA) [4], the most widely used method of probabilistic topic modeling. LDA models topics as multinomial distributions of words.
2. **Measuring Topic Similarity**: We compare several methods for measuring topic similarity so that we can use the best method to find similar topics. We look at six popular similarity metrics and compare them in terms of log likelihood of data.
3. **Constructing Topic Chains**: A topic chain is a sequence of similar topics through time. Using the topic similarity metric, we look for similar topics within a sliding time window and add links between two similar topics to construct topic chains.
4. **Long-Term Topics and Temporary Issues**: After constructing the topic chains, we can identify *long-term topics* such as the stock market, *temporary issues* such as the Olympics. We can also identify *focus shifts* in *long-term topics*.

4 Topics

The first step in our analysis is finding topics in the corpus. Because we are looking at news data which are sequential by nature, we divide the corpus into several time slices, and for each time slice, we find a set of topics that are most salient in the documents within the time slice. We first describe the topic model we used for finding the topics, then we describe our dataset and the topics found in it.

4.1 Latent Dirichlet Allocation

LDA [4] is a widely used method for probability topic modeling. LDA is a generative model that models a document using a mixture of topics. In the generative process, for each document d, a multinomial distribution θ_d over topics is randomly sampled from a Dirichlet with parameter α, and then to generate each word, a topic z_n is chosen from this topic distribution, and a word, w_n, is generated by randomly sampling from a topic-specific multinomial distribution ϕ_{z_n}. A topic-specific multinomial distribution ϕ_{z_n} is also randomly sampled from a

Table 1. Four topics discovered by LDA for the news dataset. Topics are randomly chosen and are represented by top ten probability words. Topic 1 is about "soccer game", topic 2 is about "market" and "business", topic 3 is about "smart phones", and topic 4 is about "research". Each topic is a multinomial distribution over words.

Topic 1		Topic 2		Topic 3		Topic 4	
Top words	Probability	Top words	Probability	Top words	Probability	Top words	Probability
game	0.030	growth	0.035	Apple	0.024	research	0.078
player	0.026	business	0.034	smartphone	0.018	professor	0.042
league	0.025	recovery	0.031	internet	0.017	science	0.018
coach	0.023	crisis	0.026	iphone	0.016	doctorate	0.017
soccer	0.016	prospect	0.024	mobile phone	0.013	discovery	0.016
season	0.012	policy	0.023	Google	0.012	analysis	0.012
leader	0.011	investment	0.020	computer	0.011	technology	0.010
competition	0.011	strategy	0.018	usage	0.010	universe	0.010
advance	0.007	market	0.016	advertise	0.010	plant	0.009
pro	0.007	consume	0.015	information	0.008	experiment	0.009

Dirichlet with parameter β. From the generative process, we obtain the likelihood of a document:

$$p(\mathbf{w}, \mathbf{z}, \theta_d, \Phi|\alpha, \beta) = \prod_{n=1}^{N} p(w_n|\phi_{z_n})p(z_n|\theta_d) \cdot p(\theta_d|\alpha) \cdot p(\Phi|\beta).$$

The Dirichlet parameters α and β are vectors that represent the average of the respective distributions. In many applications, it is sufficient to assume that such vectors are uniform and to fix them at a value pre-defined by the user, and these values act as smoothing coefficients.

4.2 Corpus

We collected over 130K news documents from the Web editions of three major Korean newspapers[1] between 2009-07-01 and 2010-04-10. Each news outlet covers a wide range of topics such as politics, economy, sports, entertainment, and culture, and show their own perspectives on cultural and social phenomena.

For the topic modeling task, we refined each document using a Korean morpheme analyzer and part-of-speech (POS) tagger provided by ETRI[2]. In the Korean language, each word can be broken down into morphemes. The morphemes are the smallest meaningful units, and each morpheme has a POS tag associated with it. Most of the morphemes do not carry semantic meaning but are instead used as syntactic markers, and almost every verb, adverb, and adjective can be broken down into morphemes with a noun token and one or more syntactic markers.

After preprocessing the documents as described, we divided the corpus into 28 time slices, ten days each. The average number of documents in each time

[1] http://www.yonhapnews.co.kr/, http://www.donga.com/, http://www.hani.co.kr/
[2] http://www.etri.re.kr

slice is 4,715, and the average number of unique words in each group is 13,611. We extracted 50 topics with LDA for every time slice using Gibbs sampling. To reduce the effort of estimating hyperparameters, we used symmetric Dirichlet priors. More specifically, for α and β, we adopted the commonly used values of 0.1 and 0.01 respectively. We set the number of topics to be 50 for one time slice, so the total number of topics is 1,400 for the entire corpus. We randomly chose 4 topics from the corpus and show them in Table 1, each topic represented with the words that have the highest probabilities in that topic.

5 Topic Similarity

To construct topic chains, we need to measure the similarity between a pair of topics. In previous topic modeling research where topic similarity must be measured, cosine similarity [15] and Kullback-Leibler (KL) divergence [16] are frequently used without any formal validation. There exist, however, several well-known metrics that can be used to measure topic similarity, so we compared them to see which metric would be best for our purpose. We considered six metrics and evaluated each metric using the negative log likelihood of corpus.

5.1 Six Metrics of Topic Similarity

A topic, ϕ_i, is a multinomial distribution over the vocabulary, but it can also be viewed as a ranked list of words, or a $|W|$ dimensional vector, where each dimension i is a probability of w_i in that topic. A topic can also be represented by a set of topic words–words with a probability over a threshold. These various perspectives allow the following metrics for measuring similarity between topic ϕ_i and topic ϕ_j:

- **Cosine Similarity** measures the similarity between two vectors by finding the cosine of the angle between them.
- **Jaccard's Coefficient** measures the similarity and diversity of two sets. It is defined as the size of the intersection divided by the size of the union of two sets.
- **Kendall's τ Coefficient** measures the association between two ranked lists.
- **Discounted Cumulative Gain(DCG)** measures the effectiveness of the ranked results of a web search algorithm.
- **Kullback-Leibler Divergence** is a non-symmetric measure of the difference between two probability distributions p and q.
- **Jensen-Shannon Divergence** is the symmetric variation of KL divergence.

Each metric considers a different aspect of the relationship between two topics. Kendall's τ and DCG consider the ranks of words within a topic. KL divergence and JS divergence consider the divergence of multinomial topic probabilities, and lower divergence would indicate higher similarity between two topics. Cosine similarity measures the angle of two vectors, and Jaccard's coefficient

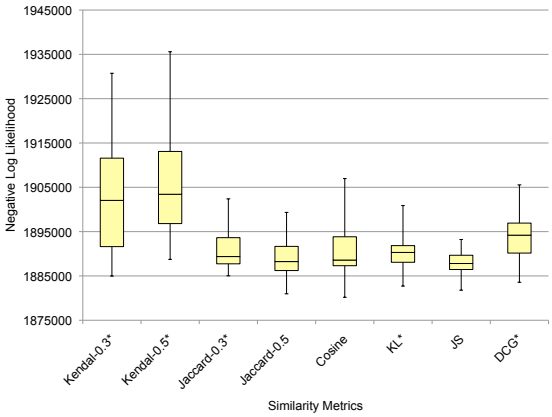

Fig. 1. Comparison of negative log likelihood for six similarity metrics using a boxplot. Negative log likelihood was computed for the corpus using the set of topics where five topics were substituted with the five most similar topics from another time slice, identified by each of the six similarity metrics. A better similarity metric gives a lower negative log likelihood. JS divergence and Jaccard's coefficient with 0.5 cumulative probability mass achieve better performances than the other metrics. An asterisk (*) next to a metric indicates statistically significant differences between the metric and JS divergence using t-test, $p < 0.01$.

looks at the association between two sets. Jaccard's coefficient must use a partial set of words because it looks at the intersection and the union of the two sets of words that represent the topics. We use the top probability words that contribute to the cumulative probability mass, which is a parameter that must be set. We also use a partial set of top probability words for Kendall's τ. This is because Kendall's τ is equally affected by the differences among high probability words and the differences among low probability words, but the words that have low probabilities in both topics should not contribute to the similarity score as much.

5.2 Comparing the Metrics

We compared the six metrics with the negative log likelihood of the corpus which measures how well the model explains the corpus. Starting from a set of topics extracted for a time slice, we substitute five topics with the topics from another time slice that are found to be most similar according to each of the six metrics to form six modified sets of topics. By comparing the negative log likelihoods using the modified sets of topics, we can see which metrics found the most similar topics. The process is as follows:

1. Train LDA for two consecutive time slices to get two sets of topics $\Phi^{t-1} = \{\phi_1^{t-1}, \phi_2^{t-1}, \phi_3^{t-1}, ...\phi_k^{t-1}\}$ and $\Phi^t = \{\phi_1^t, \phi_2^t, \phi_3^t, ...\phi_k^t\}$.

2. Compute the similarity score between ϕ_i^{t-1} and ϕ_j^t for every i, j.
3. Select top five pairs of similar topics from the two topic sets.
4. Substitute the original topics $\Phi^t = \{\phi_1^t, \phi_2^t, \phi_3^t, ...\phi_k^t\}$ with the five most similar topics from $t-1$. So the $\Phi_{new}^t = \{\phi_1^t, \phi_i^{(t-1)}, \phi_3^t, ...\phi_k^t\}$, where i is a one of the five most similar topics from the previous time slice.
5. Finally, using Φ_{new}^t, calculate the log-likelihood of data at time t.

To evaluate the metrics, we selected the first two consecutive time slices, and then trained LDA on each time slice 30 times. Using these 30 pairs of LDA results, we calculated the similarities of all topic pairs, replaced the most similar topics, and computed the negative log likelihoods. As Figure 1 shows, JS divergence and Jaccard's Coefficient produced the lowest log likelihood scores, which we interpret to mean they performed the best among the six metrics.

As we noted before, Jaccard's coefficient and Kendall's τ use a subset of the vocabulary–top probability words that contribute to a cumulative probability mass. The average size of the set of words with probability mass 0.5 is 39.56, and 0.3 is 13.58. The results show that Jaccard's coefficient can find similar topics at probability mass of 0.5, using only the top 40 words. Kendall's τ does not show good performance compared to Jaccard's coefficient although they use the same set of words. This result indicates that the ranking of top probability words does not matter much in judging topic similarity. DCG does not perform well for this topic similarity task even though it is a good metric of comparing ranked results in information retrieval (IR). This is because the typical results of IR include relevance scores, but the topics found by LDA do not have analogous scores to be used in place of relevance scores.

We further tested Jaccard's Coefficient with various probability masses. However, selecting a proper probability mass can be corpus-dependent. Hence, we conclude that JS divergence is best in terms of performance and generality, so we use JS divergence as the topic similarity metric in the rest of the paper.

6 Topics and Issues

Using the similarity discussed in the previous section, we construct topic chains to understand the topic trends in the main stream news. In this section, we discuss the construction of topic chains and associated parameters, interpretation of long topic chains, and the characteristics of short topic chains.

6.1 Constructing Topic Chains

We construct topic chains by finding similar topics within a certain time window. We use two parameters, similarity cut and sliding window, and follow this process:

1. Calculate the similarity between topic ϕ_i^t and topic ϕ_j^{t-1} for all topics at time $t-1$.

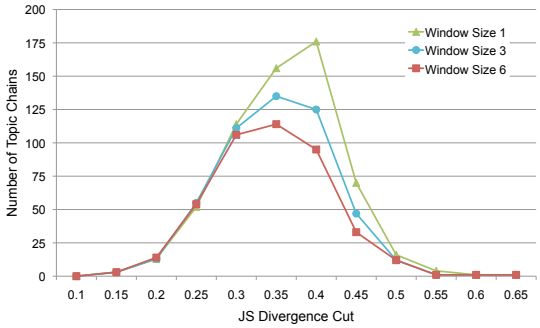

Fig. 2. Number of topic chains with different similarity cuts using JS divergence. The number of topic chains is significantly changed at JS divergence of 0.4. We chose JS divergence of 0.4 to construct topic chains.

2. If there are one or more topics such that $sim(\phi_i^t, \phi_j^{t-1})$ is greater than the similarity cut, we make links between all such topic pairs, and move to the next topic ϕ_{i+1}^t.
3. If there are no similar topic pairs, we calculate similarity between ϕ_i^t and Φ_{t-2}
4. Repeat, going back one more time slice, until one or more similar topics are found, or the time gap between the two time slices exceeds the sliding window size.

The two parameters, similarity cut and the window size, play important roles in determining the characteristics of the topic chains constructed. We discuss each of them below.

Similarity Cut There is no standard similarity cut at which we can say two topics are similar, so we construct several topic chains, varying the similarity cut and looking at the effect on the resulting topic chains. Figure 2 shows how the number of topic chains changes with similarity cut using JS divergence. We define the size of a topic chain to be the number of topics in that chain, and we count topic chains whose size is greater than one. We also experimented with various sizes of the sliding window. If we set the JS divergence cut to a large value, then all topic nodes would be disconnected, and the total number of topic chains of size greater than one would be 0. Conversely, if we set the JS divergence cut to 0, then all topic nodes would be connected, and the number of topic chains would be 1. As Figure 2 shows, the number of topic chains changes significantly at 0.4. To see the relationship between JS divergence values and the similarity of two topics in a qualitative way, we can look at pairs of topics and the JS divergence values. From the qualitative analysis and the analysis of the number of topic chains, we decided that 0.4 is an appropriate threshold of JS divergence for constructing topic chains.

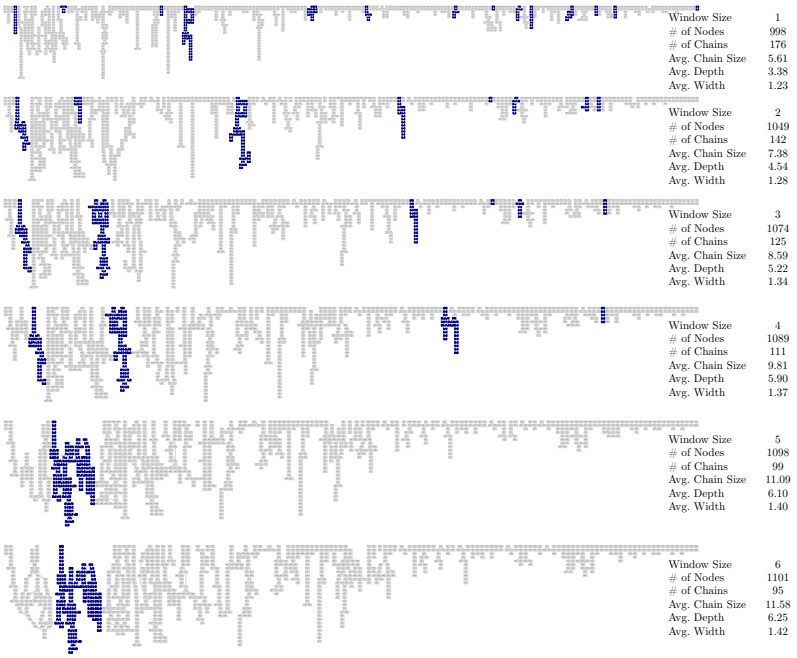

Fig. 3. Six sets of topic chains constructed with sliding windows of sizes one to six. For each set of topic chains, every topic chain starts at the same vertical position. Within each topic chain, topics are temporally ordered, the oldest (first) topic at the top and going down to the most recent topic at the bottom. The number of nodes indicates the total number of topics that are connected with one or more similar topics. Blue nodes are those that belong to the largest topic chain in the last set of topic chains, constructed with the sliding window of size six. Blue nodes start out in the first set of topic chains as small topic chains, and they agglomerate as the size of sliding window increases. The full-size figure is available at *http://uilab.kaist.ac.kr/research/topic-chain/*

Sliding Window Size The size of the sliding window is also an important factor for constructing the topic chains. If we use a sliding window of size one, it means that we only consider the previous time slice to find the similar topics for the topics of the current time slice. However, this Markov assumption is not generally helpful, as similar topics can appear over non-consecutive time slices, so proper consideration of the sliding time window is needed to capture these topic trends.

We vary the size of the sliding window from one to six and observe the changes in the resulting topic chains. Figure 3 shows the topic chains of size greater than one for the various window sizes with their descriptive statistics. First, the number of nodes indicates the total number of topics that belong in topic chains. This number excludes singleton topics and shows how many topics,

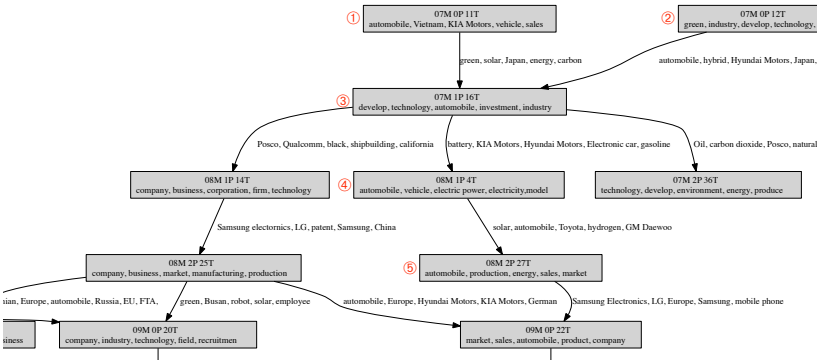

Fig. 4. Detecting focus shifts using difference of a word probability along the topic chain. Each rectangle represents a topic node, and contains top probability words. Edges connect two similar topics within a sliding window of size six, and the words next to the edge are the named entities whose probabilities are changed the most between two topics. xxM yyP zzT represent month, period, and topic number, respectively.

out of 1,400 total, are matched with one or more similar topics within the time window. The number of nodes increases at a faster rate from window size one to four and at a slower rate from window size four to six, and through that, we can see that similar topics do not necessarily appear in consecutive time slices.

Other graph characteristics also change with the size of the sliding window as shown in Figure 3. The total number of topic chains decreases as we increase the window size. This means many of the distributed small topic chains merge as the size of the sliding window increases. This is further evidenced by increases of both the average chain size and the average chain depth. The width of topic chain is the maximum number of topics of the same time slice in that chain. Unlike other increasing characteristics, the average width of the topic chain remains stable throughout the size of the sliding window. This is expected because topics of the same time slice represent different aspects of mainstream news.

Figure 3 illustrates how similar topics agglomerate as we vary the size of the sliding window. We painted in blue nodes of the largest topic chain at a sliding window of size six. We also painted the same nodes at the other sizes of the sliding window. At the window size of one, there are fourteen separate topic chains painted in blue. These chains join together to form larger chains as we increase the size of the sliding window.

6.2 Focus Shifts

When we construct topic chains, we find that there are long topic chains and short or singleton topic chains. Long topic chains tend to cover very general topics such as politics, economy, and sports, and we call them *long-term topics*.

Interpreting Long-Term Topic Chains Looking at a long-term chain is like looking at a section of the newspaper. Many of the long-term topic chains could be labelled as "politics", "business", or "sports", and the topics in those chains reflect a wide variety of subjects within those general news categories. There are also long-term topics, such as H1N1, which are more specific news items but last for a long time. Our topic chains contain more helpful information for interpreting these long-term topics. For example, you can look at the "H1N1" topic chain and read off when the topic first emerged and when it disappeared. You can also see that the topic evolved from talking about "swine flu", to "travel", to "vaccinations" and "deaths".

Named Entities in Topic Evolution Looking at the topic chains, where each node is shown with the top probability words for each individual topic, we can see the general evolution of the topic chain, but it is difficult to interpret the evolution to see what happened. This is because the words that represent the individual topics may be too general and occur in many topics throughout the topic chain. For example, words like *season, home run, game*, and *coach* are always top probability words in a topic chain about baseball. Those frequently occurring top words tell us what the general topic trend is, but it may be more interesting to see how the focus shifts for each topic within the chain.

To identify the words that can help to understand the focus of the topic chain at each time slice, we hypothesized that the words tagged as named entities–people, places and organizations–would be good discriminating words of the different focuses within the topic chain. We illustrate these named entities with the most changes in probabilities in Figure 4. Each rectangle represents a topic with the top five probability words. An edge connects two similar topics, and the words next to the edge are the named entities that change the most between the two topics. For example, topics 1 and 3 are both about the automobile industry, but the named entities *green, solar, Japan*, and *energy*, show that the focus is on green energy for topic 3. We can indeed find a related news article from the time period of topic 3 with the headline "Toyota makes eco-friendly solar car". Also we can see the evolution of the topic from 2 to 3. Topic 2 represents the general green (environmental-friendly) industry. By incorporating the focus words *automobile, hybrid, Hyndai Motors,* and *Toyota* this topic evolves into the topic of environmental-friendly automobiles, topic 3. From topic 3 to 4, the electric car and its battery problems received attention from news, and from 4 to 5, other alternative sources of energy, *solar* and *hydrogen* became the focus.

6.3 Short Topic Chains and Singleton Topics

We discussed long topic chains in the previous section, but short topic chains–chains of two or three topics, or singleton topics–are important for two reasons.

First, most of the short topic chains are about *temporary issues*. If a topic lasts over a long period of time, it would become part of a long topic chain. That means singleton topics and short topic chains are likely to be about temporary

Table 2. Examples of single node topic chains. First to sixth topics are temporary issues. First issue refers to the missile launch from North Korea, second issue is related to the death of Michael Jackson, fifth issue is related to the romance of Korean top actor and actress, and sixth issue is talking about Arbor Day on April 5. These are typical cases of temporary issues. However, the last example is not a coherent topic.

Date	Topic
0P 07M 2009Y	North Korea, missile, launch, range, UN Security Council, ship, navy, East sea, ballistic missile
0P 07M 2009Y	Jackson, family, funeral, cherish the memory of, Michael Jackson, son, LA, publish, report, death
0P 10M 2009Y	melamine, dry milk, region, environment, investigation, food, pollution, mercury, produce
2P 12M 2009Y	flight, airport, passenger, airplane, search, terror, time, security, explosion, aircraft
0P 01M 2010Y	Hyesoo Kim, actor, 2010, ski, Haejin Ryu, once, 4, soul, colleague, lover
0P 04M 2010Y	tree, recover, park, culture, movement, development, environment, ecology, forest, designation
0P 02M 2010Y	Obama, Republicans, Jeju island, game, Jeju, golf, White house, Woods, gamers, budget

issues, and we can see that is true for the examples of singleton topics and short chains listed in Table 2. Topics such as the death of Michael Jackson, reinforcing airport security at the end of the year, and romance between top actors do not last for a long time and can be considered as temporary issues.

Second, some of the singleton topics are useless. When we extract topics with LDA, the results do not consist of only meaningful topics. Sometimes LDA extracts topics that are not understandable as coherent topics. For example, the last topic in Table 2 is not a coherent topic. Constructing topic chains leaves bad results of LDA to be isolated as singleton topics. Conversely, topics that form long topic chains tend to be coherent. Evaluation of topics found by LDA is an on-going challenging research problem[17], so our topic chain framework may offer one solution of evaluating topics found in a sequential corpus. We will explore this in future work.

7 Discussions

In this paper, we proposed a framework for analyzing a corpus of news articles over a contiguous time period. Our framework discovers topics from the corpus, constructs topic chains using a topic similarity metric, identifies long-term topics and temporary issues, and detects focus shifts within each topic chain. An important contribution in this work is a comparison of various topic similarity metrics. We looked at six commonly used metrics and compared them using the negative log likelihood of corpus.

A secondary use of the topic chains is as an analysis tool to evaluate the quality of topics by a topic model. Most of the work on probabilistic topic modeling typically assume that the latent space is semantically meaningful, and so the topics are not systematically evaluated. In this work, we found that most of the topics that belong to long topic chains are semantically meaningful, whereas singleton topics are less coherent. Further analysis of the relationship among topics in the sequential corpus may find an effective way to analyze semantic meaningfulness of the topics.

References

1. Allan, J.: Introduction to topic detection and tracking. Topic Detection and Tracking, Event-based Information Organization (2002) 1–16
2. Blei, D., Lafferty, J.: Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning (2006)
3. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research (2003) 993–1022
5. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. (2004)
6. Feng, A., Allan, J.: Finding and linking incidents in news. Proceedings of the sixteenth ACM Conference on Information and Knowledge Management (2007)
7. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of new topics. Proceedings of the 24th Annual International Conference on ACM SIGIR Research and Development in Information Retrieval (2001)
8. Blei, D., Lafferty, J.: Topic models. Text Mining: Theory and Applications (2009) 71–93
9. Wang, X., Zhang, K., Jin, X., Shen, D.: Mining common topics from multiple asynchronous text streams. Proceedings of the Second ACM International Conference on Web Search and Data Mining (2009)
10. Bolelli, L., Ertekin, Ş., Giles, C.: Topic and trend detection in text collections using latent dirichlet allocation. Advances in Information Retrieval (2009)
11. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2009)
12. Shahaf, D., Guestrin, C.: Connecting the dots between news articles. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2010)
13. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery and Data Mining (2005)
14. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. Proceedings of the 15th International Conference on World Wide Web (2006)
15. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P.: Detecting topic evolution in scientific literature: how can citations help? Proceeding of the 18th ACM Conference on Information and Knowledge Management (2009)
16. Newman, D., Asuncion, A., Smyth, P.: Distributed algorithms for topic models. The Journal of Machine Learning Research **10** (2009) 1801–1828
17. Chang, J., Boyd-graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems (2010)