**Topic Chains for Understanding a News Corpus**

Dongwoo Kim and Alice Oh

CICLing, 2011

Summary submitted by *Anunaya Srivastava*

Given a sequential news corpus this work constructs topic chains, identifies long-term and short-term topics and detects focus shifts. The author presents a framework for probabilistic topic modelling for uncovering the meaningful structure and trends of important topics and issues hidden with the news archives on the web. The author also compares 6 similarity metrics and discusses how they affect topic chain formation.

Some previous works on topic detection and tracking covers only location-based specific events and do not model how topics appear and disappear over time. Few models discuss the appearance and disappearance of topics, but then they assume that the topics stay the same over time. This work covers how topics emerge, evolve with time and then disappear. Author considers a broader definition of an event i.e. not only an earthquake in Gujarat is an event but also prolonged inflation. The former is called a short-term event and the later is called long-term event. Previous techniques do not cover long-term events.

Approach followed by author –

1. **Topic extraction using LDA**

LDA is the most widely used method of probabilistic topic modelling. The author extracted 50 topics using LDA for every time slice using Gibbs sampling. Detailed discussion on LDA is omitted here and will be covered in a different report.

1. **Measuring topic similarity**

Author compares the similarity of the topics using the following similarity metrices –

1. **Cosine Similarity:** It measures the similarity between two vectors by finding the cosine of the angle between them.
2. **Jaccard’s Coefficient:** It 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.
3. **Kendall’s τ Coefficient:** It measures the association between two ranked lists.
4. **Discounted Cumulative Gain (DCG):** It measures the effectiveness of the ranked results of a web search algorithm.
5. **Kullback-Leibler Divergence:** It is a non-symmetric measure of the difference between two probability distributions p and q.
6. **Jensen-Shannon Divergence:** It is the symmetric variation of KL divergence.

Author used negative log likelihood for the comparison between the metrics. Compute the similarity of all possible topic pairs for two consecutive time slices. Then replace the top 5 most similar topics, and computer the negative log likelihoods. A better similarity metric gives lower negative log likelihood. Author found that JS divergence gives the best result and used it to get similar topics.

1. **Construct topic chains**

Author uses two parameters – similarity cut and similarity window – to construct topic chains and follow the process below.

1. Calculate the similarity between topicφti and topicφ t−1j for all topics at time t−1.
2. If there are one or more topics such that sim(φti,φt−1j) is greater than the similarity cut, we make links between all such topic pairs, and move to the next topic φti+1.
3. If there are no similar topic pairs, we calculate similarity between φti and Φt−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.

**Similarity cut:** It is the threshold value for similarity between 2 topics. If low value of similarity is set, all topics will seem to be connected. If high value is set, all topics will seem to be disconnected.

**Sliding window:** It is the number of time slices in the past till which similarity between the articles is calculated. Varying the window size shows that similar topics do not necessarily appear in consecutive time slices. Also, number of topic chains decrease as we increase window size, which means many small topic chains merge as window size increase. It is evidenced by increase in average chain size and average chain depth.

1. **Focus shift**

Long-term topics are identified by capturing the general words that occur throughout the chain. On the other hand, one can discover short-term topics and how they emerge and disappear by looking at the focus shifts. A topic chain exhibit different focusses for each individual topic in the chain. Shift in the focus can be used to infer the appearance and disappearance a topic.

To determine the focus shift we need to calculate the word probability that has changed the most between two topics. Author calculates the probability of named entities such as people, places, organizations etc. which help to understand the focus in the topic according to author’s hypothesis. Consider the diagram below. 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.

