DISCOVERING EVENT EVOLUTION GRAPHS FROM NEWSWIRES

Christopher C. Yang and Xiaodong Shi

**Overview:**
- In Topic Detection and Tracking (TDT) news stories are often organized into a flat hierarchical structure where inter-cluster relationships are missing.
- Modeling the event evolutions of news topics and presenting them in a graph structure can be useful in various applications:
  - Direct users through the news topic in information browsing.
  - Integrate with automatic summarization techniques and graphical interfaces to provide a graphical web news infomedia.
- We propose to represent the event evolutions of news topic using an event evolution graph, which is a directed graph with its vertices as events and its edges as event evolution relationships.

**Event Evolution and Event Evolution Graph:**
- We define *event evolution* as the directional dependencies or relatedness, which exhibit the track of event development, between two events inside a same news affair; the relationship between such two events is called *event evolution relationship*.
- An *event evolution graph* is defined as a directed acyclic graph (DAG) \( G = (V, L) \) consisting of events as the nodes \( V = \{e_1, e_2, \ldots, e_j\} \) and event evolution relationships as the directed edges \( L = \{(e_i, e_j)\} \) between nodes.

**Modeling Event Evolution Relationships:**
- Previous researches, e.g. event threading, uses the average pairwise story similarity as the measurement of event evolutions.
- Event threading neglects the properties of events and simply treats event as an aggregate set of news stories.
- We propose to measure the event content similarity between two events by calculating the cosine similarity of the event term vectors, which is then used to represent the textual content of each news story.
- The event term vector of event \( e_i \) is computed as the average of the document term vectors of stories that belong to \( e_i \). TF weights are used instead of traditional TF-IDF weights.
- The event content similarity between events \( e_i \) and \( e_j \) is defined as:

\[
\text{conf}(e_i, e_j) = \frac{0}{\text{cs}(e_i, e_j) \times tp(e_i, e_j) \times df(e_i, e_j)} \quad \text{if} \quad i \neq j \quad \text{and} \quad s_j > s_i
\]

\[
\text{cs}(e_i, e_j) = \cos \text{ine}_\text{sim}(etv(S_i), etv(S_j))
\]

where \( etv() \) is the event term vector representation of the set of stories belonging to the same event.

**Event Content Similarity:**
- We use the simple bag of words model to represent the textual content of each news story.
- The event term vector of event \( e_i \) is computed as the average of the document term vectors of stories that belong to \( e_i \). TF weights are used instead of traditional TF-IDF weights.
- The event content similarity between events \( e_i \) and \( e_j \) is:

\[
\text{cs}(e_i, e_j) = \cos \text{ine}_\text{sim}(etv(S_i), etv(S_j))
\]

where \( etv() \) is the event term vector representation of the set of stories belonging to the same event.

**Temporal Proximity:**
- Assume the timestamp of an event \( e_i \) is a timeinterval \( [s_i, e] \), the temporal distance between two events \( e_i \) and \( e_j \) as \( (s_i, s_j) \):

\[
d(t_i, t_j) = \begin{cases} s_j - s_i & \text{if} \quad s_i \leq s_j \\ 0 & \text{if} \quad s_i > s_j \end{cases}
\]

Intuitively if two events are farther away from each other along the timeline, the event evolution between them is less likely to exist.
- The temporal proximity between two events is: \( (s_i, s_j) \):

\[
\text{tp}(e_i, e_j) = e^{-\alpha \frac{d(t_i, t_j)}{T}}
\]

where \( T \) is the event horizon defined as the time-span of the entire news affair. \( \alpha \) is the time decaying weight \( (0 \leq \alpha \leq 1) \).

**Experiments:**
- Event threading model combined with Nearest Parent or Best Similarity graph model is selected as the baseline.
- When event evolution model is combined with static thresholding, it outperforms the rival models a lot. \( \alpha=0.5, \beta=0.5 \)

The Precision and Recall Curves (Interpolated to Standard 11 Levels) of the Comparative Experimental Results