CS 345A: Topic Chaining and Phrase Linking

Sandeep Sripada(ssandeep), Venu Gopal Kasturi(venuk)

1 Introduction

In this project, we implemented a technique to break down a collection of news articles into semantically coherent threads. The chaining of articles is done based on the content and temporal aspects of the news articles. The problem of computing threads was solved by using a matching based algorithm on a relevance graph. We also tried two approaches in analyzing the resulting threads to get relations between the most common phrases: (a) Timestamp based clustering to get phrase group links and (b) Matching on the graph constructed using phrases to get links. Results on approximately 3 million news articles over a period of four years show that the analysis is effective.

Related work include: Newsjunkie where the approach was to cluster data and select the minimum set of documents that convey the maximum information [2].

2 Thread formation

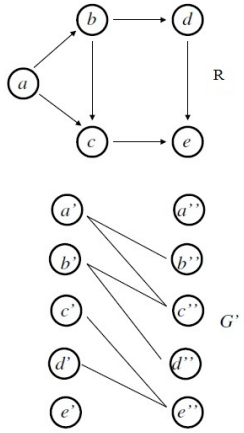
This approach is based on [1]. The new articles are preprocessed to obtain a term-document matrix M which has information about the article like article ID, timestamp, tf-idf score. The terms of the matrix are the unigrams in the articles and the score i.e. $M(D,T)$ is the amount of "presence" of the term T in the document D. This matrix will be used in the construction of a relevance graph.

2.1 Relevance graph

For each term T in the set of terms, we consider $D_T = D_1^{'}, ...$, the set of documents corresponding to the term, sorted based on the timestamps. Let w be a window parameter. We add the edge $(D_i^{'}, D_j^{'})$ if and only if $|i-j| <= w$, i.e., we add an edge between two documents if they are at most w apart. Now, as an edge can have multiple votes i.e., a multigraph on documents, we can use this information to denote the "strength" of the linking between the two documents. So after rounding the egde based on a threshold value, we pass the resulting graph to the matching based algorithm to obtain threads. The reason for rounding is to remove spurious edges caused by outliers.

Let $R^{'}$ denote the term-document relation constructed based on the query terms T. This relation is a subset of the initial term-document relation R. $R^{'} = (T^{'}, D^{'})|(T, D^{'}) \in R$; that is, $R^{'}$ consists of all the terms from the document set $D^{'}$ which was calculated based on the query terms T. The documents are time-sorted to make the subsequent computations efficient.

Figure 1: Illustration of the matching-based algorithm.



2.2 Matching based threads

The algorithm proceeds by constructing an undirected bipartite graph containing the documents as vertices. The undirected edge between documents (vertices) from one partition to the other is made only if they have a corresponding edge in the relevance graph constructed above. Figure 2.2 shows the construction of an undirected bipartite graph $G^{'}$ from the relevance graph R. The cost of this edge is set such that edges with lower cost are selected while maximizing flow and minimizing cost. In out case, we set the cost to $(-1 * R(D_i^{'}, D_j^{'}))$. Also, the capacities of each of the edge is set to 1. We added the necessary source and sink edges to use the min cost, max flow algorithm to obtain threads. Using the flow path, we implemented a union-find algorithm between the edges of the flow path to obtain vertices (documents) part of the thread.

We also tried the exact solution of calculating the min cost flow problem which is the (a,b)-threads problem. We chose the matching based solution to the exact solution because of two reasons:

- The algorithm is not memory friendly and slow.

Table 1: Threads for query terms: 'osama plane hijack world trade center'

Timestamp	Summary
2001/09/29	A Nation Challenged.
2001/10/14	Article summarizes major developments of past month in US war against terrorism.
2001/11/04	The challenge: Tasks of Reconstruction.
2001/12/16	Efforts by Hillary Clinton to persuade Congress for funds to help rebuild downtown Manhattan.
2002/01/30	Transcript of Pres Bush's State of the Union address, as recorded by The New York Times.
2002/04/09	New York Times wins record seven Pulitzer Prizes, six for its news coverage of terrorist attacks.
2002/09/11	First anniversary of Sept 11 terrorist attacks to be marked with public ceremonies.
2003/09/12	Nation pauses to remember 3,016 victims of 9/11 terrorist attacks two years ago.

- It also produced threads which were fragmented.

2.3 Implementation issues

- We had a limit on the maximum length of the thread as longer threads had topic drift.

- We also had a threshold on the set of documents considered for building the relevance graph. Only documents that had a certain number of key words were chosen to reduce the blow-up of threads because of unrelated terms. For eg: A query 'Osama world trade center' may return documents (as per the construction mentioned in 2.1) that have the term 'world' and are related to 'music' or 'arts'.

The results from a few queries are shown in Table 1 and Table 2. The window parameter here is 10 which helps in controlling the span of the articles returned. The results using this approach are quite robust and effective in generating coherent threads.

3 Phrase linking

This section describes the analysis done on the resulting threads obtained in 2. We tried two different approaches to analyze the information in the thread structure to obtain some insight into the phrase behavior.

Phrases required for the analysis were extracted using the following methods:

1. The entire content of the article was used to generate n-grams (n ranging from 4-10) and the corresponding tf-idf scores. These scores helped in picking the top phrases for a particular document.

2. All the quotations in the article were extracted and n-gram based method (as described above) was used to generate phrases and their scores.

3.1 Timestamp based document clustering

In this method, we take the threads and cluster them based on the timestamps, with documents within a time period grouped together. These groups are then further analyzed to obtain links between groups of phrases. The intuition behind this approach is that the documents in the threads within a time frame tend to talk about similar things and following these groupings along the threads would lead to interesting phrase behavior.

After clustering, each group of documents is taken and passed through the following algorithm:

```
while(no more groups)
{
     1. Get top phrases from the group
     2. Get next set of documents along the
      threads for the current group
}
```

The next set of documents are generated by taking the current group and fetching documents from threads such that the document picked succeeds the document in the current group by one hop. The set of all such documents are taken and the algorithm proceeds until next set is empty. The phrases extracted from each group are considered as linked since the documents that generate them are linked in the threads i.e. the document are ordered in a thread.

Let $T = ((D_1^1, D_2^1, D_3^1, ...), (D_1^2, D_2^2, D_3^2, ...), ...)$ represent the set of threads, where D_i^j represents a document i in the thread j.

Let $C_k = (D_i^j | D_i^j \in K)$ be the cluster of documents grouped in bin k. The documents in the new set would be the collection $(D_{i+1}^j ...)$ such that $D_i^j \in C_k$ and $D_{i+1}^j \notin C_k$.

This method gave positive results as shown in the Section 4.

Table 2: Threads for query terms: 'victims world trade center'

Timestamp	Summary
2001/10/05	Brief profiles of some of those who lost their lives in Sept 11 terrorist attack on WTC.
2001/10/10	THE VICTIMS; A Devoted Sister, a Garden Retreat and Weekends With the Parents.
2001/10/11	THE VICTIMS; A Would-Be Chef, a Game Show Winner, a Builder of Sand Castles.
2001/11/18	THE VICTIMS; Routines And Plans, And Friends And Family.
2001/12/05	A Firefighter, an Artist and a Mother Who Never Laughed Alone.
2001/12/15	Giggling Falls Silent, And Fountains Are Dry: 'My Everything' Is Gone.
2001/12/26	A Handy Bond Trader, a Kitchen Whiz Technician, a Ponytailed Doctor.
2001/12/30	Portraits of Grief brief sketches of some victims of Sept 11 attack on WTC.
2002/03/31	Portraits in Grief profiles of some of victims of World Trade Center attack.

3.2 Phrase drift using Maximum bipartite matching

We are interested in understanding how the phrases in the documents comprising the threads vary with time. Of particular interest is the way Memetracker [3] works where the evolution of phrases over time was found in a huge collection of blogs during the 2008 US Presidential election (over a 3 month span). The aim is to know how phrases influence other phrases' popularity.

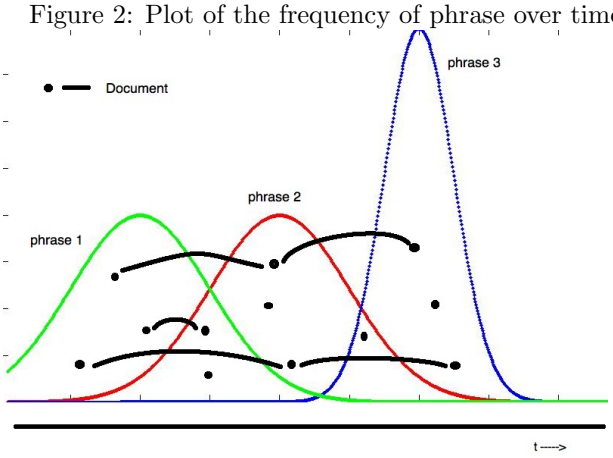
The overview of this approach is as follows:

- Obtain the phrases from the articles in the thread.
- Calculate the time to peak of each phrase.
- Build a relevance graph amongst the phrases.
- Compute the influence of each phrase over it adjacent phrase using the Maximum Bipartite matching algorithm suggested in [1].

The phrases from the documents are extracted using the n-gram method as described above. A plot of the frequency of each phrase over time is taken to obtain the 'time to peak' of each phrase. Our hypothesis is that a phrase A is a reason for a rise in the popularity of another phrase B, if there are documents D_i which contain the former and documents D_j containing the later and D_j follows D_i in the sets of threads of the news articles considered. Also, the time to peak for A should be before B. The phenomenon of a phrase rising the popularity of another is captured here by considering the temporal ordering of the documents in the threads i.e., the way in which a set of phrases are giving way for another set as the story evolves over time. It is illustrated in Figure 3.2.

A graph is constructed with the phrases as follows: Each phrase represents a node in the graph and there is an edge between two nodes, if there two documents D_i and D_j in a story which contain the phrases and the time stamp for A's peak is less than the corresponding time for B. The graph constructed using the phrases

Figure 2: Plot of the frequency of phrase over time.



would have the number of votes from A to B (since it is a multigraph) as the weight of the edge. This graph is solved to obtain the min-cost flow just as described in 2.

Implementation:

- The phrases that we have considered are the quoted phrases that are present in the articles. The hypothesis is that articles tend to quote the phrases that people make and hence we take these quotations and calculate the 4-10 grams. We then take the top-k phrases in the set of documents that are of interest.
- Similar phrases tend to add duplicity to the set of phrases we have and this disperses the TF-IDF of each phrase. We considered a bag of words representation of a phrase and calculated the Jaccard Similarity to remove duplicates. Another method that we tried was to remove phrases that have a low edit distance.
- The cost of an edge in the phrase graph is $-1 * \#votes$. This helps in our formulation to calculate the minimum cost max flow.

- Avoiding a cycle in the graph of the phrases: In the formulation of the relevance graph, we have ordered the phrases by the time for each of them to attain their peak. This can be tricky to compute as a phrase frequency can many local peaks. Hence we used the median of the time stamps of the documents which have this phrase as the 'time to peak'.

4 Results

We have performed the experiment on the New York Times dataset over 2000-2003. There were approximately 3 million news articles in the dataset and the reason for choosing the time frame was to obviously catch on the high activity during the September 11 attack.

4.1 Time based clustering

Set 1: Phrase: 1: a biography of the
Phrase: 2: reporter for the times
Phrase: 3: new york trade center
Phrase: 4: between north and south

Set 2: Phrase: 1: in a grim search for survivors among the thousands presumed
Phrase: 2: in new york city and the pentagon in
Phrase: 3: responsible pointed toward five suspects whose
Phrase: 4: mountains of rubble at what had been the world trade center
Phrase: 5: trade center yesterday in a
Phrase: 6: net for those behind the hijackers
Phrase: 7: dead in its collapse
Phrase: 8: behind the hijackers who slammed jetliners into the twin towers in

Set 3: Phrase: 1: the taliban was ready
Phrase: 2: radio shariat said the taliban
Phrase: 3: attacks in the united states
Phrase: 4: an international coalition against
Phrase: 5: possible american military strike tajikistan senior diplomats from
Phrase: 6: if washington could prove Laden's involvement in the
Phrase: 7: iran and other states hostile to the taliban met in the
Phrase: 8: was ready to hand over osama bin laden to

Set 4: Phrase: 1: of civilians witnesses and local
Phrase: 2: american bombers over tora bora
Phrase: 3: of their military forces around kandahar the united states military and
Phrase: 4: negotiations on a new posttaliban government
Phrase: 5: local officials said but an american
Phrase: 6: a village saying that the target
Phrase: 7: that surrender is the only way out of
Phrase: 8: hopeless plight meanwhile in
Phrase: 9: forces continued to send the talibans besieged defenders
Phrase: 10: that american bombs had mistakenly hit a

Set 5: Phrase: 1: weeks members of congress will
Phrase: 2: will be reminded anew
Phrase: 3: the rest of the nation
Phrase: 4: the coming weeks members
Phrase: 5: do not want their cheaper
Phrase: 6: of congress will be reminded anew that such fights often break
Phrase: 7: flowing to northerners midwesterners are eager to promote the
Phrase: 8: strike an agreement on energy issues in the coming
Phrase: 9: lawmakers of both stripes from oilproducing states think the

The results flow in the following manner: Set 1 starts with the World Trade Center, Set 2 is about the destruction caused and issues with searching for victims, Set 3 talks about the involvement of Osama Bin Laden and his troops, Set 4 talks about the bombings in the middle-east where innocent civilians were hurt, Set 5 talks about Congress and Oil.

4.2 Graph based phrase linking

The results from the second approach in calculating the links among phrases are as shown in Table 3.

5 Conclusion & Future work

In this project, we have looked at the problem of detecting the evolution of phrases in a story over time. For this we have looked at the approaches used for Topic detection & Tracking. We have suggested two hypothesis to obtain a flow of phrases in a coherent set of documents without any semantic analysis. One approach uses a time based clustering method for relating sets of phrases. The other relates a pair of phrases by the temporal presence of documents in the thread of the story. The time based clustering worked better than the second approach. The second approach failed because of using just one source, i.e. New York Times data. The second approach depended on the number of documents voting for a pair of phrases. And since we have just one source for the articles, there was drift in the story but not too many documents citing the same quotes of an event.

The quality of these results depends on two key factors - extraction of phrases from a document [4] and the representation of a given phrase. We believe that better

Table 3: Phrase links for query terms: 'osama world trade center plane hijack', method: Graph based

From	To
call back a friend of ours was on it	they said yeah thats thats what you heard
life he loved to sail he was	she was also looking through the song
father said of him after the attack on the	it from the minute he saw her
on the swing drinking coffee we had	to be in the same place a little

heuristics for considering the phrases would yield better results. It would be interesting to look at using these approaches on a huge collection of web blogs as there would be larger number of documents in a short time span which cover a story line. We can also observe a quality drift of the phrases as the story progresses. But the key would be to use the entire content of a blog rather than a representation of it with a few memes since the first step is to cluster these documents to obtain the temporally connected documents as threads.

References

[1] R. Guha, Ravi Kumar, D. Sivakumar, and Ravi Sundaram, *Unweaving a web of documents*, Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, August 21-24, 2005, Chicago, Illinois, USA.

[2] E. Gabrilovich, S. Dumais, and E. Horvitz, *Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty*, Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004), May 2004, New York, pp. 482-490.

[3] Jure Leskovec, Lars Backstrom, and Jon Kleinberg, *Meme-tracking and the dynamics of the news cycle*, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 28-July 01, 2009, Paris, France.

[4] R. Kumar, U. Mahadevan, and D. Sivakumar, *A graph-theoretic approach to extract storylines from search results*, In 10th KDD, pages 216-225, 2004.

[5] Deept Kumar, Naren Ramakrishnan, Richard F. Helm, and Malcolm Potts, *Algorithms for storytelling*, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2006, Philadelphia, PA, USA.

[6] T. Dalamagas and M. D. Dunlop, *Automatic construction of news hypertext*, In HIM, pages 265–278, 1997.