Topic Identification and Analysis in Large News Corpora

Sarjoun Doumit and **Ali Minai**
Complex Adaptive Systems Laboratory
School of Electronic & Computing Systems
University of Cincinnati
Cincinnati, Ohio 45221
doumitss@mail.uc.edu
ali.minai@uc.edu

Abstract

The media today bombards us with massive amounts of news about events ranging from the mundane to the memorable. This growing cacophony places an ever greater premium on being able to identify significant stories and to capture their salient features. In this paper, we consider the problem of mining on-line news over a certain period to identify what the major stories were in that time. Major stories are defined as those that were widely reported, persisted for significant duration or had a lasting influence on subsequent stories. Recently, some statistical methods have been proposed to extract important information from large corpora, but most of them do not consider the full richness of language or variations in its use across multiple reporting sources. We propose a method to extract major stories from large news corpora using a combination Latent Dirichlet Allocation and with n-gram analysis.

Introduction

The amount of news delivered by the numerous online media sources can be overwhelming. Although the events being reported are factually the same, the ways with which the news is delivered vary with the specific originating media source involved. It is often difficult to reliably discern the latent news information hidden beneath the news feeds and flashes due to the great diversity of topics and the sheer volume of news delivered by many different sources. Analysis of news is obviously of great value to news analysts, politicians and policy makers, but it is increasingly important also for the average consumer of news in order to make sense of the information being provided.

Latent Dirichlet Allocation (LDA) (Blei *et al.* 2003) is a probabilistic method to extract latent topics from text corpora. It considers each textual document to be generated from a distribution of latent topics, each of which defines a distribution over words. It is a powerful tool for identifying latent structure in texts, but is based on a "bag-of-words" view that largely ignores the sequential dependencies of language. This sequential information is the basis of the n-gram approach to text analysis, where preferential sequential associations between words are used to characterize text (Manning & Schultze 1999) and (Wikipedia 2012). In the present work, we used n-grams and LDA together to organize structured representations of important topics in news corpora.

The rest of this paper is organized as follows: in the next section we give an overview of relevant work, followed by a description of LDA. Then we describe our model followed by the simulation results and conclusion.

Relevant Work

There exist today many research and commercial systems that analyze textual news employing methods ranging from the statistical to the graphical, but it is still up to the news analysts or users of the system to organize and summarize the large output according to their own specific needs to benefit from the result. For example, WEIS (McClelland 1971) and (Tomlinson 1993) and CAMEO (Gerner *et al.* 2002) are both systems that use *event analysis*, i.e. they rely on expert-generated dictionaries of terms with associated weights, and parse the text to match the words from the news event to those in the dictionary. They can then map the information into a set of expert-defined classes with respect to sentiment intensity values. In the Oasys2.0 system (Cesarano *et al.* 2007), opinion analysis depends on a user feedback system rather than on experts in order to determine the intensity value of an opinion. The Oasys2.0 approach is based on aggregation of individual positive and negative identified references (Benamara *et al.* 2007). The RecordedFuture (Future 2010) and Palantir (Palantir 2004) systems also rely on experts and have at hand massive amounts of data, with inference and analysis tools that use data correlation techniques to produce results in response to specific keywords in user queries. More recently, topic chain modeling (Kim & Oh 2011) and (Oh, Lee, & Kim 2009) and (Leskovec, Backstrom, & Kleinberg 2009) has been suggested as a way to track topics across time using a similarity metric based on LDA to identify the general topics and short-term issues in the news. It is important to note that all the approaches mentioned above except topic chain models adopt query-driven and human-dependent methods to produce results.

Latent Dirichlet Allocation

There has been great interest in Latent Dirichlet Allocation ever since the publication of the seminal paper by Blei, Ng and Jordan (Blei *et al.* 2003). It is a machine learning technique that extends a previous model called *Probabilistic Latent Semantic Analysis* (Hofmann 1999) (pLSA) for reducing the dimensionality of a certain textual corpus while preserving its inherent statistical characteristics. LDA assumes that each document in a corpus can be described as a mixture of multiple latent topics which are themselves distributions over the vocabulary of the corpus. LDA assumes that documents are bags-of-words where the order of the words is not important. LDA is a generative model in that it can generate a document from a set of topics, but it can also be used as an inference tool to extract topics from a corpus of documents. This is how we use it in the work presented here.

Methods

We have been collecting and building an extensive database covering 35 online world-wide news media sources through their English-version RSS feeds (Libby 1997) to test our analysis approach. We collect all news articles from these media sources around the clock at specific intervals. A graphical representation of our news collection model is shown in figure 1.

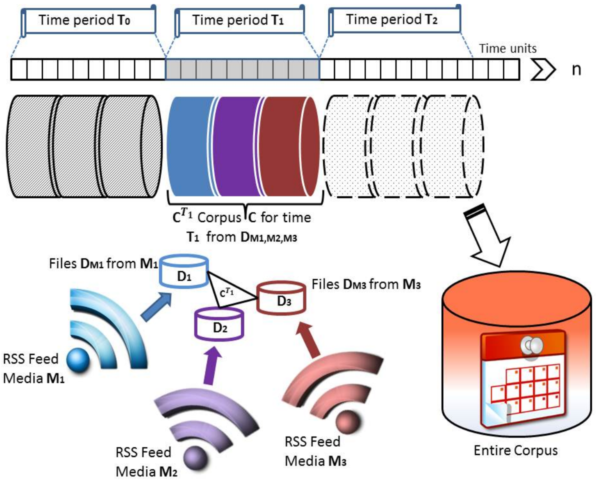

Figure 1: News collection and organization

Each RSS news item is typically just a few sentences, which poses a major challenge to any statistical model for extracting meaningful data. We compensate for this by using a large number of these small RSS news items in order to identify the significant stories that are prevalent during the time period of interest. Ambiguities created by alternative spellings for the same words are resolved by using multiple media sources *en masse*, so that the large number of words strongly correlated with the ambiguous words help in their disambiguation. Using our modified inflector-stemmer algorithm in addition to regular expressions, were were able to handle the general abbreviations prefixes and suffixes used in the text, in addition to the erroneous symbols or “words” encountered occasionally in RSS feeds. We organized the collected data from the different media in contiguous time units T_i into sub corpora, which together make up the overall corpus. This organization allows us to run our simulations on any time frame, for any media or collective media that we have. Since we are still collecting data, the overall time-frame is still growing. Once the time frame and media source(s) of interest are established, we use LDA to granularize the news documents into topics, each represented by a distribution of words. Our assumption is that LDA will be able to identify the set of important events that occurred during this time period, and this will be reflected in the generated topics. For this, we use a smoothed-LDA method based on the work of Newman (Newman 2010). There are two significant problems with the topics generated by LDA:

1. The topics discovered do not necessarily correspond to distinct stories, and typically comprise a mixture of several stories.
2. There is no structure in the topics beyond the distribution over words.

We address these problems by extracting n-grams from the topics generated by LDA, clustering them into groups corresponding to specific stories using statistical heuristics, labeling the stories based on these clusters, and organizing the terms associated with each cluster into a semantic network where the n-gram words (or concepts) are the nodes and edges indicate the strength of directed association between the words in the corpus. This provides both a labeled set of stories and an associated characteristic semantic network reflecting their structure.

Results and Discussion

To validate our approach, we tested our system on a test corpus of 400 news RSS feed stories custom-built to comprise a small number of known news topics. These were:

- The Bin Laden killing.
- The News of the World hacking scandal.
- The Prince William-Kate Middleton wedding.
- Japan’s Fukushima earthquake and tsunami disaster.

The distribution of stories is shown in Table 1.

News	Stories in Test Corpus
Bin Laden Killing	100
Japan’s Fukushima Disaster	100
Murdoch News Scandal	100
Prince William’s wedding	100
Total	400

Table 1: Test Corpus Distribution

It should be noted that the assignment of a story as belonging to a specific topic is still somewhat subjective, and it is possible that different human readers might disagree on the exact partitioning of the corpus. Figure 2 shows the results produced by the system. While 35% of the stories remained unlabeled (see below), the system was able to label the remaining 65% of the stories with 100% accuracy. The number of labeled stories from each topic are shown in Table 2.

An informal manual analysis of the detailed results indicated that stories characterized by a few salient features are

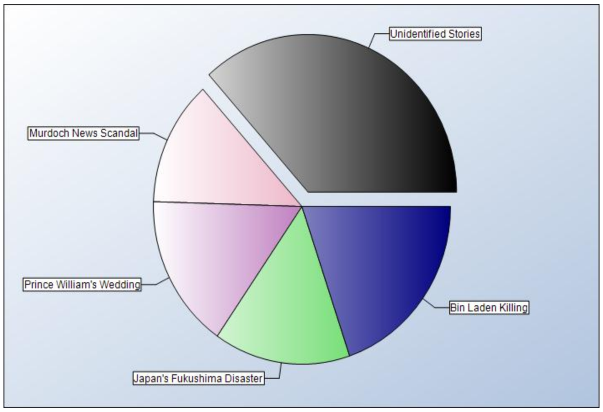

Figure 2: Test corpus identified stories

News	Labeled Stories
Bin Laden Killing	80/100
Japan's Fukushima Disaster	60/100
Murdoch News Scandal	55/100
Prince William's wedding	65/100
Total	260/400

Table 2: Labeling Performance

labeled better than complicated stories with many features. For example, in some runs (not shown, the royal wedding story was split up by the system into two stories – one about the wedding and the other about the bride's dress!

After validation on hand-crafted test sets, the method was tested on raw data from newsfeeds for the month of March 2011. Three news media sources – CNN, Fox News and the BBC – were considered separately. All three media sources produced topic labels corresponding to the Libyan uprising, the Japanese earthquake, and several other stories. However, we noticed a greater focus on the Japanese story by the two American news sources compared to the BBC. We also saw the opposite trend for the Libyan story. The semantic networks generated by the three sources for the Japanese earthquake story are shown in Figure 3 for CNN, Figure 4 for Fox News and Figure 5 for the BBC news. The size and complexity of the networks indicate the level of detail and significance each news source allocated to the story.

As can be seen in the CNN Figure 3, the word-node *japan*, when found in CNN news stories for the month of March 2011, was followed all the time by the word *nuclear* (100%) and then *plant* and *radiation* in that order. The other word-nodes in the network each had a different probability to follow in their patterns. It is interesting to see a somewhat similar pattern for the Fox News semantic network in Figure 4 where *japan* was followed by *nuclear* (50%) and *plant* (11.11%) but quite different than the BBC's network in Figure 5. Although the total number of all news stories collected from BBC was 17,350 and just 2,027 for CNN and 5,573 for Fox News, the focus of BBC for March 2011 was more on the Libyan crisis and the Ivory Coast presidential crisis, which were less significant in the American news

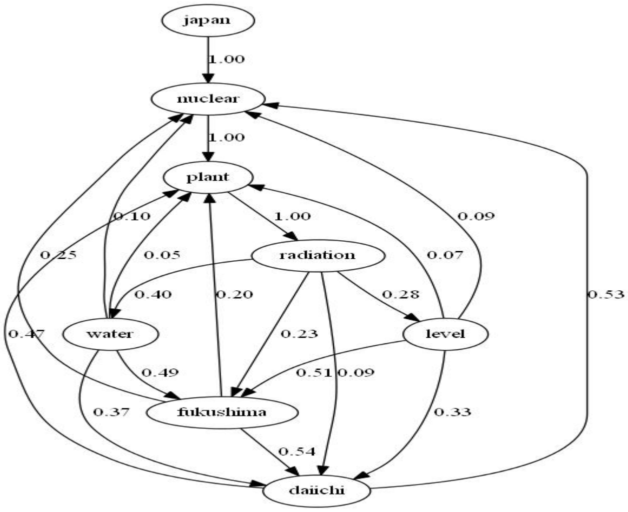

Figure 3: Japan's Earthquake - CNN March 2011

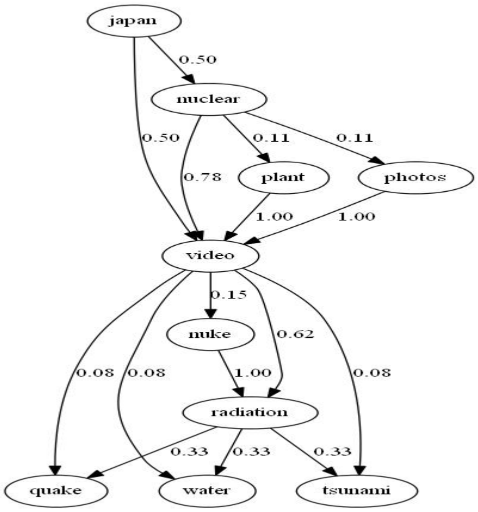

Figure 4: Japan's Earthquake - Fox News March 2011

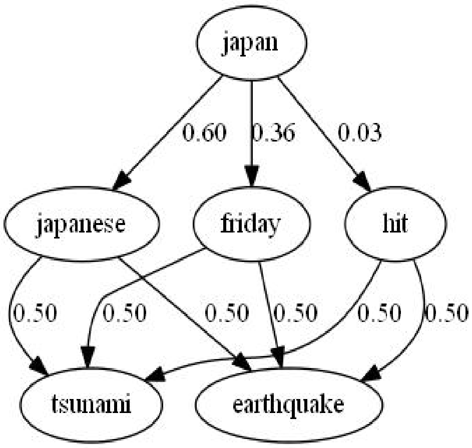

Figure 5: Japan's Earthquake - BBC March 2011

media (this was well before NATO engagement in Libya). Indeed, the semantic network extracted from BBC for the Libya story is too complex to be shown here! It was also evident that the American news media's coverage of the Japan story was richer in content and more diverse than that of the BBC. Another interesting finding, reflecting the inescapable Zipfian nature of lexical distributions, is that the n-grams rank frequencies in all cases had power law shapes, as shown in Figure 6 for the BBC.

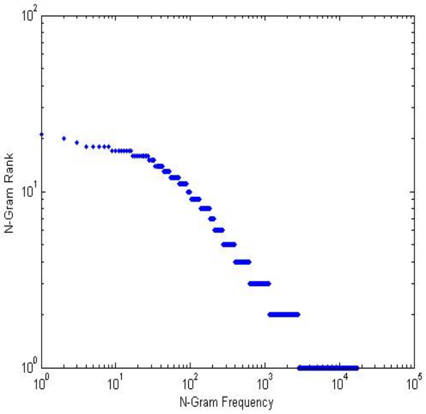

Figure 6: N-Gram Rank-Frequency - BBC January 2010

Conclusion

In this paper, we have provided a brief description of a method we are developing for the automated semantic analysis of on-line newsfeeds. The preliminary results shown indicate that the method holds great promise for deeper analysis – perhaps including longitudinal analysis – of news, which will be valuable to both professionals and the public.

References

Benamara, F.; Cesarano, C.; Picariello, A.; Reforgiato, D.; and Subrahmanian, V. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *International AAAI Conference on Weblogs and Social Media (ICWSM)* 203–206.

Blei, D.; Ng, A.; Jordan, M.; and Lafferty, J. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cesarano, C.; Picariello, A.; Reforgiato, D.; and Subrahmanian, V. 2007. The oasys 2.0 opinion analysis system. *International AAAI Conference on Weblogs and Social Media (ICWSM)* 313–314.

Future, R. 2010. Recorded future - temporal & predictive analytics engine, media analytics & news analysis. [Online; accessed 22-November-2010].

Gerner, D.; Abu-Jabr, R.; Schrodt, P.; and Yilmaz, . 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association of Foreign Policy Interactions*.

Hofmann, T. 1999. Probabilistic latent semantic analysis. *Uncertainty in Artificial Intelligence, UAI99* 289–296.

Kim, D., and Oh, A. 2011. Topic chains for understanding a news corpus. *12th International Conference on Intelligent Text Processing and Computational Linguistics(CICLING 2011)* 12.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 497–506.

Libby, D. 1997. Rss 0.91 spec, revision 3. *Netscape Communications*.

Manning, C. D., and Schultze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

McClelland, C. 1971. World event/interaction survey. *Defense Technical Information Center*.

Newman, D. 2010. Topic modeling scripts and code. *Department of Computer Science, University of California, Irvine*.

Oh, A.; Lee, H.; and Kim, Y. 2009. User evaluation of a system for classifying and displaying political viewpoints of weblogs. *AAAI Publications, Third International AAAI Conference on Weblogs and Social Media*.

Palantir. 2004. Privacy and civil liberties are in palantirs dna. [Online; accessed 10-December-2010].

Tomlinson, R. 1993. World event/interaction survey (weis) coding manual. *Department of Political Science, United States Naval Academy, Annapolis, MD.*

Van Rijsbergen, C.; Robertson, S.; and Porter, M. 1980. New models in probabilistic information retrieval. *British Library Research and Development Report* 5587.

Wikipedia. 2012. N-gram — wikipedia, the free encyclopedia. [Online; accessed 13-March-2012].