Visualizing News as Entity-Topic Interactions

Maya Ramanath, Amitabha Bagchi, Rahul Goyal and Ravee Malla
Indian Institute of Technology
New Delhi, India
{ramanath, bagchi, cs5080222, cs5080224}@cse.iitd.ernet.in

ABSTRACT

Exploring news, both real-time and archival, is a complex task. One needs to go through a series of related articles and derive an underlying story by combining these articles meaningfully. Moreover, for understanding a chain of events, one needs to go navigate through a series of articles, each of which provide only part of the complete picture. For a user without prior knowledge of a story, it is unclear as to which articles would be ideal as a starting point. We describe & evaluate ESTHETE, a news browsing interface to address this problem. It allows users to browse through archive news documents augmented with useful information like the important entities in the article along with the topics that were being talked about at that time. In addition, the tool aggregates articles and presents them as themes which highlights the important stories throughout the period. The interface uses a previously developed graph generation algorithm as a back-end to mine & score actor relationships modeling them as transformations. The interface is aimed at making this process of news discovery useful for the user. We describe the various components that go into building this interface and all the preprocessing that is done on the articles. We report the results of user evaluations on predetermined news exploratory tasks that were done on the tool, comparing them to other news aggregating services.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

News Corpora Exploration, Interface design, Entity Extraction, News Summarization, Topic Detection

1. INTRODUCTION

News browsing is one of the primary uses of the internet. With the growth of the internet and it's users, there has been a virtual explosion of the amount of information that is available in a variety of domains. However, it has become increasingly difficult to streamline and organize the search trajectory to find exactly what one is looking for. Traditionally, search[1] has been with success used to filter out web documents that are relevant to a query. Multiple parameters are considered while considering the relevance of a document to a given query. However, this approach doesn't give satisfactory results for news documents. There are various unique features that make a news document more difficult to rank against a given query. Some of these features are

- A timestamp when the news article was published
- A set of real-world events that are talked about in the article
- A set of articles that were published in the past that this article references & builds upon
- A set of entities that are described

It is unclear as to which feature must be assigned higher weight. For eg. if a user queries ”News on Bill Clinton”, is the user is looking for most important events that Bill Clinton featured in (US Elections 1992, 1996, World Politics) or some more recent but less important news that involve him? Also, if the articles lacked an underlying structure, news from US Elections would appear interleaved Bill Clinton's philanthropic endeavours. Instead, a user would want to be able to filter out the topics that he is interested in. A basic requirement is to display news on a timeline which to reflect actual world happenings through the news articles covering them.

In addition, there are multiple news sources for an event and any news story develops over a period of time. Many times, a single article is not enough to give a reader the complete context of the underlying happenings in the real world. For a single article, there is a need to summarize it to give a sense of the ideas being talked about. For a set of articles, one needs to connect them according to the entities that occur so that a user only browses through the relevant articles.

2. MOTIVATION

The primary problem that we are trying to tackle is the following. Given an news information need, what set of articles should be retrieved and how should they be presented,

[1] using notable search engines like Google

so that the user through an iterative process, can navigate through them, building more context about the real-world event. For this we note that the two primary attributes of any news article are the entities (person, organization, place, etc) that appear in them, and the topics that are talked about in the article. Hence, we have designed an interface which augments news articles with the additional information of the relevant entities and the topics. In addition, these articles are arranged in time, so that users can easily navigate to a particular point in time, and look at how various articles are linked together.

3. DATA PREPROCESSING

3.1 News Corpus

Our news articles were taken from the New York Times 20-years archive dataset. We have focussed on the year 2000, since it was an election year, and we wanted to study how a complex real-world event like an election can be made simple to interpret with our interface. Articles related to the elections were primarily found in the U.S. section of the daily. We have also studied the articles of the Sports section. In addition to the dataset, we also crawled the NY Times website to get the online web editions of these articles since they contained rich annotations like keywords & entities.

3.2 Entity Extraction

The collected articles were then queried to OpenCalais[2], a web tool developed by Reuters. The tool accepts documents which are upto 100,000 characters in length as POST requests, and replies back with all the entities that it found in the document. These are grouped under an entity classes like *person*, *organization*, etc. There is a fixed set of entity classes that OpenCalais can recoganize. Along with each entity tagged, OpenCalais also assigns a score to reflect the importance of the entity in the article. We use this score to threshold and take five most important entities. The response from OpenCalais for all these articles were stored as entities for the file.

3.3 Topic Detection

Along with the entities for an article, we also need the list of topics that were talked about in the article. Topic Detection & Tracking (TDT) is a well studied problem[6, 1], and a lot of past work has focussed on news event detection over time[7, 5]. In the past, standard document labelling techniques of PLSA[4] and LDA[2] have been applied with success, and we resorted to do the same. Hence, we divided our news corpus into slabs of 20 days, and ran LDA over these articles. As a result, we get a mixture of topics and for every article, a probability of associating that article to a topic. We assigned each article the topic label that it associated the most with. Using, we associate a chain of topics to articles appearing successively in time.

3.4 News Summarization

Looking at entities and topics related to an article gives a good idea of what the article may be about. However, to be able to draw out more context, a user would like to look at the article itself or atleast a good summary. For this, we incorporated tested out various document summarizing tools. We finally decided to Text Compactor [3] which is based on Open Text Summarizer[4]. We sent all our documents as POST requests to this tool and specified the degree of summarization desired.

3.5 Transformations among Entities

Our work implements and builds upon the work by Choudhary et al.[3] which formalizes the notion of presenting news articles as a directed graph. Given a set of articles, one can construct a graph in the following way

- Assign a node for every article labelled by the various entities[5] tagged in the article
- Model interactions between articles in time as transformations of the entities appearing in them. These transformations can be *Create*, *Continue*, *Cease*, *Merge*, *Split*
- Score transformations using metrics defined
- Use the transformations to draw edges between nodes if important transformations occur amongst entities common to corresponding articles. For eg. a CONTINUE(A) transformation for an entity A between 2 nodes causes the 2 nodes to be joined by an edge

Using the above idea, we can mine important relations among actors over time in the form of transformations happening among them. Moreover, the strength of the transformation gives a sense of the strength of the interaction. The score is presented on the interface by a varying color code.

–MORE ABOUT TRANSFORMATIONS CAN BE SAID HERE–

4. SYSTEM DESCRIPTION

Our interface is built on HTML using Javascript and PHP on the server-side. All data preprocessing steps were carried offline and a database was built to store the results. We query the database to obtain the relevant information. –A SYSTEM DIAGRAM–

5. USER EVALUATION

We have designed a set of common news retrieval tasks which we show can be done with greater ease using our interface. A set of users, both domain experts and novices were asked to solve these tasks using our tool and we report on the results.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] J. Allan. Introduction to topic detection and tracking. In J. Allan and W. B. Croft, editors, *Topic Detection and Tracking*, volume 12 of *The Information Retrieval Series*, pages 1–16. Springer US, 2002.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] http://opencalais.com

[3] http://textcompactor.com/

[4] http:/libots.sourceforge.net/

[5] The paper uses the term *actor* for an entity

[3] R. Choudhary, S. Mehta, A. Bagchi, and R. Balakrishnan. A framework for exploration of news corpora by actor evolution and interaction, 2007.

[4] T. Hofmann. Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAIâĂŹ99*, pages 289–296, 1999.

[5] D. Kim and A. Oh. Topic chains for understanding a news corpus. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 163–176, Berlin, Heidelberg, 2011. Springer-Verlag.

[6] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple semantics in topic detection and tracking. *Information Retrieval*, 7:347–368, 2004. 10.1023/B:INRT.0000011210.12953.86.

[7] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 198–207, New York, NY, USA, 2005. ACM.