

StatNews

A platform for statistical analysis of news media

by Hisham Zarka
Advisor: Laurent El-Ghaoui

Introduction

The StatNews project seeks to build a tool that can help answer questions about media through the use of quantitative statistical techniques. The tool is intended to be consumed primarily by social scientists investigating word associations and trends in the media.

Social scientists are often interested in asking questions about the associations of certain words in the media, or how different media sources compare to one another in their treatment of topics. The treatment of the word 'iraq', for example, has interesting dimensions both across time and source. Sociologists may be interested in comparing the treatment of 'iraq' in 'conservative' sources such as Fox News against its treatment in 'liberal' sources. Social scientists may also be interested in examining the dimension of time, and seeing how the treatment of 'iraq' has differed during the contemporary invasion from its treatment during the Gulf War.

Current techniques used by social scientists often lack quantitative rigor. A common approach involves introspecting many articles from a time period of interest and qualitatively determining the associations held in the news media. More robust approaches attempt to quantify the number of occurrences of a word in a time period or the words that cooccur most with a word of interest. Both of these approaches scale poorly and fail to provide an interesting level of quantitative depth. Cooccurrence analysis, which identifies the words which most often occur alongside a word of interest, tends to pick out words which occur frequently in the dataset as a whole that are usually not of interest.

We are interested in developing and implementing analysis techniques which are:

- worthwhile
- easily interpretable
- statistically sound

Our hope is that the tool can be used by those interested in studying the media, particularly with academic and research settings.

StatNews Platform

The particular work done as part of the VIGRE program has primarily focused on building the framework within which we hope to implement the above tool. There are many aspects of such a tool that need to be in place before even the simplest of analyses can be run. These components include scraping and indexing the data, and we currently operate a small-scale scraper that collects data from sources of interest, both from blogs and the traditional news media.

The analysis techniques present thus far as part of our tool include occurrence queries (which calculate the number of occurrences of a word or phrase in a corpus of interest), cooccurrence queries (which calculate the words which most often cooccur with a word or phrase of interest), as well as analyses based on Bayes' and TFIDF techniques.

Because we are interested also in the ways in which results vary with time, each of the analyses supports a rolling time horizon mode of operation in which the query is repeated for successive intervals in a specified date range.

The current analyses are described below. It is our hope to introduce more rigorous analyses as the platform matures.

Occurrence Analysis

Occurrence is the simplest type of analysis. It simply counts the number of occurrences of a word in a given corpus. This type of analysis is useful in e.g. tracking the spread of an idea or concept across media sources and, with the time-varying element, across time.

Cooccurrence Analysis

Cooccurrence analyses find the words which most often cooccur with a word of interest. For example, a cooccurrence analysis of the word 'gay' might show that it most commonly appears next to the word 'marriage' or 'rights'.

Bayes' Analysis

This type of analysis uses a naive Bayesian classifier to help gauge the extent to which certain words positively or negatively predict the presence of a word of interest. Our technique is not a traditional classification setting, and our application of Bayes is similarly non-traditional.

In particular we take from the Bayes' classifier the measure of significance, a log-likelihood ratio which, which generally is a real number indicating the extent to which a particular word positively or negatively predicts the word of interest. The result of this analysis is then the words which are most predictive of the query word in the data set.

Future Work

One of the primary goals of the platform was to make it easy to extend with additional types and methods of analysis. New statistical analyses should be easy to integrate into the system and have available the infrastructure and gigabytes of media data which we have collected thus far.

With the infrastructure completed, we hope to continue working to implement new types of analyses on the system and refining the algorithms that are currently implemented.

The most rigorous statistical analysis currently implemented is based on a Bayes' classifier, but we would like to implement additional classification algorithms such as Support Vector Machines. Additionally, we would like to improve the entity and phrase recognition capabilities of the system such that it is more capable of tracking interesting content.