

Blog Analysis - Trends and Predictions

Srinivasan Ramaswamy

srini@ischool.berkeley.edu

Applied Natural Language Processing Project Report

Abstract

Weblogs are becoming the grassroots publishing medium of the current electronic publishing era. This project attempts to explore the abundant information available in the form of blogs and try to apply various Natural Language Processing algorithms to find out interesting interpretations of the available information. It aims at predicting the trends and other inherent latent information present in the blogs with various NLP techniques. It has shown surprisingly interesting results with the sample data and hence it has great potential to predict and present various interesting information.

1. Introduction

The advent of World Wide Web has made information sharing easier and increased availability of information. In spite of the simplicity of creation of web pages using HTML, it remained as a skill, which requires some technical knowledge. But the birth of Blogs brought a digital revolution to information and publishing content. The exponential growth of blogs confirm this fact.

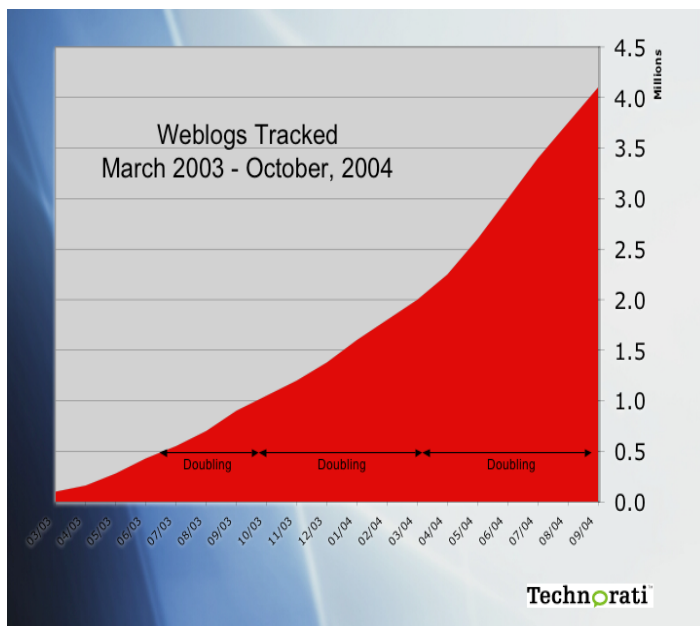
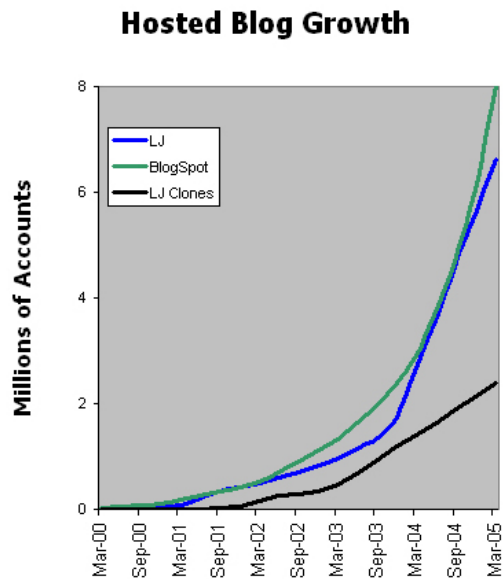


Figure 1. Blog Growth



According to Technorati[1], a major blog search engine, it currently indexes 60 million blogs. And there could be many blogs which Technorati is not indexing. As quoted in the Technorati web site “ There are 50 million blogs, some of them should be good”.

In general weblog is defined as a webpage with a set of dated entries, in reverse chronological order, using a web publishing software tool. As the number of blogs started increasing it became difficult for users to keep track of regular updates in blogs. The RSS feeds and the feed readers have completely changed this scenario, and in turn increased the participation in the blogosphere.

The Perseus[2] survey on some of blogs on twenty leading blog hosting services reveal some interesting facts and figures about the growth rate of blogs. They have also presented an analysis of population distribution figures for gender and age in the representative sample of hosted blogs. We could clearly notice that the blog counts are increasing exponentially in the recent years.

These information can be used for Market analysis and research. Some of thee pioneers in e-commerce Amazon, Netflix, etc are using automatic recommendation system, by deriving the recommendation from the customers shopping history. Research has shown that analysis of blogs can predict more interesting results as people write their personal reviews and opinions of shopping, movies, books, etc.

Some of the most popular weblog search engines like Technorati and Blogpulse present latest trends , by a automated trend analysis on the blogosphere. Though not restricted to blogs, Google Trends which presents a comparison of trends between search key terms is also shows the increased research in trend analysis.

3. Corpus Preparation

The initial step in corpus preparation is a harvest of active blogs from the Internet. As corpus preparation is a very important step in NLP, a lot of effort has been given to this. A analysis can present useful results only when its done on a large and diverse collection of data . Hence a large corpus in necessary for even a simple analysis. The most popular way of collecting blog corpus is indexing all the blogs. But in the current scenario there are 60 million blogs present in the Internet and indexing the whole web is a very complicated task. And there are numerous blog domains which hosts blogs like Blogspot, LiveJournal, Xanga, MSN Journals, etc. Collecting enormous data demands very high system requirements. Hence this project deals only with certain kind of blog directories which offers feeds for their recent updated blogs.

3.1 Corpus Aggregator

In the beginning, attempts had been made to crawl and index certain blog domains using the web crawling software “Nutch”. But due to heavy requirement of processing power, memory and time for a indexer, later a different approach of collecting the corpus from RSS feeds is followed. It is designed to work on blog directories such as Radioblog [3], Salon [4], etc which constantly publishes the latest updated blogs hosted on their domain. The aggregator reads the directory and get the list of blogs updated recently, and later it extracts the content from the XML feed of that particular blog. It can also collect data directly from feeds. This feed

aggregator uses Universal Feed Parser [5] to parse the feeds in different format. Its a python module to parse syndicated feeds. It can handle RSS 0.90, Netscape RSS 0.91, Userland RSS 0.91, RSS 0.92, RSS 0.93, RSS 0.94, RSS 1.0, RSS 2.0, Atom 0.3, Atom 1.0, and CDF feeds. It also parses several popular extension modules, including Dublin Core and Apple's iTunes extensions.

3.2 Data Format

Though the whole content of the blog is parsed we are interested only in certain information. This project considers only the title and description and the date of each blog post for its analysis. They are collected in a XML format, as its easy to process later. The corpus is collected based on date and hence a separate file exist for each and every date.

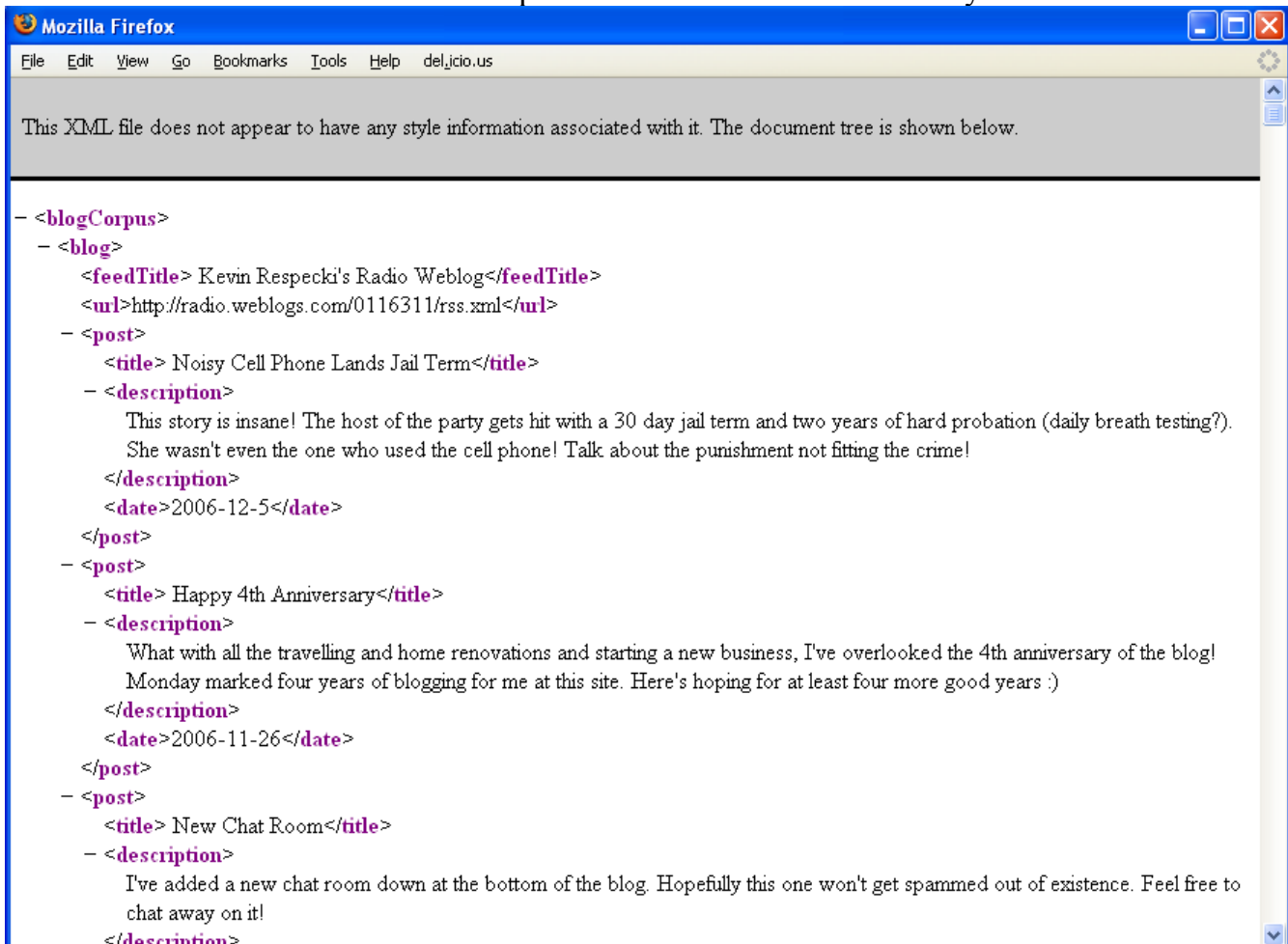


Figure 2. Sample Corpus

The feed aggregator is ran for a few days and it collected data from around 500 blogs. It garnered data of volume 8MB from the blogs along with the timestamp. The data format is designed in such a way that any kind of text analysis can be made form this corpus. This corpus can be improved with some annotated information, but is considered for the future improvements.

4. Blog Analysis

Once we have a large amount of data we can apply numerous NLP techniques to discover the patterns and trends present in the data. The interesting fact is that automated trend mining reveals lot of hidden patterns, which is not evidently known.

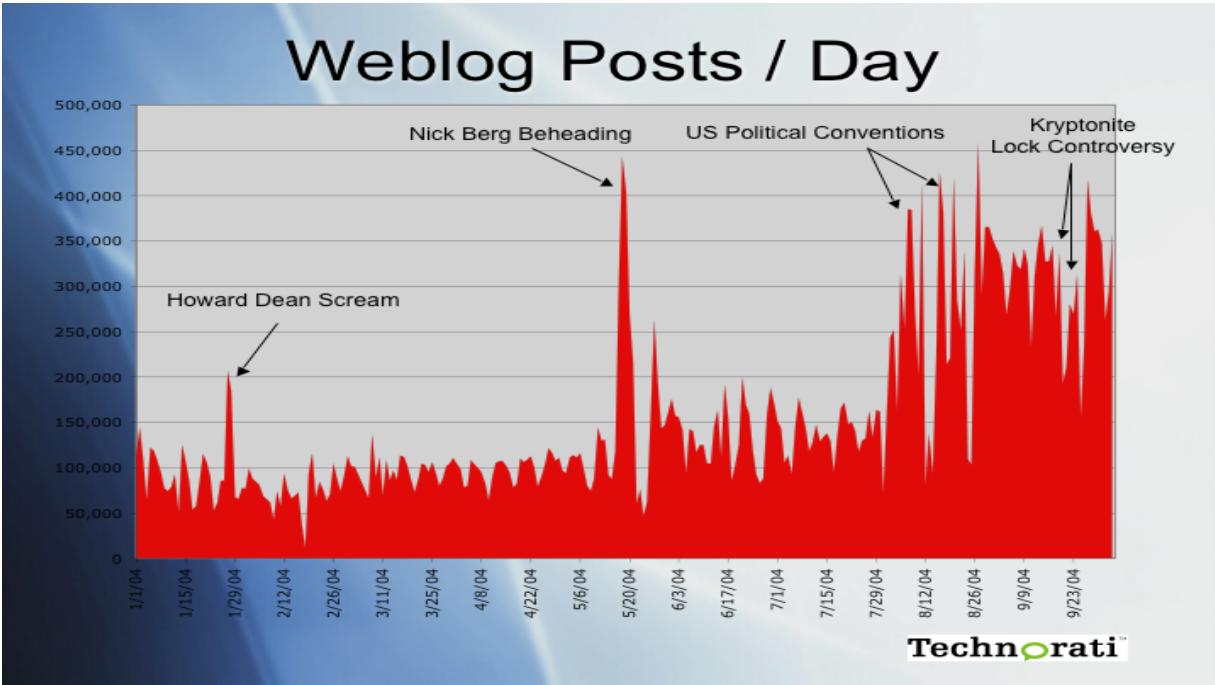


Figure 3. Blog Analysis predictions

4.1 Trend Analysis

As we have a bunch of data we can search for keywords and determine the distribution of those words through out the corpus. When a keyword is given it is searched in the whole corpus for the distribution over time with respect to the date and a XML output data is created for further processing. Currently the period has been defined as months, but the design is more scalable that it can be easily modified so that the period can be obtained as an input.

The output XML file is passed to a trend plotter and it plots the trends of the given keywords over the period of analysis. The trend plotter can produce updated graphs whenever the output XML changes. To have an automated trend plotter, the project implemented a separate chart module which made use of a library Cewolf [6] (which is based on JFreeChart). This is implemented as a servlet, as it can be hosted in the Internet and it can produce updated charts as the corpus gets updated.

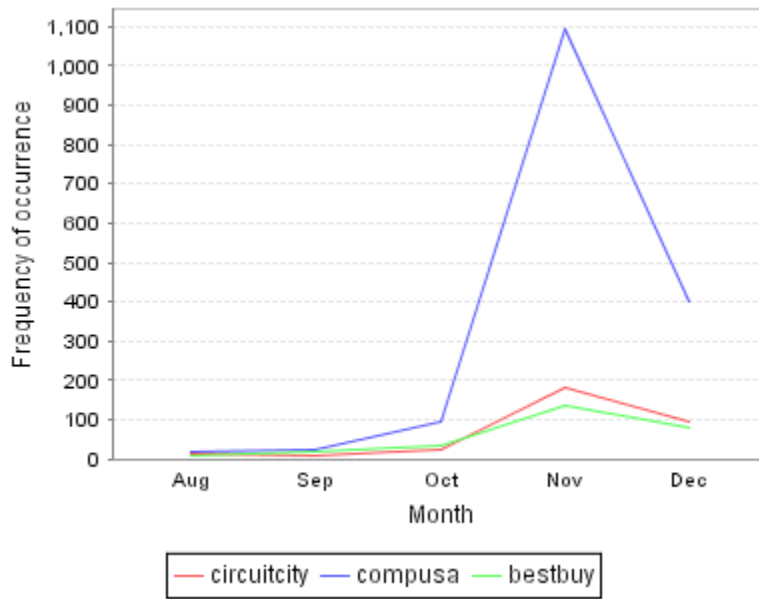


Figure 4. Trend Analysis of famous shops

Even this simple analysis reveals lot of information. For example, when a trend search is done on the sample corpus, over the keywords “thanksgiving” and “sales”, the sales was moderate before thanksgiving and during November the sales reached a high peak, due to the thanksgiving sales, and it gradually decreased after that.

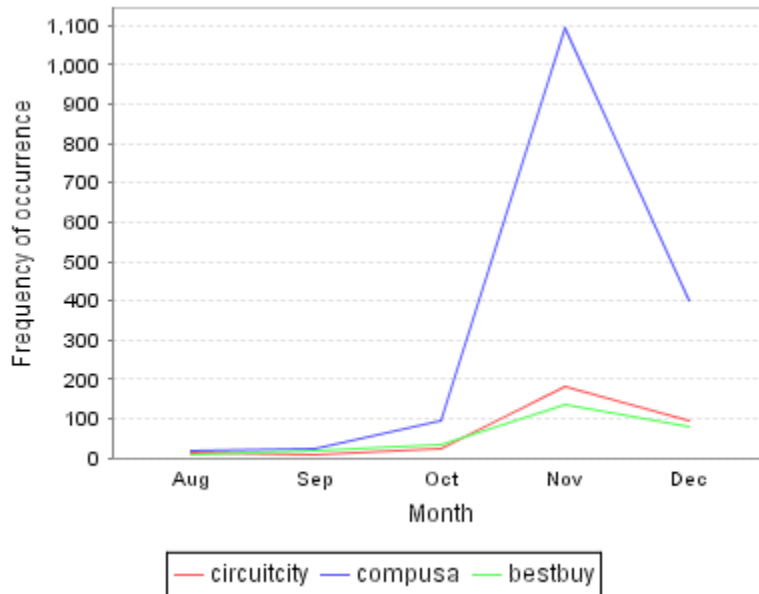


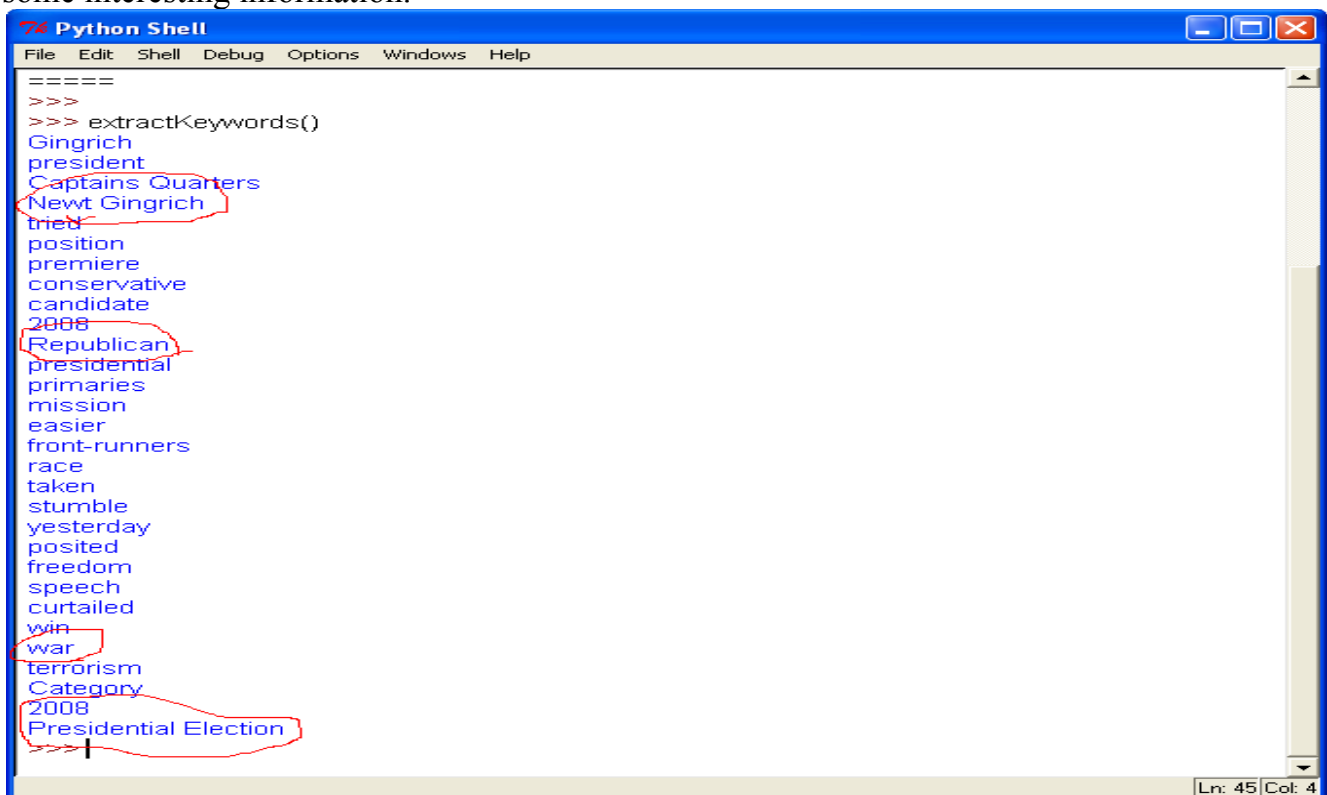
Figure 5. Trend Analysis of Thanksgiving sales

Looking at another search over the keywords “circuitcity”, “compusa” and “bestbuy” it could be predicted that there was some sales season during November as all the three shops climbs up in the plot. But circuit city achieved a peak than the other two shops. As the corpus is

limited it cannot be trusted entirely. But given a diverse and huge corpus we can predict better trends. When we predict the trends over a deeper period like “over the days”, it would reveal finer details, such as the days when the sale really attained a maximum and the days over which the sale was low.

4.2 Keyword Extraction

Blogs represent diverse topics and contains lot of textual information and hence a keyword extraction would reveal some facts about the data we are analyzing. Currently weblogs are becoming the '*sine qua non*' of political movements. During the 2004 US Elections, blogs was one of the major medium of campaign. In this keyword extraction most of the frequently occurring stop words are eliminated and the other keywords are selected based on their frequency distribution. A keyword extraction done on a subset of the corpus revealed some interesting information.



```
Python Shell
File Edit Shell Debug Options Windows Help
=====
>>>
>>> extractKeywords()
Gingrich
president
Captains Quarters
Newt Gingrich
tried
position
premiere
conservative
candidate
2008
Republican
presidential
primaries
mission
easier
front-runners
race
taken
stumble
yesterday
posited
freedom
speech
curtailed
win
war
terrorism
Category
2008
Presidential Election
>>>
```

Figure 6. Keyword Analysis of Blogs

The circled keywords are good reflections of what people are talking about in main stream politics. The keywords “war” and “terrorism” are widely discussed in current scenario politics. The keywords “2008” and “Presidential Election” occurred together which revealed to me that there is a presidential election coming up in 2008. As this naïve extractor reveals such useful information, we can generate various kinds of information with a large corpus and sophisticated algorithm.

4.3 Phrase Extraction

Another interesting technique in data mining is phrase extraction. As blogs represent the opinion, reviews, ideas and voice of people extracting phrases would provide further insights into the current trends. The phrase extraction has been implemented to identify key prepositional and other types of phrases. And moreover phrases reveal more information about a particular topic as it includes a wider context.

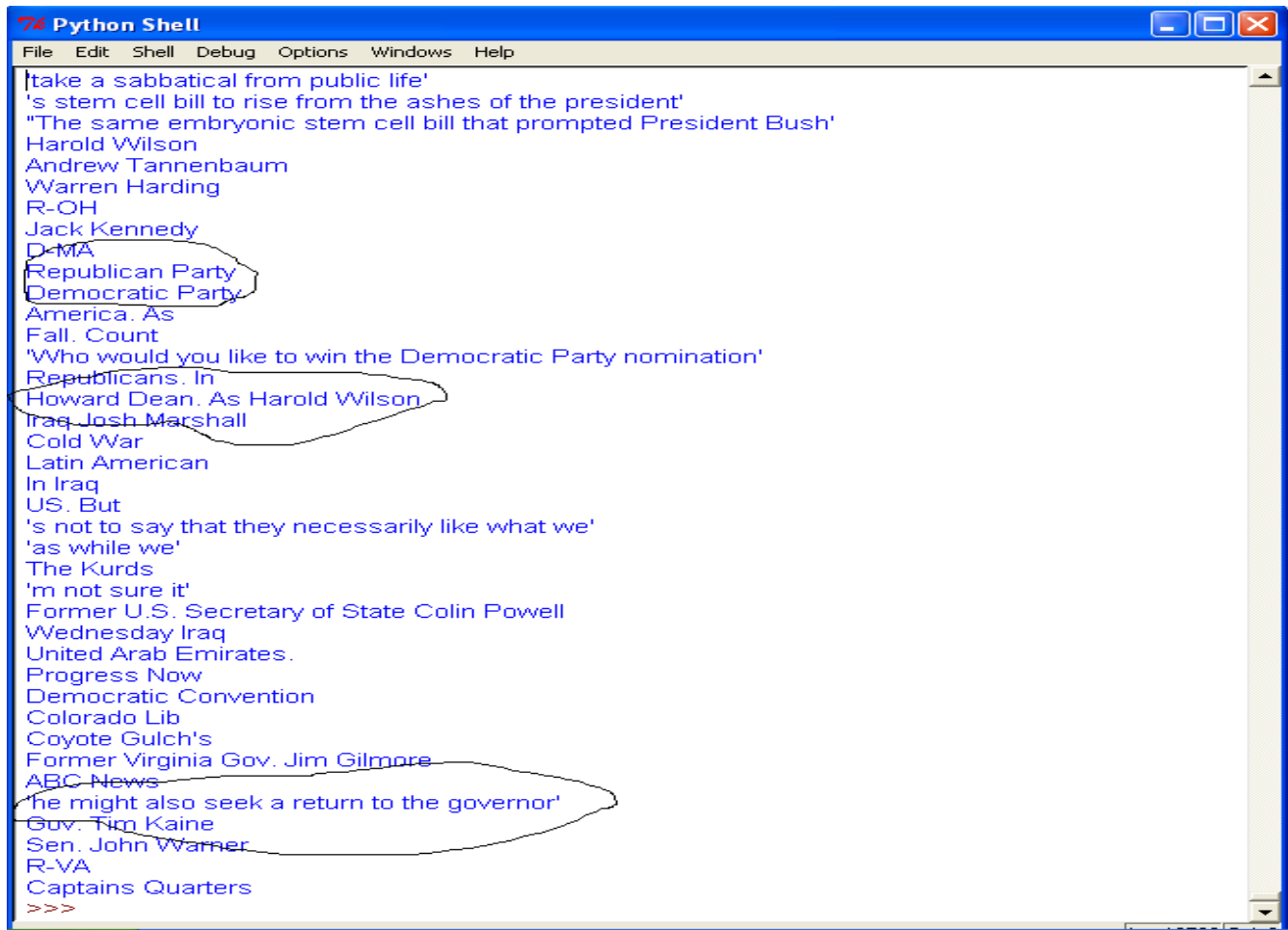


Figure 7. Key Phrase Analysis in blogs

In the above example “Republican Party” , “Democratic Party” shows the two major parties and the topics discussed around it. The phrase “he might also return to the governor” also reveals some info about Gilmore contesting for governor.

Phrase extraction is a very interesting topic and it can be fine tuned by giving more weights to the phraseness and informativeness to each phrases and selecting the phrases with high score for them. This in itself is a research topic, which has seen a good improvement in the recent past.

4.4 Named Entity Recognition

After analyzing for keywords and phrases, it would be interesting to do a Named Entity Recognition. NER is a suitable technique to extract information from reviews as it extract key person, location and entity. To perform this experiment a NER tool developed by Stanford Natural Language Processing Group [7] is used. It provides us option to select from a choice of classifiers to be used. A sample data is experimented with CRF classifier and the results are good. In the experiment shown below “Heather .A. Wilson” and “New Mexico” are Names and entities recognized in one sentence and we know that he is a republican candidate who won the elections in New Mexico. So it makes a good attempt in finding the location and thiewr key places and person. More inferences can be made if we could combine these phrase extraction and NER extraction in a specific and efficient way.

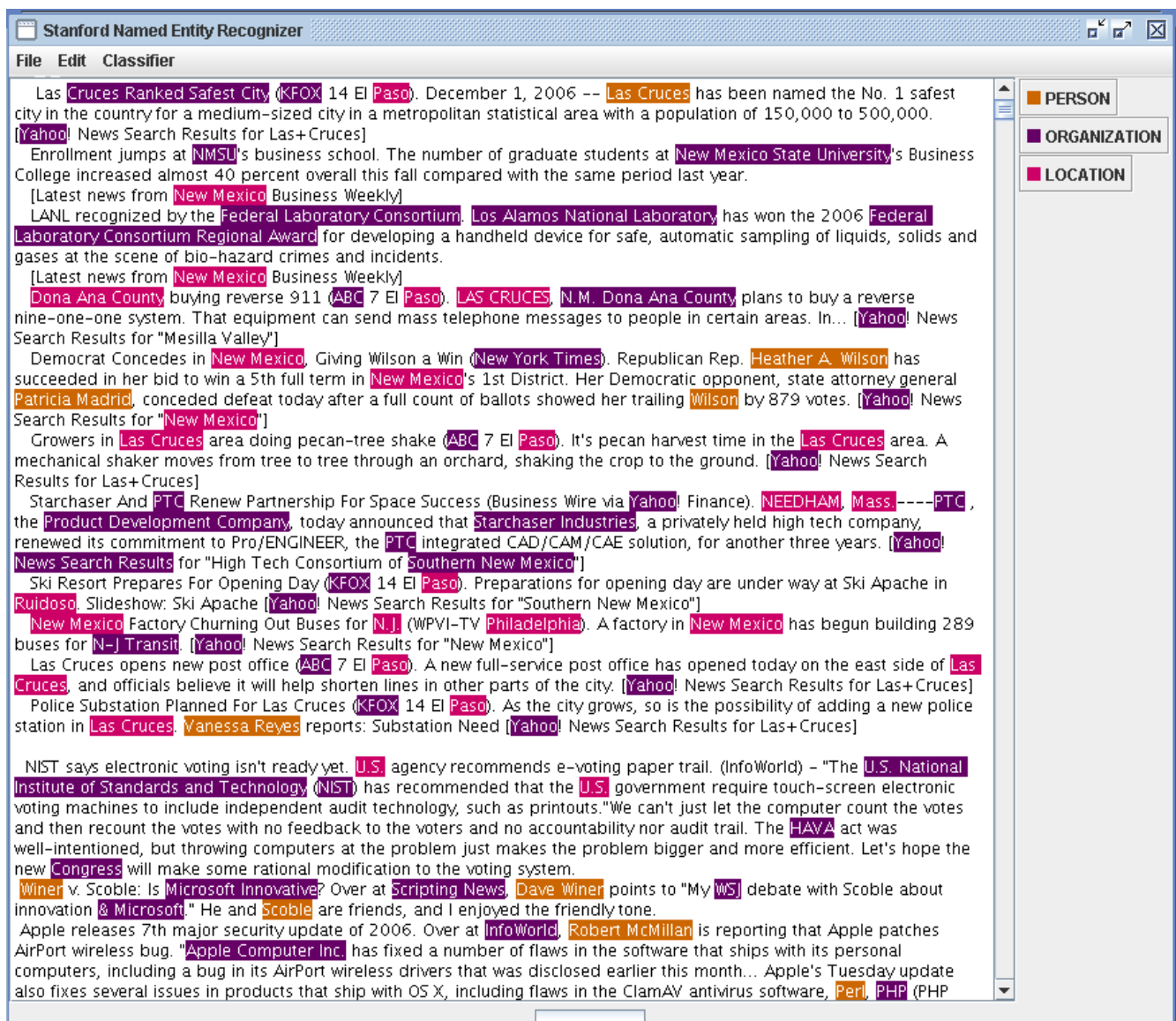


Figure 8. Named Entity Recognition Analysis in Blogs

5. Future Work

This project has a multidimensional growth for the future. There are plans for improving each and every phase of the project. I would like to improve the blog aggregator so that it could collect vast amount of data from diverse sources and store them in a compressed format. Further it should be made as a automated aggregator which runs the aggregator automatically over a constant period of time to collect data. And some additional information can be tagged if we are interested in any other extra information.

All the algorithms implemented should be improved to get accurate results. The Trend Analysis can be combined with clustering/categorization to explore whether it reveals any additional information. Further interesting analysis which could be done are cross analysis of Profile vs posts of an individual user and prediction of near future with trend analysis and verifying it with the predictions. Moreover if other invasions of blogosphere such as Moblog, Photoblogging are included then it would add up to the diversity of information and would represent a collective sample of social network.

6. Conclusion

Weblogs are fascinating domain for data mining and NLP, which offers extremely wide opportunity for research. Already lot of research is in progress in this field and some of the leaders in industry like Technorati, Blogpulse, are doing excellent trend analysis and predictions which shows us the visualization of the latest trends , topics, discussion and opinion of people around the world. There should be more research done on the way it could be used. And scaling the technology upto the pace in which the blogosphere is growing is highly challenging. Hence its an area with lot of research questions to be answered.

7. References

- [1] Technorati, <http://www.technorati.com/>
- [2] Perseus Web surveyor, <http://www.perseus.com/blogsurvey/geyser.html>
- [3] Radio Userland Community, <http://radio.weblogs.com/>
- [4] Salon Radio Community , <http://blogs.salon.com/>
- [5] Universal Feed Parser, <http://feedparser.org/docs/introduction.html>
- [6] Cewolf, <http://cewolf.sourceforge.net/new/index.html>
- [7] Stanford natural Language Processing Group, <http://nlp.stanford.edu/software/CRF-NER.shtml>
- [8] Wikipedia, www.wikipedia.org
- [9] Natalie S. Glance, Matthew Hurst and Takashi Tomokiyo, BlogPulse: Automated Trend Discovery for Weblogs
- [10] Ian Fischer, Elias Torres, Harvard University, A Distributed Blog Search Platform
- [11] Tomoyuki Nanno, Toshiaki Fujiki, Yasuhiro Suzuki, Manabu Okumura
Automatically Collecting, Monitoring, and Mining Japanese Weblogs
- [12] Gilad Mishne and Maarten de Rijke, MoodViews: Tools for Blog Mood Analysis