System for Spatio-Temporal Analysis of Online News and Blogs

Angelo Dalli
University of Malta

angelo.dalli@um.edu.mt

ABSTRACT

Previous work on spatio-temporal analysis of news items and other documents has largely focused on broad categorization of small text collections by region or country. A system for large-scale spatio-temporal analysis of online news media and blogs is presented, together with an analysis of global news media coverage over a nine year period. We demonstrate the benefits of using a hierarchical geospatial database to disambiguate between geographical named entities, and provide results for an extremely fine-grained analysis of news items. Aggregate maps of media attention for particular places around the world are compared with geographical and socio-economic data. Our analysis suggests that GDP per capita is the best indicator for media attention.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; J.4 [**Social and Behavioral Sciences**]: Economics and Sociology;

General Terms

Algorithms, Theory, Performance, Design, Economics

Keywords

Geolocation, disambiguation of geographical named entities, media attention, news, blogs, social behavior, spatio-temporal

1. INTRODUCTION

Online news and, more recently, blogs, are increasingly becoming one of the most popular destinations for Internet users, slowly increasing their influence to levels approaching those of traditional media [1]. Media attention and popular attention shifts continuously as new events happen in the world. Generally, media attention influences popular attention, although the reverse also occurs to a lesser degree. Attention in this context can be conveniently defined as the number of documents on a given subject, which is the same definition used by Zuckerman in his seminal paper on Media Attention profiles [5].

Existing news and blogs classification systems such as Google News and Blog Pulse usually focus on topic classification or keyword frequency tracking over time [6,7]. Previous work on georeferencing texts with a geospatial aware NER system has addressed the issues of spatial grounding of geographical entities and geographical name disambiguation [9,10,11,12,13] utilizing input from text analysis or Internet IP and DNS data [14]. However, almost none of these systems have comprehensively accounted for the temporal aspect associated with geographical named entities. This work (partly supported by EPSRC grant EP/C536762/1 and a small grant from Linguamine) presents a system, cpGeo, which analyses mentions of different geographical locations over time in news texts and blogs, creating real-time maps of shifting attention profiles that are convenient for highlighting current hot spots and determining what places capture the most attention in the world over time. Our interest is to find a set of indicators that are indicative of the level of attention enjoyed by different places in the world, and hence, the people living at those places.

2. ANALYSIS SYSTEM

The spatio-temporal analysis system, called cpGeo, is made up of five components (a high-performance web crawler, distributed storage, knowledge extraction, geospatial processor and data mining system) that allow download and analyze millions of items in a highly scaleable manner and generate summary reports.

cpGeo currently downloads between 18,000 to 21,000 news items every day from around 6,000 sources. Our level of coverage is rapidly approaching 100% of all online news published on the Internet everyday. In order to have adequate coverage of news from earlier years, we supplemented the news through the LDC English Gigaword corpus [14]. Blog entries are also being downloaded at a rate of around 156,000 blog items a day from around 90,000 authors. The cpGeo knowledge extraction subsystem performs various tasks related to basic text document processing and knowledge extraction. Documents are indexed and processed through a custom-built multiple document summarisation system that automatically detects document duplicates and merges highly similar documents together. Named entities and related events are identified and extracted to a temporal database. The geospatial processor identifies and disambiguates references to geographical locations around the world, and can produce graphical GIS-like presentations of its output results. The clustering and data mining system, utilizes a multivariate cluster analysis technique with an integrated rule learning model [18,19,20] and a small database of world knowledge that is used to interpret results correctly.

3. GEOSPATIAL PROCESSING

The cpGeo geospatial processor has three main components namely, the Multilingual Geospatial Database, a Multilingual Geographical Named Entity Recognizer (GNER) and a Disambiguation Module. The database has entries in 139 different languages and 3 main hierarchical levels covering 251 countries, 4,815 administrative regions and 7,574,966 individual place names and features. The database allows us to take into account the element of time and the fact that place names sometimes change over time.

Figure 1 shows the spatial database coverage, with shaded regions representing recognized geographical locations in the world. Some regions have lower coverage density (this is for example apparent for India, which has high population density but lower coverage in the spatial database).

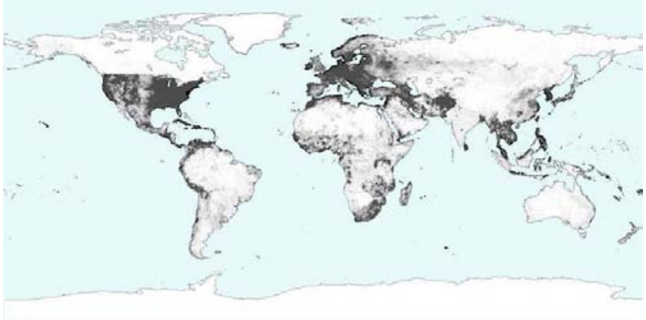

Figure 1. Spatial Database Coverage.

The geospatial database also contains additional information such as the latitude and longitude using the WGS84 geodetic reference system, feature type (populated place, street, etc.), relative importance, and aliases. The GNER uses the feature type information to determine the reliability of entries in the geospatial database, making it possible to identify, for example that "Lascaux", "Lascaux Cave", "Cave of Lascaux", and "La Grotte de Lascaux" refer to the same location. The GNER also has a geographic anaphora resolver, enabling it to know, for example, that "Bay Area" is referring to "San Francisco Bay Area".

Surprisingly, there are many duplicated place names in the world (around 10% to 25% of all place names have some duplicate elsewhere, depending upon the region or country). The GNER uses a mixture of heuristics and statistics to successfully disambiguate between duplicates. Geographical proximity and relative importance of other place names mentioned in the same context are considered in the disambiguation process.

The GNER also uses the knowledge extraction system to determine whether ambiguous names should be classified as person names or place names (e.g. to determine whether "Washington" is referring to the city, state or surname), enabling it to successfully resolve ambiguities in over 98% of cases. cpGeo achieved an F-measure of 0.9965 compared to 0.904 for the Perseus system [13].

4. EVALUATION AND RESULTS

We have evaluated the cpGeo system on our main news items database spanning from 1994 to 2005 (a total of 4,197 days). On average, every day had mentions of around 500 unique location names with 16,500 mentions of geographical named entities. Figure 2 shows the output of the system for 1 January 2000. Generally, when viewed on a global scale, the map changes slowly, although spikes and changes occur rapidly on local scales. The cpGeo system also keeps aggregate statistics of all place names mentioned together with their frequency, thus building up a map that indicates the regions in the world that are receiving the most media attention (as shown in Figure 3). In the United States it is apparent that North-East states receive more attention than other states, with the exception of California. In Europe, the UK and Belgium also receive more attention, while in Asia, Japan gets mentioned most frequently (with China catching up).

The aggregate maps can be useful in predicting the background level of attention that a particular region usually receives, providing better means of identifying spikes and anomalies instead of using simple threshold or rate increase methods. Aggregate maps represent a probability density function for the amount of news coverage likely to occur at any particular location in the world.

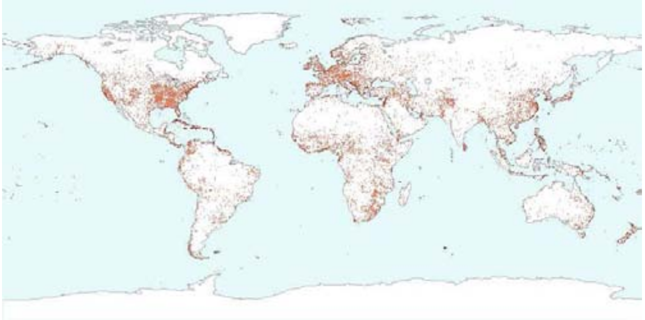

Figure 2. System Output for 1 January 2000.

Our results show that the top 80 mentioned place names consistently dominate the daily global news, generating more than 50% of all mentions on average. The top 3 daily place names generate around 11% of all mentions in the news.

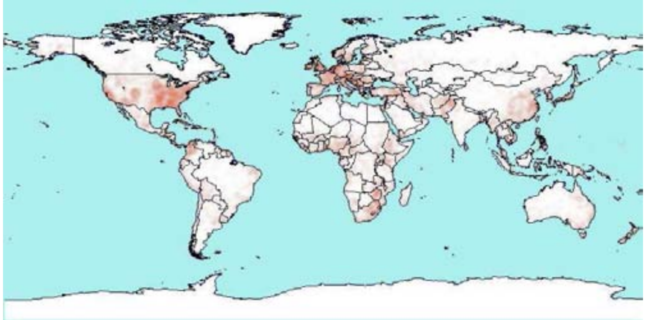

Figure 3. World Average News Media Attention.

We have also evaluated the cpGeo system on a small geographical scale using a four year collection of news about the smallest EU member state of Malta. Based on this evaluation we have determined that the cpGeo system can produce accurate results at a global resolution of around 3m x 3m.

Various statistical indicators were examined in an attempt to find correlations between statistical indicators and the cpGeo media attention ratings, with Number of Unresolved International Disputes, GDP Per Capita, and Number of Internet Users being the top three indicators that correlate with media coverage with GDP per capita being the most significant indicator of media attention.

The clustering system also produced 26 distinct clusters of countries based on these indicators, showing that for poorer countries, the secondary determining indicator for media attention is their number of disputes, while for richer countries the secondary determining factor is the number of Internet users. Thus, poorer countries are often in the news whenever they are involved in some armed conflict or dispute, and are most likely to be portrayed in a negative fashion.

The cpGeo system results also show that certain countries are abnormally represented in the media and blogs. The top 5 most over-represented countries in the world (with respect to their population levels) are the Holy See (Vatican), Monaco, Liechtenstein, Iceland and Luxembourg while the bottom 5 most under-represented countries are India, China, Brazil, Indonesia and Pakistan. There is a huge disparity between the top and bottom countries, for example, each Liechtenstein citizen gets the same average media coverage equivalent to 4,800 Pakistanis.