

# Mining News & Blogs Spatio-Temporal Data

Angelo Dalli, NLP Research Group  
University of Sheffield, United Kingdom

## STATISTICAL INDICATORS

The statistical indicators considered were:

- Population
- GDP Per Capita (Normalised to US Dollars)
- Military Expenditure as % of GDP
- Number of Unresolved International Disputes
- Industrial Production Growth Rate %
- Combined Male and Female Literacy Rates
- Area in Square Km
- Number of Internet Users

The following table shows the correlation between the indicators and media attention rating:

Statistical Indicator	Correlation
Population	0.157573683
GDP Per Capita	0.247307383
Military Expenditure (% GDP)	-0.030440339
International Disputes	0.317276794
Industrial Growth Rate %	0.038067369
Literacy	-0.009887567
Area	0.156561024
Internet Users	0.298783697

The data was filtered further by eliminating all countries or territories having a population numbering less than the Holy See (Vatican). This filtering process eliminated statistical skew that would be introduced by outliers such as the Pitcairn Islands or Johnston Atoll.

## CLASSIFICATION BANDS

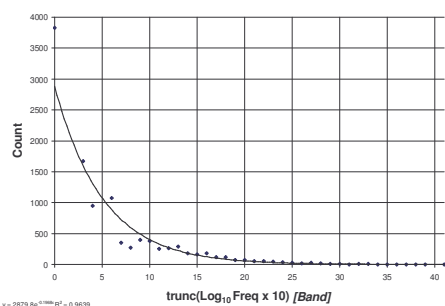
The clustering system first grouped indicators into different bands, representing high, medium and low value levels (apart from the GDP per capita which was split into four levels). The table below shows the classification bands used in the final iteration. The numbers next to each band show how the clustering system referred to each band (i.e. band 1, 2, 3, etc.).

Ind.	High 4	Med. High 3	Medium 2	Low 1
G	\$21,400 \$58,900	\$10,700 \$21,399	\$5,400 \$10,699	\$400 \$5,399
D	6-15	-	2-5	0-1
I	159 M 7.44 M	-	1.86 M 7.44 M	0 1.86 M

## GEOSPATIAL FREQUENCY DISTRIBUTIONS

An important part of our research was to analyse the frequency distribution of place names mentioned in the texts. Our evaluation results show that the top 80 mentioned place names on average, dominate the global news for that day, generating more than 50% of all mentions. The top 3 place names every day generate around 11% of all mentions in the news.

We split the frequency distribution into different frequency bands in order to find out more about the geographical named entity frequency distribution. We counted the number of entries falling into each logarithmic frequency band. The graph below shows a plot of the number of entries in each frequency band, which is almost perfectly fitted by an exponential curve.



The situation is slightly different for blog articles, with blogs having wider geographical coverage and less bias towards a few locations. In general the frequency distribution curve for blog articles has a shallower rising curve with a longer tail. The distribution for both news articles and blog entries is similar (almost identical), apart from the fact that blog entries seem to give less importance to one particular place and mention more obscure places than the mainstream media. This result is intuitive considering the more varied background of blog authors and the wider breadth of topic coverage in blogs as compared with mainstream news articles. Results from our blog database indicate that around 9% more names are mentioned in blog entries while the top 3 place names generate only 8% of all mentions, compared with 11% for news.

The frequency distribution was also compared over time to check for variances over time, but the distribution seems to be highly stable over time with exceptions happening during abnormal periods of time when spikes occur (e.g. on 11 September 2001).

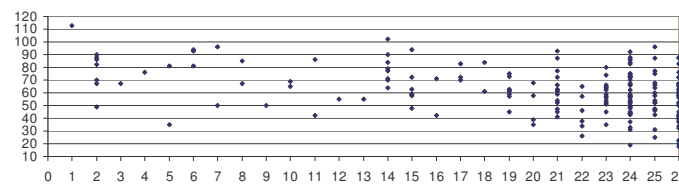
The spatio-temporal processing of news and blogs produces a wealth of information for new insights into news coverage and media attention profiles. We were interested in answering three questions: whether there is a particular pattern governing news coverage for a particular country; whether there are some statistical indicators that can be used to predict the likely amount of news coverage given to a particular country; and whether there are regions in the world that are over-represented or under-represented. The cpGeo clustering and data mining system proved invaluable in answering these questions.

The clustering and data mining system has a number of external knowledge bases, including statistical data about different countries in the world extracted from the CIA World Fact Book, which is assumed to be an authoritative source on global statistics. Although the cpGeo system can generate statistics at a very fine level of detail (approximately on a 3m by 3m level on a global scale), the statistics available are generally on a country basis.

**The cpGeo system assigns a rating for average media attention for any geographical region, which has been normalised on a scale from 0 to 100 (with 100 denoting the maximum attention possible received).**

The first step was to determine if there are significant correlations between the cpGeo media attention rating and existing statistical indicators (shown on the left). Only three indicators (**GDP Per Capita (G)**, **Number of Unresolved International Disputes (D)**, and **Number of Internet Users (I)**) have some kind of correlation with the media attention rating, and the rest of the indicators were ignored.

The second step was to produce rankings for each of the three indicators which we shall abbreviate here to GDP, Disputes and Internet. The clustering system was then used to split this ranked information into clusters using the classification bands shown on the left (which were automatically calculated). The classification bands were used to create 26 clusters of countries having similar characteristics in their indicators and showing how these characteristics relate to their media attention rating. A more coherent grouping now emerged relating the indicators to the media attention rating, with precisely defined minimum and maximum ratings according to the cluster number.



Scatter Plot of Media Attention Rating by Cluster Number

The following 26 clusters were obtained (the cluster number is arbitrary). The rating for GDP, Disputes and Internet Use (G/D/I) is shown next to each cluster number:

- Cluster 1 (4/3/3) – USA
- Cluster 2 (4/2/3) – UK, Japan, Spain, France, Australia, Canada, Taiwan
- Cluster 3 (2/3/3) – Malaysia
- Cluster 4 (3/2/3) – South Korea
- Cluster 5 (1/3/3) – Indonesia, Georgia
- Cluster 6 (4/1/3) – Germany, Netherlands, Italy
- Cluster 7 (2/2/3) – China, Thailand
- Cluster 8 (4/2/2) – Denmark, Singapore
- Cluster 9 (3/1/3) – Poland
- Cluster 10 (2/3/2) – Brazil, Russia
- Cluster 11 (3/2/2) – Israel, Argentina
- Cluster 12 (2/1/3) – Costa Rica
- Cluster 13 (1/3/2) – Equatorial Guinea
- Cluster 14 (4/1/2) – Belgium, Austria, Norway, Sweden, Hong Kong, Switzerland, New Zealand, Finland
- Cluster 15 (2/2/2) – Colombia, Turkey, Chile, Iran, Peru, Romania, Dominican Republic
- Cluster 16 (4/2/1) – United Arab Emirates, Brunei
- Cluster 17 (3/1/2) – South Africa, Czech Republic, Portugal
- Cluster 18 (1/2/2) – Philippines, Egypt
- Cluster 19 (3/2/1) – Latvia, Greece, Cyprus, Lithuania, Croatia, Hungary, Slovakia, Saudi Arabia
- Cluster 20 (1/3/1) – Papua New Guinea, Ethiopia, Cameroon, Guyana
- Cluster 21 (4/1/1) – Ireland, Luxembourg, Liechtenstein, Monaco, Gibraltar, Isle of Man, Andorra, Iceland, San Marino, Jersey, Bermuda, Qatar, Cayman Islands
- Cluster 22 (2/2/1) – Venezuela, Algeria, Namibia, Turkmenistan, Botswana, Kazakhstan
- Cluster 23 (3/1/1) – Slovenia, Puerto Rico, Bahamas, Estonia, Kuwait, Malta, Macau, Antigua and Barbuda, Mauritius, Uruguay, Oman, Barbados, Bahrain
- Cluster 24 (1/2/1) – Serbia and Montenegro, Nepal, Madagascar, West Bank, Syria, Lebanon, Gaza Strip, East Timor, Yemen, Bangladesh, Zambia, El Salvador, Nigeria, Azerbaijan, Benin, Angola, Cambodia, Sudan, Albania, Burma, Morocco, Solomon Islands, Tajikistan, Nicaragua, Suriname, Honduras, Chad, Democratic Republic of the Congo, Burundi, Guatemala, Niger, Mozambique, Djibouti, Armenia, India, Zimbabwe, Sierra Leone, Cote d'Ivoire, Liberia, Mauritania, Central African Republic, North Korea, Iraq
- Cluster 25 (2/1/1) – Bosnia and Herzegovina, Ukraine, Mexico, French Guiana, Panama, Belize, Tunisia, Fiji, Gabon, Trinidad and Tobago, American Samoa, Samoa, Bulgaria, Macedonia, Reunion, Libya, Belarus, Anguilla
- Cluster 26 (1/1/1) – Sri Lanka, Cuba, Mongolia, Somalia, Ghana, Pakistan, Jordan, Lesotho, Vietnam, Togo, Kyrgyzstan, Jamaica, Kenya, Malawi, Swaziland, Senegal, Ecuador, Burkina Faso, Maldives, Tanzania, Guinea, Mali, Rwanda, Uganda, Uzbekistan, Cook Islands, Guinea-Bissau, Moldova, Republic of the Congo, Eritrea, Bolivia, The Gambia, Afghanistan, Bhutan, Comoros, Cape Verde, Paraguay, Haiti, Laos

These clusters do correspond to intuitive groupings, especially when seen in the light of the media attention received by these countries. For example, cluster 24 groups countries that receive attention mainly through disputes. Cluster 2 countries that receive attention because of their good performing economies with the occasional dispute. In terms of media attention, the most under-represented countries are those in cluster 26.

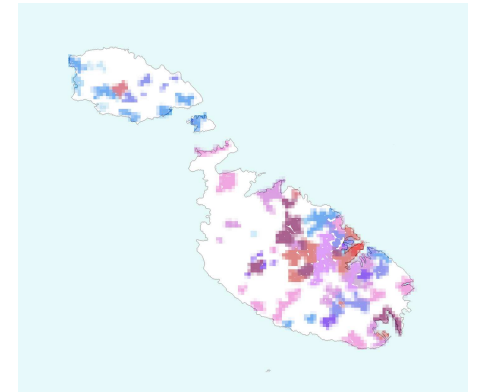
The data mining system was also used to create decision trees that can determine the minimum and maximum expected media attention ratings for a particular country based solely on the three G/D/I indicators for that country. From the decision trees it is evident that for poorer countries (i.e. those with low GDP per capita, G = 1 or 2) the secondary determining indicator for media attention is their number of disputes, while for richer countries (G = 3 or 4) 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 likely to be portrayed in a negative fashion.

## GAME THEORY AND THE BEST STRATEGIES FOR COUNTRIES TO ATTAIN MORE MEDIA ATTENTION

It is interesting to note that the decision trees can be used to predict different strategies for every country to increase (or decrease) its own media attention rating by looking at how the minimum expected rating changes if that country's G/D/I band changes (representing a change in wealth generation, disputes or internet/technology usage respectively) using simple game theory concepts. For example, China (which is in G/D/I band 2/2/3) can expect to increase its news rating by almost 90% if it increases its GDP per capita levels from the present \$5,600 to around \$10,700 (i.e. go into G/D/I band 3/2/3). On the other hand, for example, the UK can expect to increase its news rating by around 31% (and probably be seen in a more positive light) by decreasing the number of international conflicts, going from G/D/I band 4/2/3 to 4/1/3. Another interesting observation is that no matter what strategy a country's government adopts, the best way for poorer countries to gain more media attention is to increase their GDP per capita.

## EVALUATION ON A SMALL SCALE

After evaluating the cpGeo system on a global scale, we also evaluated the system on a small scale. For the small scale evaluation we chose to test the system on the smallest EU member state of Malta, which is made up of three small islands in the Mediterranean totalling just 316 square km in size. The small size of the country, together with the ready availability of copious amounts of Maltese news in English made Malta an ideal choice for small scale evaluation and testing. The small scale evaluation was based on four years of articles downloaded from the Times of Malta, ensuring comprehensive news item coverage in the database from mid-2001.



Results for Malta, Gozo and Comino

We noted the fact that having three islands of progressively smaller sizes made the evaluation more interesting, with the islands being Malta (32km by 14km), Gozo (14km by 7.5km) and Comino (3.2km by 2km). There are also some duplicated place names (e.g. Rabat in both Malta and Gozo – roughly 22km apart), providing a tough challenge for GNER disambiguation.

The cpGeo system produced a highly detailed map of Malta (see Figure 8) that accurately outlined most of the cities, towns and villages on the islands. The map was visually checked by four local Maltese persons who all agreed that the results do correspond to their intuitions about what places are mentioned most in the news. The level of detail in the figure above is incredibly high, with even obscure locations on the tiny island of Comino being precisely recognized and marked on both the map and in the data sets produced for evaluation. The different colour gradations used to mark hotspots in different areas are also more easily discernible at this high zoom level.

## GEOGRAPHIC DISAMBIGUATION

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 cpGeo system keeps track of other geographical named entities mentioned in the document and calculates their actual physical distance between them. Spatial location information allows the GNER system to flag "Manchester, Lancashire, UK" and "Manchester, Missouri, USA" as possible candidates. If the article mentions "England" or "United Kingdom" or "London" then it is quite probable that the article is referring to "Manchester, UK". The system works this out because London (UK) is much closer to Manchester in the UK than the one in Missouri. Generally, many texts contain enough redundant information to make precise disambiguation possible.

Sometimes there are occasions where place names in more than one country are mentioned and the above algorithm may still not give a completely resolved answer. For example, if the US President is visiting Arizona and makes a speech about the Iraq war, the system may not be able to decide whether Baghdad means Baghdad, Arizona (it is actually a small town in Yavapai county, Arizona) or Baghdad, Iraq. In this case, the system will determine the exact location of the place names using the following simple rules:

- Is there an explicit reference to an administrative region containing the location? (e.g. Baghdad, Iraq) If an explicit reference is found, this rule will override the other two rules.
- Check the locations of other place names mentioned. Is more than one region applicable? (e.g. if Najaf is mentioned, then the probability of Baghdad, Iraq will be higher)
- Check the populated place classification rating of the ambiguous place names and whether this place name is the name of a country or administrative region itself. (e.g. in this case Baghdad will score higher as it has a population of 5.7 million compared to the 1,500 living in Baghdad, Arizona)