Exploring a Large News Collection Using Visualization Tools

Tiago Devezas[1, 2] tdevezas@fe.up.pt
José Devezas[2] jld@fe.up.pt
Sérgio Nunes[1, 2] ssn@fe.up.pt

INESC TEC[1] and DEI[2], FEUP, University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

Abstract

The overwhelming amount of news content published online every day has made it increasingly difficult to perform macro-level analysis of the news landscape. Visual exploration tools harness both computing power and human perception to assist in making sense of large data collections. In this paper, we employed three visualization tools to explore a dataset comprising one million articles published by news organizations and blogs. The visual analysis of the dataset revealed that 1) news and blog sources evaluate very differently the importance of similar events, granting them distinct amounts of coverage, 2) there are both dissimilarities and overlaps in the publication patterns of the two source types, and 3) the content's direction and diversity behave differently over time.

1 Introduction

Finding valuable information in large collections of data can resemble looking for a needle in a haystack. An effective way to address this problem is the use of data visualization tools to explore datasets [Kei01]. The presentation of abstract data through interactive visual tools leverages human perceptual abilities and enhances cognitive performance, thus promoting discovery and sensemaking. In this paper, we present three distinct visualization tools for exploring large news collections, and apply them to the Signal Media One-Million News Articles Dataset[1], a collection of one million news and blog articles.

We show three use cases that highlight how these tools allow the investigation of distinct dimensions of the data. The first case evaluates how the hierarchy of importance given to a set of select global events, manifested through the amount of coverage, varies between news and blog sources. The second investigates the publication patterns of both source types during 24-hour and seven-day weekly cycles. The third use case studies the variation of topical diversity for news and blogs over time and employs a visualization tool developed specifically for this work. To develop this tool, an analysis was conducted to identify the topic vectors representing the directions followed daily by the articles' contents, compute a diversity score, and measure the topic diversity over time for news and blogs.

2 Corpus Characterization

The Signal 1M Dataset is comprised of one million articles published by 93,345 distinct media sources of two types: news and blogs. An analysis of the articles' media type reveals that 18,533 sources published exclusively news articles, 74,333 sources published only blog stories, and 479 had documents of both types. As for the article count by media type, nearly three-fourths were news (734,488 or 73.4%) and one-fourth blog items (265,512 or 26.6%). Thus, despite its lower amount, news sources were responsible for the publication of the majority of articles.

Even though the publication period extends from Jul 2nd 2015 to Sep 30th 2015, the majority of the articles were published between Sep 1st 2015 and Sep 30th 2015 (987,248 or 98.7%). Of these, 734,488 (74.4%)

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://research.signalmedia.co/newsir16/signal-dataset.html

were news articles and 265,512 (26.9%) blog articles. The highest number of articles published by a single source was 192,228 and the lowest amount, a single article. Regarding the overall distribution of articles, the majority of the sources (91,693 or 98.2%) published 100 articles or less, 1,565 sources (1.7%) published between 101 and 1000 articles, 85 (0.09%) between 1,001 and 5,000, one (0.001%) between 5001 and 10000, and one between 10,000 and 20,000 articles.

The topic analysis conducted for each media type stream (see Section 5.3.2) found that the top five n-grams, based on the TF-IDF score of the topic vectors, were 'south africa', 'pope francis', 'total volume table', 'high school football', and 'college football' for news articles, and 'star wars', 'school district', 'syrian refugees', 'executive director', and 'kansas city' for the blog document set.

3 Visualization of Large News Archives

The visualization and analysis of large volumes of news content is an emerging field of research [KBMK10]. The ThemeRiver application [HHN02] was one of the first efforts in this domain. It provides an interactive visualization of thematic changes across a large set of news documents over time. It uses a metaphor of a river to assist in the recognition of relationships, trends and patterns in the data. Themes are displayed as colored streams whose width — the measure of its strength — varies as it flows across time from left to right. A similar river-like visual metaphor is employed by the NewsLab system [GLYR07], which allows exploratory analysis of the temporal variation of themes, and their hierarchical structure, from a large collection of news videos.

Krstajić et al. [KBK11] present CloudLines, a visualization technique to display a compact view of multiple time series, each showing a sequence of related events and event episodes (high density sequences of events). The relative importance of events is conveyed through variations in the clusters' opacity and size. The system also permits fine-detailed analysis of individual event data points.

The complexities of visualizing the dynamics of news data streams are addressed by Krstajić et al. [KBMK10]. The system displays the evolution of news in real-time by converting the stream into threads comprised of similar articles. In addition to showing recent threads, the system computes the threads' relevance on the fly — based on the items' age and their relationships — to determine which threads to keep on screen and which ones to remove.

The development of news stories and their relationships through time is also explored by Story Tracker [KNAMK13]. The application represents the evolution of stories over time, and how they merge and split. Story clusters are displayed as rectangles whose size corresponds to the number of articles and include labels for the story title and the most important keywords. Related clusters have the same color, are edge-connected, and can be zoomed to the level of the individual articles that compose them.

The NewsStream service [NGSM15] provides several interactive tools to visually explore a continuously updated collection of financial articles, published via the RSS feeds of multiple news and blog sources. The system displays occurrences and co-occurrences of financial and geographic entities in the news, the related sentiment, a summary of the linked content through tag clouds, and temporal country co-occurrence networks displayed on a world map.

4 The MediaViz Platform

The MediaViz platform [DNR15] aims to assist in gaining insight from a large archive of news through interactive visualization tools. It comprises two components. The first is a back-end application that fetches and stores articles published via the RSS feeds of multiple online news sources and provides access to the data through an API. The second is a client application which retrieves the data provided by the API and allows its exploration through interactive visualization tools. Our approach is based on open technologies and was built with extensibility in mind: the client application is decoupled from the back-end so it can be configured to work with different datasets with minimal effort. For this paper, we stored the Signal 1M Dataset in a relational database and built a simple API. No major modifications were required for the existing visualization tools to work with the new API. However, a new tool was developed to explore topic diversity over time for news and blog articles. A fully functional demo is available online[2].

5 MediaViz Visualization Tools

Rather than focusing on individual sources, we opted to explore the two types of media sources that comprise the corpus — news and blogs —, as they allow a macro-level analysis and comparison of the dataset.

5.1 Variations in Coverage

The dynamics of the coverage that each source type granted to different themes over time are displayed by the Keywords tool. Users can insert multiple search terms and see how many articles (in absolute terms or

[2] http://irlab.fe.up.pt/p/mediaviz/newsir/

as a percentage of all articles published on the respective day) with those keywords were published daily during the selected period. Additional context can be obtained by clicking the data points, which displays a list of all related articles. Each list item includes the title, summary, publication date and the source's name, and can be clicked to display the full text.

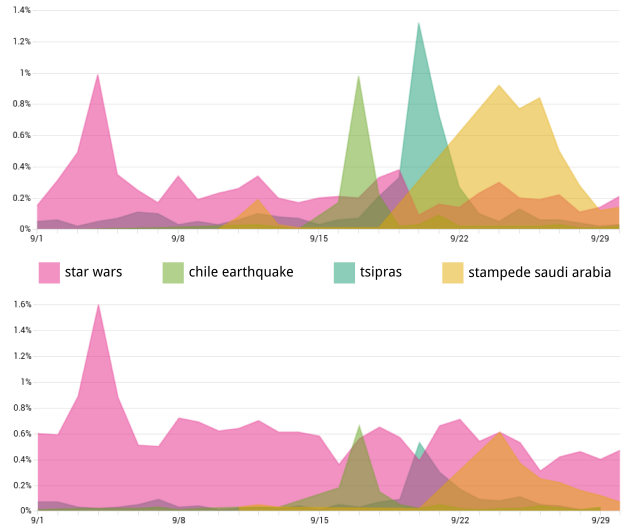

Figure 1: MediaViz Keywords tool. Top: Daily percentage of articles published by all news sources containing the given terms. Bottom: Daily percentage of articles published by all blog sources containing the same terms.

Figure 1 displays the daily percentage of articles published between Sep 1st 2015 and Sep 30th 2015 by each source type with the terms 'star wars', 'chile earthquake', 'tsipras', and 'stampede saudi arabia'. These particular terms were chosen because they are related with some relevant global events — identified after consulting several online resources — that took place on September 2015. The visualization's peaks highlight the selected events: the merchandise for the latest Star Wars movie was released on Sep 4th; an earthquake in Chile which led to the evacuation of millions of people took place on Sep 16th; on Sep 20th Alexis Tsipras was reelected as Prime Minister of Greece after resigning and calling for a snap election; and, on Sep 24th, hundreds of people died after a stampede during the annual pilgrimage to Mecca, in Saudi Arabia. As shown in Figure 1, the attention given to these events differed greatly between the two source types. News sources (top), gave similar attention to each event, while in blogs (bottom), the primacy belongs to articles mentioning Star Wars.

5.2 Publication Patterns

The Sources tool allows the comparison of publication patterns (count and percentage of articles) for multiple sources according to distinct temporal granularities: weekly, monthly and 24-hour cycles. To have comparable results, publication times are converted to the UTC time standard. The ability to compare several sources in the same screen can thus provide meaningful perspectives regarding their production cycles. This can be seen in Figure 2. News sources published a higher percentage of articles than blogs during business days, a behavior that is reversed during the weekend. While this pattern might be expected, given the particularities of each media type, the Sources tool quantitatively shows that such assertion is indeed true.

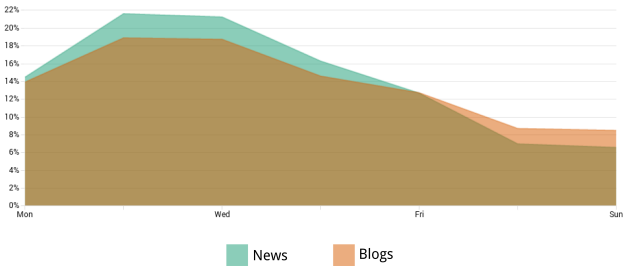

Figure 2: MediaViz Sources tool. Percentage of articles published by both source types for each day of the week.

When looking at a 24-hour cycle, news and blog sources exhibit similar patterns. As Figure 3 displays, publications follow a typical working schedule: the most active publication period occurs between 08:00 and 16:00 UTC and then gradually decreases. One possible explanation for this overlap is the growing professionalization and influence of blogs, which often compete with traditional news sources for online eyeballs. The most significant difference between the two patterns, the news sources' peak at 07:00, can be potentially explained by the publication of early morning news.

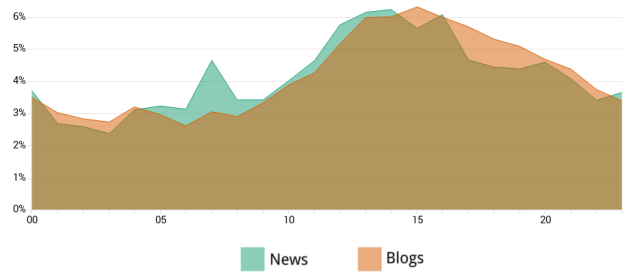

Figure 3: MediaViz Sources tool. Percentage of articles published by both source types during a 24-hour cycle.

5.3 Diversity Explorer

The Diversity Explorer tool was developed specifically for this work. Below we describe our strategy for detecting topics and measuring topical diversity between the news and blog streams.

5.3.1 Topic Detection

Our topic detection strategy was based on the clustering of text documents using n-grams of size $n = 2$ (bigrams) and $n = 3$ (trigrams) as features. The base strategy consisted of, for a given day, transforming each document into a bag of n-grams and then running k-means [HW79] using the n-gram frequencies as features. The value of k was selected based on the Silhouette method [Rou87], by testing successive values of $k \in [2, 15]$ for a random sample of 100 or less documents — in case less than 100 documents were available. Constraining the value of k, indirectly enforced the number of topics to range between 2 and 15. The result of this process was a set of k topics, represented by the centroid of each cluster and associated with the documents for each day.

Prior to the clustering phase, and in order to ensure performance, we reduced the number of features by removing n-grams that were over 99.6% sparse, i.e., features with more than 99.6% zeros, that were less useful in distinguishing documents, were simply discarded. The sparsity threshold of 99.6% was determined empirically, by experimenting with the largest daily document set and ensuring that the number of features would not explode (99% decrease from 1,834,310 to 350 features for the largest daily document set), but also with smaller daily document sets to ensure that the number of features would not be too small (nearly 0% decrease for daily document sets with less than 100 documents). After completing the feature reduction process, we repeated the previously described clustering process for the smaller matrix, obtaining k topic vectors that illustrated the different directions of followed contents in daily news.

5.3.2 Measuring Topic Diversity

In order to measure topic diversity within a corpus, we took the topic vectors for a given day and did an element-wise aggregation based on the maximum weight of each n-gram. This resulted in a set of daily vectors, describing the overall topical direction of news and blog articles per day.

Our approach to measuring topic diversity was based on a combined distance metric between all n-gram daily vectors, for a given corpus — the more distant the topics are from every other topic, the higher the diversity. We computed the normalized cosine distances X for each pair of n-gram daily vectors, separately for the news and blog corpora. Next, we calculated the mean and standard deviation for the obtained values, and combined the mean $E[X]$ and standard deviation $\sigma(X)$ into a diversity score, as described in Equation 1.

$$score(X) = E[X] - 2 \times (F(E[X]; 0.5, 1/50) \times 0.5) \times (1 - E[X]) \times (E[X] \times \sigma(X)) \quad (1)$$

$$F(x; \mu, s) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}} \quad (2)$$

The idea was for the variance to affect the mean cosine distance in the following way: for a low mean, a low variance would result in a small increase, while a high variance would result in a large increase; for a high mean, a low variance would result in a small decrease, while a high variance would result in a large decrease. For example, given a mean cosine distance of 0.9, with a 0.9 standard deviation, we know that there are several values below the mean and that, since we are using a normalized cosine distance, its maximum is one. Thus, it makes sense that we would decrease (negative sign) the diversity score with the intuition that a subset of documents would be less diverse among themselves than average. On the other hand, for a mean cosine distance of 0.1, it would only make sense to increase (positive sign) the value based on the standard deviation. To determine sign, we took advantage of a logistic distribution (Equation 2), centered on $\mu = 0.5$ and scaled to $s = 1/50$. We used this as a sign function by shifting the result by -0.5 and multiplying by 2, which gave us a value in the interval $[-1, 1]$ with a sigmoidal behavior. We then combined the mean and standard deviation to obtain the absolute value of increase or decrease, and multiplied it by the sign function.

We repeated this process for news, blogs, and the concatenated n-gram daily vectors of both corpora, for an overall topic diversity measurement. This resulted in a diversity score between zero and one, where zero meant that all the topics were exactly the same, while one meant that all the topics were completely distinct. Based on our results, topics have, overall for the combined samples, a diversity score of 0.970, a value that is as high as 0.986 for blogs, and as low as 0.976 for news. Topic diversity is similarly high in either case, despite blogs having a slightly higher diversity score.

5.3.3 Exploring Diversity Over Time

We also measured topic diversity over time, for small temporal windows, comparing news and blogs. Figure 4 shows the resulting diversity score for a sequence

of 5-day windows starting at the given date (x-axis), from Sep 1st to Sep 30th 2015, with news in green and blogs in red. As we can see, both corpora have a diversity behavior that is similar over time, with the exception of the temporal windows from Sep 15th to Sep 19th 2015. Correlation between the two diversity score distributions is 28.9% for the whole month of September, but raises to 69.3% when ignoring the period of 15–19 Sep. We calculated the differences between diversity scores over time and found that the temporal window starting at Sep 19th 2015 represented the largest break in consistency between news and blogs, with a difference in diversity of 0.205.

We analyzed the n-grams of the topics, for each corpus, within this temporal window. For the news corpus, we found 111 unique n-grams out of 175 total n-grams, meaning that 63.43% of the n-grams are unique, which indicates a high diversity. On the other hand, for the blog corpus, we found 64 unique n-grams out of 164 total n-grams, meaning that 39.02% of the n-grams are unique, which indicates a low diversity. This is consistent with our diversity score. We also calculated the Jaccard index for the set of n-grams of each corpora, for the Sep 19th 2015 temporal window, finding that 15.89% of the total number of unique n-grams appears in both news and blogs.

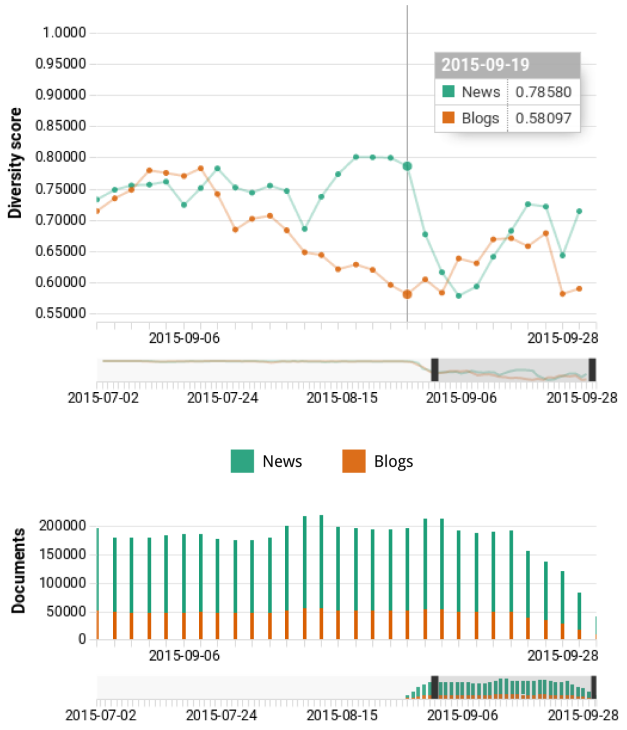

Figure 4: MediaViz diversity explorer. Top: diversity over time for windows of 5 days, starting at the given date. Bottom: number of documents for windows of 5 days, starting at the given date.

6 Conclusion

In this paper we presented the exploration of the Signal 1M Dataset, which comprises a large collection of news and blog articles, using distinct visualization tools. The visual analysis of the corpus provided interesting perspectives that would be much more difficult to obtain without the assistance of such tools. The Keywords tool allowed us to see that news and blog sources granted different levels of importance to a given set of keywords related with major global events that took place on September 2015. It was also evident, using the Sources tool, that the temporal publication patterns of these two media behaved differently — blogs published a higher percentage of content during the weekend than news sources —, but also in a similar fashion — both sources followed an identical curve during a 24-hour cycle. Finally, through the Diversity Explorer tool, we were able to visualize variations in the dynamics of topical diversity over time for each media type's content stream.

Acknowledgements

Project 'NORTE-01-0145-FEDER-000020' is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[DNR15] Tiago Devezas, Sérgio Nunes, and María Teresa Rodríguez. MediaViz: An interactive visualization platform for online media studies. In *Proceedings of the 2015 International Workshop on Human-centric Independent Computing*, pages 7–11. ACM, 2015.

[GLYR07] Mohammad Ghoniem, Dongning Luo, Jing Yang, and William Ribarsky. Newslab: Exploratory broadcast news video analysis. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 123–130. IEEE, 2007.

[HHN02] Susan Havre, Beth Hetzler, and Lucy Nowell. Themerivertm: In search of trends, patterns, and relationships. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[HW79] J A Hartigan and M A Wong. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[KBK11] Miloš Krstajić, Enrico Bertini, and Daniel A Keim. Cloudlines: Compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432–2439, 2011.

[KBMK10] Miloš Krstajić, Enrico Bertini, Florian Mansmann, and Daniel A Keim. Visual analysis of news streams with article threads. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, pages 39–46. ACM, 2010.

[Kei01] Daniel A Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[KNAMK13] Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A Keim. Story tracker: Incremental visual text analytics of news story development. *Information Visualization*, 12(3-4):308–323, 2013.

[NGSM15] Petra Kralj Novak, Miha Grcar, Borut Sluban, and Igor Mozetic. Analysis of financial news with newsstream, technical report IJS-DP-11965. *CoRR*, abs/1508.00027, 2015.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.