Analyzing and Comparing On-Line News Sources via (Two-Layer) Incremental Clustering

Francesco Cambi, Pierluigi Crescenzi, Linda Pagli

Analyzing and Comparing On-Line News Sources via (Two-Layer) Incremental Clustering

Francesco Cambi[1], Pierluigi Crescenzi[2], and Linda Pagli[3]

1 **Bridge Consulting S.r.l., Via L. Rosellini, 50127 Firenze, Italy**
fcambi@bridgeconsulting.it
2 **Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy**
pierluigi.crescenzi@unifi.it
3 **Dipartimento di Informatica, Università degli Studi di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy**
linda.pagli@unipi.it

Abstract

In this paper, we analyse the contents of the web site of two Italian news agencies and of four of the most popular Italian newspapers, in order to answer questions such as what are the most relevant news, what is the average life of news, and how much different are different sites. To this aim, we have developed a web-based application which hourly collects the articles in the main column of the six web sites, implements an incremental clustering algorithm for grouping the articles into news, and finally allows the user to see the answer to the above questions. We have also designed and implemented a two-layer modification of the incremental clustering algorithm and executed some preliminary experimental evaluation of this modification: it turns out that the two-layer clustering is extremely efficient in terms of time performances, and it has quite good performances in terms of precision and recall.

1998 ACM Subject Classification H.2.8 Database Applications: Data mining, H.3.3 Information Search and Retrieval: Clustering, H.3.5 Online Information Services: Web-based services

Keywords and phrases text mining, incremental clustering, on-line news

Digital Object Identifier 10.4230/LIPIcs.FUN.2016.9

1 Introduction

The web is a huge source of data, which are produced by companies, institutions and individuals, and, most of the times, are available for free. The use we can make of all this information is limited just by our imagination. In the last few years, for example, there has been a quite significant amount of research devoted to the analysis of the so-called on-line social networks. One of the most recent examples of such analysis is the adaptation of the well-known "six degrees of separations" phenomenon to the Facebook network: it has, indeed, has been observed that, in the case of this on-line social network, the degrees of separation are less than four [5]. In this paper, instead, we show how the information available on the web sites of news agencies and newspapers can be used in order to answer several questions about the news on-line system, such as the following ones.

- *What are the most relevant news*? Clearly, answering to this question in an automatic way implies defining the notion of relevance, that is, a news score function. In this paper, we propose one possible definition (based on the position occupied by the news in the web sites) and apply it to all the news collected from mid October until mid December of

the last year. It turns out that *the news with the maximum score is the shooting down of an American drone in Turkey*. Among the news with the highest score, there is one concerning a case of homicide. The score pattern of this latter news is quite interesting since it has three main peaks: one peak corresponds to the disappearance of the victims, the second peak corresponds to the rising of the hypothesis of homicide, and the third peak corresponds to the identification of possible suspects. We believe that it is interesting to analyse this kind of patterns because they help us to better understand the evolution of a news.

- *What is the average lifespan of a news*? Similar questions have already been posed and answered in different contexts. For instance, a network of hundreds of media sites have been recently looked into, and it has been found that most posts had a shelf life of 2 days, with the median stretching out to about 2.5 days [11]. Closer to our analysis are the results presented in [2], where the authors study the diffusion patterns of news articles from several popular news sources, determine the lifespan of a news article on Twitter by the time difference between the last and first tweet posted containing the URL to that article, and show that for most of the news media companies, about 45% of their articles survive beyond 18 hours. In this paper, we analyse almost 10000 articles, grouped into approximately 7000 news, and we observe that *the average lifespan of a news is a little bit less than one day*. The news with the longest lifespan concerns vaccination campaigns in Italy (with a lifespan greater than 19 days), while one of the many news with the shortest lifespan concerns the rescue of two bear cubs in Laos (with a lifespan of at most one hour).
- *Are different news sources really different?* Once again, answering this question implies defining a notion of similarity among the on-line news sources. In this paper, we propose such a definition, and we apply it to the six on-line sources we have analysed. Interestingly enough, it turns out that *the two news agencies are quite similar*, *one newspaper seems to echo the two news agencies*, while the other *thre newspapers are equally and significantly distant from the two news agencies*, but in opposite directions.

In order to answer the above questions, we have developed a web-based tool which allows us to perform the following operations.

- *Web scraping*. The tool can download the articles (formed by the title, the abstract, and the text) from the web site of two Italian news agencies, and of four popular Italian on-line newspapers. Almost 10000 articles have been downloaded from mid October 2015 to mid December 2015. These articles have been appropriately processed in order to be subsequently analysed in terms of their similarity.
- *Clustering*. As we said, we have defined a similarity measure between two articles, which is based on quite standard text analysis methods. By using these similarity values, a simple incremental clustering algorithm has been implemented: this algorithm basically insert a new article into the cluster with the highest average similarity, if this average similarity is greater than a specific threshold (otherwise a new cluster is created containing only the new article). This incremental clustering algorithm (which is different from previous incremental clustering approaches applied to news detection such as the one analysed in [1, 3]), performs very well in terms of precision and recall results. However, the time performance of the algorithm degrades as the number of articles (and, hence, of clusters) increases. To improve the time performances, we have designed and implemented a two-layer variation of the incremental clustering algorithm (which is also different from similar two-layer approaches such as the one used while analysing Twitter messages in [12]). In this variation, a centroid is associated with each cluster, and the centroids are

clustered by making use of the same approach (that is, by using the average similarity values). Once the best cluster of centroids has been determined, the original incremental clustering is applied to the corresponding clusters. As far as we know, this two-layer approach is different from all previous techniques, which have been used to cluster news. Even if its performances in terms of precision degrade of about 10% the time performance of the new approach are drastically better: indeed, clustering approximately 3000 articles requires less than two seconds, while the original approach required more than 2 minutes.

- *Statistical analysis.* In order to answer the first two questions listed above, we have defined a score function of news, which is based on the position of the articles, included in the same news, in the front page of the web site of the on-line news sources. The position in the front page (along with variation of the Kendall tau measure) has been used also to define similarity measure between different newspapers, and, hence, to answer the third question.

Apart from the fact that our clustering algorithms are different from previous approaches used in news detection (even if, due to the huge literature in the field of incremental clustering, they are similar to several other clustering approaches), as far as we know, this is the first time that such techniques are used to analyse and explicitly compare on-line news sources, on the ground of the order in which news appear in the web sites of the news sources themselves.

The paper is structured as follows. In Section 2, we describe the web scraping performed by our tool, and we present some preliminary descriptive statistics. In Section 3 we introduce our incremental clustering algorithm in order to detect news from articles. In Section 4 we present the results about the score and the lifespan of news, and about the comparison of the six analysed news on-line sources. In Section 5 we describe the two-layer variation of the clustering algorithm. Finally, in Section 6 we conclude and propose some possible directions for future research.

2 The Web Scraping Process

Web scraping is a technique to extract data from web sites, by making use of software libraries that simulates surfing on the web. Fortunately, most of the on-line news sources use well-defined XML schemata and tag systems, which makes the content independent of the graphical rendering of the web site. This feature (which is usually part of what is called the "web semantic") allowed us to quite easily download the contents of the articles which were present on the main web page of the analysed news sources. For instance, in the left part of Figure 1 it is shown part of the main page of one of the most popular Italian on-line newspaper. The XML schema used by this page easily allowed us to detect the main article column (rounded in green), and, within this column, the main article sub-column (rounded in red). In the figure, we have three articles, whose content can then be downloaded by using the hypertext link associated with the title of the article. In the current implementation of our tool, we have then decided to ignore the articles appearing in other sub-columns, since in the vast majority of the cases these articles seem to be of relatively smaller importance. However, it would not be too difficult to change the web scraping module of our tool in order to let it download also these articles.

In the vast majority of the cases, the content of an article consists of three components: the title, the abstract, and the text. In the example shown in the right part of Figure 1, the title is rounded in orange, the abstract is rounded in red, and the text is rounded in green. Once again, the XML schema used by the web site of the on-line news source allows us to easily detect these three components (by also eliminating all types of content we are not

Figure 1 The structure of the main page of a news source (left) and of an article (right).

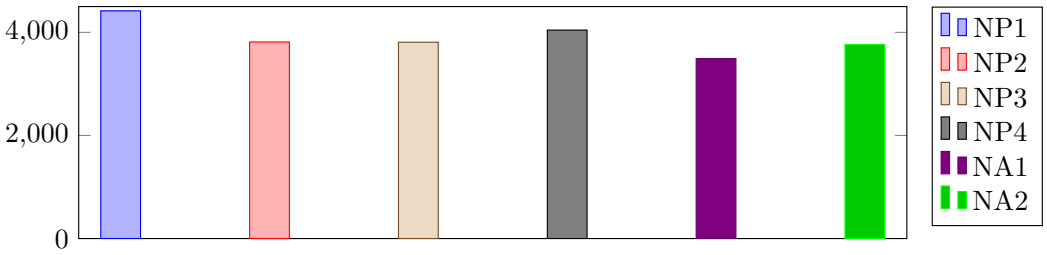

Figure 2 The number of articles downloaded from mid October 2015 until mid January 2016.

interested in, such as video and pictures). After having downloaded the title, the abstract, and the text of an article, this information is stored in a MySQL database, in order to be used for the news detection process that we are going to describe in the next section. This web scraping process has been started at the beginning of October 2015, and it is still in action: currently, we have downloaded more than 23000 articles.

2.1 Some descriptive statistics

As we just said, we have downloaded 23335 articles (during approximately five months) from the web sites of two news agencies NA1 and NA2, and of four newspapers NP1, NP2, NP3, and NP4. These articles are quite uniformly distributed among the six news sources, as it is shown in Figure 2. Note that the download of these articles has been done hourly (almost every day), but, clearly, we have avoided to download again articles that had been previously downloaded. In other words, the number of daily downloaded articles is not just 24 times the number of articles included in the home page of the web site of a news source, since the same articles can remain on the home page itself for several hours. Indeed, in Figure 3 we show the minimum, maximum, and average numbers of distinct daily downloaded articles from each of the six news sources. Once again, these numbers seem to uniformly distributed:

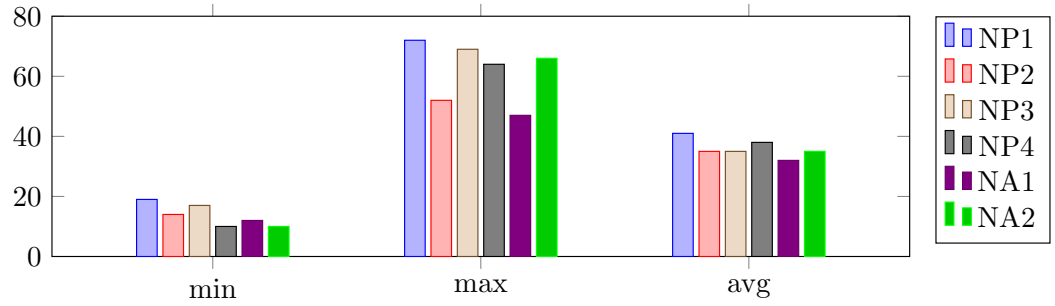

Figure 3 The minimum, maximum, and average number of daily downloaded articles from mid October 2015 until mid January 2016.

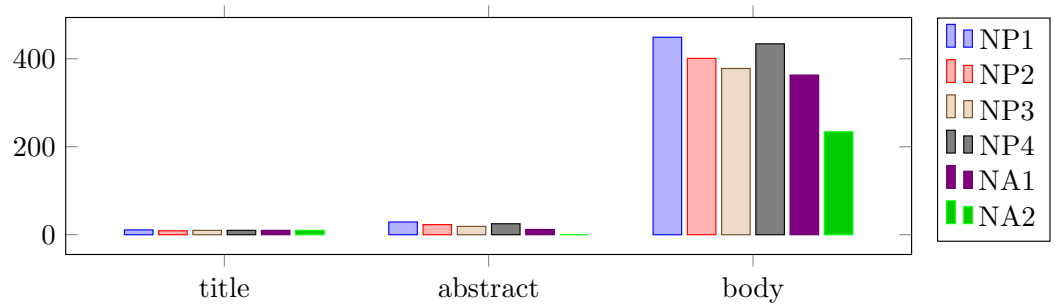

Figure 4 The average lengths of titles, abstracts, and bodies of the six news sources.

this suggest that, on the average, approximately 40 new articles are produced every day in the main area of each home page. The activity of a news source will be better analysed in Section 4.1, in which we will try to measure the degree of variability of the home page of a news source in terms of news (that is, clusters of articles), instead that in terms of articles.

Another simple statistics that can be immediately computed, once a sufficiently large dataset of articles has been downloaded, is the word length of the articles themselves. This number does not really differs among the six news sources (apart form the second news agency that seems to produce, on average, shorter articles). This, indeed, can be deduced from Figure 4, where the average lengths of the titles, the abstracts, and the bodies of the six news sources are shown. More precisely, it seems that, on average, a title (respectively, an abstract, and a body) contains 10 (respectively, 20, and 400) words. The news agency NA2, however, uses shorter body lengths and no abstract at all. This uniformity among the article lengths can be, maybe, justified because of the standards normally used within the printed news system (mainly due to the limited number of pages of a newspaper): however, it is quite surprising that similar standards are also used within the on-line news system, where, in theory, no limit is a-priori existing on the number of pages an article should utilise. It is also worth noting that, if we consider the maximum word length (instead of the average one), the first newspaper NP1 is clearly producing much longer articles: indeed, its maximum article word length is 7216 which is almost twice the second longest maximum length (that is, 3730 in the case of the fourth newspaper NP4).

We have also computed the frequency of the words used in the title or in the abstract of an article. In particular, in Figure 5 we show the frequencies of the names of the leaders of the main Italian parties. As expected, Renzi (with almost 1200 occurrences), leader of *Partito Democratico* and Italian Prime Minister, is by far the most frequent name, followed by Berlusconi (283 occurrences), leader of *Forza Italia*, and Alfano (244 occurrences), leader

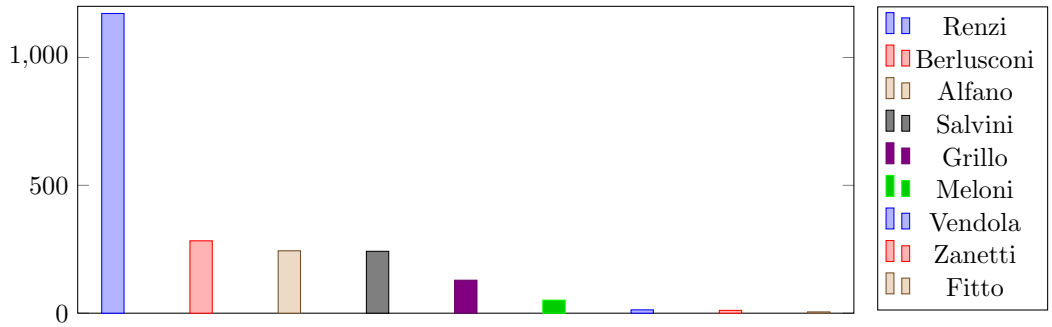

Figure 5 The number of occurrences of the name of the leaders of the Italian parties in the title or in the abstract of an article.

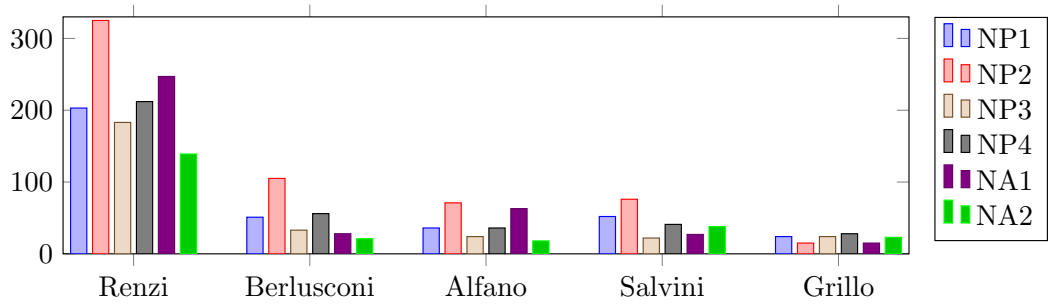

Figure 6 The number of occurrences of the name of five leaders of Italian parties in the title or in the abstarct of an article, depending on the news source.

of *Nuovo Centro-Destra*. More surprisingly, Grillo (129 occurrences), a popular comedian and blogger who is the leader of *Movimento 5 Stelle*, is only fifth in this ranking, preceded also by Salvini (242 occurrences), leader of *Lega Nord*. Interestingly enough, the second newspaper NP2 is systematically the one that most frequently includes the names of the first five leaders into the title or the abstract of an article, apart from the case of Grillo (as it is shown in Figure 6). On the other hand, the second news agency NA2 is the news source which almost always includes the name of the leaders less than the other news sources: in this case, however, we should again observe that this agency does not use abstracts at all.

Finally, we also tried to verify somehow whether the following statement taken from Wikipedia corresponds to reality: "the major news agencies generally prepare hard news stories and feature articles that can be used by other news organisations with little or no modification" [13]. To this aim, we made use of the Levenshtein distance between two strings (that is, the minimum number of insertions, deletions and substitutions required to change one string into the other) [9] in order to define a similarity measure between two string as follows: the difference between the length of the longest string and the Levenshtein distance, divided by the length of the longest string. We then computed for a (small) sample set of articles the similarity of their texts with the texts of the articles of the first news agency NA1. As a result of this comparison, we found a quite high similarity between the articles. The maximum found similarity has been, indeed, equal to 0.96. The two articles reaching this value reported about the first American democratic party debate, and their length was 4527 and 4555, respectively. This implies that the two articles were almost identical.

3 The Clustering Algorithm

As we said in the introduction, we focused our attention on news, instead of articles. Indeed, several different articles can refer to the same news, either because they are present on different web sites, or because the news evolved and new articles concerning the news itself have been produced. For this reason, we had to design a clustering of the articles into clusters corresponding to news. To this aim, we first have to define a similarity measure between articles, and subsequently apply a clustering algorithm on the ground of the similarity values.

3.1 The similarity measure

In order to define a similarity measure between articles, we made use of quite standard text mining techniques (see, for example, the first chapter of [8]). First of all, we cleaned the contents of the articles by eliminating the so-called "stop words", that is, the several hundred most common words in Italian that do not carry any significance by themselves (such as, for example, the articles). Successively, we applied a stemming algorithm for reducing derived words to their word root so that related words map to the same root, even if this root is not in itself a valid word (note that algorithms for stemming have been studied in computer science since the late sixties [10]). Finally, we identified the *keywords* of an article a, that is, the words to be used in order to measure the similarity of a with other articles, as the words appearing in its title $\texttt{title}(a)$ and in its abstract $\texttt{abstract}(a)$, and the ones appearing in its text $\texttt{text}(a)$ starting with a capital letter. In the following, we will denote by $\texttt{key}(a)$ the set of keywords of article a. For each set of articles A and for each word w, the *inverse document frequency* of w with respect to A is defined as

$$\texttt{idf}(w, A) = \log\left(\frac{|A|}{|\{a \in A : w \in \texttt{text}(a)\}|}\right).$$

Moreover, the *term frequency* of w with respect to an article $a \in A$ is defined as

$$\texttt{tf}(w, a) = \frac{\texttt{occ}(w, a)}{|a|}$$

where $\texttt{occ}(w, a)$ denotes the number of occurrences of w in $\texttt{text}(a)$, and $|a|$ denotes the number of words in $\texttt{text}(a)$. For each pair of articles a_1 and a_2 in the set A, we can then define their TF-IDF vectors as follows. Let $\texttt{key}(a_1, a_2) = \texttt{key}(a_1) \cup \texttt{key}(a_2) = \{k_1, \ldots, k_m\}$ be the ordered set of keywords of either a_1 or a_2. Then, for any i with $1 \leq i \leq m$, we define

$$\texttt{tfidf}_{a_1,a_2}[i] = \texttt{tf}(k_i, a_1) \cdot \texttt{idf}(k_i, A)$$

and

$$\texttt{tfidf}_{a_2,a_1}[i] = \texttt{tf}(k_i, a_2) \cdot \texttt{idf}(k_i, A).$$

Finally, the *cosine similarity* $\texttt{cosim}(a_1, a_2)$ between a_1 and a_2 is the cosine value of the angle formed by the two vectors $\texttt{tfidf}_{a_1,a_2}$ and $\texttt{tfidf}_{a_2,a_1}$. Formally, it is defined as

$$\texttt{cosim}(a_1, a_2) = \frac{\sum_{i=1}^{m} \texttt{tfidf}_{a_1,a_2}[i]\texttt{tfidf}_{a_2,a_1}[i]}{\sqrt{\sum_{i=1}^{m} \texttt{tfidf}_{a_1,a_2}[i]^2}\sqrt{\sum_{i=1}^{m} \texttt{tfidf}_{a_2,a_1}[i]^2}}.$$

Note that this value is as close to 1 as the two vectors are similar, that is, as the two articles have similar TF-IDF vectors.

3.2 The incremental clustering

By using the above defined similarity measure, we have designed and implemented the following quite simple incremental clustering algorithm. Assume that a set A of articles has already been clustered into a set N of n news or clusters. Indeed, N can be seen as a function $N : A \to [n]$, where $[n]$ denotes the set of integers between 1 and n. Assume also that a new article x has been downloaded and, hence, has to be classified in either one of the already existing n clusters or in a new one. For each cluster c with $c \in [n]$, let $A_c = \{a \in A : N(a) = c\}$. We then define the similarity between x and c as the average of the similarities between x and all articles in c. More formally,

$$\texttt{sim}(x, c) = \frac{1}{|A_c|} \sum_{a \in A_c} \texttt{cosim}(x, a).$$

Let c^* be the cluster c for which $\texttt{sim}(x, c)$ is maximum: if $\texttt{sim}(x, c^*)$ is greater than or equal to a given threshold τ, then $N(x) = c^*$ (and x is assigned to cluster c^*), otherwise $N(x) = n + 1$ (and a new cluster is created).

3.2.1 Threshold estimation

In order to apply the above clustering algorithm, we have to determine the threshold τ. To this aim, we have manually clustered the set T containing the first 3000 downloaded articles, and we then have determined the threshold value that produced the best results. In particular, given two clustering N_1 and N_2 with n_1 and n_2 clusters, respectively, we can define the matching function $M_{N_1,N_2} : [n_1] \to [n_2]$ as follows:

$$M_{N_1,N_2}(c_1) = \max_{c_2 \in [n_2]} \texttt{jac}(c_1, c_2)$$

where $\texttt{jac}(c_1, c_2)$ denotes the Jaccard index of c_1 and c_2, defined as $\frac{|N_1(c_1) \cap N_2(c_2)|}{|N_1(c_1) \cup N_2(c_2)|}$. In order to evaluate the clustering N_τ obtained by using threshold τ and applied to T, we have first computed the function M_{N_τ,N^*}, where N^* is the manual cluster, and subsequently computed the recall and precision values of each cluster in N_τ. The *precision* of a cluster c is defined as

$$\texttt{pre}(c) = \frac{|\{a \in T : a \in M_{N_\tau,N^*}(c)\} \cap T_c|}{|T_c|},$$

while the *recall* of a cluster c is defined as

$$\texttt{rec}(c) = \frac{|\{a \in T : a \in M_{N_\tau,N^*}(c)\} \cap T_c|}{|\{a \in T : a \in M_{N_\tau,N^*}(c)\}|}.$$

The recall and precision values can be combined by obtaining the F-measure, which is defined as follows:

$$\texttt{F}(c) = 2\frac{p(c) \cdot r(c)}{p(c) + r(c)}.$$

The *average weighted F-measure* of a clustering N_τ is then defined as the average value of the F-measure weighted with respect to the sizes of the clusters. Formally,

$$\texttt{F}(N_\tau) = \frac{\sum_{i=1}^{n_\tau} \texttt{F}(c_i)|c_i|}{\sum_{i=1}^{n_\tau} |c_i|}$$

where n_τ denotes the number of clusters in N_τ and $|c_i|$ denotes the size of the i-th cluster (that is, $|c_i| = |\{a \in T : N_\tau(a) = i\}|$). The threshold τ has then been determined by selecting the one who produced the higher average weighted F-measure. It turned out that the best value of τ is equal to 0.35: with this value of τ the average weighted F-measure is equal to 0.89.

4 Analyzing and Comparing the On-Line News Sources

Once the articles have been clustered, and, hence, the news have been detected, we can now answer to several interesting questions concerning the on-line news system.

4.1 Activity of a News Source

A *front page* $f = (t, L)$ of a news source is a time stamped ordered list L of news, which are the news appearing in the home page of the web site of the news source at time t. We would like to measure what is the degree of variability of the front pages of the same news source during the day. To this aim, we have to compare two ordered lists of news, which is equivalent to comparing two top k lists for similarity/dissimilarity. We then decided to make use of the averaging Kendall distance analysed in [6], which is a modification of the Kendall's tau metric between permutations (see the textbook [7]) for the case when we only have the top k members of the ordering. Given two ordered lists L_1 and L_2 of distinct numbers taken from a domain D, for each pair $(x, y) \in D \times D$, the contribution $K_{x,y}(L_1, L_2)$ of x and y to the averaging Kendall distance of L_1 and L_2 is defined as follows (in the following we denote by $a <_{L_i} b$ the fact that a precedes b in L_i).

1. If x and y belong to both L_1 and L_2, then $K_{x,y}(L_1, L_2) = 0$ if they are in the same order in both lists, otherwise $K_{x,y}(L_1, L_2) = 1$.
2. Else if x and y belong to L_i, for $i \in \{1, 2\}$, and either x or y belongs to L_{3-i}, then $K_{x,y}(L_1, L_2) = 0$ if $x <_{L_i} y$ and x belongs to L_{3-i} or if $y <_{L_i} x$ and y belongs to L_{3-i}, otherwise $K_{x,y}(L_1, L_2) = 1$.
3. Else if x belongs to L_i, for $i \in \{1, 2\}$, and y belongs to L_{3-i}, then $K_{x,y}(L_1, L_2) = 1$.
4. Else if x and y belong to L_i, for $i \in \{1, 2\}$, and neither x nor y belongs to L_{3-i}, then $K_{x,y}(L_1, L_2) = \frac{1}{2}$.
5. Else $K_{x,y}(L_1, L_2) = 0$.

The *averaging Kendall distance* $K(L_1, L_2)$ of L_1 and L_2 is then equal to the sum of the contributions of all possible pairs $(x, y) \in D \times D$. Given two front pages $f_1 = (t_1, L_1)$ and $f_2 = (t_2, L_2)$ we then define their distance as the averaging Kendall distance of the corresponding lists L_1 and L_2 of news. Since we have downloaded articles every hour, we have then computed the distance between the front page taken at a given time and the front page taken one hour later (that is, t_2 is equal to t_1 plus 3600 seconds). In Figure 7 we show the plot of these distances computed for the front pages of one news agency (upper part of the figure) and of one newspaper (lower part of the figure). In both cases, it is quite evident that there is a periodic behaviour in updating the web site: however, it is also clear that while the web site of the news agency is updated quite uniformly, the updating process of the web site of the newspaper is significantly concentrated in a given moment of the day (which is, quite obviously, the morning).

4.2 Distances between Different News Sources

By using the averaging Kendall distance, we can also define a distance between two different news sources (note that, in the previous case, we have used the averaging Kendall distance to define a distance between two different front pages of the same news source). Indeed, let $t_1, \ldots, t_n$ be the time instants in which we have downloaded the articles and let $f_i^S = (t_i, L_i^S)$ be the front page of the news source S taken at time t_i. The distance between two news

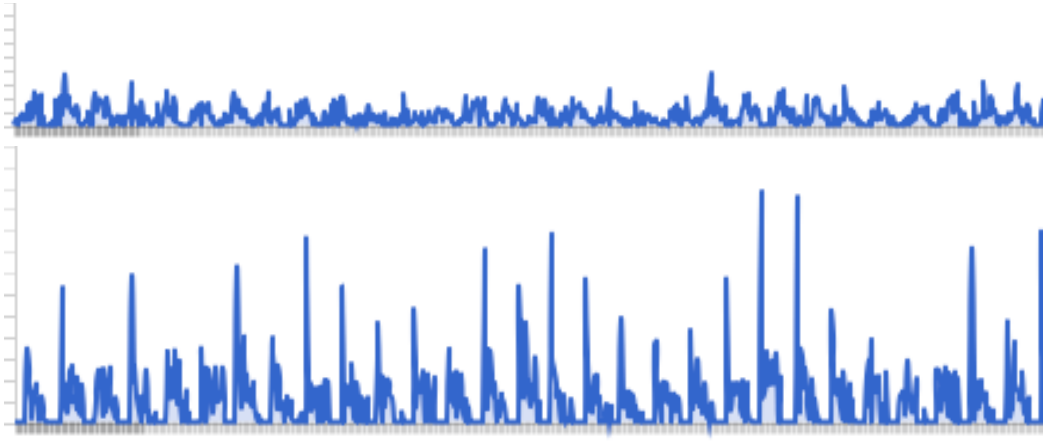

Figure 7 Analysing and comparing the activity of two different news sources.

sources S_1 and S_2 at time t_i is then defined as $K(L_i^{S_1}, L_i^{S_2})$, and the *distance* $d(S_1, S_2)$ the two news sources is the average of all such distances, that is,

$$d(S_1, S_2) = \frac{1}{n} \sum_{i=1}^{n} K(L_i^{S_1}, L_i^{S_2}).$$

Once we have computed these distances, we have then applied the well-known multi-dimensional technique [4] in order to plot the news sources on a two-dimensional plane. The result is shown in Figure 8. As it can be seen from the figure, the two news agencies (in black and grey) are in the centre of the plot and quite close each to the other. One of the four newspapers (in green) is also in the centre of the plot and quite close to both the two news agencies: this suggest that this newspaper acts more as an echo of the news agencies. The other three newspapers, instead, are quite far from the two news agencies and almost equidistant from them (even if at opposite sides): this might imply that these newspapers do indeed elaborate the new produced by the news agencies and present them in the front pages in different ways.

4.3 News Score

The front pages can also be used to determine the score of a news. Indeed, it is well known that a web reader usually reads a web page from top to bottom and from left to right. Hence, we can assume that an article, which appears first in the list of a front page, should receive a higher score than an article which appears after. We have decided to assign to each article a penalty proportional to its position in the front page. More formally, if an article a appears in position i of a front page f formed by n articles, then its *penalty* is equal to $\frac{i-1}{n}$. In other words, the score of this article is

$$\texttt{rel}(a, f) = \frac{n - i + 1}{n}.$$

We can agglomerate the score of the single articles and obtain the score of a news c with respect to a front page, as follows:

$$\texttt{rel}(c, f) = \sum_{a:N(a)=c} \texttt{rel}(a, f)\,.$$

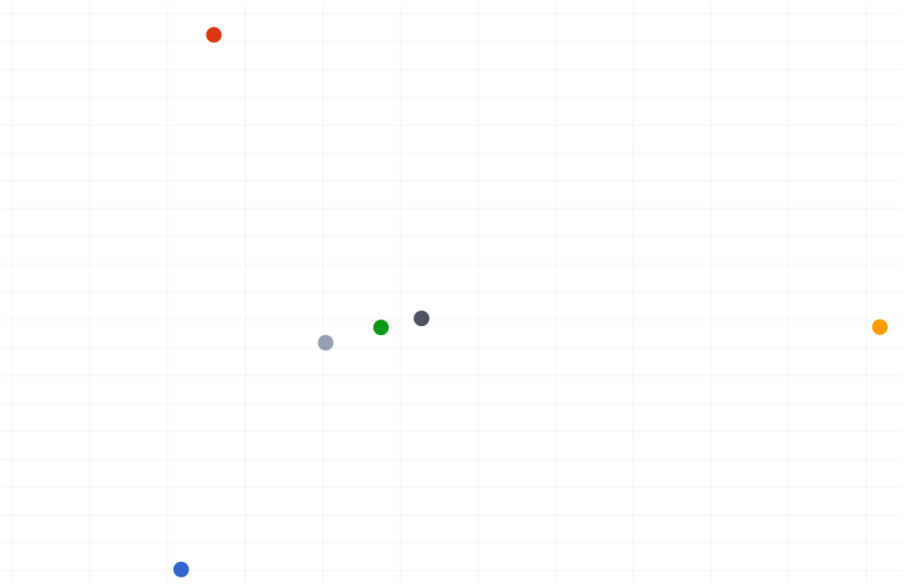

Figure 8 Plotting the news sources on the plane (the two news agencies are in black and in grey).

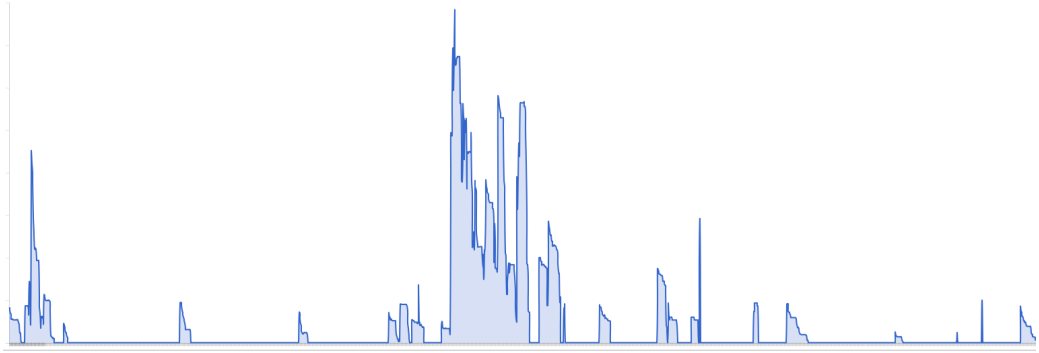

Figure 9 The score of the news concerning the shooting dwon of an American drone.

The total Score of a news is just the sum of its score with respect to all front pages. In Figure 9 we show the score values, during the last four months, of the most relevant news among the almost 7000 news obtained by clustering approximately 10000 articles. This news concerns the shooting down of an American drone in Turkey. As we can see from the figure, the score of this news has two main peaks: one peak corresponds to the news of a shooting down, while the second peak corresponds to the discover that the object shot down was an American drone. It is interesting to analyse this kind of figures because they help us to better understand the evolution of a news. For example, in Figure 10 the evolution of a news concerning a case of homicide in Italy is shown: in this case we have several peaks corresponding, more or less, to the disappearance of the victims, to the rising of the hypothesis of homicide, and to the identification of possible suspects.

4.4 Lifespan of a News

Finally, the news detection, performed by means of the clustering algorithm, allows us to estimate the lifespan of a news. In this case, we have first to define what the lifespan is.

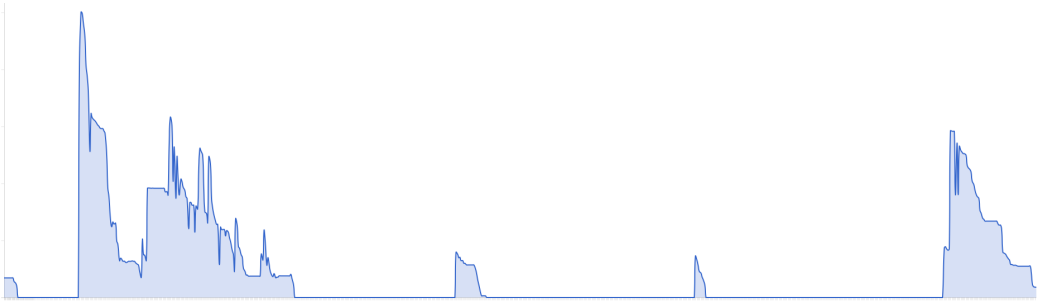

Figure 10 The score of the news concerning a case of homicide in Italy.

Indeed, it does not seem to be correct to consider the lifespan as the temporal interval between the first time the news has appeared and the last time in which it is still present in some news source, since during this interval there might have been long sub-intervals in which the news was not present at all. A typical example of this phenomenon is given by a news concerning a homicide. Typically, at the beginning the news is present for few hours, and then disappears. When some evolution in the investigation takes place, the news appears again and very fast disappears. Finally, when the killer is found, the news appears again and definitively disappears. All this can happen during a very long interval time, which cannot really be interpreted as the lifespan of the news, since during the vast majority of this interval the news was not present at all. We then decided to consider as the *lifespan* of a news as the effective time in which at least one article included in the news was present in at least one source of news. By using this definition, we computed the lifespan of the 7000 news produced by the algorithm with input the 10000 downloaded articles. It turned out that on average the lifespan of a news is approximately twenty hours: the maximum lifespan is a little bit more than 19 days, and it is reached by a news concerning vaccination campaigns inItaly, while the minimum lifespan is one hour, and it is reached, for example, by a news concerning the rescue of two bear cubs in Laos.

5 Improving the Clustering Time Performance

Despite of the very good performances in terms of precision and recall values, the incremental clustering algorithm described above has the disadvantage of degrading its performances while increasing the number of articles to be clustered. Indeed, any new article has to be compared with any other article already clustered in order to compute its average similarity with all the existing clusters. In order to improve the time performances of the clustering algorithm, we have then designed and implemented a two-layer variation of the previously described incremental algorithm. In this variation, each cluster of the first level is "represented" by its *centroid*, which is a non-existing article whose keyword set is the union of all the keyword sets of the articles in the cluster, and whose occurrence function is the average of all the occurrence functions. The centroids are themselves clustered by still using the same incremental approach (with maybe a different threshold value). More formally, the *centroid* of a set of articles A is defined as a "dummy" article m such that

$$\texttt{key}(m) = \bigcup_{a \in A} k(a)$$

and, for each $w \in \texttt{key}(m)$,

$$\texttt{tf}(w, m) = \frac{1}{|A|} \sum_{a \in A} \texttt{tf}(w, a).$$

A *two-layer clustering* is defined as a pair (N_1, N_2) of two functions $N_1 : A \to [n_1]$, where n_1 is the number of clusters in the first level, and $N_2 : C_1 \to [n_2]$, where C_1 is the set of centroids of the n_1 clusters at the first level and n_2 is the cardinality of the set C_2 of clusters in the second level. Given a two-layer clustering (N_1, N_2) and given a new article $x \notin A$, let c_2^* be the cluster $c_2 \in C_2$ for which $\texttt{sim}(x, c_2)$ is maximum: if $\texttt{sim}(x, c_2^*)$ is smaller than a given threshold τ_2, then $N_2(x) = n_2 + 1$ and $N_1(x) = n_1 + 1$ (that is, a new cluster is created both in the first and in the second level). Otherwise, the incremental clustering algorithm described in the previous section is applied to the sub-clustering of N_1 determined by the centroids included in c_2^*. If x is inserted in a cluster c_1^*, then a new centroid of c_1^* is computed and substituted inside the cluster in C_2 containing the previous centroid. Otherwise, a new cluster is created both in the first and in the second level (that is, $N_2(x) = n_2 + 1$ and $N_1(x) = n_1 + 1$).

5.1 Threshold estimation

In order to apply the above two-layer incremental clustering algorithm, we have to determine the two thresholds τ_1 and τ_2. To this aim, we have used again the manual clustering T of the first 3000 downloaded articles, and we have chosen the two values of τ_1 and τ_2 which produced the higher average F-measure. It turned out that the best value of τ_1 and τ_2 is equal to 0.19 and 0.48, respectively: with these values of τ_1 and τ_2 the average F-measure is equal to 0.8. As expected, the obtained average F-measure is lower than the one obtained in the one-layer clustering algorithm. However, the time performance improvement is quite impressive: the clustering of the 3000 articles was executed in less than 2 seconds, while the original approach required more than 2 minutes. Since the number of articles which are downloaded is increasing quite rapidly (approximately 200 articles every day), we believe that a little loss in precision is a reasonable price to be paid in order to make the clustering really efficient.

6 Conclusions

In this paper we have proposed a new tool for analysing and comparing on-line news sources. The main ingredients of this tool are a web scraping module, a news detection module based on a (two-layer) incremental clustering algorithm, and a statistical analysis module, which allowed us to answer several questions concerning the Italian on-line news system. Although the study is at the beginning and it is limited to the Italian news system, the results are quite encouraging: some of them are surprising (for instance, the average lifespan of a news), others can be seen just as curiosities, as for instance the news with minimal lifespan (that is, the rescue of two bear cubs in Laos) or which are the most relevant news according to their position in the site, compared with the truly most relevant ones. Another curiosity is to know that some of the online newspapers add very little information to their news, that result almost identical to that of the news agencies.

Apart from integrating our tool with other statistical analysis, we think that the most interesting possible future research directions concerns the possibility of improving our incremental clustering algorithm. Indeed, this can be done either by considering a k-level clustering algorithm with $k > 2$, or by using the estimated average lifespan of the news

in order to eliminate from the clustering execution all clusters (that is, news) which are sufficiently old.

Acknowledgements. We would like to thank Alice Marchetti and Andrea Marino to suggest us to include some descriptive statistics of the collection of downloaded articles. The second and third authors received additional support from the Italian Ministry of Education, University and Research, under PRIN 2012C4E3KT National research project AMANDA.

References

1 J. Azzopardi and C. Staff. Incremental Clustering of News Reports. *Algorithms*, 5:364–378, 2012.
2 D. Bhattacharya and S. Ram. Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 966–971, 2012.
3 Jon Borglund. Event-Centric Clustering of News Articles. Technical report, Department of Information Technology, University of Uppsala, 2013.
4 T.F. Cox and M.A.A. Cox. *Multidimensional Scaling (2nd ed.)*. Chapman and Hall, 2000.
5 S. Edunov, C.G. Diuk, I.O. Filiz, S. Bhagat, and M. Burke. Three and a half degrees of separation, 2016. URL: `http://research.facebook.com/blog/`.
6 R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top K Lists. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 28–36, 2003.
7 M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Edward Arnold, 1990.
8 J. Leskovec, A. Rajaraman, and J.D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2014.
9 Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
10 J.B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.
11 Parse.ly. What is the Lifespan of an Article?, 2015. URL: `http://parsely.com`.
12 G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Two-level Message Clustering for Topic Detection in Twitter. In *SNOW 2014 Data Challenge co-located with 23rd International World Wide Web Conference*, pages 49–56, 2014.
13 Wikipedia – News Agency. URL: `https://en.wikipedia.org/wiki/News_agency`.