Ranking a Stream of News[*]

Gianna M. Del Corso
Dipartimento di Informatica,
Univeristy of Pisa
delcorso@di.unipi.it

Antonio Gullí[†]
Dipartimento di Informatica,
Univeristy of Pisa, IIT-CNR
gulli@di.unipi.it

Francesco Romani
Dipartimento di Informatica,
Univeristy of Pisa
romani@di.unipi.it

ABSTRACT

According to a recent survey made by Nielsen NetRatings, searching on news articles is one of the most important activity online. Indeed, Google, Yahoo, MSN and many others have proposed commercial search engines for indexing news feeds. Despite this commercial interest, no academic research has focused on ranking a stream of news articles and a set of news sources. In this paper, we introduce this problem by proposing a ranking framework which models: (1) the process of generation of a stream of news articles, (2) the news articles clustering by topics, and (3) the evolution of news story over the time. The ranking algorithm proposed ranks news information, finding the most authoritative news sources and identifying the most interesting events in the different categories to which news article belongs. All these ranking measures take in account the time and can be obtained without a predefined sliding window of observation over the stream. The complexity of our algorithm is linear in the number of pieces of news still under consideration at the time of a new posting. This allow a continuous on-line process of ranking. Our ranking framework is validated on a collection of more than 300,000 pieces of news, produced in two months by more then 2000 news sources belonging to 13 different categories (World, U.S, Europe, Sports, Business, etc). This collection is extracted from the index of comeToMyHead, an academic news search engine available online.

Categories and Subject Descriptors

H.3.1 [**Information Storage And Retrieval**]: Content Analysis and IndexingRetrieval models; Search process; H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage And Retrieval**]: OnlineInformation Services

General Terms

Algorithms, Experimentation.

Keywords

News Engines, Information Extraction, News Ranking.

1. INTRODUCTION

In the last year there has been a surge of interest about news engines, i.e. software tools for gathering, indexing, searching, clustering and delivering personalized news information to Web users. According to a recent survey made by Nielsen NetRatings [20, 24], news browsing and searching is one of the most important Internet activities with more than 28 millions of active U.S. users in October 2004 (see Figure 1). For instance, Yahoo! News had an audience which is roughly the half of Yahoo! Web Search, a third of Google Web Search and a bit more than AOL Web Search. This is surprising enough if we consider that, for instance, Yahoo News had an audience of about 13 millions users in "the" 2002 [20]. "The Internet complements television for news coverage as it provides a different perspective and greater depth of information - statistics, pictures, interactive maps, streaming video, and analyst comments," said Peter Steyn of Nielsen/Netrating. Certainly, recent events such as SARS, War in Iraq, Terrorism Alerts and other similar dramatic events contributed to diffuse the use of online news search engines. The huge amount of news articles available online reflects the users' need for a plurality of information and opinions. News engines are, then, a direct link to fresh and unfiltered information.

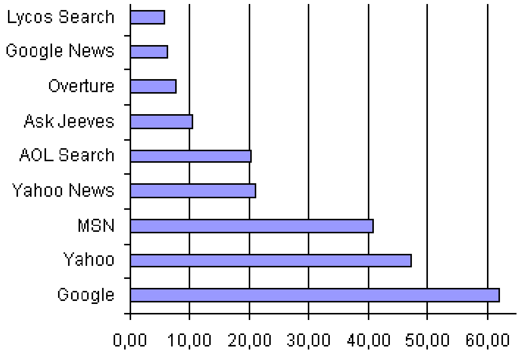


Figure 1: Comparing News and Web Search Engines (October 2004, Nielsen/Netratings).

The commercial scenario.

Many commercial news engines are already available such as Google News [22], Yahoo News [30], MSNBot [23], Findory [21] and NewsInEssence [26]. Google News retrieves news information by more than 4,000 sources, organizes it in cat-

[*] Partially supported by the GNCS-INDAM Project: "Problematiche Numeriche nel WEB"

[†] Partially supported by the Italian Registry of ccTLD.it.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.

egories and automatically builds a page with the most important news articles for each category. Besides, it clusters similar pieces of news. Yahoo news runs analogous services on more than 5,000 sources. Microsoft recently announced its NewsBot, a news engine that provides personalized news browsing according to different profiles built for each user. Findory proposes a similar personalized service, which relies on patent pending algorithms. Another important news engine is NewsInEssence, which clusters and summarizes similar news articles. A complete list of commercial news engine is given in [29]. There is no public available information about the way in which these commercial search engines rank news articles. Nevertheless, an extensive testing performed by the authors of this paper on these systems showed anecdotal evidences that they take in account several criteria such as freshness, news sources authoritativeness and replications/aggregation of pieces of news. In this paper we introduce a framework which also exploits these criteria.

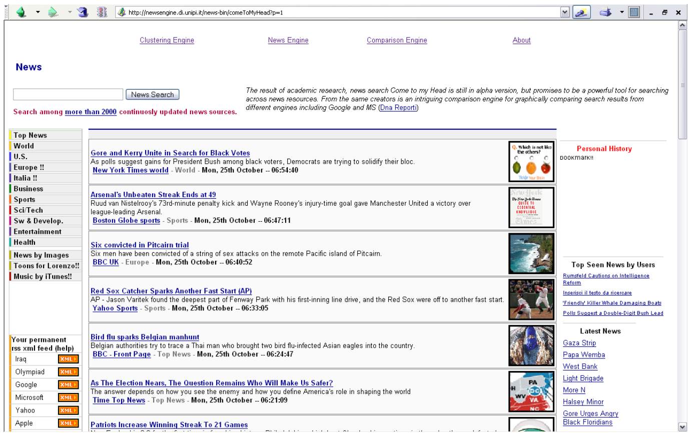
The scientific scenario.

Despite this great variety of commercial solutions for news search engines, we found just a few papers on this subject [4, 5, 7, 8, 10, 6]. NewsInEssence [4, 5] is a system for finding and summarizing clusters of related news articles from multiple sources on the Web. The system aims to generate automatically summaries of news events by using a centroid based summarization technique. It considers salient terms forming the cluster of related documents, and uses these terms to construct a cluster summary. QCS [7] is a software tool and development framework for streamlined IR. The system matches a query to relevant documents, clusters the resulting subset of documents by topic, and produces a single summary for each topic. The main goal of the above works is to create summaries of clustered news articles. In [3] a topic mining framework for news data stream is proposed. In [10] the authors study the problem of finding news articles on the web that are relevant to the ongoing stream of TV broadcast news. In [6] a tool to automatically extracting news from Web sites is proposed. In [8] is proposed and analyzed NewsJunkie, a system that personalizes news articles for users by identifying the novelty of stories in the context of stories users have already reviewed.

Mannilla et al. in [13] introduced the problem of finding frequent episodes in event sequences, subject to observation-window constrain, where an episode is defined as a partially ordered collections of events, and can be represented as a directed acyclic graph. In [2] Atallah et al. proposed an extension of [13] to rank a collection of episodes according to their significance. We remark that the concept of episode does not take into account the entities which produced the episode itself and how episodes aggregate each others. In this paper, we show that these are crucial features for ranking news stories.

The news engine.

comeToMyHead is an academic news search engine available at http://newsengine.di.unipi.it/ for gathering, indexing, searching, clustering and delivering personalized news information to Web users. This engine is a running software prototype developed by our research group to investigate many different aspects of News engines. In the context of this paper, we have used this search engine to gather a collection of news articles from many different sources over a period of two months. Our experimental settings are based on the news data collected by comeToMyHead in two months by more than 2000 news sources classified in 13 different categories, and consists of about 300,000 pieces of news. Besides, we are currently integrating the ranking strategies proposed in this paper into the production version of the engine.



Search among more than 2000 continuosly updated news sources.

The result of academic research, news search Come to my Head is still in alpha version, but promises to be a powerful tool for searching across news resources. From the same creators is an intriguing comparison engine for graphically comparing search results from different engines including Google and MS (Dna Reporti)

Gore and Kerry Unite in Search for Black Votes
As polls suggest gains for President Bush among black voters, Democrats are trying to solidify their bloc.
New York Times world - World - Mon, 25th October -- 06:54:40

Arsenal's Unbeaten Streak Ends at 49
Ruud van Nistelrooy's 73rd-minute penalty kick and Wayne Rooney's injury-time goal gave Manchester United a victory over league-leading Arsenal.
Boston Globe sports - Sports - Mon, 25th October -- 06:47:11

Six convicted in Pitcairn trial
Six men have been convicted of a string of sex attacks on the remote Pacific island of Pitcairn.
BBC UK - Europe - Mon, 25th October -- 06:40:52

Red Sox Catcher Sparks Another Fast Start (AP)
AP - Jason Varitek found the deepest part of Fenway Park with his first-inning line drive, and the Red Sox were off to another fast start.
Yahoo Sports - Sports - Mon, 25th October -- 06:33:05

Bird flu sparks Belgian manhunt
Belgian authorities try to trace a Thai man who brought two bird flu-infected Asian eagles into the country.
BBC - Front Page - Top News - Mon, 25th October -- 06:24:47

As The Election Nears, The Question Remains Who Will Make Us Safer?
The answer depends on how you see the enemy and how you define America's role in shaping the world
Time Top News - Top News - Mon, 25th October -- 06:21:09

Patriots Increase Winning Streak To 21 Games

Figure 2: The comeToMyHead News Engine.

2. OUR CONTRIBUTION

In this paper we discuss the problem of ranking news sources and a stream of news information evolving during the time. To the best of our knowledge this is the first academic paper on this subject, hence we do not have the possibility to compare our results with other ranking methods. For this reason we had to formalize the problem describing a number of desirable properties we ask to our ranking scheme (Section 3) and to introduce a suitable model for describing interactions between articles and news sources (Section 4). The ranking algorithm is obtained introducing progressively a number of constraints to match the requested properties and is validated on two intuitive limit cases, which allows us to rule out more intuitive approaches (Section 5). The final algorithm is described in Section 6. It works online by ranking each piece of news at the time of its emission. It can also influence the rank of the news sources. The complexity of our method is linear with the number of news articles still of interest at a particular time of observation.

Our ranking scheme depends on two parameters, ρ accounting for the decay rate of freshness of news articles, and β which gives us the amount of source's rank we want to transfer to each posted piece of news. We studied the sensitivity of the ranks obtained varying these parameters and we saw that our algorithm is robust, in the sense that the correlation between ranks remains high changing the decay rule and the parameter β.

A large experimentation was performed, and in Section 7 we present some of these results. The results obtained ranking news articles and news sources for each category confirm the ability of our method to recognize the most authoritative sources and to assign an high rank to important pieces of news.

The algorithms proposed in this paper aim to a general ranking schema based on unbiased factors rather then personal consideration like that topic of interest for the user or even ideology. Like in web search ranking scheme, it is possible to extend our approach introducing a personalization parameter accounting for the personal taste of the user.

3. SOME DESIDERATA

Ranking news articles is a rather different task than ranking Web pages. From one side, we can expect a smaller amount of spam since news stories come from controlled sources. When a piece of news is issued, we can have two different scenarios: the news article can be completely independent on the already published stories, or can be aggregated to a (set of) news articles previously posted. Anyway, we stress that, by definition, a news article is a fresh piece of information. For this reason, when a news article is posted there is almost no HTML link pointing to it. Therefore, HTML link based analysis techniques, such as PageRank [15], can produce a limited benefit for news ranking. In Section 4 we propose a model which exploits a *virtual linking* relationship between pieces of news and news sources based both on the news posting process and on the natural aggregation by topics between different news stories. Now, we discuss some desirable properties of ranking algorithms for news articles and news sources before presenting the algorithms designed to match these requests.

Property P1: *Ranking for News posting and News sources.* The algorithms should assign a separate rank for news articles and news sources.

Property P2: *Important News articles are Clustered.* An important news story n is probably (partially) replicated by many sources. For instance, consider a news article n originated by a press agency. The measure of its importance is also expressed by the number of different online newspapers which replicate n or extract parts of text from n. The phenomenon of citing stories released by other sources is common in the context of (Web) journals. From the news engine point of view, this means that the (weighted) size of the cluster formed around n is a measure of its importance.

Property P3: *Mutual Reinforcement between News Articles and News Sources.* We can assign different importance to different news sources according to the importance of the news articles they produce. So that, a piece of news coming from "Washington Post" can be more authoritative than a similar article coming from say "ACME press", since "Washington Post" is known for producing good stories.

Property P4: *Time awareness.* The importance of a piece of news changes over the time. We are dealing with a stream of information where a fresh news story should be considered more important than an old one.

Property P5: *Online processing.* We require that the time and space complexity of the ranking algorithm allows online processing, i.e. at some time the complexity can depend on the mean amount of news articles arriving but not on the time since the observation started.

In Section 6 we define an algorithm for ranking news articles and news sources which match the above properties. The algorithm is progressively designed ruling out easier algorithms which do not satisfy some of the above requirements.

4. A MODEL FOR NEWS ARTICLES

News posting can be thought as a continuous stream process. For dealing with it, we can exploit a window of observation. A first way to analyze the stream, is to have a window of fixed size. In this way the maximum size of observed data is constant, but we can miss the opportunity to discover temporal relationship between news articles posted at a time not covered by the current window. A second way is to use an unbounded time window of observation. Of course, by adopting this method the size of the observed data increases with the time. This is a typical situation with data streaming problems where the flow of information is so overwhelming that it is unfeasible even to store the data or to perform a single (or more than one) scan operation(s) over the data (see [14] and references therein). This is particularly true for information flows, since different news sources can post independently many stream of news articles. In Section 5.2 we propose a solution to this problem. This solution handles the data stream of news information with no predefined time window of observation. The solution takes in account a particular decay function associated to any given piece of news. The algorithms proposed turn out to be tunable, in the sense that we can change the decay parameters according to the categories in which the news posting is classified.

In the following, we introduce the model which characterizes news articles and news sources. Given a news stream, a set of news sources, and fixing a time window ω, the news creation process can be represented by means of a undirected graph $G_\omega = (V, E)$ where $V = S \cup N$ and S are the nodes representing the news sources, while N are the nodes representing the news stream seen in the time window ω. Analogously, the set of edges E is partitioned in two disjoint sets E_1 and E_2. E_1 is the set of undirected edges between S and N. It represents the news creation process, E_2 is the set of undirected edges with both endpoints in N and it represents the results of the clustering process which allows to connect similar pieces of news. The edges in E_2 can be annotated with weights which represent the similarity between two pieces of news. The nodes in S "cover" those in N, i.e., $\forall n \in N, \exists\ s \in S$ such that $(s, n) \in E_1$.

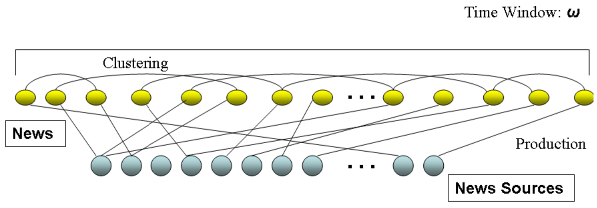


Figure 3: News Ranking Graph.

To satisfy the property (P2), we define a similarity measure among the news articles, which depends on the clustering algorithm chosen and accounts for the similarity among the news stories. Given two nodes n_i and n_j we define the *continuous similarity* measure as a real value $\sigma_{ij} \in [0,1]$, with the meaning that σ_{ij} is close to 1 if n_i is similar to n_j. A simplified version provides a *discrete similarity* measure, which holds 1 if the two news postings are exactly the same (in other words, they are mirrored) and 0 if they are different.

Let A be the (weighted) adjacency matrix associated with G_ω. We can attribute an identifier to the nodes in G_ω so that any source precedes the pieces of news. We define the matrix

$$A = \begin{bmatrix} O & B \\ B^T & \Sigma \end{bmatrix},$$

where B refers to edges from sources to news articles, and $b_{ij} = 1$ iff the source s_i emitted article n_j and Σ is the similarity matrix. Assuming one can learn similarity of sources, the matrix A can be modified in the upper-left corner incorporating a submatrix taking into account a source-source information.

An important parameter of a news engine is the amount of articles emitted in a short period of time from all the sources in a given category. This quantity, denoted by $\mathsf{newsflow}(\mathsf{t},\mathsf{c})$ for time t and category c, is subject to drastic variation over the time as a consequence of great resonance events (for instance, during the first days of November 2004 we had a peak in $\mathsf{newsflow}$ for category "U.S." due to the Presidential Election).

We remark that this model describes a framework where one can plug-in different data stream clustering algorithms (see [1, 9] and the references therein) for creating and weighting the set of edges E_2. Starting from the above model, in Section 5 we propose some ranking algorithms which progressively satisfy the properties described in Section 3, and fit the general model for representing news articles and news sources described here.

5. ALGORITHMS FOR NEWS ARTICLE AND NEWS SOURCES

To evaluate the consistence of the algorithms presented in this section, we consider some limit cases for which the algorithms should show a reasonable behavior. These limit cases allow us to refine the algorithms and match the properties described in Section 3. They are:

- `LC1`: A unique source s_1 emits a stream of independent news articles with average emission rate $1/\Delta$. We expect the source to have a stationary mean value rank μ independent of the time and the size of the observation window ω. μ should be an increasing function in $1/\Delta$.
- `LC2`: Two news sources s_1, s_2, where s_1 produces a stream of independent news articles with average rate $1/\Delta$, and s_2 re-posting the same news stream generated by s_1 with a given average delay. Essentially, the source s_2 is a mirror of s_1. Hence, the two sources should have a similar rank.

5.1 Non-Time-aware Ranking Algorithms

Any algorithm described in this section satisfies only a subset of the properties described in Section 3. Indeed, they are naive approaches that one has to rule out before proposing more sophisticated algorithms. In particular, these methods do not deal with the news flow as a data stream, but assumes that they are available as a static data set. In the next section we introduce algorithms which overcome the limit of those given here.

```
Algorithm NTA1
```

The naive approach is that a news source has a rank proportional to the number of pieces of news it generates and, conversely, that a news article should rank high if there are many other news stories close to it. Formally, denoting by $\mathbf{r} = [\mathbf{r}_S, \mathbf{r}_N]^T$ the vector of sources and news ranks, we can compute them as

$$\mathbf{r} = A\mathbf{u},$$

where $\mathbf{u} = [\mathbf{u}_S, \mathbf{u}_N]^T$ is the vector with all entries equal to one. Given the structure of A, this means that

$$\begin{cases} \mathbf{r}_S = B\mathbf{u}_N, \text{ and} \\ \mathbf{r}_N = B^T\mathbf{u}_S + \Sigma\,\mathbf{u}_N = \mathbf{u}_S + \Sigma\,\mathbf{u}_N, \end{cases}$$

that is each source receives a rank equal to the number of news articles emitted by that source, while the single piece of news has a rank proportional to the number of similar news articles.

This algorithm shows a bad behavior in the limit case `LC1`. Indeed, the rank r_{s_1} of a unique news source s_1, will increase unbounded with the number of observed news articles. Besides, algorithm `NTA1` satisfies the properties (P1) and (P2) but not (P3), (P4) and (P5).

```
Algorithm NTA2
```

The second algorithm exploits the mutual reinforcement property between news articles and news sources similarly to the way HITS algorithm [12] identifies Web hubs and authorities. Let us consider the fixed point equation

$$\mathbf{r} = A\mathbf{r}. \tag{1}$$

From the block structure of A we get

$$\begin{cases} \mathbf{r}_S = B\mathbf{r}_N \\ \mathbf{r}_N = B^T\mathbf{r}_S + \Sigma\,\mathbf{r}_N. \end{cases}$$

From equation (1), it turns out that in order to have a nonzero solution, $\mathbf{r}$ should be a right eigenvector corresponding to an eigenvalue equal to 1, but this is not true in general. In particular, this does not hold for case `LC1` and $\mathbf{r} = \mathbf{0}$ is the only solution of (1). This algorithm is also not stream oriented like the NTA1. A major difference with NTA1 is that NTA2 satisfy the properties (P1), (P2) and (P3).

It is easy to show that the class of non time-aware algorithms do not satisfy at least one of the limit cases defined in Section 5.

Moreover, the fixed time-window scheme can not explore precise temporal information within a window, and misses the opportunity to discover temporal relationship between news articles released at a time not covered by the current window.

5.2 Time-Aware Ranking Algorithms

To deal with a news data stream we have to design time-aware mechanisms, which do not use fixed time observation windows over the flow of information. The key idea is that the importance of a piece of news is strictly related to the time of his emission. Hence, we model this phenomenon introducing a parameter α which accounts for the decay of "freshness" of the news story. This α depends on the category to which the news article belongs. For instance, it is usually a good idea to consider sport news decaying more rapidly than health news.

We denote by $R(n,t)$ the rank of news article n at time t, and analogously, $R(s,t)$ is the rank of source s at time t. Moreover, by $S(n_i) = s_k$ we mean that n_i has been posted by source s_k.

Decay rule: We adopt the following exponential decay rule for the rank of n_i which has been released at time t_i:

$$R(n_i,t) = e^{-\alpha(t-t_i)}R(n_i,t_i), \quad t > t_i. \tag{2}$$

The value α is obtained from the half-life decay time ρ, that is the time required by the rank to halve its value, with the relation $e^{-\alpha\rho} = \frac{1}{2}$. In the following, we will specify the parameter ρ, expressed in hours, instead of α. Besides, we discuss how to obtain the formulation of an effective algorithm for ranking news articles and sources. We show that naive time-aware algorithms show a bad behavior in many cases, then we refine them in order to have a complete control of the ranking process.

Algorithms TA1

The first class of time-aware algorithms assigns to a news source the sum of the ranks of the news information generated by that source in the past, according to the above decay rule. The algorithms belonging to this class differs from each other only for the way of ranking each news article at the time of its first posting.

Setting to one the rank of a news article at the time of its initial posting, we have

$$\begin{cases} R(s_k,t) = \sum_{S(n_i)=s_k} R(n_i,t) \\ R(n_i,t_i) = 1. \end{cases} \tag{3}$$

Assuming that the source s_k did not post any news information in the interval $[t, t+\tau]$, we have that the variation of ranks after an elapsed time of τ is described by the two following relations

$$\begin{aligned} R(n_i,t+\tau) &= e^{-\alpha\tau}R(n_i,t), \quad t \geq t_i \qquad (4)\\ R(s_k,t+\tau) &= e^{-\alpha\tau}R(s_k,t), \end{aligned}$$

We note that this algorithm attenuates the effect of previously issued news articles, and it meets the limit case LC1. Indeed, assuming case LC1 is satisfied, for the stationary mean value μ of the rank of s_1, we have

$$\mu = \theta\mu + 1, \tag{5}$$

where $\theta = e^{-\alpha\Delta}$. From (5) we derive the mean value of the rank $\mu = 1/(1-\theta)$ in the case of a single source emitting independent news articles with average rate $1/\Delta$. We point out that this algorithm satisfies Properties (P1), (P4) and (P5) but it does not satisfy (P3) since the rank attributed to a news article does not depend on the rank of the source which posted it.

For accounting Property (P3), we can still consider equation (3), changing the rank attributed to a piece of news when it is released. For instance, we can define the rank of a news story as a portion of the rank of its source just an instant before emitting it. The algorithm becomes

$$\begin{cases} R(s_k,t) = \sum_{S(n_i)=s_k} R(n_i,t), \\ R(n_i,t_i) = c \lim_{\tau\to 0^+} R(S(n_i), t_i - \tau), \end{cases}$$

where $0 < c < 1$. As a starting point we assume $R(s_k, t_0) = 1$, however, with any non-zero initial conditions the limit case LC1 has again a bad behavior. There is no stationary mean value of the rank even for a single source s_1 emitting a stream of independent news articles. In fact, assuming μ to be the stationary mean value of $R(s,t)$, we have

$$\mu = \theta\mu + c\,\theta\mu,$$

which cannot be solved for $\mu \neq 0$.

To solve the problem, we change again the starting point in (3) to smooth the influence of the news source on the rank of the news articles. Let us set

$$R(n_i,t_i) = \left[\lim_{\tau\to 0^+} R\left(S(n_i), t_i - \tau\right)\right]^{\beta}, \quad 0 < \beta < 1.$$

The parameter β is similar to the magic ε accounting for the random jump in Google's PageRank [15]. In fact, as for the random jump probability, the presence of β is here motivated both by a mathematical and a practical reason. From a mathematical view point, the fixed point equation involving the sources, has a non null solution. From a practical point of view, by changing β we can tune how much the arrival of a single fresh piece of news can increase the rank of a news source. In fact, let t_{i-1} be the time of emission of the previous news article from source s_k, and let t_i be the time of release of n_i by s_k. If in the interval (t_{i-1}, t_i) no article has been issued by s_k, we have

$$R(s_k,t_i) = e^{-\alpha(t_i-t_{i-1})}R(s_k,t_{i-1}) + R(s_k,t_{i-1})^{\beta}.$$

For the limit case LC1 the fixed point equation now becomes

$$\mu = \theta\mu + (\theta\mu)^{\beta}$$

which has the solution $\mu = \left(\frac{\theta^{\beta}}{1-\theta}\right)^{\frac{1}{1-\beta}}$. In this model we can also deal very easily with the limit case LC2.

Algorithm TA2

We have seen that the algorithms in the class TA1 satisfy the limit cases and the Properties (P1), (P3), (P4) and (P5). However, it does not satisfy the Property (P2) since the rank of a news article is not related to the rank of similar ones. This is a desired property since if an article is known to be of interest there will be a large number of news sources which will post similar pieces of information. Therefore, a good news ranking algorithm working over a stream of information should also exploit some data stream clustering technique. Formally, this can be described as follows. Let us set the rank of a piece of news at emission time to be

$$\begin{aligned} R(n_i,t_i) &= \left[\lim_{\tau\to 0^+} R\left(S(n_i), t_i-\tau\right)\right]^{\beta} + \\ &+ \sum_{t_j<t_i} e^{-\alpha(t_i-t_j)}\sigma_{ij} R(n_j,t_j)^{\beta} \end{aligned} \tag{6}$$

where $0 < \beta < 1$. In this case the rank of an article is dependent on the rank of the source and on the rank of similar news articles issued previously whose importance has already decayed of a negative exponential factor. The rank of sources is still

$$R(s_k,t) = \sum_{S(n_i)=s_k} R(n_i,t).$$

Unfortunately, studying the behavior of this algorithm on the limit cases LC2 we obtain that a news source mirroring another, gets a finite rank significantly greater than the rank of the mirrored one.

6. THE FINAL TIME AWARE ALGORITHM: TA3

In order to fix the behavior of the formula assigning ranks to news sources and dealing with the limit case LC2, we modify "a posteriori" the rank of a mirrored source. In particular, a source which has emitted in the past news stories highly mirrored in the future, will receive a "bonus" acknowledging the importance. The final equation for news sources and news stream becomes

$$\begin{aligned} R(s_k,t) &= \sum_{S(n_i)=s_k} e^{-\alpha(t-t_i)} R(n_i,t) + \\ &+ \sum_{S(n_i)=s_k} e^{-\alpha(t-t_i)} \sum_{\substack{t_j > t_i \\ S(n_i)\neq s_k}} \sigma_{ij} R(n_j,t_j)^{\beta}, \\ R(n_i,t_i) &= \left[\lim_{\tau\to 0^+} R\left(S(n_i), t_i-\tau\right)\right]^{\beta} + \\ &+ \sum_{t_j<t_i} e^{-\alpha(t_i-t_j)}\sigma_{ij} R(n_j,t_j)^{\beta}. \end{aligned} \tag{7}$$

The rank of a news source s_k is then given by the ranks of the piece of news generated in the past, plus a factor of the rank of news articles similar to those issued by s_k and posted later on by other sources. The equation for ranking the articles remains the same (see equation 6). Note that if an article n aggregates with a set of pieces of news posted in the future, we do not assign to n an extra bonus (acknowledging a posteriori the importance of n). The idea is that we want to privilege the freshness of news posting instead of its clustering importance. However, the news source which first posted an highly aggregating article is awarded of an extra rank, because that news source made a scoop (in journalistic jargon).

This algorithm is coherent with all the desirable properties described in Section 3 but it is more complicated than those analyzed in previous sections, and it is not easy to write down a formula for the stationary mean value of the source. However, as shown in Figure 4, limit cases LC1 and LC2 are satisfied.

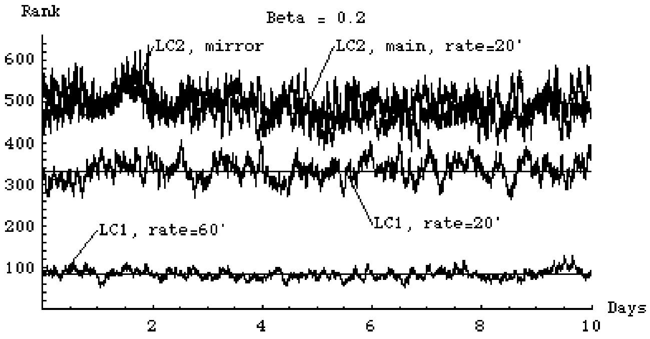


Figure 4: Simulated behavior of the limit cases LC1 and LC2 with $\beta = 0.2$. From below, the two straight lines represent the theoretical values of LC1 with a decay rate ρ of 60 min and of 20 min. There is a good agreement between theoretical and actual values of source ranks. In the upper part the ranks of two sources emitting the same news stream are plotted.

6.1 Clustering Technique

The naive clustering used in COMETOMYHEAD set $\sigma_{ij} = 1$ if n_i and n_j are the same, (i.e. they are mirrored). In our news collection, these cases where very limited. Hence, by using these values of σ_{ij} the result of news sources ranking is highly correlated with the simple counting of the posted news articles. A more significant indication can be obtained by taking a continuous measure of the lexical similarity between the abstracts of the news posting. These abstracts are directly extracted by the index of the news engine itself. In our current implementation, the news abstract are represented using the canonical "bag of words" representation. These abstracts are filtered out against a list of stop words. The lexical similarity is, then, expressed as a function of the words in common between news abstracts. We remark, that dealing with a continuous similarity measure produces a matrix Σ full and whose dimension increases over the time. Fortunately, the decay rule allows us to consider only the more recently produced part of the matrix, keeping it with a size proportional to the $\mathsf{newsflow(t,c)}$, and therefore satisfying the Property (P5).

6.2 Ranking the Events

An interesting feature of our algorithm is the possibility to analyze the behavior of the mean value of the ranks of all the sources, over the time and for each given category. This measure gives us an idea of the activity of that category and is related with particularly relevant events. In particular, we define the mean value of the rank of all the sources at a given time t, that is

$$\mu(t) = \frac{\sum_{s_k\in S} R(s_k,t)}{|S|}. \tag{8}$$

In Section 7 we discuss this mean value for a particular category.

7. EXPERIMENTAL SETTINGS

We performed our experiments on a PC with a Pentium IV 3GHz, 2.0GB of memory and 512Kb of L2 cache. For space reason, we report just the most important results. The interested reader can ask the authors for a more extensive testing. The code is written in Java and the ranking of about 20,000 news pieces requires few minutes, including the computation done by our clustering algorithm.

For evaluating the quality of results, we used the data set collected by ComeToMyHead an academic News Search engine, gathering news articles from more than 2000 continuously updated sources. The data set consists of about 300,000 pieces of news collected over a period of two months (from 8/07/04 to 10/11/04) and classified in 13 different categories (see Figure 5, 6). Each article n is uniquely identified by a triple $< u, c, s >$, where u is the URL where the news article is located, c is its category, and s is the news source which produced n. The data set is searchable online at http://newsengine.di.unipi.it.

To allow our ranking algorithm to achieve a stationary behavior, all the experiments, the measurements start from 8/17/04, discarding the first 10 days of observation.

Category	# Postings	Category	# Postings
Business	34547	Entertainment	43957
Europe	19000	Health	11190
Italia	7865	Music Feeds	690
Sci/Tech	25562	Software & Dev.	2356
Sports	39033	Toons[1]	1405
Top News	54904	U.S.	10089
World	53422		

Figure 5: How the news postings gathered in two months by comeToMyHead distribute among the 13 categories.

Category	# Sources	Category	# Sources
Business	1256	Entertainment	1970
Europe	5	Health	1080
Italia	312	Music Feeds	1
Sci/Tech	1108	Software & Dev.	17
Sports	1316	Toons	15
Top News	8	U.S.	239
World	974		

Figure 6: The number of news sources for the 13 categories (gathered by the comeToMyHead).

Sensitivity to the parameters

A first group of experiments addressed the sensitivity at changes of the parameters. We recall that our ranking scheme depends on two parameters, ρ, accounting for the decay rate of freshness of news articles, and β, which gives us the amount of source's rank we want to transfer to each news posting. As a measure of concordance between the ranks produced with different values of the parameters, we adopted the well known Spearman [16] and Kendall-Tau correlations [11]. We report the ranks computed for the category "World" with algorithm TA3, for values of $\beta_i = i/10$, where $i = 1, 2, \ldots, 9$ and for $\rho = 12$ hours, 24 hours and 48 hours. In Figure 7, for a fixed ρ the abscissa β_i represents the correlation between the ranks obtained with values β_i and β_{i-1}. From this plot we can see that Kendall-Tau correlation is a more sensitive measure than Spearman correlation, and that the algorithm is not much sensitive to changing in the parameters involved. This is a nice property since we do not have a way to establish the optimal choice of these parameters.

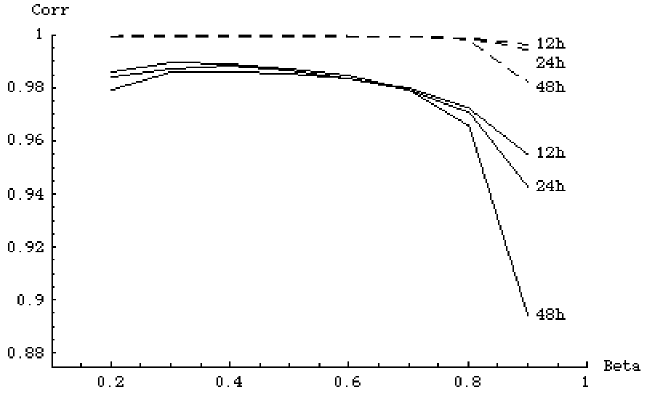


Figure 7: For the category "World", the figure represents the correlations between ranks of news sources obtained with two successive values of β differing for 0.1. The solid lines are the Kendall-Tau measure, the dashed lines are the Spearman correlation coefficients.

It is very important also to compare the source rank obtained with our algorithm with the one obtained with a simpler schema. For this reason, we compare the mean source ranks over the observed period generated with algorithm TA3 with the naive rank obtained using method NTA1. We recall that NTA1 assigns to a source a rank equal to the number of news posted. A matrix of Kendall-Tau correlation values is obtained comparing the two ranks with β varying from 0.1 to 0.9 and for ρ varying from 5 hours to 54 hours. In Figure 8 this matrix is plotted as a 3-D graph. The correlation values show how the algorithm TA3 differentiates from the naive NTA1.

Ranking news articles and news sources

The second group of experiments addresses the principal goal of the paper, i.e. the problem of ranking news articles and news sources. Figure 9 shows the evolution of the rank over a period of 55 days of the top four sources in the category "World". The two plot are obtained choosing $\beta = 0.5$ and for two choices of the half-life decay time, that is $\rho = 24$ and 48 hours. RedNova [27] results the most authoritative source, followed by Yahoo! World[30], Reuters World [28] and BBC News World [17][2]. We observed that the most authoritative sources remains the same changing both ρ and β.

In Figure 10 we report the top ten news source for the category "World" returned by our algorithm setting $\rho = 24$ hours and $\beta = 0.2$. Note that "Yahoo Politics" is considered more important than "BBC News world" due to the

[2] We remark that these ranks express the results of a computer algorithm, and they do not express any opinion of the authors of this paper.

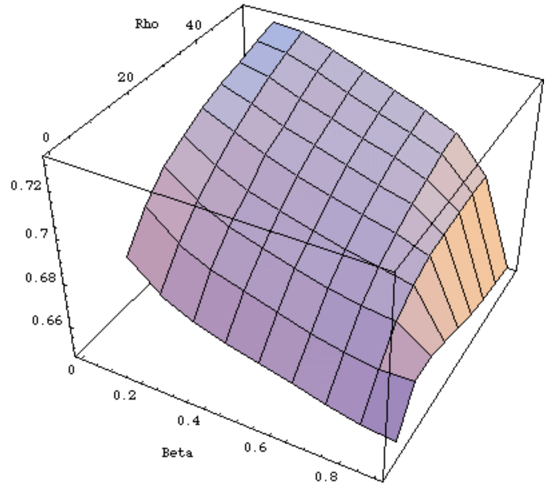


Figure 8: A 3-D plot of Kendall correlation between the news source rank vector produced by algorithm TA3, with various values of ρ and β, and the rank produced by algorithm NTA1 simply counting the news articles emitted.

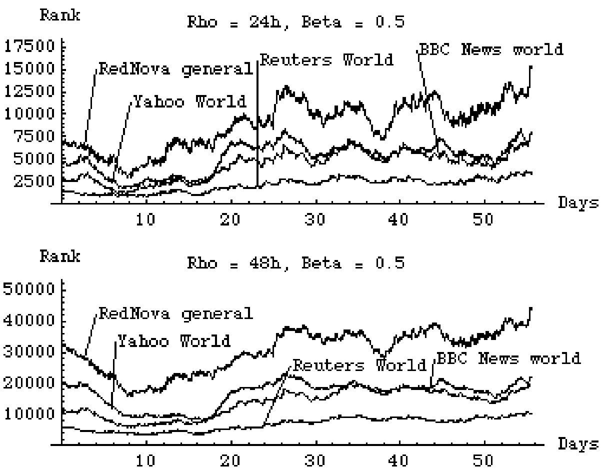


Figure 9: Top News Source for the "Word" category, with decay time $\rho = 24$h and 48h and $\beta = 0.5$. Note that for the same value of β a greater time of decay ρ gives us smoother functions and higher value of ranks. However, it does not change the order of the most authoritative sources.

importance of the news articles posted. A similar behavior is showed by the other categories, as well.

Source	# Postings
RedNova general	3154
Yahoo World	1924
Reuters World	1363
Yahoo Politics	900
BBC News world	1368
Reuters	555
Xinhua	339
New York Times world	549
Boston Globe world	357
The Washington Post world	320

Figure 10: Top ten news source for the category "World" ($\rho = 24$h and $\beta = 0, 2$). Second column contains the number of news articles posted by each news agency. Note that "Yahoo Politics" is considered more important than "BBC News world", regardless of the number of news posted.

In Figure 11, 12 we report the top ten news articles for categories "World" and "Sports", using $\rho = 24$ hours and $\beta = 0.2$. For space constraint we can not give the top news articles of the other categories present in comeToMyHead. The news posting in these tables are those which score an higher absolute rank over the period of observation. Note that our algorithm ranks any posted articles, and for top pieces of news it is common to recognize in the top list reissues of the same piece of information. The most important ranking criteria of our algorithm are freshness of news articles and authoritativeness of the news agencies.

Posted	News Source	News Abstract
10/11	RedNova general	Israeli Airstrike Kills Hamas Militant
10/11	RedNova general	Frederick Gets 8 Years in Iraq Abuse Case
10/5	RedNova general	Kerry Warns Draft Possible if Bush Wins
9/8	RedNova general	Iran Says U.N. Nuclear Ban 'Illegal'
9/12	RedNova general	Video Shows British Hostage Plead for Life
10/11	Yahoo World	Israeli Airstrike Kills Hamas Militant (AP)
9/11	RedNova general	Web Site: 2nd U.S. Hostage Killed in Iraq
9/19	RedNova general	British Hostage in Iraq Pleads for Help
9/22	Yahoo World	Sharon Vows to Escalate Gaza Offensive (AP)
9/16	Channel News Asia	Palestinian killed on intifada anniversary

Figure 11: Top ten news articles during all the observation period for the category "World" (ρ =24h and $\beta = 0.2$).

In Figure 13 and 14, are listed the top ten fresh news articles for the category "World" and "Sports" in the last day of observation. In these lists it is possible to recognize posting of news articles regarding the same event. Since these news articles are all fresh, the ranking depends essentially on the rank of the source.

Ranking the news events

In Figure 15 the function $\mu(t)$ defined in (8) is plotted over the time. The value at time t represents the mean of the ranks of the sources in the category "Sports", hence peaks may correspond to particularly significant events.

Posted	News Source	News Abstract
8/17	Reuters	Argentina Wins First Olympic Gold for 52 Years
8/18	Reuters	British Stun US in Sprint Relay
8/18	NBCOlympics	Argentina wins first basketball gold
9/9	Reuters Sports	Monty Seals Record Ryder Cup Triumph for Europe
8/18	Reuters Sports	Men's Basketball: Argentina Beats Italy, Takes Gold
10/11	Yahoo Sports	Pot Charge May Be Dropped Against Anthony (AP)
10/10	Reuters Sports	Record-Breaking Red Sox Reach World Series
8/17	China Daily	China's Xing Huina wins Olympic women's 10,000m gold
8/17	Reuters Sports	El Guerrouj, Holmes Stride Into Olympic History
8/18	Reuters Sports	Hammer Gold Medallist Annus Loses Medal

Figure 12: Top ten news articles during all the observation period for the category "Sports" (ρ =24h and $\beta = 0.2$).

Posted	News Source	News Abstract
10/11	RedNova general	Israeli Airstrike Kills Hamas Militant
10/11	RedNova general	Frederick Gets 8 Years in Iraq Abuse Case
10/11	CNN International	Israeli airstrike kills top Hamas leader
10/11	Yahoo Politics	Bush Criticizes Kerry on Health Care (AP)
10/11	RedNova general	Man Opens Fire at Mo. Manufacturing Plant
10/11	Yahoo Politics	Bush, Kerry Spar on Science, Health Care (AP)
10/11	Yahoo Politics	Smith Political Dinner Gets Bush, Carey (AP)
10/11	RedNova general	AP Poll: Bush, Kerry Tied in Popular Vote
10/11	Yahoo World	Fidel Castro Fractures Knee, Arm in Fall (AP)
10/11	Boston Globe	US Army Reservist sentenced to eight years for Abu Ghraib abuse

Figure 13: Top ten news articles the last day of the observation period for the category "World" (ρ =24h and $\beta = 0.2$), only fresh news articles are present.

Evaluating Precision

Another interesting measure is to consider the quality of ranked news articles. To perform this evaluation we consider the standard P@N measure over the news stories, defined as $P@N_{news} = \frac{|C \bigcap R|}{|R|}$ where , R is the subset of the N top news articles returned by our algorithm, and C is the set of manually tagged relevant postings. In particular, we fixed a particular time of observation over the data stream of news articles and ranked the pieces of news. Then, we asked a group of three people to manually assess the relevance on the top articles by taking in account the particular instant of time chosen and the category to which the pieces of news belong. Only the precision of the final algorithm in Section 6 has been evaluated since the earlier variations of the algorithm do not satisfy the mathematical requirements given in Section 3. In Figure 16 we report the P@N reported for the top news articles in the category "U.S." during the period of observation.

Posted	News Source	News Abstract
10/11	Yahoo Sports	Pot Charge May Be Dropped Against Anthony (AP)
10/11	Yahoo Sports	Anthony Leads Nuggets Past Clippers (AP)
10/11	NDTV.com	Tennis: Top seeded Henman loses to Ivan Ljubicic
10/11	Reuters	UPDATE 1-Lewis fires spectacular 62 to take Funai lead
10/11	Reuters Sports	Cards Secure World Series Clash with Red Sox
10/11	Yahoo Sports	Court: Paul Hamm Can Keep Olympic Gold (AP)
10/11	Yahoo Sports	Nuggets' Anthony Cited for Pot Possession (AP)
10/11	Reuters	Chelsea won't sack me, says Mutu
10/11	Reuters Sports	Record-Breaking Red Sox Reach World Series
10/11	Yahoo Sports	Dolphins Owner Undecided About Coach, GM (AP)

Figure 14: Top ten news articles the last day of the observation period for the category "Sports" (ρ =24h and $\beta = 0.2$), only fresh pieces of news are present.

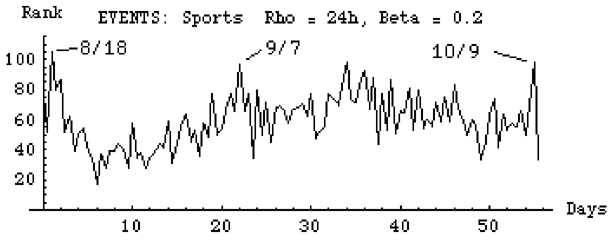


Figure 15: For the category "Sports" a plot of the function $\mu(t)$ is represented. Pecks correspond to particular significant events.

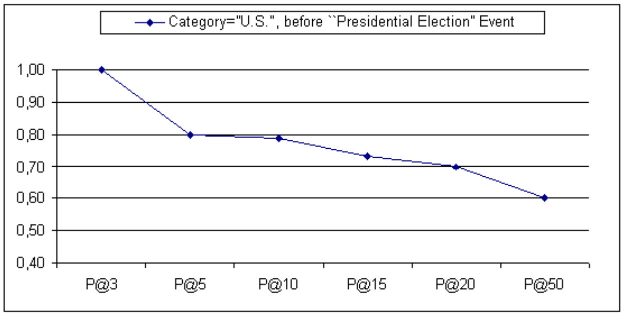


Figure 16: P@N for "U.S." during the period of observation.

8. CONCLUSION

In this paper we have presented an algorithm for ranking news articles and news sources. The algorithm has been constructed step by step ruling out simpler ideas that were not working on intuitive cases. Our research has been motivated by the large interest in commercial news engine versus the lack of research papers in this area. An extensive testing on more than 300,000 pieces of news, posted by 2000 sources over two months, has been performed, showing very encouraging results both for news articles and news sources.

The methodology proposed in this paper has a larger application than the ranking of news article and press agency. We plan to apply the ideas discussed in this paper to other classes of problems such as the problem of ranking publications, authors and scientific journals.

9. REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. Yu. Clustream: A framework for clustering evolving data streams. In *Proceedings of the 29th International Conference on Very Large Data Bases*, 2003.

[2] M. Atallah and R. Gwadera. Detection of significant sets of episodes in event sequences. In *Proceedings of the International Data Mining Conference*, pages 3–10, 2004.

[3] S. Chung and D. McLeod. Dynamic topic mining from news stream data. In *Proceedings of International Conference on Ontologies, Databases and Applications of Semantics*, pages 653–670, 2003.

[4] S. Blair-Goldensohn, D. R. Radev, Z. Zhang, and R. S. Raghavan. Interactive, domain-independent identification and summarization of topically related news articles. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 225–238, 2001.

[5] S. Blair-Goldensohn D. R. Radev, Z. Zhang, and R. S. Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the Human Language Technology Conference*, 2001.

[6] D. de Castro Reis, P. Golgher, A. da Silva, and A. Laender. Automatic web news extraction using tree edit distance. In *Proceedings of the 13th International WWW Conference*, pages 502–511, 2004.

[7] D. M. Dunlavy, J. P. Conroy, and D. P. O'Leary. Qcs: A tool for querying, clustering, and summarizing documents. In *Proceedings of the Human Language Technology Conference-NAACL*, pages 11–12, 2003.

[8] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International WWW Conference*, pages 482–490, 2004.

[9] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. In *IEEE Transactions on Knowledge and Data Engineering*, pages 515–528, 2003.

[10] M. Henzinger, B. Chang, B. Milch, and S. Brin. Query-free news search. In *Proceedings of the 12th International WWW Conference*, pages 1–10, 2003.

[11] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 430:81–93, 1938.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[13] H. Mannilla, H. Toivonen, and A. Verkamo. Discovery of frequent episodes in event sequences. In *Proceedings of the Data Mining and Knowledge Discovery*, pages 259–289, 1997.

[14] S. Muthukrishnan. Data streams: algorithms and applications. In *Proceedings of the ACM-SIAM 14th Symposium on Discrete Algorithms*, page 413, 2003.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[17] `http://www.bbcworld.com/`.

[18] `http://www.boston.com/`.

[19] `http://www.cnn.com/`.

[20] `http://www.e-marketing-news.co.uk/Aug04-Greg.html`.

[21] `http://www.findory.com/`.

[22] `http://news.google.com/`.

[23] `http://newsbot.msnbc.msn.com/`.

[24] `http://www.nielsen-netratings.com/`.

[25] `http://www.channelnewsasia.com/`.

[26] `http://www.newsinessence.com/`.

[27] `http://www.rednova.com/`.

[28] `http://www.reuters.com/`.

[29] `http://searchenginewatch.com/`.

[30] `http://news.yahoo.com/`.