A Ranking Algorithm for News Data Streams

Bhavana Mahour, Akhilesh Tiwari
Department of C.S.E. & IT, MITS, Gwalior (M.P.), India

ABSTRACT
With the advent of internet, print media such as newspapers and magazines have been moving themselves to websites providing news on the go. Also some people prefer news on the net more as pictures along with video is available in it, also we can search a particular topic. In this paper we suggest a ranking algorithm that exploits the dependency between the rank of news article, topic and source by the help of a virtual graph model. We also add the imitating or copying of new articles to revise the ranking of a source or article. Our complexity is linear and for starters we have just added the national news channels in our algorithm. To validate our algorithm we use huge manual data collecting news from national news websites.

Keywords
News Ranking, articles, source, topic, Time aware algorithm

1. INTRODUCTION
With terabytes of news adding every day on the internet and a plethora of news sources available including international news channels such as BBC, CNN and along with national news channels such as AAJTAK, ZEE NEWS moving themselves to world of internet. For the convenience of users a number of search engines are available on the internet such as Google, Yahoo news on the net but we consider there ranking criteria is based on commercial use as to who will pay more will be on the top.

“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.

In past few years, more and more people have been using the internet not just for entertainment but also reading news on the web. According to a recent survey made by Nielsen netRatings [7], the number of people browses news on the web to know what’s happening around the world. It has become part of our daily life. User easily accesses the news information on the web. There are more commercial news search engine available such as Google News [8], Yahoo News [9] NewsInEssence[10] and newsbot [12] providing news information according to their will. Google News search provide the information, aggregates headlines from news sources worldwide, group’s similar stories together and displays them according to user’s personalized interests. Yahoo News [3] search engine provide the news information with similar services and another important news search engine NewsInEssence[2] which provide the news information, clustering and summarizes similar news articles. In Liu et al. proposed an algorithm for ranking news related to some query submitted at a specified timestamp [5].

In the research area, a lot of work has been done in the field of web page ranking. Page ranking method are based on the linking structure of web graph such as Page Rank [6], HITS [8] SALSA [7] and Berberichet al. [9] many techniques are used to the page ranking algorithm. Introduce a time aware ranking method [2]. Time aware rank method differs from Page Rank method [3]. it considers web pages modified time and activity during ranking. The time aware ranking [6]. Provide fresh information on news articles thus we have to develop new Ranking methods to solve the problem. In an online news ranking method is propose to news source and related articles, copy right articles source to another source.

In this paper, the problem of ranking algorithm for news source, articles, topic and copyright articles source to another source. And also we propose a graph model to formalize the relationship between news sources, articles and source to another source. Our ranking algorithm is based on this model and reflects the mutual reinforcement between news source, articles and source to another source.

2. PROPOSED METHODS
2.1 News Properties
If we have to rank a news article, it is all the way different from ranking web pages since it contains very less or no HTML links which can be exploited as in PageRank[15] hence we have to determine the rank based on the virtual linking between topics, articles and sources. The most important properties are discussed below.

1. Ranking for News posting and News source
We have to devise an algorithm that separately calculates the rank of source and articles as well as also topics which are most read about.

2. Time awareness
The importance of a news item changes over time for the reason that we are dealing with a stream of information in which a fresh news story can be considered more important than an old one. Thinking of news as a stream of data, the rank of news will decrease as the time elapses.

3. Important News articles are clustered
It is probable that an important news article n will be (partially) replicated by many sources, considering for example, a news article n originated from a press agency. A measure of its importance can also be gauged by the number of different on-line newspapers which either replicate or extract sections of text from n. The phenomenon of using stories released by other sources is common in the context of Web journals and, as far as the news engine is concerned, this means that the (weighted) size of the cluster formed around n can be deemed a measure of its importance.

4. Mutual Reinforcement between articles and Sources
A lot of articles are generated daily and we decide the rank of the article of the same heading coming from two different sources, then the source having more credibility or have the reputation of giving good news such as “Zee News” will be
given more importance than an article by a local or state level source.

5. Source Authority
Different news web sites have different authorities. Suppose two nearly identical news articles, one is published by ZeeNews, the other is from an unknown web site. Since ZeeNews is a famous and credible news website, we would expect the news from ZeeNews having higher rank. Also news topics may overlap with each other.

6. Mutual Reinforcement of imitating Sources
We take into consideration the sources which copies the whole article or part of the article in its own website. We use this to calculate the rank of the parent source (from which the article is originating) by increasing its rank depending on the amount of its copying and also since the others are copying it; this reduces the credibility of the other sources and thus reducing their rank by considering the amount of imitation by them.

2.2 VIRTUAL GRAPH MODEL
Although there is little explicit HTML links between news web pages, we can create virtual link between source, topic, articles and copied articles. We define a virtual undirected link graph G = (V, E), where V = S U C U N U N'. Herein, (S stands for the set of source, N stands for the set of articles; C stand for the topics; and N' stand for the articles which have been copied). The virtual set E contains four disjoint partition; E1-E4. E1 represent the news creation process. E2 represent the similarity between topic and articles (news stream); when articles ai belongs to topic ti there’s an undirected edge between ai and ti. E3 it is a relation between topic and source in which an articles ai is copied by source sj this nodes indicates our copyright article also this will reduce the ranking of source sj. E4 it is the fresh article published by source sj which is copying other articles from other source.

![News virtual link graph](image)

We use weighted adjacency matrix to express the graph and denote it as

\[
A = \begin{pmatrix}
0 & P & Q \\
0 & R & S \\
0 & S & \Sigma
\end{pmatrix}
\]

Where P, Q, S refer to edges from source to topic source to article and topic to articles respectively. R is refer to edges from copy right articles and \(\Sigma\) is the similarity matrix. We can use a sub matrix at the upper-left corner of A instead of O.

2.3 STREAM CLUSTERING
We could use a clustering algorithm to cluster similar articles from topic.

2.3.1 Similarity Computation
We have used a manual approach for similarity computation. We have given our articles/topics to some individuals which give us a value of similarity on a scale of 0-10 converted into decimal number.

2.3.2 Clustering
Our proposed algorithm uses bottom up approach and expect our cluster to be highly coupled. Our topic cluster T has a constant parameter \(TH(0 \leq TH \leq 1)\) for similarity threshold between article and topic and also we merge two articles in a cluster by using threshold parameter \(H(0 \leq H \leq 1)\) by comparing the contents of the articles as well as the time difference between them t and defining them into a topic cluster.

**Proposed Algorithm**

**Input:** T - a set of individual cluster of topics \(T = \{T_1, T_2, T_3, \ldots, T_n\}\). incoming articles \(A\) articles threshold \(H\), topic threshold \(TH\), Time difference between articles \(t\)

1. For all incoming articles \(a_i \in A\), \(A = \{a_1, a_2, a_3, \ldots, a_k\}\) set of articles.
2. For each \(T_i\), check sim (\(T_i, a_i\)) \(> TH\) \(\{i = 1 \text{ to } n\}\)
3. For each \(a_i \in A\) [initialize cluster]
4. Choose \((a_i, a_j) \in A \times A\) with max sim \((a_i, a_j)\)
5. Check sim \((a_i, a_j, t) > H\)
6. Set merge \(= 1\)
7. \(a_k = a_i \cup a_j\)
8. add \(a_k\) to \(T_i\)
9. else return
10. repeat till \(T_n\)
11. end

2.4 Matrix Expressing
The naïve approach is that a news sources has a rank proportional to the number of pieces of news it generates and conversely, that a news articles should rank high if there are many other news stories close to it formally, denoting by \(r = \begin{pmatrix} r_S & r_S \end{pmatrix}\) the vector of news source, copied news source and rank of news stream. We can compute them as

\[
r = Ar
\]

\(r = \begin{pmatrix} r_S \end{pmatrix}\) where \(r_S\) represents vector news source, \(r_S\) represents vector copied news source or the news source which
3. TIME-AWARE RANKING ALGORITHM

To deal with a news data stream we have to design time aware mechanisms, which do not use fixed time observation window over the flow of information. They key idea is that the importance of a piece of news is strictly related to the time of its emission. Hence, we model this phenomenon in traducing a parameter \( \alpha \) which accounts for the decay of “freshness” of the news story. This \( \alpha \) depends on the category to which the news articles and topic belongs. For instance, it is usually a good idea to consider sport news articles related to a topic rather than sport news itself. Similarly we can deduce the rank of topics:

We denote the rank of some source \( s \) as \( R(s, t) \) at time \( t \), the rank of some topic \( c \) as \( R(c, t) \) at time \( t \) the rank of some articles \( n \) as \( R(n, t) \) at time \( t \). We use \( S(n_i) = s_j \) to represent that articles \( n_i \) is published by source \( s_j \), an use \( C(n_i) = c_j \) to show that articles \( n_i \) belongs to topic \( c_j \). This key idea is that the news stream copy other articles and \( r_n \) is the vector representing rank of news stream. Similarly, we decay the rank of some topic \( c \) according to \( (3) \):

\[
R(c, t) = e^{-\alpha(t-t_i)} R(c_i, t_i), \quad t > t_i \quad \ldots(3)
\]

Now we consider the decaying of rank as time goes by. Let \( \rho \) decay time, \( \beta \) jump probability, \( \gamma_s = 1 \) (copyright articles), \( \gamma_c = 0 \) (unique no copyright articles). Let \( s \) source, \( c \) topic, \( n \) articles \( s_{num}(c_i, s_k) \) be the number of articles published by \( s_k \) in topic \( c_i \), \( num(c_j) \) number of articles in topic \( c_j \).

3.1 Algorithm TA1

\[
\text{Rank of source } (s_k, t) = \sum_{s(n_i)=s_k} e^{-\alpha(t-t_i)} R(n_i, t_i) + \sum_{t_j > t} \gamma_{s_j} e^{-\alpha(t-t_j)} \gamma_{s_j} R(n_j, t_j) \beta \quad \ldots(4)
\]

Similarly we can deduce the rank of topics:

3.2 Algorithm TA2

\[
\text{Rank of topic } (c_k, t) = \sum_{c(n_i)=c_k} e^{-\alpha(t-t_i)} R(n_i, t_i) \beta \quad \ldots(5)
\]

3.3 Algorithm TA3

\[
\text{Rank of articles } (n_k, t) = \frac{1}{num(c_k)} R(c_k, t) \beta + \sum_{t_j < t} e^{-\alpha(t-t_j)} \gamma_{t_j} R(n_j, t_j) \beta \quad \ldots(6)
\]

4. EXPERIMENT

Our experiment is based on sub set of data collected from websites of national news channels such as AAJTAK, Zee News, India TV and NDTV for a period of 15 days. The code is written in PHP and the ranking of about 500 news pieces of information. During our experiments, we set the half-life decay time \( \rho \) to be 24 hour, random jump probability \( \beta \) to be 0.2, and the clustering similarity threshold \( H \) to be 0.7 and topic threshold TH 0.3.

Table 1. top 5 rank topics

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.26</td>
<td>Delhi election 2013</td>
<td>Zee TV</td>
</tr>
<tr>
<td>19</td>
<td>Terrorist Attack</td>
<td>STAR TV</td>
</tr>
<tr>
<td>28.66</td>
<td>Sport news</td>
<td>Zee TV</td>
</tr>
<tr>
<td>14.4</td>
<td>Natural disaster</td>
<td>NDTV</td>
</tr>
<tr>
<td>44.49</td>
<td>Politics Of Religion</td>
<td>INDIA TV</td>
</tr>
</tbody>
</table>
Table 2. top 5 rank articles

<table>
<thead>
<tr>
<th>Rank</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>134.8</td>
<td>Delhi poll: records 67% voters turnout</td>
</tr>
<tr>
<td>40.66</td>
<td>Blast in Delhi NCR kills 7 and several injured</td>
</tr>
<tr>
<td>209.9</td>
<td>2 CRPF personnel killed in terrorist attack in Jammu and Kashmir</td>
</tr>
<tr>
<td>66.9</td>
<td>Uttarakhand floods</td>
</tr>
<tr>
<td>96.89</td>
<td>2013 India and south Africa test series</td>
</tr>
</tbody>
</table>

Table 3.top 5 rank articles

<table>
<thead>
<tr>
<th>Rank</th>
<th>No. Of Articles Copied By Other Sources</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.94</td>
<td>14</td>
<td>NDTV</td>
</tr>
<tr>
<td>4.77</td>
<td>9</td>
<td>ZEE TV</td>
</tr>
<tr>
<td>0.25</td>
<td>7</td>
<td>INDIA TV</td>
</tr>
<tr>
<td>9.16</td>
<td>7</td>
<td>STAR TV</td>
</tr>
<tr>
<td>2.74</td>
<td>5</td>
<td>AAJ TAK</td>
</tr>
</tbody>
</table>

To find the precision of our algorithm we used number of clicks done by users for a particular topic/source/articles consisting of a group of 10 people and we match the results with our calculated ranking by the formula $P_s = \frac{n(S)}{K_s} \times 100$ where $n(S)$ is the total number of clicks for sources $S$ and $K_s$ is our ranking judged to be important by our algorithm. Similarly ranks of topics and articles can be calculated. By our experiments we find that the precision of topics and articles decreases as after time a lot of articles can be seen for a particular topic/news are available so the user is not restricted to accessing a particular article only.
5. CONCLUSION AND FUTURE WORK
Our algorithm calculates the ranking of sources, topics and articles altogether the news stream. The data collected from national news channels for a period of 15 days shows encouraging results and the order is also increasing linearly. Future work for our algorithm is the personalization of news in which the data of every user logged also need to be taken and then news is displayed. Similarly it can be used to rank research papers and also popularity of a particular page in social networking sites such as Facebook where people search for related pages.

6. REFERENCES


