Ranking of Evolving Stories Through Meta-Aggregation

Juozas Gordevičius
Free University of
Bozen-Bolzano, Italy
gordevicius@inf.unibz.it

Francisco J. Estrada
Thoora Inc., Toronto, Canada
francisco@thoora.com

Hyun Chul Lee
Thoora Inc., Toronto, Canada
chul.lee@thoora.com

Periklis Andritsos
Thoora Inc., Toronto, Canada
periklis@thoora.com

Johann Gamper
Free University of
Bozen-Bolzano, Italy
gamper@inf.unibz.it

ABSTRACT

In this paper we focus on the problem of ranking news stories within their historical context by exploiting their content similarity. We observe that news stories evolve and thus have to be ranked in a time and query dependent manner. We do this in two steps. First, the *mining* step discovers metastories, which constitute meaningful groups of similar stories that occur at arbitrary points in time. Second, the *ranking* step uses well known measures of content similarity to construct implicit links among all metastories, and uses them to rank those metastories that overlap the time interval provided in a user query. We use real data from conventional and social media sources (weblogs) to study the impact of different meta-aggregation techniques and similarity measures in the final ranking. We evaluate the framework using both objective and subjective criteria, and discuss the selection of clustering method and similarity measure that lead to the best ranking results.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; I.5.4 [**Pattern Recognition**]: Text Processing

General Terms

Algorithms, Experimentation

Keywords

Text Mining, Temporal Ranking

1. INTRODUCTION

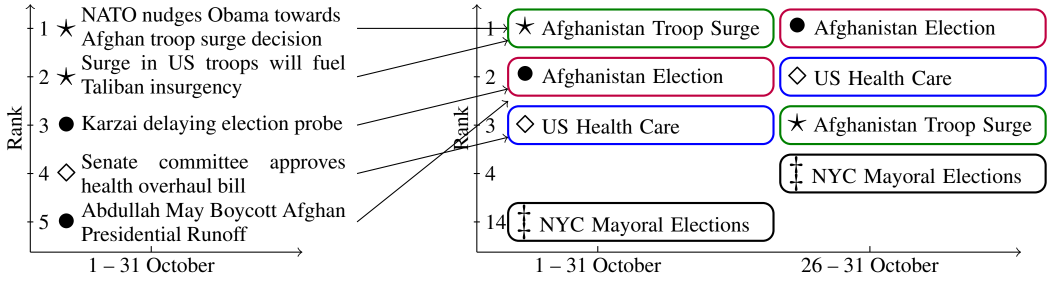
Existing technologies for online news browsing allow users to have continuous access to up-to-date news, as well as to a wide range of related opinions and comments coming from the social media. However, the larger-scale problem of aggregating, searching, and ranking the historical archives of such content has not been thoroughly addressed. Since news aggregation engines tend to construct small, highly consistent, short-lived clusters of similar news articles, the evolution of stories through multiple events that are widely spaced in time is not captured in the archives. Under these conditions, it is difficult to rank stories in the context of a user specified time interval. For example, consider a BBC article[1] in which the authors attempt to list and rank the most important stories of the past decade. The article points out that the readers who were polled to create the list missed important events such as the sequencing of the human genome. Given the large time-span, and the lack of specific query keywords, a keyword-based search would not succeed in this case either. In addition to this, the proposed ranking was subjective, and thus is likely to disagree with the actual relevance of each story. Therefore, there is a need to answer such and other similar queries in a principled and automatic way.

In this paper we show that the time interval specified in the query should affect ranks of stories whose evolution overlaps the interval. We propose a two-step framework that allows to compute such temporally sensitive ranking. First, the *metastory mining* step employs a clustering method to create metastories by aggregating a set of story clusters, similar to the way news aggregators place stories together. The generation of metastories is a challenging problem because even though the central theme of each metastory is the same, the actual content of the news articles within can vary widely. Second, the *ranking* step, selects the metastories overlapping the user-specified time interval and produces their ranking. Ranking is done through link analysis. However, unlike traditional PageRank [6] which is based on hypertext-links, we construct implicit links using content-based similarity between all pairs of active metastories. We exploit the symmetry of similarity measures to avoid the expensive power iteration traditionally used to compute the PageRank vector.

Consider the example of news story aggregation and ranking process in Figure 1. We use story clusters containing news articles as well as related blogs provided by a state-of-the-art platform for news and social media aggregation called Thoora. Five out of 700 individual story clusters valid during October 2009 are shown in Figure 1(a). As indicated by the arrows, they can be aggregated into metastories corresponding to important events. This metastory aggregation is performed once, over the set of all existing story clusters. The ranking step determines the relative ordering of the metastories with regard to a time interval specified by the user. Figure 1(b) shows two rankings for different query intervals: October 1-31, and October 26-31, 2009. Observe that though metastories

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

[1] `http://news.bbc.co.uk/2/hi/8409040.stm`


Rank
1 ★ NATO nudges Obama towards Afghan troop surge decision
2 ★ Surge in US troops will fuel Taliban insurgency
3 ● Karzai delaying election probe
4 ◇ Senate committee approves health overhaul bill
5 ● Abdullah May Boycott Afghan Presidential Runoff
1 – 31 October
Rank
1 ★ Afghanistan Troop Surge
2 ● Afghanistan Election
3 ◇ US Health Care
4
14 NYC Mayoral Elections
● Afghanistan Election
◇ US Health Care
★ Afghanistan Troop Surge
NYC Mayoral Elections
1 – 31 October
26 – 31 October

(a) Individual stories

(b) Metastories

Figure 1: Effect of the query time interval on the ranking results

are the same, the rankings differ due to time-dependent nature of the ranking process.

2. PREVIOUS WORK

Topic Detection and Tracking (TDT) is related to our problem of metastory detection. The goal of TDT is to automatically identify topics within a set of documents, and to keep track of these topics as the document set evolves. Most of the existing TDT algorithms focus on the online discovery of new topics from recent data and on the tracking of developments within known topics [2, 1, 12, 18]. The critical difference between TDT and our approach is in the granularity of the clustering. We find, in line with Leskovec et al. [17], that clusters in TDT are too general to represent individual stories evolving through time, and that a much finer clustering is needed. In addition, we merge related stories into metastories without any restriction on their temporal distance, and unlike Allan et al. [2] we do not use time decay in our similarity measure. Finally, while standard TDT methods work on reasonably clean news document data, our initial story components contain news documents and blogs that do not adhere to any standards either of content, language, or structure. This complicates the metastory generation process beyond what standard TDT was designed to handle.

Regarding the document ranking, our framework is closely related to PageRank [6, 5] and HITS [13], two widely used approaches for web-page ranking based on hyperlink analysis. There have been a few attempts to rank documents replacing traditional hyperlinks with links based on similarity scores. In contrast to [23, 14], we assume no prior hyperlinks between stories and use solely the content similarity. Furthermore, we argue that using a symmetric similarity measure to construct the links is more intuitive and appropriate than asymmetric Kullback-Leibler divergence proposed by Kurland et al. [16, 15]. We show in this work that this property of the similarity measure allows very efficient evaluation of PageRank.

3. THE FRAMEWORK

Here we formally describe the ranking framework. Let W be a keyword dictionary. A *story* s is defined as a pair $s = \langle f_s, T_s \rangle$, where f_s is the term frequency vector whose entries correspond to particular terms $w_i \in W, i = 1, \ldots, |W|$ and represent the number of times the term w_i occurs within s. $T_s = [b_s, e_s]$ is the *lifespan* of s represented as an interval bounded by two time points, b_s and e_s. Two lifespans, T' and T'', overlap (or intersect) if $T' \cap T'' \neq \emptyset$. In our particular application scenario, a story refers to a cluster of news and blog articles and its lifespan overlaps all the documents in the story.

From an input set of stories, S, a clustering algorithm constructs disjoint and non-empty subsets $S_1, \ldots, S_k \subseteq S$, i.e. $\bigcup_i S_i = S$ and $S_i \cap S_j = \emptyset$ for all $i \neq j$. Then a *metastory* m is a triplet $m = \langle S_m, f_m, T_m \rangle$, where $S_m \subseteq S$. Its term frequency vector f_m is an aggregation over a set of term frequency vectors, i.e. $f_m = \sum_{s \in S_m} f_s$. Its lifespan $T_m = [b_m, e_m]$ refers to the maximum time-interval covering the entire set of lifespans associated with all stories in S_m, $b_m = \min_{s \in S_m} \{b_s | b_s \in T_s\}$ and $e_m = \max_{s \in S_m} \{e_s | e_s \in T_s\}]$.

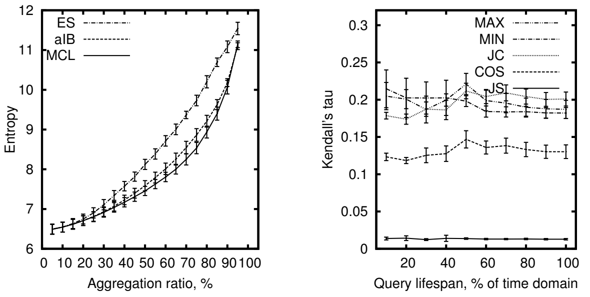
We can rank a set of metastories, M, using a content similarity function $sim : M \times M \rightarrow [\epsilon, 1]$. Assuming that $\epsilon > 0$ is the minimum similarity between metastories ensures that the graph constructed through the similarity values is fully connected. Without loss of generality, we also assume that frequency vectors represent proper probability distributions, i.e. they are normalized to have unit mass. Then, the *rank* of metastory $m \in M$ is

$$rank(M, m) = \sum_{m' \in M} \frac{sim(m, m') \cdot rank(M, m')}{\sum_{m'' \in M} sim(m', m'')} \qquad (1)$$

According to this equation, a metastory should be ranked high if it is similar to other highly ranked metastories. Computing the ranks is equivalent to finding the stationary distribution of a random walk over the graph whose nodes correspond to metastories, and whose edges are the normalized similarities [6]. While in general it is a computationally expensive task, the following lemma that follows from [7] shows that for symmetric similarity measures the computation is straightforward and efficient.

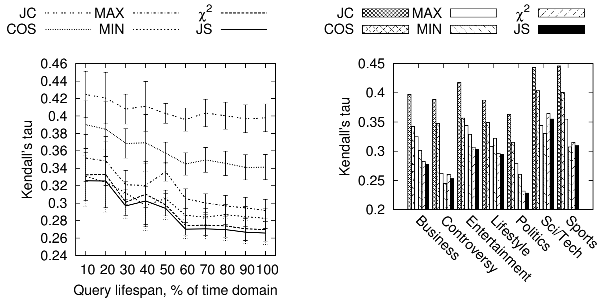
LEMMA 1. *For a set of metastories, M, similarity measure sim s.t. $\forall m, m' \in M$ $sim(m, m') = sim(m', m)$, and some constant λ we have $rank(M, m) = \lambda \sum_{m' \in M} sim(m, m')$.*

Now we incorporate the temporal dimension into the ranking of metastories. Intuitively, the rank of a metastory that intersect with the lifespan specified in the user query should depend only on those member stories that also intersect the query. This way the temporal aspect associated with a specific query allows us to capture the evolution of stories as illustrated previously in Figure 1. Let $Q = [b_Q, e_Q]$ be the user specified lifespan. For each metastory $m \in M$ we construct its reduction m' that only contains stories overlapping the query, i.e. $S_{m'} \subseteq S_m \wedge \forall s \in S_{m'}$. Accordingly, its term frequency vector is $f_{m'} = \sum_{s \in S_{m'}} f_s$ and the lifespan starts at $b_{m'} = \min_{s \in S_{m'}} \{b_s | b_s \in T_s\}$ and ends at $e_{m'} = \max_{s \in S_{m'}} \{e_s | e_s \in T_s\}$. Let M_Q be the set of all the reduced and non-empty metastories, i.e. $M_Q = \{m'_i | m_i \in M \wedge S_{m'_i} \neq \emptyset \ \wedge i = 1, \ldots, k\}$. Then, the *query dependent rank* of a metastory $m \in M$ is equal to the rank of its reduced version



(a) (b)

Figure 2: (a) Clustering entropy as a function of aggregation ratio; (b)Kendall's tau distance between the rankings produced by χ^2 and other similarity measures as a function of query length



(a) (b)

Figure 3: Similarity-based ranking vs. social impact

in the set of all the reduced and non-empty metastories:

$$rank_Q(M,m) = \begin{cases} rank(M_Q, m'), & \text{if } m' \in M_Q, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

4. EXPERIMENTAL STUDY

To design an optimal temporal ranking framework we must, first, select a clustering algorithm that produces good metastories, then we must choose the similarity measure that ranks the metastories best. Our experimental study provides sufficient information regarding these two choices and illustrates the usefulness of our approach. The data we collected from Thoora spans 56 days starting from 15 September 2009, and is classified into seven categories: Business, Entertainment, Controversy, Lifestyle, Politics, Science & Technology, and Sports. For each category we create a separate dataset that records the 50 biggest stories for each day yielding on average 700 unique story clusters per dataset. All the experiments are replicated on each dataset and average values as well as standard errors are reported.

Using the *expected entropy* [3] measure we compare the quality of clustering results obtained from three different clustering methods: efficient graph-based segmentation (ES) [10], Markov Clustering (MCL) [22] and the agglomerative Information Bottleneck (aIB) [20]. These methods represent a wide range of clustering paradigms and have been successfully applied in different fields[10, 8, 11, 3, 20]. For the ES and MCL algorithms we use the χ^2 distance as it has been shown to provide close to optimal performance at a low computational cost [19]. Figure 2(a) shows the expected entropy each algorithm yields as a function of the aggregation ratio. The lower this ratio, the more clusters there are in our system and vice versa. Lower entropy values suggest better clustering quality, small error bars indicate that the entropy results are stable across all categories. The results of this experiment provide us with a principled way to choose the best clustering method among those studied. In the context of our particular problem we have observed that aggregation ratios over 30% tend to produce metastories that are too general and do not correspond to a well defined world event. We will, therefore, set the aggregation ratio to a fixed value of 20%. At this ratio MCL and aIB produce equivalent result, yet aIB provides a much finer control of the level of aggregation. Therefore, we will use aIB as the clustering method in all the subsequent experiments.

Having selected the clustering algorithm to produce the metastories, we turn to the analysis of the ranking results. In order to see whether the choice of the similarity measure influences the final outcome we measure the normalized Kendall's tau distance [9], K, between the rankings produced using the following symmetric measures that have shown good performance in clustering and retrieval tasks: Jensen-Shannon (JS) [20], χ^2, minimum (MIN) and maximum (MAX) overlap, Jaccard Coefficient (JC), and Cosine (COS) similarity [4]. Distances given by χ^2 and JS are converted to similarities by dividing them by maximal distance value and subtracting resulting normalized values from 1. When query lifespans are fixed to cover 100% of the time domain χ^2 and JS rankings are the closest ($K = 0.013$). All the other rankings are significantly distant with JC and MAX being the furthest apart ($K = 0.326$). Fig. 2(b) depicts the average distance between χ^2 and rankings by other measures as a function of query lifespan. It is apparent that the distance values do not depend on the query, therefore, it is the similarity function that influences the final ranking independently of the time interval specified in the query and, therefore, must be chosen in a principled way to yield the best ranking result.

Next, we construct a ground truth ranking based on the social impact of a story measured by the count of blogs related to it. We assume that the more important a story, the more people will blog about it. Fig. 3(a) shows the distance between the framework rankings and the ground truth as a function of the size of the query window. Fig. 3(b) shows the average result separately for each category. Overall, we observe that JS and χ^2 perform best and the other measures lag behind. This is reasonable and agrees with the conclusions from [19]. The average Kendall's tau distance for JS is 0.2914 which is remarkable given that the ranking relies exclusively on the similarity scores with no access to social impact information. This verifies the validity of our ranking approach for situations in which the social impact factor is not available. It is worth noting that the differences in performance over different categories may be indicative of particular properties of stories in each category that may be exploited for ranking purposes. This remains a topic of future work.

We further validate our framework using the Amazon Mechanical Turk[2] (AMT). This service facilitates human interaction for answering arbitrary questions. We use the AMT to generate subjective ground-truth rankings to use in the evaluation of the ranking framework. For each category we randomly and uniformly sample a subset of 15 metastories and choose one representative title for each of them. Pairs of titles are shown to 11 different AMT users who are asked to choose the one that should be ranked higher. Intuitively, if a title gets more votes than any other title in its category, it should be ranked first. We construct ranked lists from the pairwise

[2] http://www.mturk.com

Table 1: Kendall's tau distance between rankings produced by different measures

	Average	Std. Err.		Average	Std. Err.
χ^2 vs AMT	**0.410**	0.019	χ^2 vs SI	0.312	0.041
JS vs AMT	0.413	0.019	JS vs SI	**0.304**	0.041
MAX vs AMT	**0.398**	0.024	MAX vs SI	0.374	0.054
MIN vs AMT	**0.410**	0.024	MIN vs SI	0.308	0.054
COS vs AMT	0.419	0.027	COS vs SI	0.369	0.037
JC vs AMT	0.433	0.022	JC vs SI	0.419	0.054
SI vs AMT	0.516	0.028			

votes using the Schulze method [21] and further refer to them the *AMT lists*. Table 1 shows the distances between framework, social impact (SI) and AMT list rankings. Note that we only compare ranks of the metastories that are part of the AMT lists. χ^2, MAX, and MIN give the best overall ranking with regards to the AMT and the standard error across different categories varies the least for χ^2 and JS indicating that these two measures are the most consistent. This is in agreement with the results reported previously in Figure 3(a). An interesting observation is that the distance between BC and the AMT lists is larger than the distance between any distance measure and either BC or the AMT lists. This could very well be caused by cultural and environmental differences between AMT users and the blogger community. However, it is encouraging to verify that in both cases the similarity measures do a good job of ranking metastories. As a final test, we compared rankings obtained using PageRank and HITS implemented as suggested in [16]. While we do not show the full results due to the lack of space, we note that PageRank provides slightly but consistently better ranking at better computational cost.

5. CONCLUSIONS AND FUTURE WORK

We have presented a framework for the ranking of evolving stories through meta-aggregation. Metastory ranks are obtained in two steps. First, related stories are clustered into metastories. We explored the issue of selecting an optimal clustering method for this task, and showed that the MCL algorithm yields the best results. Second, the metastories are ranked using implicit links constructed based on their content similarity and the time interval specified by the user. The choice of similarity function for ranking was a significant factor determining the overall quality of the ranking result. Interestingly, the ranking quality does not depend on the lifespan of the query, but varies across data categories instead. We discussed the efficient computation of metastory ranks and showed that overall the Jensen-Shannon distance measure provides the best and the most stable prediction of the social impact a metastory would have within a specific historical context. Our results confirm the validity of the framework and its usefulness to determine a time-dependent rank of the metastories in a completely unsupervised manner. Future work will focus on scalability issues, including the efficient computation of metastory representatives, and the incremental computation of ranking results.

6. REFERENCES

[1] Gediminas Adomavicius and Jesse Bockstedt. C-TREND: Temporal cluster graphs for identifying and visualizing trends in multiattribute transactional data. *IEEE Trans. Knowl. Data Eng.*, 20(6):721–735, 2008.

[2] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, Umass Amherst, and James Allan. Topic detection and tracking pilot study, 1998.

[3] Periklis Andritsos, Panayiotis Tsaparas, Renée J. Miller, and Kenneth C. Sevcik. Limbo: Scalable clustering of categorical data. In *EDBT*, pages 123–146, 2004.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, 1999.

[5] Pavel Berkhin. Survey: A survey on pagerank computing. *Internet Mathematics*, 2(1), 2005.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[7] C. S. Chennubhotla. *Spectral Methods for Multi-Scale Feature Extraction and Data Clustering*. PhD thesis, University of Toronto, Dept. of Computer Science, 2004.

[8] Ouzounis C.A. Enright A.J., Van Dongen S. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[9] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

[10] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[11] Oktie Hassanzadeh, Fei Chiang, Hyun Chul Lee, and Renée J. Miller. Framework for evaluating clustering algorithms in duplicate detection. *PVLDB*, 2(1):1282–1293, 2009.

[12] Jon Kleinberg. Temporal dynamics of on-line information streams. In *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2004.

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

[14] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. Blogrank: Ranking weblogs based on connectivity and similarity features. In *Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications, (AAA-IDEA)*, page 8, 2006.

[15] Oren Kurland and Lillian Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. *CoRR*, abs/cs/0601045, 2006.

[16] Oren Kurland and Lillian Lee. Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In *SIGIR*, pages 83–90, 2006.

[17] Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[18] Jakub Piskorski, Hristo Tanev, Martin Atkinson, and Erik Van der Goot. Cluster-centric approach to news event extraction. In *New Trends in Multimedia and Network Information Systems*, pages 276–290. 2008.

[19] Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84(1):25–43, 2001.

[20] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *NIPS*, pages 617–623, 1999.

[21] N. Tideman. *Collective Decisions and Voting: The Potential for Public Choice*. Ashgate Publishing, 2006.

[22] Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.

[23] Gu Xu and Wei-Ying Ma. Building implicit links from content for forum search. In *SIGIR*, pages 300–307, 2006.