Towards Characterization of Actor Evolution and Interactions in News Corpora

Rohan Choudhary[1,⋆], Sameep Mehta[2], Amitabha Bagchi[1], and Rahul Balakrishnan[1]

[1] Indian Institute of Technology,New Delhi, India
[2] IBM India Research Lab, New Delhi, India

Abstract. The natural way to model a news corpus is as a directed graph where stories are linked to one another through a variety of relationships. We formalize this notion by viewing each news story as a set of actors, and by viewing links between stories as transformations these actors go through. We propose and model a simple and comprehensive set of transformations: *create, merge, split, continue,* and *cease*. These transformations capture evolution of a single actor and interactions among multiple actors. We present algorithms to rank each transformation and show how ranking helps us to infer important relationships between actors and stories in a corpus. We demonstrate the effectiveness of our notions by experimenting on large news corpora.

1 Introduction

Browsing news websites and searching for relevant news forms a major portion of a user's interaction with the web. With the presence of efficient and accurate search engines, it has become extremely simple for a user to find news of interest. However, the amount of online news data available makes it difficult and time consuming for the user to logically arrange and read the news. Therefore, there is a strong need to organize the data in a manner that allows the user to extract meaningful information quickly. Moreover the user must be presented news items in a manner which captures the interrelated nature of news items in an evolving news corpus. Simply arranging news items in order of their timestamps is not enough.

Kerry says President would cut retiree payouts: " That's up to $500 a month less for food, for clothing, for the occasional gift for a grandchildren." *Kerry* warned on Sunday as he addressed elderly and middle-aged worshipers at a black church in Columbus, Ohio, bringing to the fore a major issue in the election that he has rarely touched on. Kerry's comments on *social security* came as he headed to Florida for a voter turnout push timed to Monday's start of early voting.

News Story 1

The Topic Detection and Tracking (TDT) [1] research initiative was formed in 1998 to address such issues in news organization. A topic is defined as a cluster of news stories connected by a seminal event. For example, the US elections 2004 is a topic and all the news stories connected with it are labeled as being inside the topic. Nallapati et.al. [5] presented an algorithm to discover dependencies between news stories by

⋆ rohan@cse.iitd.ernet.in

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 422–429, 2008.

taking into account the content of the news. For example, in US Elections 2004 topic, stories about Bush are related to each other and stories about Kerry are related to each other. The news items can now be arranged as a graph such that each node represents one news item and each edge captures both kinds of dependencies between two news stories: textual and temporal.

Same-Sex Marriages: Bush Backs Ban in Constitution Pres *Bush* backs constitutional amendment to ban *same-sex* marriages; holds marriage cannot be separated from its 'cultural, religious and natural roots' without weakening society

News Story 2

These algorithms were based on the key assumption that *a single theme is associated with each news item*. However, this assumption does not hold true in many cases. For example, a news item discussing Bush's health care policy indeed has two themes/actors *Bush* and *Health Care*. Going beyond just a simple multiplicity of actors is the fact that the interrelationship between actors is major feature of a news corpus, and it is a feature that users look for, implicitly or explicitly. Keeping this in view our key contention is this: *Actors interact and these interactions provide valuable cues which can be used to discover useful parts, patterns and properties of the news corpus.* We define five key types of evolutions/transformations which actors can undergo. These are *create, merge, split, continue,* and *cease*. Some of the transformations inside a news corpus are more important than others. Based on this idea, we provide quantitative metrics to measure importance of any transformation. The usefulness of these transformations is demonstrated by the empirical observation that top ranked transformations in-fact, correspond to important events, stories and actors in a news corpus.

The Final Debate: The mission of Wednesdays night presidential debate was to engage *George W. Bush* and *John Kerry* in a discussion of domestic issues. True, both men hewed to their talking points and tried harder to score cheap shots than to offer clear explanations. But its hard to believe that anyone who watched with attention didn't come away with a good handle on who John Kerry and George W. Bush are, what they believe, and how they would approach running the country

News Story 3

The focus of this article is to characterize the interactions among actors and propose quantifying measures for them. We do not approach the problem of identifying actors, instead we depend on the algorithms proposed by Mei et al. [3] to identify actors/themes. To reiterate,the key contributions of this paper are: i) We present an actor based view of news corpora and posit an interaction graph of actors as the appropriate organizational framework for these corpora. ii) We define, discover and rank the key transformations that capture the evolution of a single actor and its interactions with other actors. We also empirically show how the ranking aids a user in retrieving important and interesting aspects of the news corpus.

We have also proposed an automatic interaction graph generation algorithm. The algorithm enforces the top transformations that are mined from the news corpus. Due to shortage of space, we have not detailed the algorithm in this paper. Interested Readers are directed to [2] for details of the algorithm.

2 An Actor Based View of News Corpora

In this section we present and develop an actor based view of news corpora. We first define interaction graphs which form the basic structure of a news corpus organized by actors, then we study the transformations these actors undergo in the interaction graphs.

The Interaction Graph

The basic structure we proceed with is an *interaction graph* which is a major improvement on the structure proposed by Nallapati et. al. [5]. In our interaction graph each story is represented by a node. The actors present in a story are enumerated inside the node. Links may be established between news stories having common actors. Edges connecting two stories are annotated with actors common to the two stories. It is our contention that this is a natural and satisfactory way of organizing a news corpus being presented to a human user. For expository purposes consider a news corpus consisting of three news items: S1, S2 and S3(temporally ordered). Relevant actors in each news item are identified and marked. An interaction graph of these stories is shown in Figure 1. Stories 2 and 3 are linked because of the presence of a common actor, i.e., *Bush*. The actors ***Bush*** and ***Kerry*** both are present in Story 3. The presence of edges from Story 1 and Story 2 to Story 3 implies that previously non co-occurring actors appeared together in story 3.We call such a transformation a *merge* of two actors. Similar definitions hold for other transformations. Once all the transformations have been discovered we score them to ascertain their significance using a scoring procedure that takes into account stories in the temporal neighborhood.

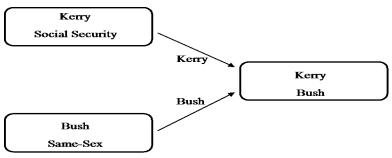

Fig. 1. Interaction Graph for the three stories

We would like to clarify again that we have not discussed the interaction graph generation algorithm in this paper. For details of the algorithm, the reader is encouraged to look at our technical report. [2].

Basic Notations

Given a news corpus consisting of D news items with respective time stamps $\{t_1, t_2 \ldots, t_D\}$, where $t_i \leq t_{i+1}$ D_i represents i^{th} news item with a time stamp of t_i. Associated with each news item D_i is an actor vector K_i of length n_i, $\{K_i^1, K_i^2, \ldots K_i^{n_i}\}$. A word or a phrase appearing in the news corpus is considered an actor if it occurs repeatedly in a time period. This vector can be derived by using the theme extraction algorithms proposed by Mei and Zhai [3]. These actors are subsets of salient themes across a topic. $G^l = (V^l, E^l)$ denotes a news graph till time t_l. Whenever there is no ambiguity we denote the graph simply by G. Each node represents a unique news item, i.e., $|V_l|$ is same as the number of news items collected till t_l and vertex V_i represents news item D_i. A direction edge $e_{(i,j)}$ from node(news items) V_i to V_j implies that $t_i < t_j$ and there is overlap between the corresponding actor vectors, i.e., $K_i \cap K_j \neq \phi$ We maintain the list of actors associated with such an edge in $K_{i,j}$. Also let $C_{t_j}^{t_l} = \cup_{i=j}^{l} K_i$ represent the set of all the actors discovered in the time window $[t_j, t_l]$.

Actor transformations

We now develop a framework for extracting information from news corpora:*actor transformations*. We contend that the interaction between news stories can be modeled as one of five fundamental transformations that one or more actors involved in those news stories undergo. These five transformations are *create, merge, split, continue* and *cease*. We assume that G^l and other variables, as defined above are available to us. The definitions below then serve as a way of mining the transformations at news story D_i. We now formally define these transformations. Figure 2 shows a sample interaction graph for US election 2004. The numbers inside the node establish a temporal order (not continuous dates) and the annotation on the edge represents the common actors. We will require the following functions for this formalization and the other measures we define in later sections:

Membership Testing Function: The declaration of this function is **BOOL IsMember(List, A)**. The function returns TRUE if $A \in List$ else it returns FALSE.

Set Intersection Function: The declaration of this function is **List SetIntersect (List$_1$, List$_2$)**. This function returns a list of actors common in both $List_1$ and $List_2$.

Set Union Function: The declaration of this function is **List SetUnion(List$_1$, List$_2$)**. This function returns a list of actors present in either $List_1$ or $List_2$.

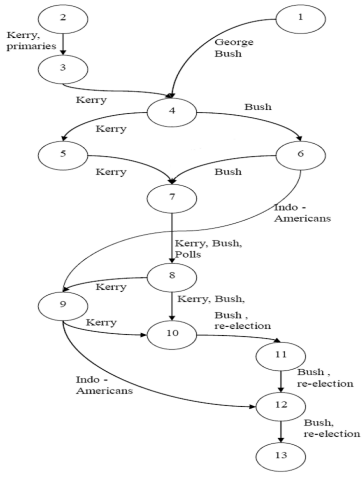

Fig. 2. Example of an Interaction Graph

Merge Actors $\underline{A}$ and $\underline{B}$ are marked as merged at D_i if the following conditions hold:

Condition 1- $\underline{A}$ and $\underline{B}$ are present in K_i.

Test: IsMember(K_i,A) = T $\wedge$ IsMember(K_i,B) = T.

Condition 2- Both $\underline{A}$ and $\underline{B}$ never co-occur in an edge to this news story D_i.

Test: $\nexists$ j $< (i)$ IsMember($K_{j,i}$,A) = T $\wedge$ IsMember($K_{j,i}$,B) =T

In figure 2 actors in node 1($\underline{Bush}$) and 3 ($\underline{Kerry}$) merge at node 4.

Split Actors $\underline{A}$ and $\underline{B}$ are marked as split at D_i if:

Condition 1- $\underline{A}$ and $\underline{B}$ co-occur at t_i

Test: (IsMember(K_i,A) = T $\wedge$ IsMember(K_i,B) =T)

Condition 2- There is a news story D_k such that there is an edge from Story D_i to D_k and only actor $\underline{B}$ is present in the Story D_k.

Test: $\exists$ k $> i$ IsMember($K_{i,k}$, A) = F $\wedge$ IsMember($K_{i,k}$, B) = T.

Condition 3- There is a news story D_j such that there is an edge from Story D_i to D_j and only actor $\underline{A}$ is present in the Story D_j.

Test: $\exists$ j $> i$ IsMember($K_{i,j}$, A) = T $\wedge$ IsMember($K_{i,j}$, B) = F.

Condition 4- There is no news news story D_j such that there is an edge from Story D_i to D_j and both actor $\underline{A}$ and $\underline{B}$ are present in the Story D_j.

Test: $\nexists$ j $> i$ IsMember($K_{i,j}$, A) = T $\wedge$ IsMember($K_{i,j}$, B) = T.

Please note that swapping A and B in the above conditions also constitutes a valid split. An example of split can be seen at node 4 because the co-occurring actors now occur individually at node 5 and node 6.
Create An actor A is marked as created at D_i if:
Condition 1 A is present in K_i
Test: IsMember(K_i,A) = T
Condition 2 There is no news story D_j such that there is an edge from D_j to D_i and A is present in K_j
Test: $\nexists$ j $< (i)$ IsMember($K_{j,i}$,A) = T
Indo-Americans and *Polls* was created at node 6 and node 7 respectively.
Continue An actor A is marked as continued at D_i if:
Condition 1 A is present in K_i
Test: IsMember(K_i,A) = T
Condition 2 There is a news story D_j such that A is present in K_j and there is an edge from D_j to D_i
Test: $\exists$ j $< (i)$,IsMember($K_{j,i}$,A) = T
Polls continued at (7,8) whereas *Bush* was present at (1,4,6,7,8,10,11,12,13).
Cease An actor A is marked as ceased at D_i if:
Condition 1 A is present in in K_i
Test: IsMember(K_i,A) = T
Condition 2 There is no news story D_j such that A is present in K_j and there is an edge from D_i to D_j
Test: $\nexists$ j $> (i)$,IsMember($K_{i,j}$,A) = T
Indo-Americans and *Polls* ceased to exist after node 12 and node 8 respectively.

We would like to emphasize that each news story or an actor can be involved in multiple transformations.For example between node 11 and node 12 in Figure 2 *Bush* is continuing as well as merging with *Indo-Americans*.

3 Ranking Transformations

The actor transformations described in the previous section can be used to gain insights into the data and extract useful information about the structure, evolution, key events, and storylines of a topic. However, in a typical large news corpus, we expect to discover a number of key transformations. To extract useful information, the user would have to iterate through all the transformations and find the important ones. This iterative process will soon become cumbersome and error prone. Therefore, one major challenge is to rank the discovered transformations.

In this section we, first, define a set of metrics to quantify the importance of an actor and co-occurrences of two or more actors. These metrics are then used to rank the transformations. Recall that $List_A^{[t_1,t_2]}$ denotes list of all the news stories in the time interval $[t_1, t_2]$ containing actor A and $N^{[t_1,t_2]}$ represents total stories in the interval $[t_1, t_2]$.

Strength: Strength of A during time interval $[t_1, t_2]$ is: $Strength_A^{[t_1,t_2]} = \frac{|List_A^{[t_1,t_2]}|}{N^{[t_1,t_2]}}$
This metric captures the fraction of news stories in which an actor appears during a given time interval. $Strength_A^{[t_1,t_2]} = 1$ implies that all the news stories contain A

and therefore A is regarded as a very important actor in the specified time period. This metric is used to rank individual actors. The definition can be extended to calculate the collective strength of a set of L actors as: $Strength^{[t_1,t_2])}_{(A_1,A_2...A_L)} = \frac{|\cap_{i=1}^{L} List_{A_i}^{[t_1,t_2]}|}{N^{[t_1,t_2]}}$

Coupling: Coupling between A and B during time interval $[t_1, t_2]$ is given by: $Coupling^{[t_1,t_2]}_{(A,B)} = \frac{|List_A^{[t_1,t_2]} \cap List_B^{[t_1,t_2]}|}{|List_A^{[t_1,t_2]} \cup List_B^{[t_1,t_2]}|}$

This metric measures co-occurrence of A and B in the given time period, i.e, how many news stories contain both A and B. $Coupling^{[t_1,t_2]}_{(A,B)} = 1$ implies that all the news stories in the given time period which contain A (B) also contains B (A) which implies a high and therefore an important coupling.

Next, we discuss how these metrics are used to rank the transformation. In this discussion we will be using P to denote a retrospective window i.e. P is the number of previous time steps (news stories) that are taken into account. Similarly, F denotes a future window i.e. F is the number of subsequent time steps (news stories) that are taken into account. The reader is encouraged to read our technical report [2] for the motivation behind the measures.

Importance of Split Transformation: A split transformation between A and B at time t is considered important if i) $Strength^{[t-P,t]}_{(A,B)}$ is high and ii) $Coupling^{[t,t+F]}_{(A,B)}$ is low. Using these two conditions, score of a split is given as: $\frac{e^{Strength^{[t-P,t]}_{(A,B)}}}{e^{Coupling^{[t,t+F]}_{(A,B)}}}$

Importance of Merge Transformation: A merge transformation between A and B at time t is considered important if i) $Strength_A^{[t-P,t]}$ and $Strength_B^{[t-P,t]}$ is high and ii) $Coupling^{[t-P,t]}_{(A,B)}$ is low. Using these two conditions, the score of a merge is given as: $\frac{e^{Strength_B^{[t-P,t]}} \times e^{Strength_A^{[t-P,t]}}}{e^{Coupling^{[t-P,t]}_{(A,B)}}}$

Importance of Continue Transformation: : Continuation of concept vector $K_{i,j}$ from story D_i to D_j is important if $Strength^{[t-H,t+F]}_{K_{i,j}}$ is high. The score simply is collective strength of $K_{i,j}$ in $[t - H, t + F]$.

Importance of Create Transformation: Creation of A at time t is considered important if $Strength_A^{[t,t+F]}$ is high. F denotes the number of future time steps (news stories) which should be considered to ascertain the quality of create transformation. The score is simply its strength in $[t, t + F]$.

Importance of Cease Transformation: Cessation of A at time t is considered important if $Strength_A^{[t-P,t]}$ is high. The score is simply the strength in $[t - P, t]$.

4 Experimental Results

Due to lack of space we provide detailed experiments only on FIFA World Cup, 2006 and US Presidential Elections, 2004. In our technical report [2], we have discussed experiments on other datasets and more interesting inferences.

FIFA World Cup, 2006: The first dataset FIFA World Cup 2006, consists of 459 news stories published between 02/06/2006 and 15/07/2006 by www.rediff.com.The main actors of the topic are the teams and some of the well reported players. We mined the transformations from the complete FIFA dataset. Next, we assigned scores to the transformations and picked the top 14 merges. The stories associated with these 14 transformations are shown in Table 1. The first column shows the stories according to their rank (in decreasing order) and the second column shows the same transformation arranged by time (decreasing). As evident from the list all the major stories received high score. These results strengthen our belief that the ranking procedure is indeed useful and that the user can be provided top stories based on score. The user can then explore any of these stories in more detail. Similarly, the top two creations discovered are: *Zidane's Head Butt* and *Fan Clashes*.

Table 1. Top ranked Merges in FIFA

Germany v/s Portugal	Italy v/s France
Portugal v/s England	Germany v/s Portugal
Italy v/s France	France v/s Portugal
Italy v/s Germany	Italy v/s Germany
Argentina v/s Germany	Brazil v/s France
Brazil v/s France	England v/s Portugal
France v/s Portugal	Argentina v/s Germany
England v/s Ecuador	Italy v/s Ukraine
England v/s Sweden	Spain v/s France
Sweden v/s Germany	Brazil v/s Ghana
Spain v/s France	Italy v/s Australia
Brazil v/s Ghana	Germany v/s Sweden
Italy v/s Australia	England v/s Sweden
Italy v/s Ukraine	Germany v/s Ecuador

US Elections 2004: This dataset consists of 389 news stories published between 02/02/04 and 15/11/04 by nytimes.com. The key actors of the topic are *Bush*, *Kerry* and important election issues like *abortion* and *social security*. We again mined the top transformations from the corpus and ranked them using our measures. Table 2 shows the abstract of new stories where top 8 creations occurred in this corpus. The actual actors are also noted in the table. The size of future window F is taken as 8 days. We can see that the top creations actually correspond to the major stories and events inside the topic.

5 Related Work

Topic detection and tracking has been a popular research topic in the areas of text mining, information retrieval and organization. Interested readers are pointed to an excellent survey in [1] The need for having a temporal structure within a topic was identified by Nallapati et al. [5]. The authors proposed a directed acyclic graph where each node represented an event and each edge represented a dependency between the two nodes. Although we also work on directed acyclic graph, the nodes in our graph are the individual news stores. Also, in their work, the focus was on generating the graph. In this paper, we use properties of the graph to draw interesting inferences about the topic.

Table 2. Synopsis of the top ranked creations in US Election 2004 corpus

Date	Story and Creation of Actor
30/08	Republican Convention kicks off (convention)
13/04	Iraq issue starts coming up (Iraq)
06/07	Kerry chooses Edwards as running mate (Edwards, running mate)
14/05	Issue of same-sex marriage (same-sex marriage)
28/07	Issue of economy during democratic convention (economy)
28/07	Issue of global terrorism at democratic convention (terror)
01/08	Republicans challenge Kerry's Vietnam records (Vietnam)
13/05	Ralph Nader wins endorsement of Reforms Party (Nader)

The problem of discovering evolutionary theme patterns from text was first identified by Mei and Zhai [3,4]. The authors defined notion of theme across a time period and salient themes across the whole topic. The evolution of a theme was captured, however, the interaction between themes was not accounted for. The algorithms proposed in [3] can be used for detection of the major actors of a topic. Mei and Zhai [4] also demonstrated that a document can belong to multiple contexts. This is very similar to our modeling of each news story as an interaction of major actors which belong to that story. In their seminal work, Silver and Wang [6] enumerated the key transformations which a three dimensional scientific feature can undergo. Recently Spiliopoulou et al. [7] presented similar transformations to capture and monitor evolving clusters. Both these algorithm defined a customized overlap (intersect) function to derive the relationships. Our algorithms use set intersection algorithm.

6 Conclusions

In this article we presented definitions and algorithms for discovering the key transformations which actors in a news corpus can undergo. The intuition behind our approach is that each news story encompasses multiple themes/actors. Each individual actor evolves over time and simultaneously interacts with other actors. These interactions point to interesting and important parts of a news corpus. To reduce the number of transformation which the user has to evaluate, we outlined a scoring procedure to rank the transformations. We empirically showed that the transformations with high score typically point to the important stories in the corpus by discussing the results on two large datasets.

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: Final report. In: DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (2006)
2. Choudhary, R., Mehta, S., Bagchi, A., Balakrishna, R.: A framework for exploring news corpora by actor evolution and interaction. IBM Research Report- RI07004 (2007)
3. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: KDD 2005: 11th ACM SIGKDD international conference on Knowledge Discovery and data mining, pp. 198–207 (2005)
4. Mei, Q., Zhai, C.: A mixture model for contextual text mining. In: KDD 2006: 12th ACM SIGKDD international conference on Knowledge Discovery and data mining, pp. 649–655 (2006)
5. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: CIKM 2004: 13th ACM International Conference on Information and Knowledge Management, pp. 446–453 (2004)
6. Silver, D., Wang, X.: Volume tracking. In: VIS 1996: 7th conference on Visualization, pp. 157–164 (1996)
7. Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: Monic: modeling and monitoring cluster transitions. In: KDD 2006: 12th ACM SIGKDD international conference on Knowledge Discovery and data mining, pp. 706–711 (2006)