R107004 May 2007 Computer Science

IBM Research Report

A Framework for Exploration of News Corpora by Actor Evolution and Interaction

Rohan Choudhary
Dept. of Computer Science and Engineering
Indian Institute of Technology
New Delhi - 110016. India.

Sameep Mehta
IBM Research Division
IBM India Research Lab
Plot No -4 Phase 2 Block-C, Vasant Kunj
New Delhi - 110070, India

Amitabha Bagchi
Dept. of Computer Science and Engineering
Indian Institute of Technology
New Delhi - 110016. India.

Rahul Balakrishnan
Dept. of Computer Science and Engineering
Indian Institute of Technology
New Delhi - 110016. India.

Abstract

We present a general framework for modeling and exploration of news corpus. 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 as well as interactions among multiple actors. We present metrics to assign a score to each discovered transformation. These scores quantify the importance of individual events and aid in ranking the transformations. We show how ranking helps us infer important relationships between actors and stories in a corpus. Next, the derived transformations and associated ranking is used to generate a news graph. To handle the large size of the graph we propose summarization scheme which again leverages the derived transformations. Finally, we propose a interface which aid user to explore the corpus in a interactive fashion and finds the information of interest in an iterative manner. 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. Simply arranging news items in order of their timestamps is not enough. The content of the story has to be central to any system that enhances a user's news reading experience on the web.

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. However, this definition does not provide information about dependencies among stories. To alleviate this problem Nallapati et. al. [17] presented an algorithm to discover dependencies between news stories by 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.

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 evolutions/transformations which actors can undergo. These are *create, merge, split, continue,* and *cease*.

The natural way of organizing a news corpus is as an directed graph wherein the nodes represent the stories and the edges capture the relationship amongst the stories. The direction of the edge represents the time dimension. We use the above mentioned transformations and the quantification metrics to construct a news graph. The fundamental idea behind our work is this: *quantitatively strong transformations give us important qualitative insights into the news corpus.* This leads us to contend that any method *must include the quantitatively highest ranked transformations in the news graph it constructs.*

This intuitive understanding, if captured by an automated algorithm, can aid users tremendously in extracting important information efficiently. However, the specific application of news browsing is not the ideal candidate for complete automation. Each user reads and disseminates news in her own unique fashion. The users may have different needs, interests, and inclination. Its is implausible (if not impossible) to comprehensively incorporate all such information in an automated algorithm. User profiling algorithm help solve this problem to some extent but not completely. Therefore, we contend that the users should be provided with an interactive visual interface which aids her in focusing on parts of news of interest and help find relevant information efficiently. The toolkit provides each user the capabilities to perform a focused and goal oriented search. It supports the well-established *zoom, filter* and *details–on–demand* paradigm used for information visualization.

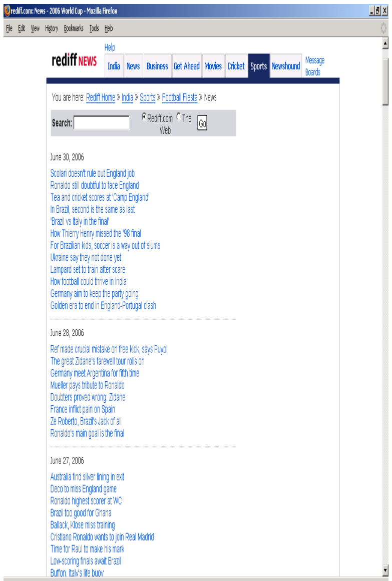
While visualizing the news graphs we noticed that even for a small news corpus of 200 stories (therefore 200 nodes) it becomes difficult to find useful information. This is primarily due to large number of edges. Since we take into account both temporal and textual information while generating news graphs a news story may be linked to multiple other stories. The view becomes cluttered and browsing became un-principled, which defeats this whole excises. Therefore, the visualization interface is boot strapped with a summarization step. Our news graph structure naturally facilitates summarization of the news corpus. We present an algorithm which looks at the strength of individual story and evolution of the relationship among the stories and aggregates them together if they are not particularly important. The refined (or summarized) graph is much easier to display and handle but may not provide all the information needed by the user. However, the user can expand the summarized nodes using the visual interface by using the mouse. This feature provides user the capability to analyze the news corpus in a multi resolution fashion.

> 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 in Ohio. 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

> 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



You are here: Rediff Home » India » Sports » Football Fiesta » News

June 30, 2006

Scolari doesn't rule out England job
Ronaldo still doubtful to face England
Tea and cricket scores at 'Camp England'
In Brazil, second is the same as last
'Brazil vs Italy in the final'
How Thierry Henry missed the '98 final
For Brazilian kids, soccer is a way out of slums
Ukraine say they not done yet
Lampard set to train after scare
How football could thrive in India
Germany aim to keep the party going
Golden era to end in England-Portugal clash

June 28, 2006

Ref made crucial mistake on free kick, says Puyol
The great Zidane's farewell tour rolls on
Germany meet Argentina for fifth time
Mueller pays tribute to Ronaldo
Doubters proved wrong: Zidane
France inflict pain on Spain
Ze Roberto, Brazil's Jack of all
Ronaldo's main goal is the final

June 27, 2006

Australia find silver lining in exit
Deco to miss England game
Ronaldo highest scorer at WC
Brazil too good for Ghana
Ballack, Klose miss training
Cristiano Ronaldo wants to join Real Madrid
Time for Raul to make his mark
Low-scoring finals await Brazil
Buffon, Italy's life buoy

Figure 1: Snap shot of a typical news archive webpage.

To reiterate, the major contributions of this paper are :

1. We present an actor based view of news corpora and posit an interaction graph of actors as the appropriate organizational framework for these corpora..

2. We define and characterize key transformations that capture the evolution of a single actor and its interactions with other actors.

3. We present a scoring mechanism which assigns quantitative importance to each transformation. We also show how the ranking of transformations aids a user in retrieving important and interesting aspects of the news corpus. Additionally, we use the transformations to derive relationships between actors.

4. We propose an automatic news graph generation algorithm. The algorithm works by including quantitatively important actor transformations that are mined from the news corpus. The algorithm is fast and can process a news corpus of 400 stories in 5

5. We present methods to refine/summarize the news graphs. The summarization part examines the sequences of transformation to remove/aggregate parts of graph. The refined graph is much easier to display.

6. We develop a visual interface to display the news graphs. The user can interact with the graph to find the information efficiently. We also provide filters which can be used to prune "unwanted" parts of the graph and focus on stories which are more appealing to the user.

7. Finally, we evaluate our algorithm on several large news corpora. We demonstrate the usefulness of our visual toolkit for effective goal oriented search.

The rest of the paper is organized as follows. We provide the motivation and background in section 2. Section 3 describes the notations, formal problem definition, key transformation and respective mining algorithms. The ranking procedure and other derived structural/temporal relationships are presented in Section 4. The graph generation algorithm is presented in Section 5. The graph summarization algorithm is presented in section 6 In Section 7, the visualization interface is presented. We present our results along with the user study in section 8. In Section 9, we discuss related work most pertinent to this work. Finally, section 11 concludes the paper with a brief summary of our contributions.

2 Motivation and Background

Before presenting the framework, we define and motivate interaction graphs which form the basic structure our work.

2.1 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. [17]. 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. This notion of an interaction graph captures the way in which a human newsreader organizes relationships between news items in her head. 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.

> 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 didnt 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

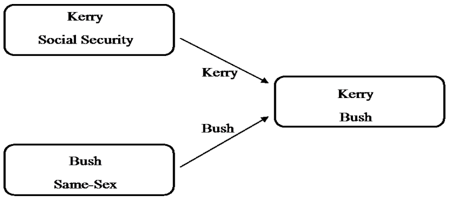


Figure 2: Interaction Graph for the three stories

An interaction graph of these stories is shown in Figure 2. Stories 2 and 3 are linked because of the presence of a common actor, i.e., *Bush*. The actors <u>*Bush*</u> and <u>*Kerry*</u> 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. The top few stories can then be extracted and arranged chronologically to get an overall view of the data. The transformations are also used to derive several interesting and meaningful relationships between actors.

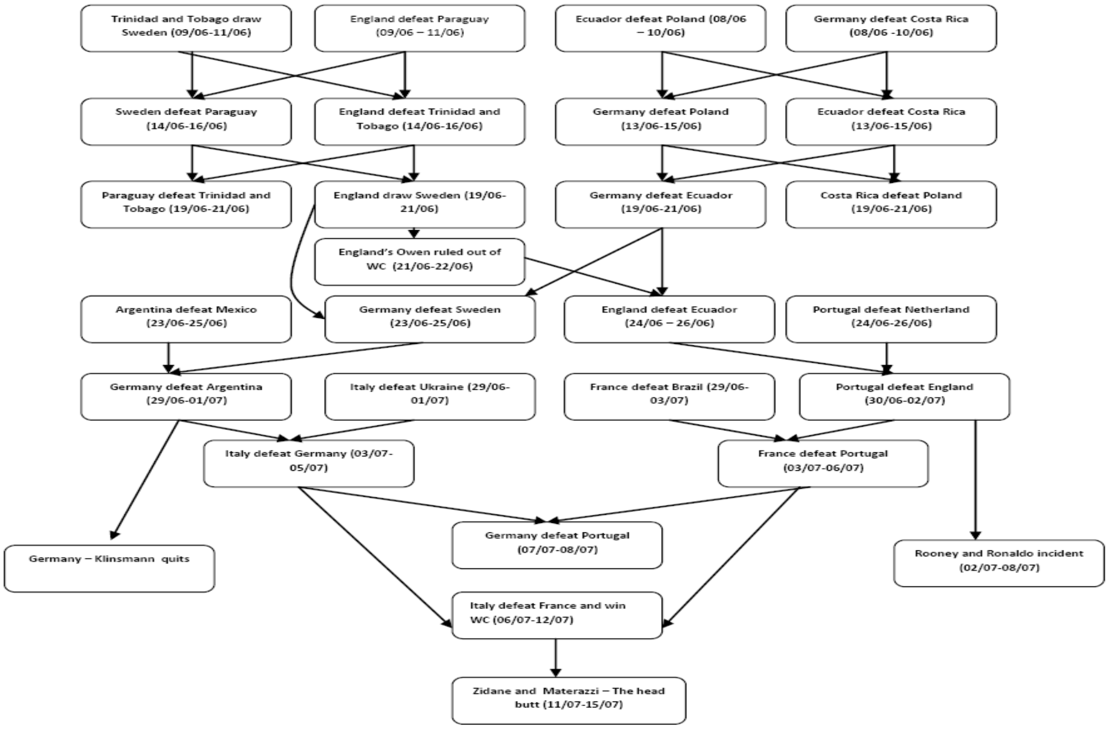

Trinidad and Tobago draw Sweden (09/06-11/06)
England defeat Paraguay (09/06 – 11/06)
Ecuador defeat Poland (08/06 – 10/06)
Germany defeat Costa Rica (08/06 -10/06)
Sweden defeat Paraguay (14/06-16/06)
England defeat Trinidad and Tobago (14/06-16/06)
Germany defeat Poland (13/06-15/06)
Ecuador defeat Costa Rica (13/06-15/06)
Paraguay defeat Trinidad and Tobago (19/06-21/06)
England draw Sweden (19/06-21/06)
Germany defeat Ecuador (19/06-21/06)
Costa Rica defeat Poland (19/06-21/06)
England's Owen ruled out of WC (21/06-22/06)
Argentina defeat Mexico (23/06-25/06)
Germany defeat Sweden (23/06-25/06)
England defeat Ecuador (24/06 – 26/06)
Portugal defeat Netherland (24/06-26/06)
Germany defeat Argentina (29/06-01/07)
Italy defeat Ukraine (29/06-01/07)
France defeat Brazil (29/06-03/07)
Portugal defeat England (30/06-02/07)
Italy defeat Germany (03/07-05/07)
France defeat Portugal (03/07-06/07)
Germany – Klinsmann quits
Germany defeat Portugal (07/07-08/07)
Rooney and Ronaldo incident (02/07-08/07)
Italy defeat France and win WC (06/07-12/07)
Zidane and Materazzi – The head butt (11/07-15/07)

Figure 3: Summarized interaction graph for FIFA World Cup 2006.

To motivate the need for such a graph consider, figure 1 which presents a snapshot of a typical news archive webpage (FIFA in this case). The snapshot shows the archive of FIFA World Cup 2006. The stories are grouped by date of publication and the archive can be searched by keyword. Such an interface provides very limited functionality. A very simple and common task of finding related stories becomes extremely hard. The user has to read the headlines and guess if the stories are related. Then, she can click on the story, read it and the process has to be repeated till she is satisfied. There is no explicit information which will provide cues, visual or otherwise, to the user to aid in this task. On the other hand the news graph presented in this work encodes such cues in the directed edges and the labels of each node. Figure 3 shows a highly summarized news graph generated from the stories in FIFA archive. It is evident that each node captures important event. The directed edges capture relationships among different news stories. For example, a user can visually ascertain that the (in)famous head butt is related to the FIFA final. Similarly, simply traversing the graph from a node provides the timeline of the actors in the node. In section 8 we pose common questions which will interest a user and outline the process of answering the questions from the news graph. To reiterate, *we strongly believe that a news graph representation provides valuable cues for effective and efficient news browsing.*

Figure 4 presents the overall framework for analysis and visual exploration of news corpora. In this paper we focus on the development of web based interactive visual interface. Our philosophy does not allow us to assume we know what the user needs, instead we provide the user all the information we have, organizing it in an intuitive and easy-to-navigate way. However, understanding that users do not want to process large amounts of information at the same time we have developed summarization strategies. These strategies require minimal amount of extra processing because they leverage already discovered information like the key actors and critical events.

It was our contention that an annotated directed graph structure is the best way to represent an evolving news corpus. Each node in this graph corresponds to a news story. Each node was annotated with *actors* which can be thought of as key player or themes/sub-themes associated with at story. We did not attempt to solve the problem of detecting actors ourselves. Instead we used a simplified version of the method described by Mei et. al. [15].

With the concept of actor at hand, we discover the key transformations. The transformations are

- **Merge:** Two previously non co-occurring actors are marked as merged at time t_i if they appear together in a news story published at t_i.
- **Split:** Two actors are marked as splitted at t_i when they appear together in story at t_i and in separate news stories after t_i.
- **Continue:** An actor appearing in consecutive stories is marked as continued.
- **Create:** The first occurrence of an actor is marked as creation.
- **Cease:** The last occurrence of an actor is marked as cessation.

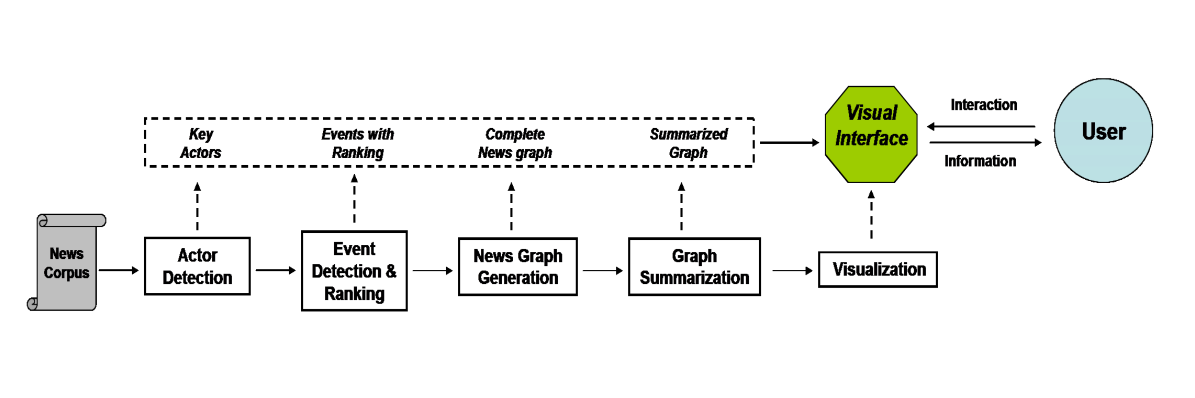


Figure 4: Schematic Overview of the proposed framework.

Next, we develop two metrics *Strength* and *Coupling*. *Strength* measures the importance of an individual actor whereas *Coupling* captures the importance of relationship among actors. These two metrics are then combined to assign a single score to each of the above mentioned transformations.

With these two metrics available we constructed news graphs by the following simple process. On receiving a fresh news story

- We first detected the actors present in it,
- Next, we linked it up to all the news stories occurring in a window before its arrival.
- We found all the actor transformations it was part of.
- We ranked these actor transformations
- We retained the edges corresponding to highly ranked actor transformations and removed the other edges. In our proposed interface the user can iteratively control this step by controlling what fraction of the top ranked transformations are retained.

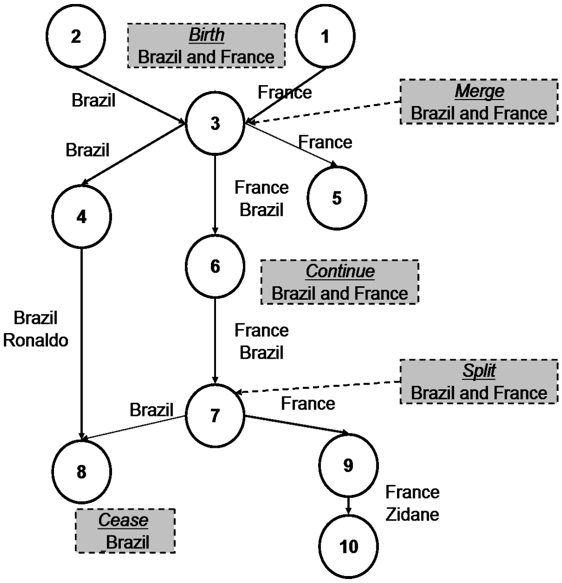


Figure 5: An example news graph highlighting the key transformations. Dotted boxes show the transformation along with the actors involved in it.

Figure 5 shows part of the FIFA news graph generated by our algorithm. The nodes represent the stories where number in node specify the temporal ordering. The labels on edges are the common actors between nodes. The shaded rectangles show the transformations. For example, actors, Brazil and France, appeared separately in stories 1 and 2 and co-occurred at node 3. We mark this as a merge transformation.

Next, we summarize the news graph. The thought process behind the development of our summarization algorithm is similar to the ideas behind temporal reasoning algorithms (c.f. Allen's seminal work [3]). A simple example which captures the flavor of our methods is as follows: "A occurs before B, B occurs before C" implies that "A occurs before C". If event B is not important and can be removed, then we can consider the implication itself as a summary.

Finally, we develop visual interface where the user can interactively explore the news graph. We show how common news based query can be easily answered by using the interface

Prior to this work, several researchers have used the same set of transformations in different areas for a variety of tasks. Silver and Wang [20] were the first one to enumerate the set of key transformations which three dimensional scientific features can undergo. Spiliopoulou et al. [21] presented similar transformations to capture and monitor evolving clusters. Similarly, Asur et al. [4] employed the events to understand evolving graphs. Yang et al. [23] presented algorithms to discover different types of spatio-temporal patterns for scientific datasets (proteins) and pointed towards the possibility of knowledge discovery by capturing the evolution and interaction of patterns again by using same set of transformations. However, none of the previous work assigned any quantitative scores to the events. Moreover, we are not aware of any previous efforts on the critical event aware summarization and visualization schemes.

3 An actor based view of news corpora

In this section present define the key transformation and associated algorithms to mine them.

3.1 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 [15]. These actors are subsets of salient themes across a topic.

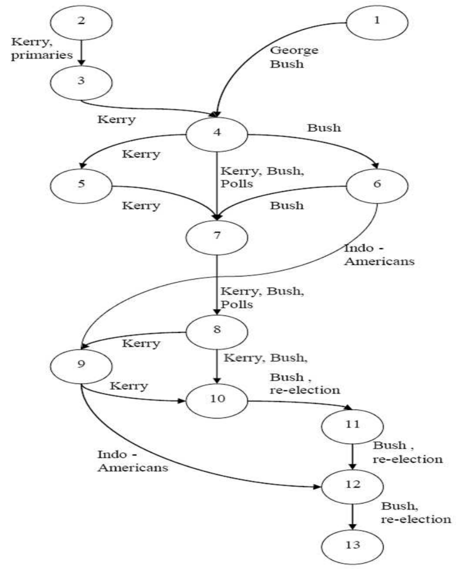


Figure 6: A small part of the interaction graph generated from US Elections 2004 News Corpus

$G^l = (V^l, E^l)$ 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 $[t_j, t_l]$.

3.2 Actor transformations

We now develop a framework for extracting information from news corpora:*actor transformations*. It is our contention 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 now formally define these transformations. Figure 6 shows a small part of the graph generated 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 between nodes. Please note, only the edges connecting the nodes present in the selected sub-graph are drawn, other edges are not shown in this figure. This sub-graph is chosen just to explain the transformations. We present other parts in later sections. The definitions below also serve as a way of mining the transformations at a news story D_i in the Graph. 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 $List_2$.

Merge Two 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: $not\exists$ j $< i$ IsMember($K_{j,i}$,A) = T $\wedge$ IsMember($K_{j,i}$,B) =T
Condition 3- There is a news story D_j such that there is an edge from Story D_j to D_i and only actor $\underline{A}$ is present in the Story D_j.
Test: $\exists$ j $< i$ IsMember($K_{j,i}$, A) = T $\wedge$ IsMember($K_{j,i}$, B) = F.
Condition 4- There is a news story D_k such that there is an edge from Story D_k to D_i and only

actor $\underline{B}$ is present in the Story D_k.
Test: $\exists$ k $< i$ IsMember($K_{k,i}$, A) = F $\wedge$ IsMember($K_{k,i}$, B) = T.
In figure 6 actors in node 1(*Bush*) and 3 (*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 $\underline{B}$ are present in K_i.
Test: (IsMember(K_i,A) = T $\wedge$ IsMember(K_i,B) =T)
Condition 2- There is no 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: *not*$\exists$ j $> i$ IsMember($K_{i,j}$, A) = T $\wedge$ IsMember($K_{i,j}$, 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 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.
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 $\underline{A}$ is marked as created at D_i if:
Condition 1 $\underline{A}$ 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 $\underline{A}$ is present in K_j
Test: *not*$\exists$ j $< i$ IsMember($K_{j,i}$,A) = T
Indo-Americans and *Polls* was created at node 6 and node 7 respectively.

Continue An actor $\underline{A}$ is marked as continued at D_i if:
Condition 1 $\underline{A}$ is present in K_i
Test: IsMember(K_i,A) = T
Condition 2 There is a news story D_j such that $\underline{A}$ 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 $\underline{A}$ is marked as ceased at D_i if:
Condition 1 $\underline{A}$ is present in K_i
Test: IsMember(K_i,A) = T
Condition 2 There is no news story D_j such that $\underline{A}$ is present in K_j and there is an edge from D_i to D_j
Test: *not*$\exists$ 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 can be involved in multiple transformations. New actors can be created while old actors can cease within the same news story. Even an actor can be part of multiple transformations between two consecutive news stories. For example between node 11 and node 12 in Figure 6 *Bush* is continuing as well as merging with *Indo-Americans*.

In order to use these transformations to extract information from a news corpus it is important to quantify how strong or weak they are. In the next section we provide measures for the strength of each type of transformation and discuss how we can use these relative strengths to infer relationships between actors.

4 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 order to extract this information we need to quantify these transformations. In this section we define a set of metrics for transformation strength and then go on to demonstrate how the ability to rank transformations relative to each other can help us extract information about actor interrelationships.

To motivate and demonstrate our ideas we use three datasets: US Elections 2004 (TDT topic category 1), the Clinton-Lewinsky Scandal 1998 (TDT topic category 2), and the O.J. Simpson Murder Case 1995 (TDT topic category 3). We also refer to topics from other TDT topic categories wherever relevant.

4.1 Quantifying Transformations

In a typical 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. This ranking is essential to the design of nice multi-resolution visual interfaces (c.f. Section 10) that conform to Shneiderman's mantra [5]: *zoom, filter and details on demand*. We define metrics to quantify the importance of an actor and co-occurrences of two actors. These metrics will then be employed 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 $\underline{A}$ and $N^{[t_1,t_2]}$ represents total stories in the interval $[t_1,t_2]$.

Strength: Strength of $\underline{A}$ during time interval $[t_1,t_2]$ is:

$$Strength_A^{[t_1,t_2]} = \frac{|List_A^{[t_1,t_2]}|}{N^{[t_1,t_2]}} \tag{1}$$

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 $\underline{A}$ and therefore $\underline{A}$ is regarded as a very important actor in the specified time period. This metric is used to rank individual actors. In the US election news corpus, for example, we found that $\underline{Bush}$ and $\underline{Kerry}$ are main actors. Studying the change of strength of an actor over time provides valuable information about the evolution of the actor. A decrease in strength implies decrease in $\underline{A}$'s importance e.g. $\underline{Dean}$ was marked as a very important actor (evident in Figure 22) at the start of the campaign. However, the strength decreased over time (corresponds to Dean's bowing out of the presidential race). Similarly, a periodic behavior may point to seasonal trends. This is very well exhibited by news stories pertaining to movie reviews which show a periodic peak on Fridays. The definition can be extended as below:

Strength: The collective strength of a set of L actors is given by:

$$Strength_{(A_1,A_2...A_L)}^{[t_1,t_2])} = \frac{|\cap_{i=1}^{L} List_{A_i}^{[t_1,t_2]}|}{N^{[t_1,t_2]}} \tag{2}$$

Coupling: Symmetric coupling between $\underline{A}$ and $\underline{B}$ during time interval $[t_1,t_2]$ is given by:

$$Coupling_{(A,B)}^{[t_1,t_2]} = \frac{|List_A^{[t_1,t_2]} \cap List_B^{[t_1,t_2]}|}{|List_A^{[t_1,t_2]} \cup List_B^{[t_1,t_2]}|} \tag{3}$$

This metric measures co-occurrence of $\underline{A}$ and $\underline{B}$ in the given time period, i.e, how many news stories contain both $\underline{A}$ and $\underline{B}$. $Coupling_{(A,B)}^{[t_1,t_2]} = 1$ implies that all the news stories in the given time period which contain $\underline{A\ (B)}$ also contains $\underline{B\ (A)}$ which implies a high and therefore an important coupling. We found that during Clinton-Lewinsky Scandal, $\underline{Clinton}$ and his lawyer $\underline{Kendall}$ remained coupled throughout the investigation whereas the interaction between $\underline{Clinton}$ and $\underline{Linda\ Tripp}$, another key player, was much smaller. Again analyzing this metric over time is also very helpful in understanding the trends of interaction. The above equation captures symmetric coupling. This can be extended to define asymmetric coupling. Asymmetric coupling between $\underline{A}$ and $\underline{B}$ w.r.t. to $\underline{A}$ is given as:

$$Coupling_{(A,B|A)}^{[t_1,t_2]} = \frac{|List_A^{[t_1,t_2]} \cap List_B^{[t_1,t_2]}|}{|List_A^{[t_1,t_2]}|} \tag{4}$$

Similar to *Strength*, these definitions can be also extended to capture collective coupling of L actors.

Next, we discuss how these metrics are used to rank the transformation. We also provide the rationale for this ranking procedure and examples to justify it. 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.

Importance of Split Transformation: A split transformation between $\underline{A}$ and $\underline{B}$ 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:

$$\frac{e^{Strength^{[t-P,t]}_{(A,B)}}}{e^{Coupling^{[t,t+F]}_{(A,B)}}} \tag{5}$$

Rationale: The first condition implies that the actors involved in a split should have occurred together frequently in recent history. The second condition enforces that the future interaction between them should be minimal. Collectively, these conditions imply that *a split is more important if two strong highly interacting actors split and continue without interactions.* Using this measure a very high score was assigned to split of Democratic candidates like *Kerry* and *Dean* at the end of election primaries. After the primaries there was very little interaction between these two.

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

$$\frac{e^{Strength^{[t-P,t]}_{B}} \times e^{Strength^{[t-P,t]}_{A}}}{e^{Coupling^{[t-P,t]}_{(A,B)}}} \tag{6}$$

Rationale: The first condition implies that the actors involved in merge should themselves be important. The second condition enforces that the past interaction between them should be low. Collectively these conditions imply *a merge is more significant if two strong non(low)-interacting actors merge.* Note that the merge score is not dependent on the evolution of these actors in the future. This is because our proposed score captures the non-interacting aspect of actors and therefore any merge involving these actors is important. Using this score we found that in the OJ Simpson murder case the first interaction between the *defense lawyers* and the *prosecution* is a much more significant event than the recurring interactions/merges later during the court trial.

Importance of Continue Transformation: : Continuation of actor 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 $[t-H, t+F]$.
Rationale: : If the collective strength of some actors is high during a period, then this would correspond to a group of stories, in which the actors co-occur. The stories can be related to an important event if their number is significant. Using this score we found that , in US Elections 2004, during August, the continuation of *Vietnam* and *Kerry* got a high score, because news about the Kerry's Vietnam issue occurred regularly in August 2004.

Importance of Create Transformation: Creation of $\underline{A}$ at time t is considered important if $Strength^{[t,t+F]}_{A}$ 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 $[t, t+F]$.
Rationale: The motivation behind this score is simple. By definition, an actor is important if it appears in a lot of stories in the near future. Using this score we found that the creation of actor *Guards* was assigned a high score in the Clinton-Lewinsky scandal because related news of their testimony occurred regularly for the next few days.

Importance of Cease Transformation: Cessation of $\underline{A}$ at time t is considered important if $Strength^{[t-P,t]}_{A}$ is high. The score is simply the strength in $[t-P, t]$.
Rationale: A actor which ceases to exist is important only if it appears frequently in the recent past. Using this score we found that in the O. J. Simpson case *prosecution* was an important actor which ceases to exist after the prosecution rested its case. Just prior to cessation, it was an important actor appearing in almost every story.

Comment on P and F: Our ranking procedure use stories from the recent past, P, and near future, F. The use of these parameters enables the algorithms to find important transformations which are also temporally local. If the whole time interval of the news corpus is used, even the most important transformations will get a low score and therefore, will not be presented to the user so we need to focus in on a time interval of reasonable length.

Moreover, using P and F provides a natural way of obtaining important transformations in a multi-resolution fashion. For example, initially the user can set high values for P and F and discover important transformations. In the subsequent passes the user can decrease P and F and find transformations which are important but not as strong as the ones found in previous iteration.

We would like to emphasize that the goal of this exercise is to not to automatically present the user with the most useful information. We contend that this goal cannot be met because every user reads and assimilates news in a different fashion. Moreover, while browsing news archives, the user might be looking for information which is not marked important by a completely automatic algorithm. Our final aim is to provide a browsing environment to engage the user and present the information in a more meaningful and informative fashion. The user guides the algorithm by setting the parameters and in-turn the results of the algorithm guide the user to change the parameters.

We contend that the measures proposed above allow us to infer actor interrelationships. We now elaborate on this claim.

4.2 Establishing Actor Interrelationships

In this section, we define various relationships which exist between $\underline{A}$ and $\underline{B}$ in a time interval $[t_i, t_j]$. The relationships are derived by using the discovered transformations and above defined metrics. The proposed relationships are:

- **Orthogonal Actors:** $\underline{A}$ and $\underline{B}$ are considered orthogonal actors if (i) the number of merges involving $\underline{A}$ and $\underline{B}$ during $[t_i, t_j]$ is small and (ii) $Coupling_{(A,B)}^{[t_i,t_j]}$ is small. These actors can provide useful information if they were also involved in a split operation before the time period. An example of such actors is $\underline{\textit{Toll}}$ and $\underline{\textit{Relief}}$ in a natural disaster news corpus.

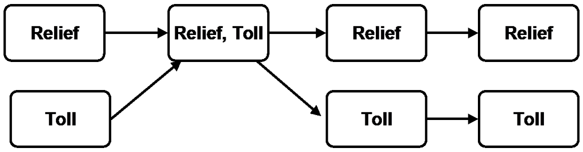


Figure 7: Example of orthogonal Actors.

- **Aligned Actors:** $\underline{A}$ and $\underline{B}$ are considered aligned if (i) the number of merges involving $\underline{A}$and $\underline{B}$ during $[t_i, t_j]$ is small and (ii) $Coupling_{(A,B)}^{[t_i,t_j]}$ is high. Intuitively, the conditions imply that the actors are co-occurring in many stories without going through a series of merges and splits. We found $\underline{\textit{prosecution}}$ and $\underline{\textit{defense}}$ as aligned actors in many time periods during Simpson trial.

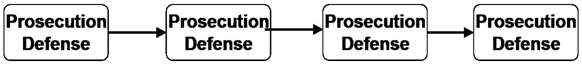


Figure 8: Example of Aligned actors

- **Chain Actors:** $\underline{A}$ and $\underline{B}$ are considered as chain actors (i) if the number of merges involving $\underline{A}$ and $\underline{B}$ during $[t_i, t_j]$ is high and (ii) if the number of splits involving $\underline{A}$ and $\underline{B}$ during $[t_i, t_j]$ is high. This case represents a scenario when the two actors are interacting in an on and off manner. This relationship usually emerges in a bilateral event like a sports series between two teams. After each game there are stories in which both teams co-occur. However, after some time the stories start to focus on individual teams dealing with team selection, strategy and practice for the next game.

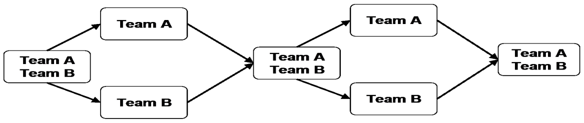


Figure 9: Structural Pattern of interaction between chain actors

- **Sub Actor:** $\underline{B}$ is considered a sub-actor of $\underline{A}$ if $Coupling^{[t_i,t_j]}_{(A,B|B)}$ is very high. The condition implies that in majority of stories where $\underline{A}$ occurs $\underline{B}$ also occurs. A simple example of such relationship is *Kerry* being sub actor of *Elections*, *Democrats* and *President*.

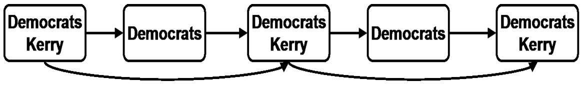


Figure 10: Example of sub-actors

5 Graph generation algorithm

In this section we describe an efficient method for creating news graphs. We create this graph in an evolving fashion i.e. we assume that the graph has been built for the first t news stories that have arrived and explain how to connect the $t+1$st news story to the preexisting graph. The basic idea is that the appearance of this $t+1$st news story means that there is a set of important transformations that the actors in it have undergone, and these transformations must be represented in the news graph we construct. We now discuss this algorithm in detail.

Notation. 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 a actor vector K_i of length n_i, $\{K_i^1, K_i^2, \ldots K_i^{n_i}\}$. $G^l = (V^l, E^l)$ denotes a news interaction graph till time t_l. Whenever clear 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_j ¡ t_i and there is overlap between the corresponding actor vectors, i.e., $K_i \cap K_j \neq \phi$. $C^{t_l}_{t_{l-H}} = \cup^{t_l}_{i=t_{l-H}} K_i$ represent the set of all the actors discovered in the time window $[t_{l-H}, t_l]$.

We also define a proximity measure between any two news stories D_i and D_j,

$$\text{proximity}(D_i, D_j) = \frac{(K_i \cap K_j)e^{-\alpha \cdot |t_i - tj|} Sim(D_i, D_j)}{(K_i \cup K_j)} \tag{7}$$

Sim(D_i,D_j) can be calculated using any standard document similarity measure. In this paper we used TFIDF as feature vector and simple L_2 distance metric. Our proposed proximity measure takes into account the overlap of actors between stories. It also considers the time elapsed between stories and also incorporates traditional distance metrics. We will use this proximity measure later in our algorithm.

The algorithm. In Figure 11, we give a high level view of our algorithm. The algorithm begins by attempting to figure out which are the important transformations for the newly arrived story.

Algorithm Step 1. Temporarily connect all the nodes in V_l^H to D_{l+1}. Use the transformation mining algorithms presented earlier to find all the transformations involving D_{l+1}. Score these transformations using the quantitative measures. Identify the set T of *important* transformations i.e. those with a transformation score greater than a user-defined threshold θ.

Algorithm Step 2. Disconnect the connections made in Step 1. These were made so that we could mine the important transformations and form the set T.

Algorithm Step 3. Form a set of connections A from V_l^H to D_{l+1} which contains all the transformations of T.

It is clear that step 3 can be done in many ways and that the correct connections should be made from stories which have a greater proximity to D_{l+1}. In order to achieve this we formulate step 3 as a weighted set cover problem. In the weighted set cover problem we are given a universe U of elements with $|U| = n$ and a family of sets $F = (S_1 \ldots S_m)$, where $S_i \subset U$, and c_i is the cost associated with S_i, pick sets in F that cover U and have the minimum possible cost. In our case, let $U = (C^{t_l}_{t_{l-H}} \cap K_{l+1}))$. F is the same as V_l^H and the elements inside each set S_i in F are the actors K_i covered by the corresponding story D_i in V_l^H. Since each S_i in our set system corresponds to a news story, we use the inverse of proximity(D_i, D_{l+1}) at the weight of S_i.

Step 3 is executed by then finding $A \subset V_l^H$ consisting of news stories corresponding to the sets forming the minimum cost weighted cover for U. The weighted set cover problem is NP Complete. So we used a well known approximation algorithm [10] which gives a $O(LogD)$ approximation where $|S_i| \leq D \forall$ sets S_i in F.

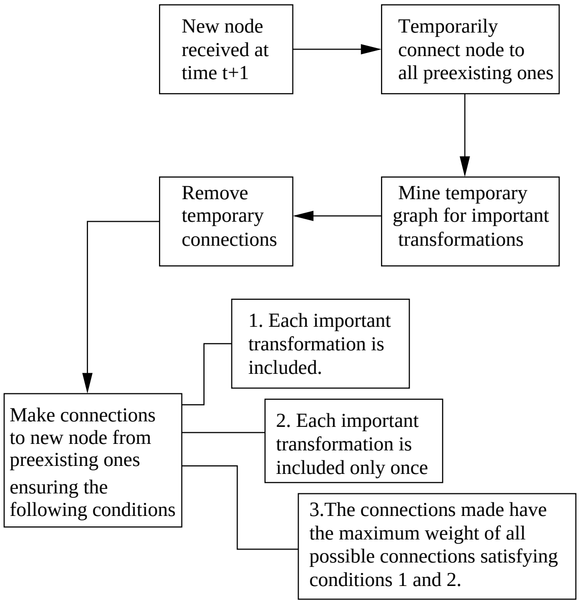

New node received at time t+1
Temporarily connect node to all preexisting ones
Mine temporary graph for important transformations
Remove temporary connections
Make connections to new node from preexisting ones ensuring the following conditions
1. Each important transformation is included.
2. Each important transformation is included only once
3.The connections made have the maximum weight of all possible connections satisfying conditions 1 and 2.

Figure 11: A schematic representation of the algorithm.

	Merge$^{i+1}_{A,B}$	Split$^{i+1}_{A,B}$	Create$^{i+1}_{A}$	Cease$^{i+1}_{A}$	Continue$^{i+1}_{A}$
Merge$^{i}_{A,B}$	X	Continue	X	No Change	No Change
Split$^{i}_{A,B}$	Remove	X	X	No Change	No Change
Create$^{i}_{A}$	No Change	No Change	X	Remove	No Change
Cease$^{i}_{A}$	X	X	Remove	X	X
Continue$^{i}_{A}$	No Change	No Change	X	No Change	Continue

Table 1: Summarization based on Temporal Ordering of the Transformations

Note that the output of a simple weighted set cover algorithm may not satisfy our requirement that all the transformations in T must be represented. To enforce this condition we have to do a little more work. We modify the universe U by adding dummy elements corresponding to transformations in T :

(a) Merge: $T_i = [Merge^{l+1}_{X,Y}]$. Corresponding to a merge of actors X and Y in T, we add the elements X_i and Y_i to U. We add X_i to a set $S_j \in F$ if $X \in K_j$ and $Y \notin K_j$.

A similar definition holds true for Y_i. This addition ensures that the set A that is finally chosen has 2 such news stories D_i and D_j such that one of them contains actor X and the other contains actor Y , but both of them do not contain actor X and Y together. This would always result in a merge of actors X and Y at news story D_{l+1}.

(b) Split: $T_i = [Split^{j}_{X,Y}]$. On adding D_{l+1} a split of actors X and Y may occur at D_j. This could happen, for example, if D_j has a forward edge to a node containing only actor X and D_{l+1} contains actor Y. If this split is important then to enforce it we add the dummy actor XY_{Split} to the D_j and to D_{l+1}. Now, in order to cover this actor, the set cover solution will make an edge from D_j to D_{l+1} thereby enforcing the split.

(c) Continuation: $T_i = [Cont^{l+1}_{X,Y}]$ Corresponding to a continue of X and Y at D_{l+1}, we add element XY_i to U. We add XY_i to a set $S_j \in F$ if $X \in K_j$ and $Y \notin K_j$. Similar to split, this addition ensures that the set A contains a news story D_i which has both X and Y. This would then ensure a continuation of X and Y from D_i to D_{l+1}

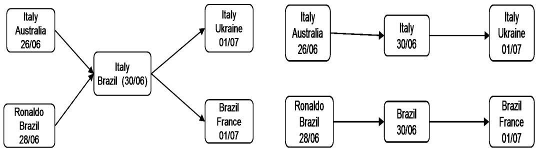
6 Summarization

In this section, we present our summarization algorithm. The method looks at two successive time instants, t and $t + 1$. Depending on the types of the two transformation observed in these time instants we make our summarization decisions. Table 1 presents various scenarios. The row and column heading of the table denote the transformations used at t_i and t_{i+1} to construct the graphs.

Note that Merge$^i_{A,B}$ represents merging of actors A and B at time t_i. A sequence of k transformations is denoted by $[T^i_{Actors}, T^{i+1}_{Actors}, \ldots T^{i+k}_{Actors}]$, where T is one of the key transformations, $Actors$ denotes the list of actors involved in the corresponding transformation and $i < i+1 < \ldots < i+k$.

Symbol X in a cell implies that the corresponding sequence of transformations is invalid. For example, sequence $[Merge^i_{A,B}, Merge^{i+1}_{A,B}]$ is not a valid transformation because two actor A and B merging at t_i cannot merge again at t_{i+1}. This fact is captured by X in cell $(1,1)$. Similarly, $No\ Change$ marks the cases when no refinement is possible. For example, $[Merge^i_{A,B}, Cease^{i+1}_A]$ sequence cannot be summarized/refined. We keep such transformations unchanged. Next, we explain remaining sequences and corresponding refinements with motivation and examples derived from FIFA dataset.

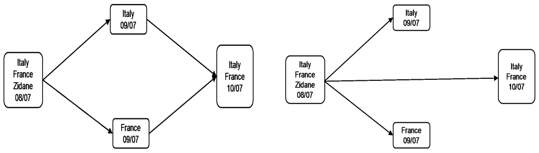
- $[Merge^i_{A,B},\ Split^{i+1}_{A,B}]$ - Figure 12(a) pictorially shows this case. The story observes that Italy and Brazil were potential finalists. In this scenario, the merge transformation is replaced by two continue events: $Continue^i_A$ and $Continue^i_B$. The sequence denotes the interaction of two actors for extremely short period of time. The interaction is captured by the merge event at t_i. However, at t_{i+1} the actors again appear in different non-overlapping news stories. Such short lived interactions are not very useful[1]. Figure 12(b) pictorially shows the resultant graph.



(a) (b)

Figure 12: Example demonstrating the merge followed by split transformation and associated changes to the graph.

- $[Split^i_{A,B}, Merge^{i+1}_{A,B}]$ - This case is exactly opposite to the previous one. Here, two co-occurring actors split for one single time instant and then again co-occur. Figure 13(a) shows one such example from FIFA dataset. The stories in this subgraph were published during the days of the final between Italy and France. The news stories with both Italy and France correspond to stories about final, while stories with individual actors correspond to team specific aspects. Rather than showing this as a merge and a split, one can transform this into a a continuity of Italy and France. Please note that the Italy and France at t_i cannot be represented by a single node because they don't co-occur at t_i (cause of split transformation). Figure 13(b) depicts the new graph.



(a) (b)

Figure 13: Example demonstrating the split followed by merge transformation and associated changes to the graph.

- $[Create^i_A, Cease^{i+1}_A]$ - This sequence points to an actor which was observed in only one news story. Such short lived actors (valid for one time instant) are extremely non-informative in news stories. Figure 14(a) shows such an case where news about Christiano Ronaldo appeared in one story and then there was no other follow up story. Therefore, the actor was removed from this news story. Please note that actor can be important in some other parts of the corpus (which Christiano indeed was) and will not be removed from those parts.

- $[Cease^i_A, Create^{i+1}_A]$ - This sequence essentially points to re-creation of an actor. The actor was present at t_{i-1}, ceased at t_i and is created (again) at t_{i+1}. In such cases, we drop both,

[1] Please note this contention holds in the context of news understanding. In other domains, these short lived sporadic interactions might be extremely useful.

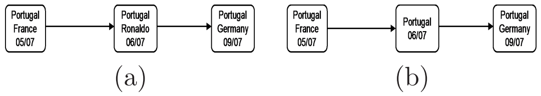


Figure 14: Example showing the case when a actors stops to exist right after its creation.

cease and create, event and the actor is assumed to continue in interval $[t_{i-1}, t_{i+1}]$. Figure 15(a) highlights such a case. In the graph, we replace the cease of Germany at the earlier node with a continuity to the later node.

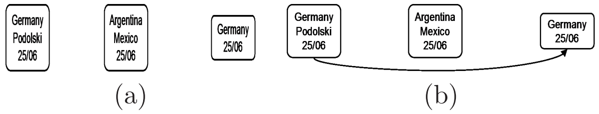


Figure 15: Example highlighting rebirth of an actor.

- $[Continue_A^i, Continue_A^{i+1}]$ - This represents the most ubiquitous case, wherein an actor continues. In such case the continuation chains are simply replaced by a single node. Figure 16 shows original and summarized graphs corresponding to stories on Zidane and Materazzi after the final. Continuous stories on Zidane and Materazzi correspond to the verbal duel between Zidane and Materazzi after the final, which continued for a few days. Such stories are summarized to form an *aggregate node*.

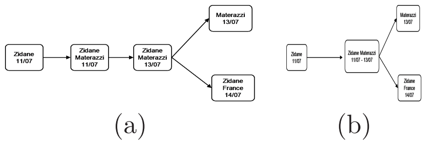


Figure 16: Example of continuation and associated summarization.

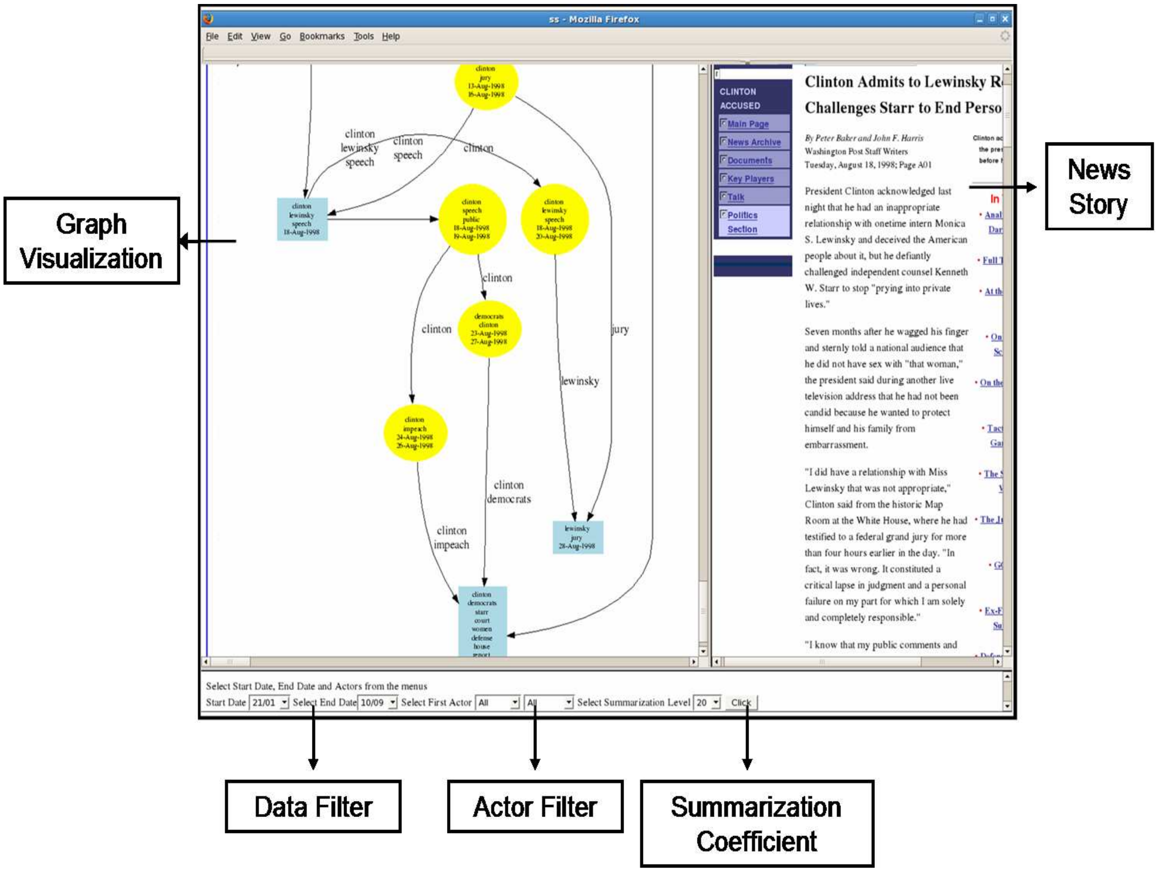
Please note that first case increases the number of nodes which is contrary to the idea of summarization. However, both the resulting continuation chains will be summarized by single node. We would like to point that the sequences are handled in the order they are enumerated above. The primary reason for such an ordering is that the first four types of sequences once summarized can produce a *Continue- Continue* sequence however, summarizing *Continue- Continue* sequences cannot result in other four sequences. Therefore, with such an ordering these new sequences will be handled naturally when we process the sequences in last category. The last category sequences are processed recursively to discover and summarize continuation chains of length ≥ 3. Other sequences are processed linearly.

For all the summarized nodes appropriate edges are added as governed by the graph generation methodology. Finally, the actor list of a summarized nodes is taken as the union of actors of corresponding single nodes.

7 Visualization

In previous sections we described our algorithms for news graph generation and summarization. Even though the proposed algorithms generate high quality graphs, the size of the graphs impedes effective and efficient dissemination of news articles. Apart from the size, other limiting factor are preference, interest and prior knowledge of the individual reading the news. These factors differ considerably from one individual to another which make automatic selection of ‘”important” news implausible (if not impossible). To handle these challenges we developed a visual system wherein the user can interact with the generated news graph, filters the stories and browse the news in a principled fashion. Essentially, the visual system enables the user to perform a focused search and also provides the capabilities for exploring the corpus. We used publicly available Webdot interface of Graphviz[2] for developing this toolkit.

[2] http:www.graphviz.org


ss - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Graph Visualization
News Story
Data Filter
Actor Filter
Summarization Coefficient
clinton jury 13-Aug-1998 15-Aug-1998
clinton lewinsky speech
clinton speech
clinton
clinton lewinsky speech 18-Aug-1998
clinton speech public 18-Aug-1998 19-Aug-1998
clinton lewinsky speech 18-Aug-1998 20-Aug-1998
clinton
democrats clinton 23-Aug-1998 27-Aug-1998
clinton
jury
lewinsky
clinton impeach 24-Aug-1998 25-Aug-1998
clinton democrats
clinton impeach
lewinsky jury 28-Aug-1998
clinton democrats starr court women defense house report
CLINTON ACCUSED
Main Page
News Archive
Documents
Key Players
Talk
Politics Section
Click

Clinton Admits to Lewinsky R

Challenges Starr to End Perso

By Peter Baker and John F. Harris
Washington Post Staff Writers
Tuesday, August 18, 1998; Page A01

President Clinton acknowledged last night that he had an inappropriate relationship with onetime intern Monica S. Lewinsky and deceived the American people about it, but he defiantly challenged independent counsel Kenneth W. Starr to stop "prying into private lives."

Seven months after he wagged his finger and sternly told a national audience that he did not have sex with "that woman," the president said during another live television address that he had not been candid because he wanted to protect himself and his family from embarrassment.

"I did have a relationship with Miss Lewinsky that was not appropriate," Clinton said from the historic Map Room at the White House, where he had testified to a federal grand jury for more than four hours earlier in the day. "In fact, it was wrong. It constituted a critical lapse in judgment and a personal failure on my part for which I am solely and completely responsible."

"I know that my public comments and

Select Start Date, End Date and Actors from the menus

Start Date 21/01 Select End Date 10/09 Select First Actor All All Select Summarization Level 20

Figure 17: A snap shot of our visual toolkit. The presented graph is generated for Clinton-Lewinsky dataset.

In this section, we describe the key components of the visual system. Figure 17 shows a snapshot of our interface. The graph is generated from Lewinsky-Clinton corpus. The key components of the interface are :

- **Graph Visualization Window -** This window displays the final generated and refined news graph. Different type of nodes are differentiated by using different shapes. The square shaped nodes denote a single news story whereas a ellipse shaped node represents an aggregate/summarized node. Each simple node displays the date when the story was published and also the actors. Similarly, an aggregate node displays the time interval and actors.

- **News Story Window -** This window displays the text of new stories. The user can click on a simple node and the corresponding story will be displayed. The actors in the selected news story are marked in this shown text. However, when an aggregate node is clicked, the headlines of all stories inside the aggregate node are shown. The user can click and see the actual stories.

- **User Interactions -** The user interface is designed such that it adheres to *zoom, filter* and *details–on–demand* paradigm as proposed by Schinderman [5]. The user can use mouse wheel or up down key to increase/decrease the size of the displayed graph. This feature is extremely useful when the size of graph is very large and user wants to focus on one specific time interval or a key event in the corpus. We provide the capability of filtering the data based on time interval and actors. The user can specify the start and end date to display the graph formed by new stories in the specified interval. Similarly, user can select actor(s) and the graph pertinent to selected actors will be displayed. Both these filters can be used simultaneously to further drill down in that dataset. Finally, we provide details on demand. The user can click on any simple node and the corresponding story is displayed in this window. The news story also has the corresponding actors marked and hyper linked. Clicking the the hyper links will display the graph corresponding to that actor in graph visualization window. However, clicking on the aggregate node will display the headlines in news story window which can then be clicked

Dataset	Source	Number of Stories	Start Date	End Data
FIFA World Cup 2006	www.rediff.com	459	June/1/2006	July/15/2006
US Elections 2004	www.nytimes.com	389	February/2/2004	November/25/2004
Clinton Lewinsky Saga	www.washingtonpost.com	914	January/21/1998	September/10/1998
Tsunami Dataset	www.rediff.com	270	December/26/2004	March/25/2005

Table 2: Dataset Description

Node#	Synopsis
1	Germany,Italy set for showdown
2	History on France's side
3	Portugal and France seek place in final
4	Scolari says Portugal can turn the page
5	Italy defeats Germany in Semifinal
6	Fabio Cannavaro's role in the win
7	Germans stunned by dramatic world cup exit
8	France defeat Portugal in semi-final
9	Zidane aims to say adieu with World Cup win
10	Beckenbauer praises Klinsmann
11	History suggests tight final in prospect
12	Christiano Ronaldo on Player of the Year list
13	Italy in spectacular form
14	Final will be a tactical battle
15	Germany defeat Portugal for third place
16	Italy defeat France to win World Cup
17	Zidane and Materazzi : Head Butt
18	Italy coach Lippi resigns
19	Klinsmann quits as Germany coach
20	Zidane and Materazzi : War of Words continues

Table 3: Synopsis of the stories shown in Figure 3

to browse actual news story. Moreover, the user can examine the graph in a multi-resolution fashion by choosing the summarization coefficient (SC) from a drop down list. This feature uses the sorted (in ascending order) ranking of transformations. From this sorted list, top SC% transformations are selected and corresponding edges and nodes are removed from the graph. We show the usefulness of this feature in Section 8 for goal oriented search.

8 Experimental Results

In this section we demonstrate the use of our toolkit on different news corpora. Table 2 provides the description of the datasets used for evaluation. We contend that the representation of news corpus using a directed graph is an effective way of answering common user queries about the set of documents. We have designed and conducted two user studies to quantitatively measure the quality of the graph as well to show the effectiveness of the visual toolkit towards the overall goal of information extraction. We also present the timing results of the graph generation and summarization algorithms. Table 2 provides the description of the datasets used for evaluation.

8.1 FIFA World Cup 2006

Figure 3 contains a small part of the interaction graph for the news corpus corresponding to the FIFA World Cup 2006. The part shown is composed of stories published between 03/07/06 and 12/07/06. The nodes are numbered taking into account the temporal ordering between the stories. Table 3 enumerates the headline for each node in the FIFA graph. The main actors of the topic are the teams and some of the players that were well covered by the news source used. Characterizing stories with actors makes the graph easy to understand since each relationship is marked with the corresponding actors. Also, in this special case, the graph looked very similar to a knockout graph during the later stages of the tournament.

Relationship of Actors: The relationship of two teams : *France* and *Italy* is shown in Figure 18. Till the finals, these two teams were orthogonal (Phase 1). During the days leading to the final and after it, stories were published about their clash in the final. During this phase(Phase 2), they were aligned with each other. After the final (Phase 3), they again became orthogonal to each other.

Ranked Transformation: 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 4. 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). Top two creations discovered are: *Zidane's Head Butt* and *Polish and German*

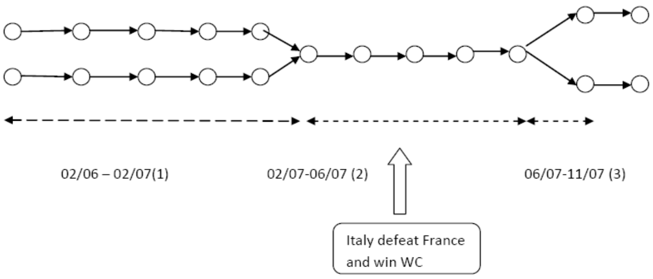


Figure 18: Relationships between Italy and France at different time periods.

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

Table 4: Top ranked Merges in FIFA 2006 corpus

Fan Clash. In this corpus, after every merge at round-robin stage a split occurs and every knock out match is followed by cessation of the losing team. Therefore, we do not explicitly enumerate these events here. 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.

Strength over time: Figure 19 shows the strength of different teams over time. Note the high peak present for France and Italy at end of the plot. This corresponds to the final being played between these two countries. The strength of other teams decrease over time and eventually comes very close to zero. The decrease (eventually coming down to zero) in strength marks the team being knocked out of the tournament.

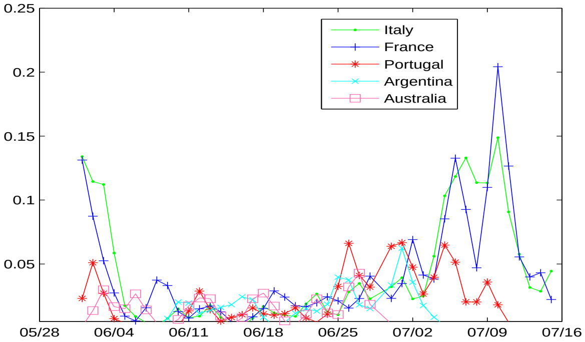


Figure 19: Strength over time for various teams in FIFA.

8.2 US Election 2004 News Corpus

We present a small part of the graph in Figure 20. This part is composed of stories published between 30/07/04 and 28/08/04. The key actors of the topic are Bush and Kerry, the main candidates in the presidential race. As evident from the figure Cheney is focused on Iraq war whereas

Node#	Synopsis
1	Cheney on Iraq
2	Kerry delivers speech at democratic convention
3	Bush campaigning in a Republican bastion
4	Kerry and Edwards reach Albuquerque
5	Bush campaigning in the southwest. Focus on Iraq and jobs
6	Cheney talking about Iraq while Kerry reacts in Las Vegas
7	Kerry attacks Bush about the ads attacking him about Vietnam
8	Bush reacts to Kerry criticism while admitting mistakes about Iraq
9	Cheney rakes up Iraq again

Table 5: Synopsis of the stories shown in Figure 20

Date	Story
25/02	Kerry starts winning a sequence of primaries (Kerry, primaries)
01/03	Discussion on Same-Sex marriage in the Democrat primaries (Same-sex marriage and Primaries)
20/08	Kerrys response to Bush on Vietnam Issues (Kerry and Bush)
15/10	Abortion comes up as an issue in the Bush-Kerry debates (Abortion and Debates)
12/09	GOP draws criticism from Kerry on Arms Ban (Kerry and Bush)
08/03	Battle for Florida hots up (Kerry and Bush)
12/08	Bush mocks Kerry on Vietnam (Kerry and Bush)
22/05	Bush visits Louisiana , where Democrats are campaigning (Bush and Democrats)
29/05	Kerry doubles his attack on Bush for Iraq War (Kerry and Bush)
30/10	Ruling by two federal courts to GOP (courts and GOP)

Table 6: Top ranked merges in US Election 2004 Corpus

Bush and *Kerry* are arguing over multiple issues. For example, node 7 and 8 represent stories related to campaign about *Kerry* and his views on *Vietnam*. In table 5, we present the synopsis of the stories.

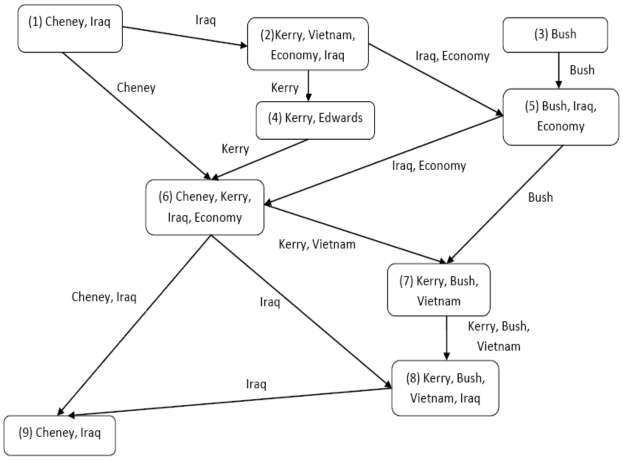


Figure 20: A small part of the interaction graph generated from US Elections 2004 News Corpus

Relationship of Actors: Figure 21, shows different relationships between Kerry and Bush during the elections. In the US elections 2004, *Kerry* and *Bush* were orthogonal during the early months when *democrat* primaries were going on. Once *Kerry* won the primaries, *Kerry* and *Bush* started interacting. This was the chain part of their interaction. In the days leading to the election, *Kerry* and *Bush* were involved in successive debate sessions. In that phase of the elections, both were completely aligned. Later on, after *Bush* won the *election*, *Kerry* became a sub-actor of *Bush* as most of the stories involving *Kerry* also talked about the *election* victory of *Bush*.

Ranked Transformations: Table 6 shows the abstract of the top 10 merges (scores in decreasing order) and corresponding dates identified in this corpus. The actors involved in merge are noted at the end of each headline. We notice that most of the major merges involve *Kerry* and *Bush*. This is because their strength is very high throughout the topic and thus whenever there is less interaction between them, then a new merger is ranked very high.

Table 7 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. There are some important creations that were not ranked in the top 8 creations. This was because, after being born, the actors were not immediately significant. But their significance rose over the next

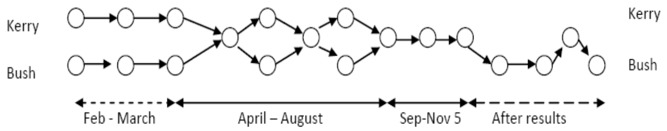


Figure 21: Relationships between Bush and Kerry at different time periods.

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 Kerrys Vietnam records (Vietnam)
13/05	Ralph Nader wins endorsement of Reforms Party (Nader)

Table 7: Synopsis of the top ranked creations in US Election 2004 corpus

few days. When the future window was changed to 20 days, more interesting creations were found. An example is the Kerry- Bush debate . The debates started in September end and continued till mid October. While the creation of Debate does not occur in the top 8 results with future window as 8 days, it appears in the top 5 results with F=20. Finally, table 8 shows top cessation of actors found by using the retrospective window (P) size as 20 days.

Strength Over Time: Figure 22 shows how strength of various actors vary over time. *Bush* and *Kerry* are clearly the strongest actors in this corpus. An extremely high peak was observed for a small time period for *Dean*, but it ceased soon. *Abortion* shows intermittent peaks, corresponding to debates/speeches delivered by candidates. *Vietnam* was a strong actor for a small time period (with Kerry replying to Vietnam related remarks).

8.3 User Study

We designed a user study to measure precision and recall of our algorithms. We sought help of 6 individuals. Each individual evaluated graphs generated from every dataset. However, owing to large number of stories in each corpus, it was impossible to get the users to evaluate the complete news graphs. Therefore, we adopted the following methodology. For each user, we randomly selected one node from the graph and noted the corresponding story's date of publication. Next, we selected 10 stories published after that date. The subgraph formed by these 11 nodes is generated and presented to the user along with the text of actual news stories. Users were asked to read the news stories. For each edge in the automatically generated graph, user were asked to mark the edge as *Correct* or *Incorrect*. The *incorrect* edges capture False Positives (FP), i.e., the edges which were formed by our algorithm but not deemed important by the user. Finally, we asked the user to add edges which they think are important but missed by the algorithm. We treat this set of these edges as False Negatives (FN). This process was repeated for each user and for each dataset. A node is selected randomly to start the process so that each user can evaluate different parts of the news graph. Table 13 shows the compilation of our results. The edges column represents the number of edges that existed between the 11 selected stories. Finally, table 14 shows the average precision and recall values for each of the datasets. We are able to achieve high values for both the measures.

Date	Story
03/03	Edwards bows out of Presidential Race (Edwards)
04/03	End of primaries (primaries)
15/10	End of third debate session (debates)
23/02	Howard Dean bows out of Presidential Race (Dean)
05/09	End of republican convention (convention)
17/06	Ralph Nader excluded from Presidential Debates (Nader)

Table 8: Top ranked cessation in US Election 2004 corpus.

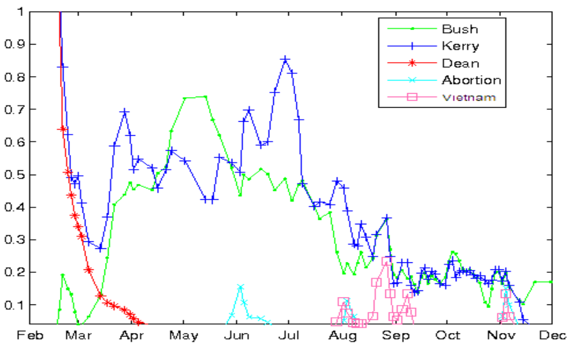


Figure 22: Strength over time for various actors in US Election 2006

	FIFA 2006			US Elections 2004			Clinton-Lewinsky			Tsunami		
User ID	Edges	FP	FN	Edges	FP	FN	Edges	FP	FN	Edges	FP	FN
User 1	14	4	5	16	3	2	11	2	0	15	5	1
User 2	15	5	4	14	3	0	12	0	2	15	5	5
User 3	15	5	4	14	3	1	20	6	3	7	5	1
User 4	12	2	5	18	7	2	12	3	1	13	4	1
User 5	10	2	4	20	5	1	15	4	1	10	6	2
User 6	3	FP	4	17	5	3	17	6	3	8	1	6

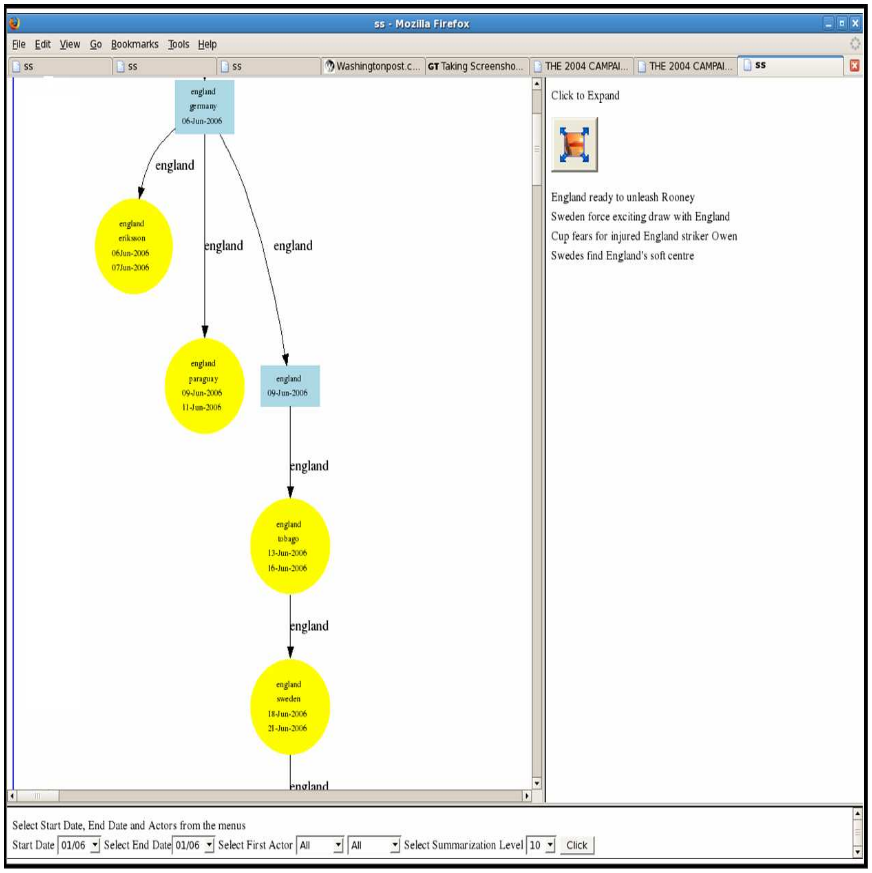
Table 9: Result of the user study

Dataset	Precision	Recall
FIFA World Cup 2006	.7406	.6887
US Elections 2004	.7373	.8902
Clinton Lewinsky Saga	.7586	.8641
Tsunami Dataset	.6666	.7647

Table 10: Average precision and recall values.

Dataset	Time(in minutes)
FIFA World Cup 2006	5
US Elections 2004	5
Clinton Lewinsky Saga	9
Tsunami Dataset	4

Table 11: Timing results for graph generation and summarization algorithms.


ss - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
ss
ss
ss
Washingtonpost.c...
Taking Screensho...
THE 2004 CAMPAI...
THE 2004 CAMPAI...
ss
england germany 06-Jun-2006
england
england eriksson 06Jun-2006 07Jun-2006
england
england
england paraguay 09-Jun-2006 11-Jun-2006
england 09-Jun-2006
england
england tobago 13-Jun-2006 16-Jun-2006
england
england sweden 18-Jun-2006 21-Jun-2006
england
Click to Expand

England ready to unleash Rooney
Sweden force exciting draw with England
Cup fears for injured England striker Owen
Swedes find England's soft centre

Select Start Date, End Date and Actors from the menus
Start Date 01/06 Select End Date 01/06 Select First Actor All All Select Summarization Level 10 Click

Figure 23: Snap shot for world cup

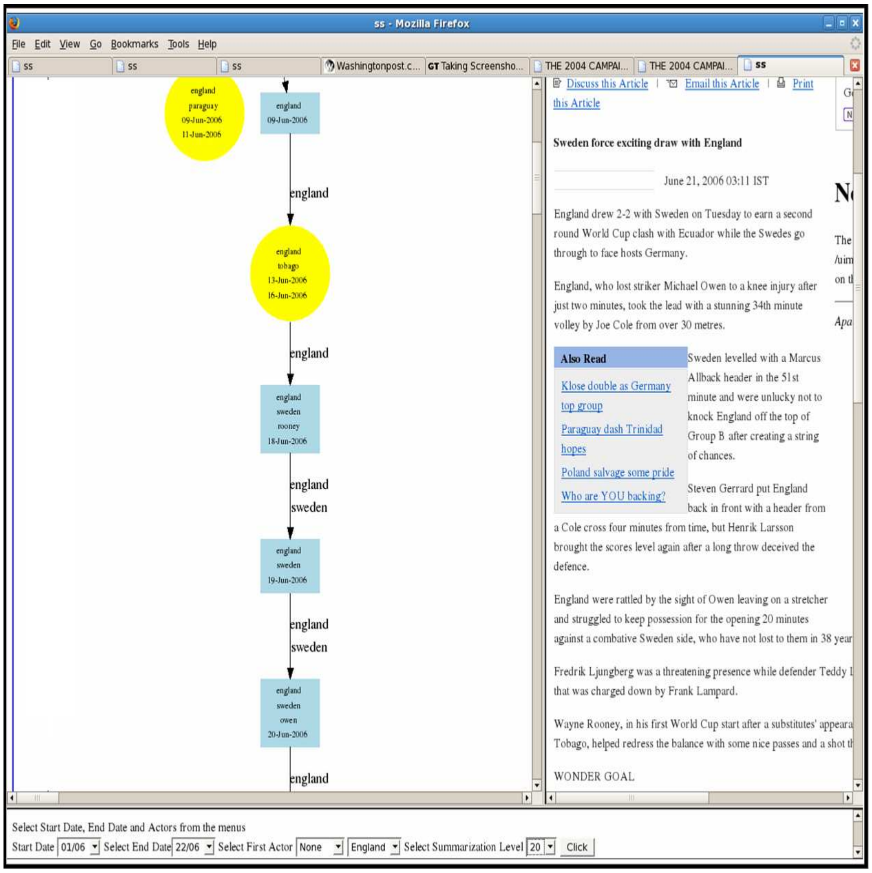
8.4 Timing

The proposed algorithms are extremely efficient. For the Clinton–Lewinsky dataset with 914 stories, the algorithm took only 9 minutes to generate and summarize the graphs. Table 15 shows results for all datasets. The experiments were performed on a system with 1.6 GhZ processor and 256 MB of main memory

- **Actor Based Query:** *Find all teams which England played against in FIFA World Cup 2006?*

  Process - The analysis starts by filtering the original graph based on actor England. The snapshot of the graph obtained is displayed in Figure 23. Every node with two teams as actors (depicted on the node) corresponds to one match. We can visually ascertain that news stories about England and Paraguay appeared between June/9/2006 and June/11/2006. The teams actually played on June/10/2006. Clicking on a aggregate node provides the details of the news stories about that match. For example, Figure 24 shows the resulting graph when the node England–Sweden is expanded. The node England-Germany dated June/6/2006 corresponds to news about the arrival of English team in Germany. We find that the nodes in the summarized graph correspond to the matches england played. There are nodes corresponding to other important transformations that England took part in.

- **Time Based Query Q2:** *Find all the major events in Clinton-Lewinsky Saga from January*



Sweden force exciting draw with England

June 21, 2006 03:11 IST

England drew 2-2 with Sweden on Tuesday to earn a second round World Cup clash with Ecuador while the Swedes go through to face hosts Germany.

England, who lost striker Michael Owen to a knee injury after just two minutes, took the lead with a stunning 34th minute volley by Joe Cole from over 30 metres.

Also Read

Klose double as Germany top group

Paraguay dash Trinidad hopes

Poland salvage some pride

Who are YOU backing?

Sweden levelled with a Marcus Allback header in the 51st minute and were unlucky not to knock England off the top of Group B after creating a string of chances.

Steven Gerrard put England back in front with a header from a Cole cross four minutes from time, but Henrik Larsson brought the scores level again after a long throw deceived the defence.

England were rattled by the sight of Owen leaving on a stretcher and struggled to keep possession for the opening 20 minutes against a combative Sweden side, who have not lost to them in 38 year

Fredrik Ljungberg was a threatening presence while defender Teddy

that was charged down by Frank Lampard.

Wayne Rooney, in his first World Cup start after a substitutes' appeara

Tobago, helped redress the balance with some nice passes and a shot th

WONDER GOAL

Select Start Date, End Date and Actors from the menus

Start Date 01/06 Select End Date 22/06 Select First Actor None England Select Summarization Level 20 Click

Figure 24: Snap shot when one of aggregate node is expanded

to September

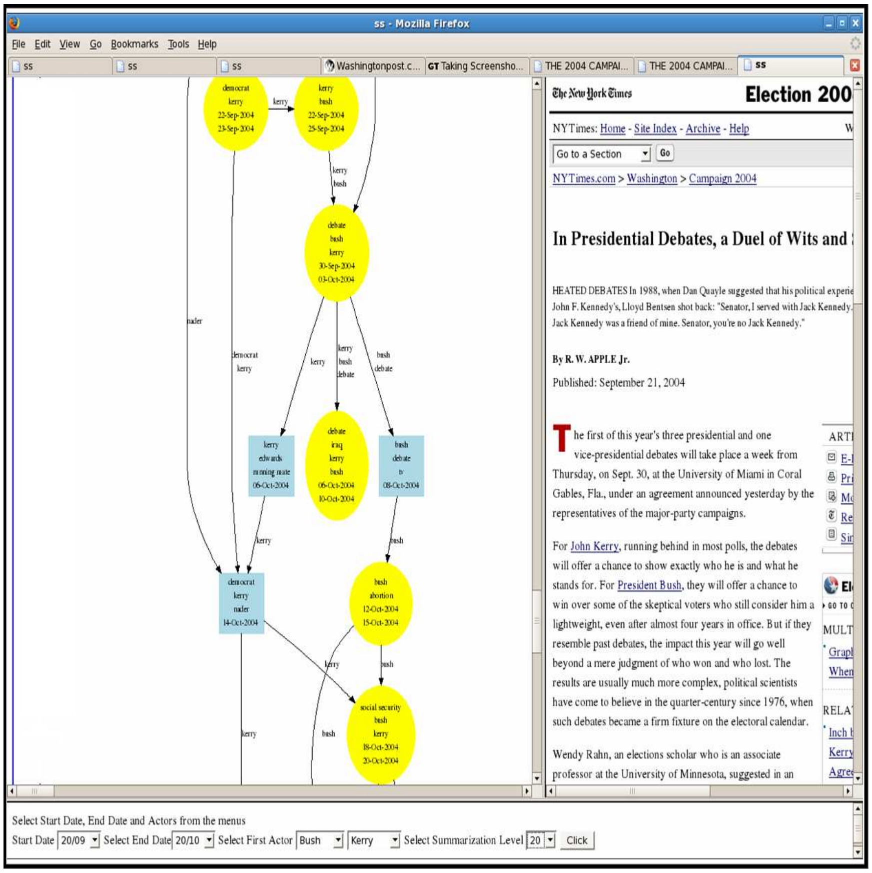
Process - Figure 17 captures the snapshot of the summarized graph. To start with, we keep the summarization at 90%. We find that many of the major events like canceling Paula Jones Case (April/1998) to Clinton Acceptance (August/1998) have corresponding nodes in the graph. On decreasing the summarization level, smaller events like Quitting of Ginsburg also start appearing as nodes in the graph. Note that the advantage of this type of summarization over other corpus summarization methods is that it presents a timeline based view representing the evolution of the topic. Also the summarized graph has more information than a simple timeline since it captures dependency among the major events in the news topic.

- **Actor and Time based Query Q3:** *Find important interactions between Bush and Kerry at different periods during US elections 2004?*

  Process- This was achieved by selecting Bush and Kerry in the visualization interface and selecting different periods during the elections. We selected different periods and studied the interactions between the two. We found that Kerry and Bush did not interact much between February/2004 and April/2004 since Kerry was busy with his own primaries. There were, however, important merges involving Kerry and Bush on February/25/2004 and March/7/2004 when it became clear that Kerry would win Democrat primaries At this time, the major issue was same-sex marriage. During the period between April and August, Kerry and Bush debated about Iraq and Vietnam war. During September/2004 and October/2004, Kerry and Bush appeared together in most of the news stories with the major issues like economy, Iraq and same–sex marriage were discussed. Figure 25 shows the associated snap shot of the toolkit.

- **Story based Query Q4:** *Find other stories related to the story about damage and relief work in Tsunami dataset?*

  Process- The story in focus consisted of information on damage statistics and relief work at Andaman and Nicobar Islands near Indian mainland, published on December/30/2004. We worked with the summarized graph to get dependencies. To find the immediate dependencies



In Presidential Debates, a Duel of Wits and

HEATED DEBATES In 1988, when Dan Quayle suggested that his political experie
John F. Kennedy's, Lloyd Bentsen shot back: "Senator, I served with Jack Kennedy.
Jack Kennedy was a friend of mine. Senator, you're no Jack Kennedy."

By R. W. APPLE Jr.

Published: September 21, 2004

The first of this year's three presidential and one vice-presidential debates will take place a week from Thursday, on Sept. 30, at the University of Miami in Coral Gables, Fla., under an agreement announced yesterday by the representatives of the major-party campaigns.

For John Kerry, running behind in most polls, the debates will offer a chance to show exactly who he is and what he stands for. For President Bush, they will offer a chance to win over some of the skeptical voters who still consider him a lightweight, even after almost four years in office. But if they resemble past debates, the impact this year will go well beyond a mere judgment of who won and who lost. The results are usually much more complex, political scientists have come to believe in the quarter-century since 1976, when such debates became a firm fixture on the electoral calendar.

Wendy Rahn, an elections scholar who is an associate professor at the University of Minnesota, suggested in an

Select Start Date, End Date and Actors from the menus
Start Date 20/09 Select End Date 20/10 Select First Actor Bush Kerry Select Summarization Level 20 Click

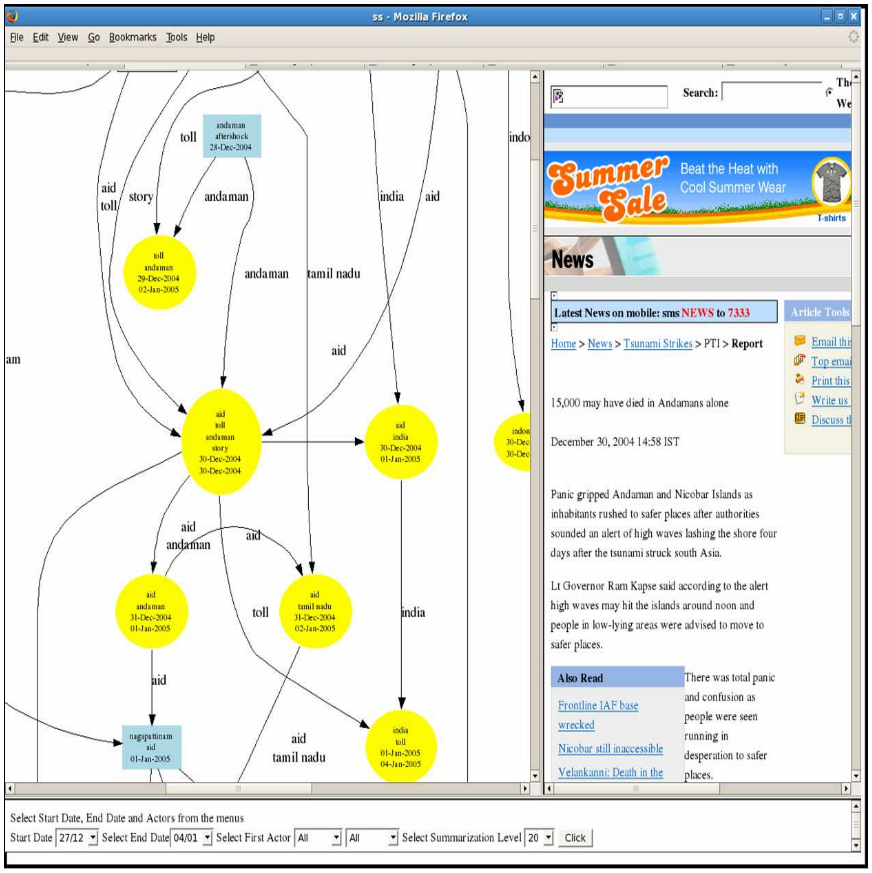
Figure 25: Snap shot for US election dataset.

we selected the subgraph between December/28/2004 and January/1/2005. The immediate neighbors of the story are marked by examining the nodes directly connected to the node in question. Major stories following this story (temporally) included stories on general toll figures in India and updated figures on Andaman Islands. The stories preceding it included stories on the damage statistics and relief work at Nagapattinam in Tamil Nadu, India and stories about on aftershocks in Andaman and Nicobar islands. By increasing the time window to December/27/2004 to January/4/2005 and considering paths from this story instead of edges, we extracted stories on toll and relief statistics. Figure 26) shows the snapshot for this period.

- **Past Reference based Query Q5:** *Find the stories needed to understand Zidane winning the golden ball?*

  Process: News stories often contain references to earlier related stories. A user, while browsing such a story, might like to read the previous related stories. For example, in figure 27, the user is reading the story on Zidane winning the golden ball. The story contains a reference to the match between France and Spain, where Zidane scored some goals (highlighted by the cursor) which resulted in his golden goal. In order to reach that particular story, a user just needs to filter the graph using actors France and Spain. This yields a new graph in the graph pane, with nodes just corresponding to stories covering both France and Spain. On clicking one of the nodes, a story related to their match opens in the story pane, as can be seen in figure 28. Thus the filters provided by the toolkit, are very useful in providing a smooth browsing interface for the user, where the amount of effort put in founding news stories of interest is very less.

The key difference between Q4 and Q5 is that Q4 only looks in immediate temporal neighbors, both past and future, of a story. The related stories in close temporal neighborhood are typically linked by edges. On the other hand Q5 can be used to find a story in distant past where no path exist between the stories. But they are still related. In some senses Q5 helps to find “ pre-requiste” stories.



Latest News on mobile: sms NEWS to 7333

Home > News > Tsunami Strikes > PTI > Report

15,000 may have died in Andamans alone

December 30, 2004 14:58 IST

Panic gripped Andaman and Nicobar Islands as inhabitants rushed to safer places after authorities sounded an alert of high waves lashing the shore four days after the tsunami struck south Asia.

Lt Governor Ram Kapse said according to the alert high waves may hit the islands around noon and people in low-lying areas were advised to move to safer places.

Also Read

Frontline IAF base wrecked

Nicobar still inaccessible

Velankanni: Death in the

There was total panic and confusion as people were seen running in desperation to safer places.

Select Start Date, End Date and Actors from the menus

Start Date 27/12 Select End Date 04/01 Select First Actor All All Select Summarization Level 20

Figure 26: Snap shot for Tsunami dataset.

Dataset	Time(in minutes)
FIFA World Cup 2006	5
US Elections 2004	5
Clinton Lewinsky Saga	9
Tsunami Dataset	4

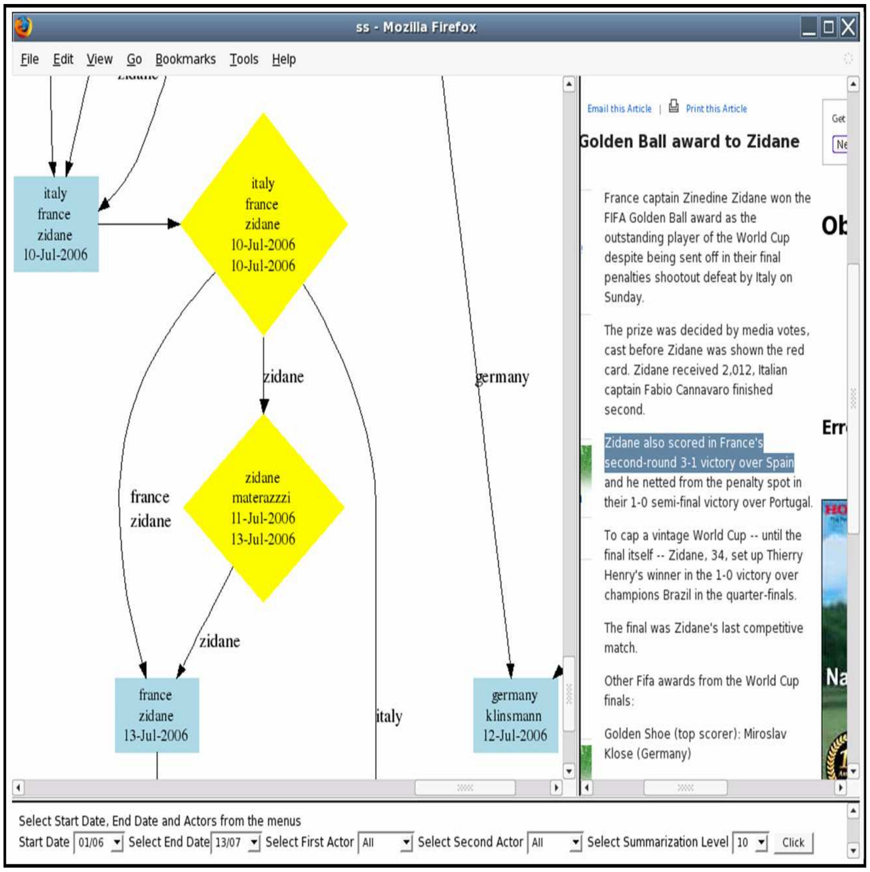
Table 12: Timing results for graph generation and summarization algorithms.

8.5 Timing

The proposed algorithms are extremely efficient. For the Clinton–Lewinsky dataset with 914 stories, the algorithm took only 9 minutes to generate and summarize the graphs. Table 15 shows results for all datasets. The experiments were performed on a system with 1.6 GhZ processor and 256 MB of main memory

8.6 User Study

We designed a user study to measure precision and recall of our algorithms. We sought help of 6 individuals. Each individual evaluated graphs generated from every dataset. However, owing to large number of stories in each corpus, it was impossible to get the users to evaluate the complete news graphs. Therefore, we adopted the following methodology. For each user, we randomly selected one node from the graph and noted the corresponding story's date of publication. Next, we selected 10 stories published after that date. The subgraph formed by these 11 nodes is generated and presented to the user along with the text of actual news stories. Users were asked to read the news stories. For each edge in the automatically generated graph, user were asked to mark the edge as *Correct* or *Incorrect*. The *incorrect* edges capture False Positives (FP), i.e., the edges which were formed by our algorithm but not deemed important by the user. Finally, we asked the user to add edges which they think are important but missed by the algorithm. We treat this set of these edges



Golden Ball award to Zidane

France captain Zinedine Zidane won the FIFA Golden Ball award as the outstanding player of the World Cup despite being sent off in their final penalties shootout defeat by Italy on Sunday.

The prize was decided by media votes, cast before Zidane was shown the red card. Zidane received 2,012, Italian captain Fabio Cannavaro finished second.

Zidane also scored in France's second-round 3-1 victory over Spain and he netted from the penalty spot in their 1-0 semi-final victory over Portugal.

To cap a vintage World Cup -- until the final itself -- Zidane, 34, set up Thierry Henry's winner in the 1-0 victory over champions Brazil in the quarter-finals.

The final was Zidane's last competitive match.

Other Fifa awards from the World Cup finals:

Golden Shoe (top scorer): Miroslav Klose (Germany)

Select Start Date, End Date and Actors from the menus

Start Date 01/06 Select End Date 13/07 Select First Actor All Select Second Actor All Select Summarization Level 10 Click

Figure 27: Snap shot showing part of FIFA graph related to Zidane winning the golden ball award .

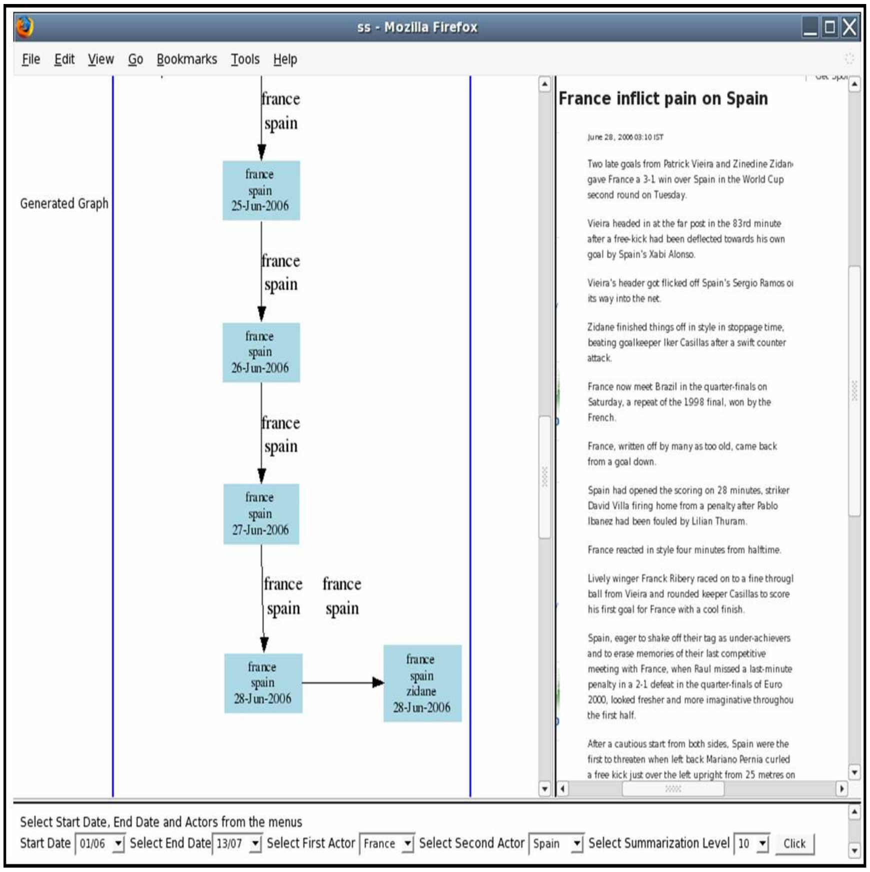
as False Negatives (FN). This process was repeated for each user and for each dataset. A node is selected randomly to start the process so that each user can evaluate different parts of the news graph. Table 13 shows the compilation of our results. The edges column represents the number of edges that existed between the 11 selected stories. Finally, table 14 shows the average precision and recall values for each of the datasets. We are able to achieve high values for both the measures.

8.7 Timing

The proposed algorithms are extremely efficient. For the Clinton–Lewinsky dataset with 914 stories, the algorithm took only 9 minutes to generate and summarize the graphs. Table 15 shows results for all datasets. The experiments were performed on a system with 1.6 GhZ processor and 256 MB of main memory

	FIFA 2006			US Elections 2004			Clinton-Lewinsky			Tsunami		
User ID	Edges	FP	FN	Edges	FP	FN	Edges	FP	FN	Edges	FP	FN
User 1	14	4	5	16	3	2	11	2	0	15	5	1
User 2	15	5	4	14	3	0	12	0	2	15	5	5
User 3	15	5	4	14	3	1	20	6	3	7	5	1
User 4	12	2	5	18	7	2	12	3	1	13	4	1
User 5	10	2	4	20	5	1	15	4	1	10	6	2
User 6	3	FP	4	17	5	3	17	6	3	8	1	6

Table 13: Result of the user study



France inflict pain on Spain

June 28, 2006 03:10 IST

Two late goals from Patrick Vieira and Zinedine Zidan gave France a 3-1 win over Spain in the World Cup second round on Tuesday.

Vieira headed in at the far post in the 83rd minute after a free-kick had been deflected towards his own goal by Spain's Xabi Alonso.

Vieira's header got flicked off Spain's Sergio Ramos o its way into the net.

Zidane finished things off in style in stoppage time, beating goalkeeper Iker Casillas after a swift counter attack.

France now meet Brazil in the quarter-finals on Saturday, a repeat of the 1998 final, won by the French.

France, written off by many as too old, came back from a goal down.

Spain had opened the scoring on 28 minutes, striker David Villa firing home from a penalty after Pablo Ibanez had been fouled by Lilian Thuram.

France reacted in style four minutes from halftime.

Lively winger Franck Ribery raced on to a fine througl ball from Vieira and rounded keeper Casillas to score his first goal for France with a cool finish.

Spain, eager to shake off their tag as under-achievers and to erase memories of their last competitive meeting with France, when Raul missed a last-minute penalty in a 2-1 defeat in the quarter-finals of Euro 2000, looked fresher and more imaginative throughou the first half.

After a cautious start from both sides, Spain were the first to threaten when left back Mariano Pernia curled a free kick just over the left upright from 25 metres on

Select Start Date, End Date and Actors from the menus

Start Date 01/06 Select End Date 13/07 Select First Actor France Select Second Actor Spain Select Summarization Level 10 Click

Figure 28: After using the filters the user can quickly find related stories which appeared 3 weeks ago.

Dataset	Precision	Recall
FIFA World Cup 2006	.7406	.6887
US Elections 2004	.7373	.8902
Clinton Lewinsky Saga	.7586	.8641
Tsunami Dataset	.6666	.7647

Table 14: Average precision and recall values.

9 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 excellent surveys in [1, 6, 12]

The need for having a temporal structure within a topic was identified by Nallapati et al. [17]. 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 also use properties of the graph to draw interesting inferences about the topic.

The problem of discovering evolutionary theme patterns from text was first identified by Mei and Zhai [15, 16]. 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 [15] can be used for detection of the major actors of a topic. Mei and Zhai [16] 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.

There have been other attempts towards providing a structure to a news topic. In TDT 2003, a hierarchical structure of a topic was proposed. However, a hierarchical structure of the topic does

Dataset	Time(in minutes)
FIFA World Cup 2006	5
US Elections 2004	5
Clinton Lewinsky Saga	9
Tsunami Dataset	4

Table 15: Timing results for graph generation and summarization algorithms.

not fully capture the temporal dependencies. Summarization of news topics [2, 13] also attempt to present a holistic view of the topic but the evolutionary behavior in not explicitly considered. We would like to note that our work is not totally orthogonal to summarization. The stories involved in top ranked transformations do provide a summary of the new corpus.

In their seminal work, Silver and Wang [20] enumerated the key transformations which a three dimensional scientific feature can undergo. Recently Spiliopoulou et al. [21] presented similar transformations to capture and monitor evolving clusters. Yang et al. [23] presented algorithms to discover different types of spatio-temporal patterns for scientific datasets (proteins) and pointed towards the possibility of knowledge discovery by capturing the evolution and interaction of patterns. All these algorithm defined a customized overlap (intersect) function to derive the relationships. Our algorithms use set intersection algorithm.

Recently Toyoda et al. [22] presented a method to understand evolution of topics using the evolving link structure of the web. Our work is significantly different from theirs, since we characterize the evolution using the major actors of the topic. .

Grobelnik and Mladenic [8] presented a system for news visualization. The authors extracted name entities from news stories. A graph is constructed with the entities as nodes and edges between entities if they co-occur in at least one of the document. The toolkit provides the functionality of searching by entity name and further exploration. However, the time dimension is completely ignored while graph generation and visual interface design.

Rennison [19] presented a visualization system which takes into account the relationships between news articles. The hierarchical structure is used for visualization. However, the relationships have to be provided by the user and the temporal relationship are not taken into account. If the user can specify the temporal relationship, the toolkit can use it. However, we believe this is a unrealistic assumption. The only temporal information an average user can provide will be time/date of story. Advanced relations involving multiple actors across time cannot be provided by the user.

Fortuna et al [7] presented algorithm for visualizing text corpus. The authors use latent semantic scaling and multidimensional scaling to find key concept in documents. Density/frequency of each concept is calculated. All the concepts are displayed on the screen with the texture color representing the density. The work, again, ignores the time dimension and from the visual interface the temporal information can be extracted in any fashion.

ThemeRiver [9] is a prototype driven developed by PNL and is very closely related to our work. The system relies on text mining methods to discover all the themes present in corpus. The themes and strengths are then plotted across time. This visual representation provides information about creation, cessation and continuation of themes very easily. However, it is not easy to visually ascertain merge and split events, which in case of a graph representation is trivial.

There has been related initiatives on arranging videos news content. However video analysis and organization is out of scope of this paper. Interested readers are pointed to [14, 11].

Research areas like temporal summarization [13, 18] of news topics also attempt to present a holistic view of the topic but stop short of defining a graph to represent it. Allan et al. [2] presented algorithms to create temporal summaries of the news topics. The authors define two metrics *useful* and *novel* to measure the importance of sentences in a news story. The focus of this work is different from our, we summarize whole news corpus while preserving important ones where Allan et al. [2] goal was to represent each story by the most useful and novel sentence.

10 Discussion and Future Work

In this paper we described visual exploration and summarization steps of our larger initiative of news organization and exploration. In future, we plan to extend the overall framework in multiple directions. At present the algorithms are geared towards analysis of news from a single source. Our main focus for the future is to extend it to handle multiple news sources. This extension requires algorithms to perform data integration. With such an integration, we would be able to extract even more important and interesting information. For example, some of the problems which can be handled are:

1. **Which news sources are typically the first one to break a story?** This can be done by checking the time delay in the creation event across different news sources.

2. **Which news source typically covers news exhaustively?** This can be done by comparing the list of actors across multiple sources.

3. **Does there exist a bias in news coverage and reporting?** This bias can be simply based on location, for example, US based news source may decide to focus on American football than on FIFA Cup, however same will not hold in news from UK. Other forms of bias can be even more interesting. For example do some sources selectively report/ignore stories based on their affiliation with some institution (can be political)?

Additionally the structural information gained from multiple sources can be used to refine our interaction graph and strengthen the validity of our inferences.

11 Conclusions

In this article we proposed a framework for exploration of new corpora. We presented definitions and very simple algorithms for discovering the key transformations: *merge, split, create, cease and continue* 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. Next, We proposed a novel algorithm to construct news graph from a news corpus. Our algorithms leverage the interactions among actors present in the news stories. The key premise behind our work is that while generating the news graph, important actors and important interactions should be preserved. To handle the large news graphs, we proposed a summarization method, which marks relatively less important or easy to understand parts of the graph and remove or aggregate them respectively. Finally, we proposed an interactive visual interface to present the graphs to the user. Our toolkit leverage the interactions among actors present in the news stories. We implemented simple filters so that user can prune uninteresting parts and perform focused search. Moreover, the user can also interactively select the summarization level and examine the same graph at different resolutions. We demonstrated the usefulness of our algorithms on several large new corpora. The toolkit was found to be very effective while performing goal oriented search. We also presented the timing results for graph generation and summarization algorithms.

References

[1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study: Final report. In *DARPA Broadcast News Transcription and Understanding Workshop*, 2006.

[2] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of news topics. In *SIGIR*, pages 10–18, 2001.

[3] James F. Allen. An interval-based representation of temporal knowledge. In *IJCAI*, pages 221–226, 1981.

[4] Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *KDD*, pages 913–921, 2007.

[5] Stuart Card, Jock Mackinlay, and Ben Shneiderman. *Information Visualization: Using Vision to Think.* Morgan Kauffman Publishers,, 1999.

[6] Jonathan G. Fiscus and George R. Doddington. Topic detection and tracking evaluation overview. pages 17–31, 2002.

[7] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. Visualization of text document corpus. *Informatica (Slovenia)*, 29(4):497–504, 2005.

[8] Marko Grobelnik and Dunja Mladenic. Visualization of news articles. In *SIKDD*, 2004.

[9] Susan Havre, Elizabeth G. Hetzler, Paul Whitney, and Lucy T. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Trans. Vis. Comput. Graph.*, 8(1):9–20, 2002.

[10] Dorit S. Hochbaum. *Approximation algorithms for NP-hard problems*, chapter 3. PWS Publishing company, 1997.

[11] Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. Topic threading for structuring a large-scale news video archive. In *CIVR*, pages 123–131, 2004.

[12] April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. A survey of emerging trend detection in textual data mining.

[13] Ravi Kumar, Uma Mahadevan, and D. Sivakumar. A graph-theoretic approach to extract storylines from search results. In *KDD*, pages 216–225, 2004.

[14] Hangzai Luo, Jianping Fan, Yuli Gao, William Ribarsky, and Shin'ichi Satoh. Large-scale news video retrieval via visualization. In *ACM Multimedia*, pages 783–784, 2006.

[15] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207, 2005.

[16] Qiaozhu Mei and ChengXiang Zhai. A mixture model for contextual text mining. In *KDD*, pages 649–655, 2006.

[17] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *CIKM*, pages 446–453, 2004.

[18] Norman Papernick and Alexander Hauptmann. Summarization of broadcast news video through link analysis of named entities. In *In Link Analysis Workshop, AAAI*, pages 53–61, 2005.

[19] Earl Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *ACM Symposium on User Interface Software and Technology*, pages 3–12, 1994.

[20] Deborah Silver and Xin Wang. Volume tracking. In *IEEE Visualization*, pages 157–164, 1996.

[21] Myra Spiliopoulou, Irene Ntoutsi, Yannis Theodoridis, and Rene Schult. Monic: modeling and monitoring cluster transitions. In *KDD*, pages 706–711, 2006.

[22] Masashi Toyoda and Masaru Kitsuregawa. Extracting evolution of web communities from a series of web archives. In *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 28–37. ACM Press, 2003.

[23] Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. Mining spatial object associations for scientific data. In *IJCAI*, pages 902–907, 2005.