Multi-Perspective Linking of News Articles within a Repository

Arpit Khurdiya, Lipika Dey, Nidhi Raj and Sk. Mirajul Haque
Tata Consultancy Services, Innovation Labs, Delhi, India
{arpit.khurdiya, lipika.dey, nidhi1.r, skm.haque}@tcs.com

Abstract

Given the number of online sources for news, the volumes of news generated are so daunting that gaining insight from these collections become impossible without some aid to link them. Semantic linking of news articles facilitates grouping of similar or relevant news stories together for ease of human consumption. For example, a political analyst may like to have a single view of all news articles that report visits of State heads of different countries to a single country to make an in-depth analytical report on the possible impacts of all associated events. It is likely that no news source links all the relevant news together. In this paper, we discuss a multi-resolution, multi-perspective news analysis system that can link news articles collected from diverse sources over a period of time. The distinctive feature of the proposed news linking system is its capability to simultaneously link news articles and stories at multiple levels of granularity. At the lowest level several articles reporting the same event are linked together. Higher level groupings are more contextual and semantic. We have deployed a range of algorithms that use statistical text processing and Natural Language Processing techniques. The system is incremental in nature and depicts how stories have evolved over time along with main actors and activities. It also illustrates how a single story diverges into multiple themes or multiple stories converge due to conceptual similarity. Accuracy of linking thematically and conceptually linked news articles are also presented.

1 Introduction

News analysts have to analyze large volumes of news received from different sources to predict the political, financial or social impacts of the events reported in these. Story-centric organization aids in effective assimilation of news content by providing consolidated view of a collection of contextually-related news articles. Some online news providers use Ontology based tagging systems to help in semi-automated linking of related news articles while new articles are uploaded. News aggregators like Google news provides a single view of an incident by grouping relevant news articles gathered from multiple sources. But the grouping is restricted to articles received in the recent past only. Analysts often need to analyze current news in the context of past events, which are no more available on the web. Online News providers also do little to characterize a story in terms of its key components. They do not show how it has evolved over time, its spatial span and most importantly the highlights of the different phases of evolution. Story representations can also be used to emphasize contextual or logical similarities of apparently disconnected information.

In this paper we present a method for identification, tracking and characterization of major stories within a news repository. Our work emphasizes both on story identification as well as story representation. A story is represented graphically as a connected set of time-stamped nodes that are linked using directed edges. Each node is associated to a set of news articles. A story is characterized by the key entities, their roles and associated actions for each phase in a story. These elements are automatically extracted from the articles and used to generate an element called *story highlight*. Action evolution and the changing roles of entities become evident through the highlights. A major distinguishing feature of the present work is the ability to identify diverging and converging points which indicate thematic twists and turns within stories and also identify conceptually related stories.

The proposed system uses Latent Dirichlet Allocation (LDA) based topic extraction as the basis for identifying story contents. Topic correlation for topics extracted over different time periods is used to track stories over time. Natural Language Processing (NLP) techniques like entity extraction and verb classification have been deployed along with Maximum Entropy based analysis to identify the key components for characterizing stories. We have evaluated the present system by comparing its performance with Google news presentation. Google news presents news stories in a clustered fashion, sorted by date and relevance. We show that the proposed system can not only group articles in a similar fashion, but it can additionally present added dimensions like differing contexts within the same story or linking similar stories arising out of different contexts.

The rest of the paper is organized as follows. In section 2 we have presented a brief overview of related work in tracking stories from news articles. It may be noted that most of

these have focused only on linking conceptually similar news stories and very little focus have been given towards representing a story at multiple levels of granularity. In section 3, we present complete details of the proposed system. Section 4 presents results and performance analysis of the proposed system.

2 Earlier Work

One of the prime challenges of news analytics today is the large volumes of content that is generated constantly all around the world. Information linkage has become absolutely important to deal with this vast deluge of information. Event detection from media sources like video or images has received a lot of attention in the past. Event extraction from news stories has also attracted attention from text mining researchers.

[Uramoto and Takeda, 1993] had proposed a graph-based method to link news articles based on their content. Linking was based on document similarity. [Ichiro et al., 2003] had proposed a method to link news videos based on topic segmentation. Semantic and chronological relations were used in conjunction with topics to track related co-occurring events. In [Zhai and Shah, 2005], the authors had proposed a concept-tracking method which links news videos with the same topic across multiple sources. Semantic linkage of news stories was based on a combined correlation of visual content and spoken words. Frame-based similarity measures among key-frames and normalized text similarity measures had been used to identify and rank similar stories. [Liu et al., 2007] presented methods for event detection and tracking changes from news stories that discuss events of interest to support decision making for business. Event-trends are detected using association rule-mining for pre-defined event feature sets. Event changes are identified through analysis of association rules induced over consecutive time-periods. In [Lin and Liang, 2008], a news-story summarization method based on topic detection and tracking was presented. This method called SToRe (Story-line based Topic Retrospection) used WEBSOM[1], an extension of Growing Hierarchical Self-Organizing Map (GHSOM[2]) to cluster similar news reports and then performed event tracking within these topics. Though this work was very similar to ours, however, there are some basic differences in our approach. The earlier work had assumed that (1) every story collected for reviewing a news topic belongs to a set of events ε, and (2) a story only describes a single event in ε, both of which are very restrictive, both of which need not be true in real-life. The proposed system has no such assumptions and works for more generalized situations.

3 Story Building and Characterization from News Articles Stored in a Large Repository

Conceptually, story tracking has similarities with event tracking within media collections. Our story representation framework is developed along similar lines as the 5W1H based event representation framework described in [Xee et al. 2007]. Events have been defined as real-world occurrences that unfold over space and time. Though there has been substantial work in event extraction and representation from multimedia documents, most of these have substantial reliance over human-provided tags attached to the content. The key challenge lies in automated identification of textual elements that can represent key elements of a story. The proposed story tracking framework is a fully automated one, though can be edited by human editors if desired.

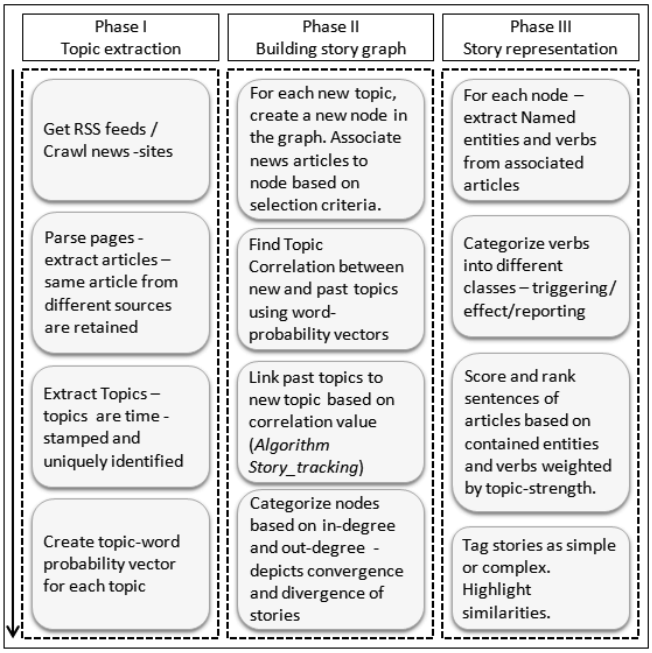

Figure 1: Overview of story building and characterization

In the proposed framework for story-centric news analysis, story detection is primarily based on topic discovery and correlation within news repositories. Topic detection is implemented using Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. Figure 1 explains the phases in story detection and tracking.

The news collection is represented as a graph, of nodes and links, termed as G (say). Story-identification, tracking and annotation are essentially focused towards building this graph G and are achieved through the following steps.

Step 1: The first step of story-tracking is to organize the news articles of a day around a set of key topics which can be derived from the entire collection of the day's news articles gathered from multiple sources, using LDA. The number of topics is provided as an input to the system and remains same for all days. A set of nodes are then created in the system corresponding to each new topic. Each node is thereafter associated to a set of news articles which are selected based on the strength of the corresponding topic in the article.

Step 2: The next step is to establish topic continuity. This is achieved by linking the newly created nodes in Step 1 to the existing story graph. Topic correlation lays the founda-

[1] http://websom.hut.fi/websom/
[2] http://www.ifs.tuwien.ac.at/~andi/ghsom/

tion for story tracking. Topics for two different days are compared using the Pearson's correlation coefficient, which has been found to provide the best similarity measure for topics [White and Jose, 2004]. A high correlation value between a pair of topics extracted for two different days is indicative of continuation of the same or related news. Section 3.2 explains how correlated topics are linked to form the story-graph.

Step 3: The final step involves adding flesh and blood to the story structure to enable easy comprehension. During this phase, the key entities and sentences are identified and extracted from the underlying news articles associated to the nodes of the graph. These elements thereafter not only depict how a story has evolved but also help in understanding the relationship among the stories. The details of how these stories are built and identified are further explained in the following sub-sections.

3.1 Topic Detection within News Repositories

LDA is a generative model in which words and documents are related through latent topics. A document collection is modeled as a distribution of topics while topics are modeled as distribution of words. Each document can be modeled as a probability distribution over a pre-defined number of topics. Since the words are observed, the document and the topic distributions, Θ and Φ, are conditionally independent. Let D represent a set of news articles collected from multiple sources over N days.

Let $d_{i,n}$ represents the i^{th} news article received on n^{th} day, where $1 <= n <= N$.

Let $T_N = \{\bar{t}_n\}$ represent the entire set of topics extracted over N days, where $\bar{t}_n = \{t_n^1, t_n^2, \ldots, t_n^p\}$ represents the topics extracted on n^{th} day. The number of topics to be extracted each day is set to p. Let r denote the number of news articles on a day. r varies for each day. The following steps depict how topics are extracted for each day.

1. Draw p multinomials Φ_p from a Dirichlet prior β, one for each topic p.
2. Draw r multinomials Θ_r from a Dirichlet prior α, one for each document d.
3. For all documents in the corpus, then for all words, w in the document:
   (a) Draw a topic z_t from multinomial Θ_r ; $p(z_t|\ \alpha)$
   (b) Draw a word w from multinomial Φ_z; $p(w|z_t, \beta)$

Using Gibb's sampling, probability of a topic z_t given a word w, $P(z_t|w)$ is approximated by means of the Monte Carlo algorithm which iterates over each word token in the text collection and estimates the probability of assigning the current word token w to each topic conditioned on the topic assignments to all other word tokens. Each topic z_t is modeled as probability distribution of words in it. Each document d_i is modeled as a topic vector of values $\{z^i_1, z^i_2, \ldots, z^i_p\}$, where z^i_k denotes the strength of topic z_k in d_i.

Once the topics are extracted, a set of nodes are created corresponding to the topics. Let $S_n = \{s_n^1, s_n^2, \ldots, s_n^p\}$ denote the set of p nodes created during this step. Node s_n^i is associated to topic $t_n^i \in \bar{t}_n$. A set of news articles is now selected to be associated to each node. Different selection criteria can be used for this step. We preset three different criteria that were experimented with. To attach a set of articles to topic node for z_k: (a) select all articles whose maximum topic strength is in z_k (b) select all articles whose maximum topic strength is in z_k and z_k is greater than a pre-defined threshold (c) select all articles whose topic strength z_k is greater than a pre-defined threshold. In section 4 we have presented how these criteria affect the performance of the system. The next section elaborates how these nodes are used to build the story graph.

3.2 Building the Story Graph

Let G denote the output graph. Initially, i.e. at day 1, G is empty. So this step simply sets $G = S_1$. Each node in S_1 denotes a story which is a collection of associated news articles.

For any other given day n, let $H = \{< s_{n,}^i t_n^i >,\ 1 \leq i \leq p\}$ denote the node-topic pairs created in the earlier step.

Let K denote the set of all nodes in G with out-degree 0. K contains all the nodes in G that depict terminating points for existing stories. Let $\bar{t}_K$ denote the set of topics associated to nodes in K. It may be noted that $\bar{t}_K$ may contain topics extracted from multiple days in the past. For each node $s_n^i \in$ H, nodes contextually similar to it are identified from K, based on topic correlation. For each element $\bar{t}_K^j \in$ K, Pearson correlation coefficient C_{ij} between t_n^i and t_k^j is computed as

$C_{ij} = \mathrm{Corr}(t_n^i, t_K^j)$, where $\mathrm{Corr}(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$, X and Y are the word-probability distributions associated to t_i^s and t_j^k respectively.

A directed link is created from each node of K corresponding to $\bar{t}_K^j$ and t_n^i, provided the correlation value is greater than a threshold.

As new set of nodes get added to G, G grows into a complex graph which is not necessarily connected. Each isolated sub-graph in G represents a story. This story may be either *simple* or *complex*. A simple story represents the evolution of a single event with a single start node, termed as *source*, and a single terminating node termed as *sink*. A complex story connects several simple stories due to their contextual or thematic overlap. Two stories are said to be *thematically similar* if they contain overlapping set of entities and concepts. For example, all stories on weather and climate of a place may be thematically related, though one set may be writing about its temperatures while another set may be discussing the effects of vagaries of weather on agriculture. A set of *conceptually* related stories on the other hand may not have common entities throughout but have higher-level of conceptual similarity. All news articles describing local political events may belong to this type. For example, news stories about electoral campaigns, announcement of new financial policies by the government and dinners hosted by the prime minister for a visiting State Head may be linked due the fact that they all report political activities. We now describe how simple and complex stories are characterized in the present framework.

Based on their in-degree and out-degree, nodes in G may be classified into:

Simple nodes - These are nodes whose maximum in-degree and out-degree is 1. These can be further classified as *source* nodes, *sink* (terminating) nodes, or *continuation* nodes. While source nodes have zero in-degree, and sink nodes have zero out-degree, simple continuation nodes have both in-degree and out-degree 1. A set of simple nodes which includes a single source and a single sink comprises a simple story.

Complex nodes - These are nodes whose in-degree, out-degree or both are greater than 1. Like simple nodes, these can also be further classified as source, sink or continuation nodes. Complex nodes act as converging or diverging points for stories. A node whose in-degree is greater than 1, acts as a *converging* point for multiple stories. A node whose out-degree is greater than 1, acts as *diverging* point for stories. A single node can exhibit both the properties. A set of simple stories, with little or no overlap in terms of contained entities, converging at a point are termed as *conceptually related* stories. A set of stories diverging from a single node indicate branching out of a single initial story into multiple *themes* or directions. The next section elaborates on how semantic content is associated to the graphical structure.

3.3 Characterizing Stories

Aim of news analysis is to identify different significant events and make a compact representation of their genesis or cause, and possible effects on other events. Since news is event-centric, it is important to identify the main actors and their roles in the events. This is done by analyzing the action verbs and identifying the key actions that trigger the event. The significance and roles of entities may change over time.

Our aim is to identify the significant named entities, causes and effects that are reported as part of the news along with their roles in key actions, also showing along the line how these roles might have changed. Generating interpretations, similar to those created manually by the analysts is out of the scope of the present work.

A news story is characterized by a set of named entities, action verbs and sentences containing these that are chosen by the system from the articles associated to the nodes. A combined visualization of these time-stamped elements illustrates the evolution of a story.

Named entities are extracted from the articles using Stanford Named Entity Recognition (NER)[3] tool. Based on their frequencies in the underlying set of articles, the importance of the entities is determined using Maximum Entropy based scoring method.

Activity recognition comprises of categorizing activity verbs, extracted from the articles using Stanford Parser[1], into three different categories. Verbs extracted from stories are classified into different classes using VerbNet[4], and then ranked using Maximum Entropy based scoring method. The first set of verbs, called the *triggering* actions can encompass a wide-variety of classes and are selected based on their frequency and significance in the story. The second category includes action verbs that encode *effects* and *reactions* caused by the triggering actions. The third category identifies reporting actions or statements made by people or organizations and are called *reporting* verbs.

VerbNet (VN) is a large on-line lexicon that organizes verbs into different classes and subclasses. The class-subclass definitions have been arrived at based on their thematic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function, as observed in large collections. Each VN class contains a set of syntactic descriptions, or syntactic frames, depicting the possible surface realizations of the argument structure for constructions such as transitive, intransitive, prepositional phrases, resultatives, and a large set of diathesis alternations. Semantic restrictions (such as animate, human, organization) are used to constrain the types of thematic roles allowed by the arguments, and further restrictions may be imposed to indicate the syntactic nature of the constituent likely to be associated with the thematic role.

The proposed news analytics system marks all verbs belonging to the classes "Other_cos" (class no. 45.4) and "calibratable_cos" (class no. 45.6-1) as *effect* verbs. Some commonly occurring *effect* verbs mined from news repositories are include drop, cause, rise, soar, affect, jump etc. These verbs are useful in identifying sentences that report impact of an event. All verbs belonging to classes "say" (class no. 37.7-1), "indicate" (class no. 78-1), "tell" (class 37.2), "talk"(class no. 37.5), "lecture" (class no. 37.11-1), and "advise" (class no. 37.9) have been considered as *reporting* verbs. All other verbs have been considered for their role as a *triggering* verb. Commonly occurring verbs in this category mined from news repositories are murder, kill, purchase, threaten, deploy, record, begin etc.

Scoring Entities and Actions using Maximum Entropy

News articles are in general quite descriptive and not all entities, or sentences in these articles are required to generate the compact comprehensive picture. The key challenge is to extract those elements from the news which can represent the content in a succinct way. Though this has similarities to document summarization, however it also differs in some major ways. Summaries do not necessarily highlight roles, causes, effects, reporting etc. that are crucial to news analysis. News articles associated to a single node contain topically similar stories collected on the same day. These articles report the same event though their contents may be somewhat different. We have opted for Maximum entropy based identification for significant entities and actions, since it is known to model the uncertainty of information in the best possible way, by modelling surprise over probability. In an evolving news story there are regular entities and actions that occur with uniform probability, and surprise entities co-occurring with the earlier ones at sudden time instants, adding twists and turns to the story. Proce-

[3] http://nlp.stanford.edu/software/lex-parser.shtml

[4] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

dure *Story_characterization* employs Maximum-Entropy based scoring to extract the significant entities, verbs and sentences from these articles. Initially the entropy-value of all entities and action verbs are computed and ranked according to their entropy value. For each verb category, the set of sentences containing them, are scored based on the entropy scores of the verbs and entities belonging to it. The score assigned to a sentence is further weighted by the topic strength of the article containing it.

Procedure: Story_characterization

1. Let S_i be a node in the graph G.
2. For each node S_i
   2.1. Let
   - NE_i denote the set of named entities associated to S_i
   - TV_i denote the set of triggering action verbs associated to S_i
   - EV_i denote the set of effect verbs associated to S_i
   - RV_i denote the set of reporting verbs associated to S_i.

   2.2. For all $e_k \in NE_i$
   2.2.1. Compute entity-frequency pair $<\{e_k, f_k\}>$
   2.3. For all elements v^j_l belonging to TV_i, EV_i and RV_i Compute verb-frequency pair $<v^j_l, f_l>$, where j=*1*,*2*, or *3* depending on whether v_l belongs to TV_i, EV_i or RV_i.
   2.4. For each entity-frequency pair and action-frequency pair
   2.4.1. Compute the probability mass function p(x).
   $p(x) = \frac{f_x}{\sum_{i=1}^{I} f_i}$ where x is either a named entity or a verb, f_x is the frequency of entity or verb x at S_i, and the denominator is the sum of frequencies of all entities or verbs of type similar to that of x.
   2.4.2. Compute entropy for each entity and action verb as $E(x)= -p(x) \ln(p(x))$.
   2.5. For each sentence m_j of S_i compute its relevance to the topic node as
   $RF(s_j) = \sum_{k=1}^{K} E(X_k) * \sum_{l=1}^{L} E(X_l) * z_i$, where K is the total number of entities and L is the total number of verbs (inclusive of all classes) present in m_j and z_i denotes the strength of topic *i* in document containing m_j.
   2.6. Sort the sentences in S_i in decreasing order of relevance.
   2.7. Select the maximum-scoring sentences that contain a verb $v_l \in EV_i$.
   2.8. Select the maximum-scoring sentences that contain a verb $v_l \in RV_i$ and has not been selected in step 2.7.
   2.9. Select the maximum-scoring sentence out of the remaining set that does not contain an effect or a reporting verb.

End_procedure

Once the relevant entities, action verbs and sentences at node level for each sub graph are extracted, the degree of mutual overlap determines the similarity of the linked stories.

4 Experiments and Results

We present some results from an implementation of the proposed news analytics system which has been deployed as a web-application on a Java platform. A number of experiments have been conducted with different sets of news sites with and without RSS feeds.

To provide an objective evaluation, we compared the system performance to Google News which clusters news articles from various sources and presents them as related stories. To do this, we collected news articles presented by Google News during the period 20th Dec, 2010 to 9th January, 2011. The entire period was divided into three batches of 8, 7 and 6 days respectively. For each batch, the links of the news articles were obtained from Google News, and thereafter the articles were downloaded from their original sources, parsed, and stored in local repositories along with their date, time and source. No information about their original grouping was used by our system.

Story detection and characterization was performed on these three repositories independently and the story groups identified by the proposed system were compared with the Google grouping of those stories. Each Google grouping of a set of articles is termed as a *Google-story*. Each connected component of the story graph returned by the proposed system is termed as an *inferred-story*. Table 1 provides a performance analysis of the proposed system in terms of its recall. Recall measures the capability of the system to retain related groups of articles together. Recall of a set of inferred-stories I with respect to a set of Google stories O is computed as:

$$R = \frac{|I \cap O|}{|O|}$$

The set I varies depending on the selection criteria chosen for assigning news articles to a node in the graph. Recall values for different criteria are shown in Table 1.

Recall Cutoff	Selection Crieteria on topic strength for articles to be associated to a node in G						
Th_k	**max**	**max>=0.3**	max>=**0.4**	max>=**0.5**	C_T>=**0.2**	C_T>=**0.3**	C_T>=**0.4**
Repository 1 (8 days, 1546 articles, 175 Google stories, 60 inferred stories)							
0.4	0.84	0.80	0.71	0.58	**0.96**	0.84	0.71
0.5	0.74	0.72	0.63	0.50	**0.85**	0.76	0.64
0.6	0.71	0.68	0.58	0.46	**0.82**	0.72	0.59
0.7	0.64	0.62	0.54	0.40	**0.77**	0.67	0.55
0.8	0.57	0.56	0.46	0.33	**0.69**	0.60	0.46
Repository 2 (7 days, 1925 articles, 268 Google stories, 51 inferred stories)							
0.4	0.87	0.78	0.63	0.49	**0.95**	0.80	0.63
0.5	0.79	0.71	0.57	0.43	**0.87**	0.73	0.57
0.6	0.77	0.69	0.56	0.41	**0.85**	0.71	0.56
0.7	0.70	0.62	0.50	0.35	**0.77**	0.64	0.50
0.8	0.66	0.56	0.43	0.29	**0.73**	0.57	0.43
Repository 3 (6 days, 1917 articles, 250 Google stories, 54 inferred stories)							
0.4	0.87	0.78	0.65	0.49	**0.96**	0.81	0.65
0.5	0.80	0.71	0.58	0.43	**0.87**	0.73	0.58
0.6	0.77	0.69	0.56	0.42	**0.85**	0.71	0.56
0.7	0.71	0.63	0.50	0.36	**0.78**	0.65	0.50
0.8	0.66	0.57	0.44	0.31	**0.72**	0.58	0.44

Table 1: Computing recall of proposed story building system

The relationship between Google-stories and inferred-stories is many-to-many, i.e. articles of one Google-story may be spread over multiple inferred-stories and vice-versa. Only

those pairs whose overlap in terms of articles is greater than a threshold T are considered for final recall computation. T is termed as the recall-cutoff. This is to ensure that an inferred-story is not considered similar to a Google-story unless it crosses T. Table 1 also summarizes the performance of the proposed system for different recall-cutoffs. It can be seen that the best average recall value of the proposed system is around 0.95 which can be attained with a topic-threshold of 0.2 and recall-cutoff 0.4. It means that 95% of the stories which were grouped together by Google news co-occur with at-least 40% of their original group members in the proposed system. The last rows in each division of table 1 show that on an average, 71% of the stories co-occur with 80% of their original group-members in the proposed system. Based on two independent human evaluators' reports, it is observed that the system links related news articles with an average precision of 83.5%.

Figure 2 shows a portion of the story graph generated for repository 1. The size of a node is proportional to the number of articles that get attached to it. This figure also illustrates how story highlights help in understanding thematically and/or conceptually related stories.

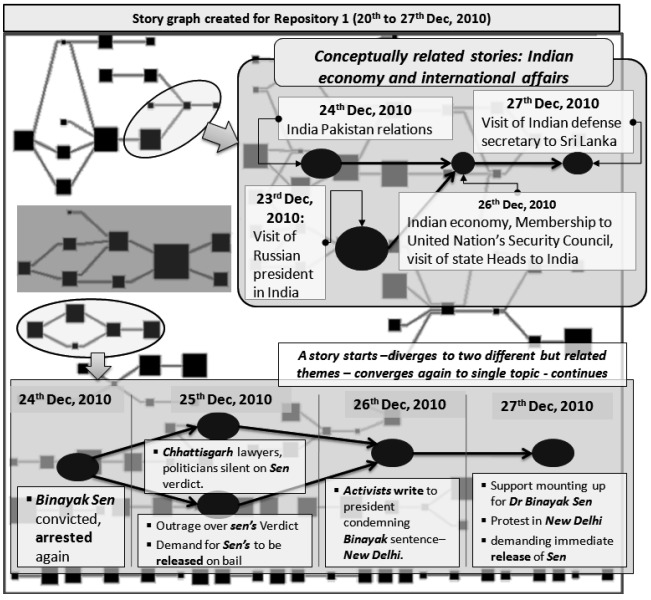

Figure 2: Sample Story Graph, with high-lighted portions blown up to show thematically related story with divergence and convergence (bottom)., conceptually related story (top right)

It is observed that the number of inferred-stories in the proposed system is always fewer than the number of independent stories seen in Google News. This is due to the fact that both thematically and conceptually related stories get linked in the proposed system, even when they are days apart. This is not so in Google News. For example, there were three Google-stories containing 68, 30 and 7 articles respectively between 20th – 27th December, 2010, all of which reported news about fasts and hunger strikes by various political leaders in connection with suicide by farmers in two different regions. The first and second groups were pertaining to the same region but referred to two different leaders. These articles were grouped under one inferred-story and were spread over 9 nodes within it, marked by the shaded region in Figure 2. These nodes show how the different fasts had originated, how the two different stories get linked and how they evolve over time and terminate.

5 Conclusions

In this paper, we have proposed a story-linking and characterization system to enable tracking of conceptually and thematically related news articles. It can characterize evolution of stories. Work on better story visualization is under progress. Further work is also being carried on to improve the performance of topic extraction using improved versions of LDA. Exploiting graph properties for better story interpretation is also being explored.

References

[Uramoto and Takeda, 1998] Naohiko Uramoto and Koichi Takeda, Method for Relating Multiple Newspaper Articles by Using Graphs, and Its Application to Webcasting, in COLING 98 - 17th International Conference on Computational Linguistics, Vol. 2, 1998.

[Ichiro et al., 2003] I. Ichiro, M.Hiroshi and K.Norio, Threading news video topics, Proc. Of 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2003.

[Zhai and Shah,, 2005] Yun Zhai and Mubarak Shah. Tracking news stories across different sources, in Proc. 13th ACM international conference on Multimedia, 2005.

[Xee et al. 2007] Lexing Xee, Hari Sundaram and M. Campbell, Event mining in multimedia streams, Proceedings of the IEEE, Vol. 96, 4, pp. 623-647, 2008.

[Westermann and Jain, 2007] Utz Westermann and Ramesh Jain, Toward a common event model for multimedia applications, IEEE Multimedia, Volume 14, 1, pp. 19-29, 2007.

[Lin and Liang, 2008] Fu-ren Lin and Chia-Hao Liang, Storyline-based summarization for news topic retrospection, Decision Support Systems, 45, pp. 473-490, 2008.

[Blei et al., 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, vol.3, pp. 993–1022, 2003.

[Liu et al., 2009] Duen-Ren Liu, Meng-Jung Shih, Churn-Jung Liau, Chin-Hui Lai, Mining the change of event trends for decision support in environmental scanning, Expert Systems with Applications 36, 972–984, 2009.

[White and Jose, 2004] Ryen W. White and Joemon A. Jose, A Study of Topic Similarity Measures, SIGIR, 2004.