Event Evolution Modeling for Efficient News Search

Mohammad Ubaidullah Bokhari
Chairman, Department of Computer Science, AMU.
Aligarh (India)

Mohd. Kashif Adhami
Research Scholar,
Department of Computer Science, AMU.
Aligarh (India)

ABSTRACT
With the advancement in internet technology, users can now easily extract large number of news stories of any ongoing incidents from existing newswires. These hundreds and thousands of news stories causes the problem of information overload. Users find difficulty in capturing the blueprint of the incident as the volume of information is too large. So it becomes necessary to organize news stories into events and learn how these events developed or evolved within the topic. This paper discusses some of the efforts made to model and discover relationships between news events, which had been a less focused area as compared to TDT research which solely focused flat hierarchical structure. A real-world example is discussed for event evolution analysis and future extensions have been proposed.

General Terms
News events, News stories, Component events, Web search engines.

Keywords
Event Evolution, Evolutionary Patterns, News Event Search.

1. INTRODUCTION
With the development of web, '*news events'* are reported by different news articles in the form of web pages. People are getting more interested in reading news articles online to find out what events occurred [11]. News events are continuously introducing 'breaking' stories in the world. News events are generally reported in unstructured text document, unlike Complex Event Processing [14] having structured events. A composite or complex news event may consist of several component events, referred to as episodes. There are usually interrelationships among these component events, i. e. they are dependent on each other. For example, the event *" DNA from 78 Germanwings crash victims found"*, contains several interrelated component events, e.g., the event " *Germanwings plane crash*" causes the happening of the event *" Investigators reached the crash site"* and "*DNA from 78 Germanwings crash victims found "* and so on. For instance, it is clear that the event "*Germanwings plane crash*" is a "milestone" component event which affects and causes the occurrence of the other component events (Fig.1). People's interest now does not stick to sole news article on an event but also the related events reported by other news articles. They are often interested in the whole picture of an event evolution or development along a timeline. Thus evoking, the need for modeling the dependence relationships among component events and determine which component events contributes in the whole event evolution or development. The current news web sites do not assist users in finding out relevant news articles easily, and they may urge to go across all these news articles in order to find out the interrelationships between component events, and identify their importance in the development of the whole event. Therefore efficiently searching events and organizing the search results in an easily syntactically understandable manner becomes necessary to avoid viewing the huge amount of news articles in time consuming manner.

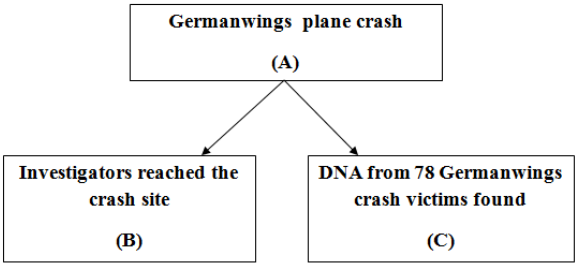

Fig. 1 Milestone event (A) causes the happening of component events (B&C).

Existing search engines returns a list of news web pages when users input event keywords as a query. Rather than organizing the results by events and relationships between events, these engines only give a ranking list of news web pages. It is time consuming and difficult for the users to procure the main picture of an event. Users need to go through these returned documents and reorganize them to extract component events and their feasible relationships.

2. EVENT EVOLUTION DESCRIPTION

2.1 Basic Definitions
In this section, first we discuss some basic definitions related to event evolution. These definitions were complied, to formal definitions given in [21], [22] and [6]. A *story* is a news article providing some information to users. An *Event* is something which occurred at some particular time and place. *Topic* refers to a collection of events strongly inter-related to each other. According to [16], a story acts as the smallest unit in the event evolution hierarchy. Generally a timestamp is

assigned to an event. A location stamp can also be assigned to an event, but for the events, like "reaction of customers around the world on apple iphone price hike", no location stamp can be assigned to it. A topic starts with a fresh seminal event, succeeded by other related events.

2.2 Event Evolution

Yang et al. [8] defined event evolution as the transitional development process of related events within the same topic. For complete topic, the event evolution should begin with a seminal event and eventually evolve to the ending event. Different chains of developing events will be produced in the event evolution as one event may cause the development of various other events or various events may jointly develop/evolve to another event. Again Yang et al. [9] formally defined the event relationship as the directional logical dependencies or relatedness between two events. Here the word 'directional' implies a MUST condition for event A to occur for evolving event B, i. e., event evolution relationship is not bidirectional and cannot be reversed (figure 2). They suggested the appropriateness of using statistical language models to capture event evolution relationships between events based on their observations and gave three criteria for modeling event evolution relationships.

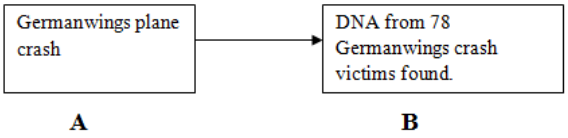

Fig. 2 Directional Logical Dependency from A to B (event A causes the happening of event B).

Events can be identified and generated by the subtask of event detection and tracking in TDT which is discussed in section 4.2. As in this paper we concentrated solely on event evolution modeling, we used manually generated events instead of using TDT techiniques. We used the query 'Germanwings plane crash' in Google News. The manually generated and annotated news events are shown in Figure 5.

Google News aggregates news from multiple sources such as ABC, New York Times, CNN etc. For component news events we extracted news from ABC News Website. Extracting news from a single news website avoid duplicated news from multiple sources. So for the query "Germanwings plane crash we collected 14 component events shown in figure 5. A sample event evolution model is shown in figure 4, which is having 14 events and 24 event evolution relationships. These event evolution relationships were identified by human annotator just to show an illustration. For example, there is an event evolution relationship from "Investigators reached the crash site" (Event 2) to "One of the Black-Boxes from the plane found" (Event 4). Following the event evolution relationships, we will find "Documents seized from co-pilot's home "(Event 9) and finally the terminal event "Investigators revealed the cause: co-pilot locked the captain and deliberately crashed the plane" (Event 13). There are many possible paths from the starting event to the terminal event through event evolution relationships. Some paths are shorter and elucidate the most significant events. The use of event evolution models also leverages sub-graphs/structures, which might be elucidating a particular aspect of the complete news affair and some users might be interested in those aspects. Table-1 uttered some evolutionary sub-paths that might be derived from the main structure. Some events are side events, such as "Reactions from leaders around the world" (Event 10). It does not evolve to other events. Through event evolution modeling, we can easily understand how the news stories are developing along a timeline. It also exhibits a graphical summary of large number of documents that are reporting the events in a news topic.

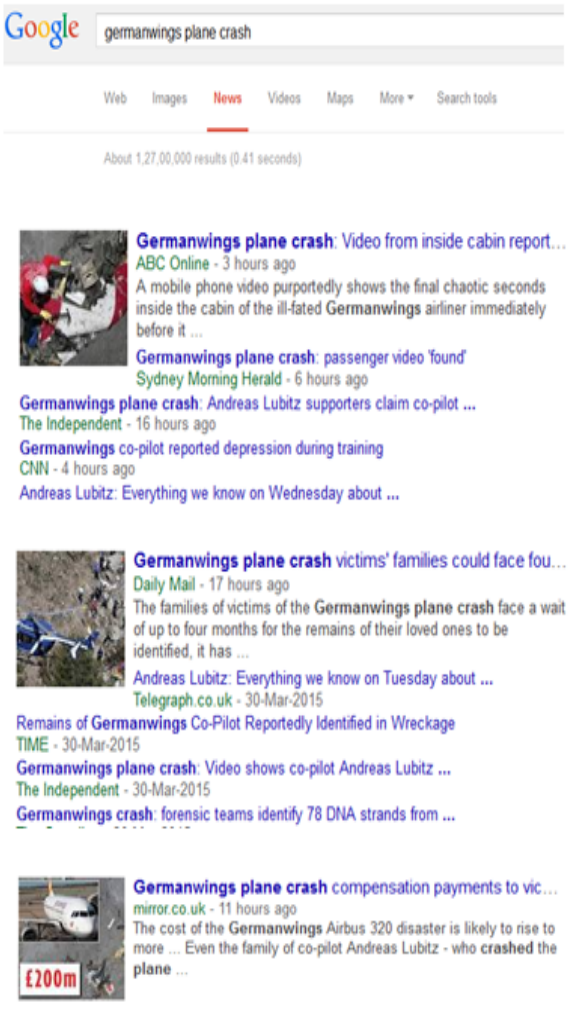

Fig. 3 Google News Results for 'Germanwings plane crash' query.

Table 1: Extracting Sub-Aspects from the main Topic.

Possible paths	**Particular aspect of the news affair**
1-2-4-6-7-9-12-13	Complete Investigations for the cause.
1-3-11-14	About passengers
1-2-5-8-12-14	About crash site.
1-3-4-5-6-8-11-12-14	Complete search process.
1-10	Political aspect
1-2-4-6-12	Investigations at crash site.
1-7-9-13	Investigations in Germany.

Event evolution modeling can be employed in various applications. A well-adept web news infomediary system can be made by integrating the event evolution models with automatic summarization and named entity recognition techniques. The summary of each event and the transition from one event to another is also shown in automatic summarization. In named entity recognition techniques, names of the person, location or organization involved in the event are extracted. Event evolution modeling helps in tracking the person, location or organization along a particular path in an event evolution model. It also supports convenient information-browsing platform for news searchers. Many interesting sub-models or patterns can be derived from the main structure to identify the set of events that does not follow the main line of the topic development.

3. RELATED WORKS

3.1 Event Recognition and Processing

The complex event processing network was enticed in the field of modeling by Luckman [14]. The conceptual model of EPN based on this idea was then devised by Sharon and Etzion [18]. Artikis et. al [3] supported intelligent resource management by presenting EP-IRM, an event processing system. This system recognized composite events from multiple sources of information. Artikis et. al [1] classified the different types of uncertainty found in event processing applications and discussed the implementations on event representation and reasoning. Wasserkrug et. al [33] presented a generic and formal framework for handling the event derivation under uncertainty.

3.2 Topic Detection and Tracking (TDT)

TDT is an active research area whose objective is to organize news documents given a surge of regularly generated fresh stories coming from a wide range of resources like text or audio in several languages. The list of tasks in TDT includes topic detection, story segmentation, fresh event detection, link detection and topic tracking [24]. The motive of the story segmentation is to identify the boundaries for cohesive text fragments. It is concerned with the audio sources and not the newswire sources [6]. Topic detection clusters stories of the identical topic together. New event detection promotes binary decision on each fresh document for discussing new topic not has been reported yet. Link detection identifies document similarity for same topic.

Several techniques have been proposed for detecting news topics and tracking fresh stories for a news topic. Allan et al. [23] regard each incoming news document as a query that was built on the earlier clustered documents to identify if the incoming news document is analogous to any of the clustered documents. A story was regarded as a first story if no identical documents can be found. For efficiency, inverted indexing was used and terms were weighted by *TF-IDF* and *surprisingness.* Yang et al. [37] applied the group-average and single-pass clustering for topic detection. Yang et al. [36], [34] have also determined a multistrategy approach that aggregated *k*NN, Rocchio and language modeling for topic tracking. Carthy [25], Yang et al. [35] and Allan et al. [23] used the NLP approach by blending lexical chains with keywords for topic tracking and exploited seven types of name entities. Chen et al. [7] built an aging theory to model the life-cycle of events by exploiting the temporal relationships between documents. Xu et al. [13] used the webcast text to detect semantic event in video.

3.3 News Summarization

News summarization also aims at automatic processing of news to retrieve useful information and reduce information overload. Newsblaster [26] clusters news stories into hierarchical structure. The lowest level units were similar to news events and larger clusters correspond to topics. NewsinEssence [15] was another summarization system which supported user formed clusters. News summarization system reduce information overload but does not keep track of topic evolution over time. There is also a plethora of research in generating structured representation of news articles as a timeline summary [32], [31] and [20].

4. EVENT EVOLUTION MODELING

4.1 Evolutionary Patterns in Text Streams

Modeling relationships between events has been less focused area and out of the scope of current TDT research. The works from Mei and Zhai [27] [28] studied a peculiar task of discovering and summarizing the evolutionary patterns of themes in a text stream. A theme may be a combination of several events or a part of an event in an interval. But their work did not capture the interrelationships of major events. Fung et. al [17] proposed an algorithm named Time Driven Documents-partition to built an event hierarchy in a user query based text corpus. Their work only modeled the composite relationships between events instead of the dependence relationships which might be more interested by

the users. Shahaf et al. [16] found a coherent chain linking of documents for users to navigate.

4.2 Discovering and Representing Stories from Graphs

The works [22], [21] and [29] focused on discovering stories from news documents and representing the content of stories by graph. Subasic and Berendt [22] proposed a method and visualization tool for mapping and interacting with news stories. In [21] they extended their previous work by defining ETP3 (evolutionary theme pattern discovery, summary and exploration) for tracking of story evolution. Ishii et al. [19] classified extracted sentences for simple language patterns in Japanese to extract casual relations. Choudhary et al. [29] gave a set of transformations to capture evolution of an actor and interactions among actors.

4.3 Event Evolution Modeling

Yang et al. [10] proposed an event evolution pattern discovery technique. They detected event episodes according to their temporal relationships instead of evolution relationships. Temporal relationships assist organizing event episodes in sequence according to their temporal order but not necessarily emulate evolution paths between events. They concentrated n terror attack incident for their case study. They defined the event evolution relationship between event A and B such that it followed three rules which were, first event A must temporally precede event B, secondly event A must be the necessary and/or the sufficient condition of event B and lastly the event evolution relationship must concide with the user information needs. They extended their work in [8] and [12]. Yang et al. [9] defined the event evolution relationships between events and proposed way to measure the event evolution relationships. Identification of the event evolution relationship between two events, in their work, was based on event timestamp, event content similarity, temporal proximity and document distributional proximity. They adopted *precision* and *recall* as evaluation measures. They illustrated the event evolution graph of a terror attack incident. For this case they collected 32 CNN news documents organized into eight events.their complete dataset consists of 10 topics. The concept of event threading was proposed by Nallapati et al. [30]. For understanding a news topic quickly it is inefficient and restrictive to organize news stories by their topics into a flat hierarchy, so they captured the rich structure of events and their dependencies in a news topic through event models. Their definition of event threading was based on content similarity relationship from preceding event to a later event. Tree structure was used to organize event threading rather than a graph. They employed a simple similarity measure between documents to array documents into events and the average document similarity to measure the content dependencies between events. Nallapati's work was extended by Feng and Allan [4]. They proposed news organization infrastructure called incident threading where text snippets passaged the threading by breaking each news story into finer fragments. In their later work, proposed a model called incident threading in [5].

5. CONCLUSION AND FUTURE WORK

In this paper, we reviewed some of the existing works on event evolution modeling. Lots of work has been done in TDT research area but event evolution modeling remained a less focused area. We discussed event evolution modeling taking a real-world example and we intend to discuss limitations in existing works and future extensions in this area.

To discover dependence relationships between two events existing works measured content similarity of events by matching the keywords of the event but there can be some keywords which are related/dependent but not identical like "classroom" and "students". The two words are dependent but existing methods treat them as no relationship. To avoid this mutual information can be adopted to measure the dependence between two terms.

Existing works considered temporal relationship and content similarity for identifying event dependence relationship, in future different approaches such as event reference relationships can also be used.

New methods can be devised to measure the degree of importance of events and rank events according to their importance degree, as people are more interested in important component events than ordinary ones. Another significant extension can be to implement a visualization tool with a sophisticated user interface and further investigation can be done on personalized news event search.

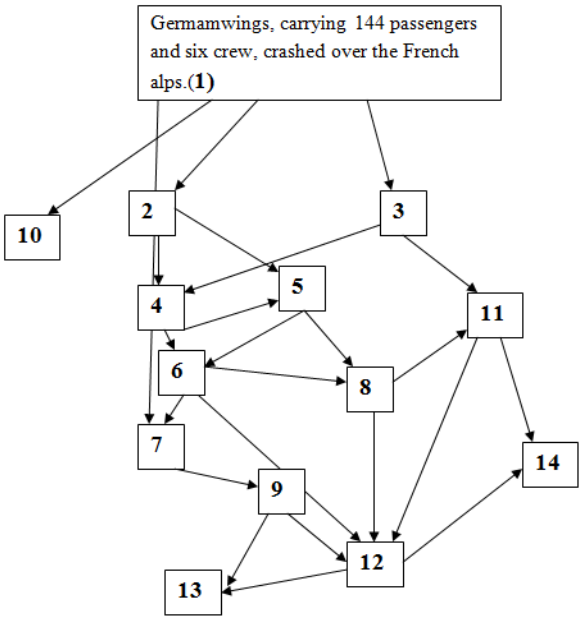

Fig. 4 A Sample Event Evolution Model for the topic "Germanwings plane crash" based on human annotation.

Table 2: A Comparative Study: Major Works on Event Evolution Modeling.

Work	Data sets	Evaluation measures	Features used for capturing dependencies	Event taxonomy	Main objectives	Disadvantage
Nallapati 2004	28 topics from TDT2 corpus and 25 topics from the tdt3 corpus.	Cluster precision (CP), Cluster Recall (CR), Dependency Precision (DP), Dependency Precision (DP) and Joint F1-measure.	Word-based features, time-ordering.	Story->event->topic	Attempted to capture the rich structure of events and their dependencies in a news topic through event models.	Time decaying similarity measure at the document level may not be appropriate for event episode discovery.
Wei et al 2009	53 events and 1,468 relevant news stories selected from TDT2 and TDT3 corpora.	Cluster Recall(CP) and Cluster Precision (CP).	Document similarity, inter-cluster similarity**.**	Story->episode->event	Feature selection metric and document representation scheme for event episode discovery.	News documents belong to one and only one episode, desirable to extent the event episode discovery technique for dealing with multi-episode news documents.
Yang et al. 2006	10 news topics and 78 news stories per topic.	Precision and recall	Temporal and document distributional proximity.	Story->event->topic	Develop the technique to identify event evolution relationships and represented the event evolution graphs.	Temporal and document distributional proximity may have some overlapping effects, i. e., cancel the effect of each other to some degree.
Yang et al. 2009	10 news topics and 78.2 stories per topic.	Precision and recall	Event content similarity, temporal proximity and document distributional proximity.	Story->event->topic	An event evolution graph is constructed for efficient news search.	Taking content dependence into consideration and measuring content dependence through simple similarity functions, are effective only when two events share enough features.
Feng and Allan 2007.	6 topics from TDT-3 corpus	$Prec_{cluster}$,$REC_{cluster}$ and $F_{cluster}$ for measuring clustering accuracy	Two-stage clustering algorithm with scenario-specific rules, global optimization framework for establishing links in incident network.	Story->event->topic	A news organization infrastructure is proposed	The data collection is small and limited, dependencies between events are binary.
Feng and Allan 2009.	Part of the GALE corpus consisting of 17 qeuries/topics and 170 news documents.	*Concentration* and *purity* scores for evaluating clustering performance	Term vectors, together with the automatically extracted main characters, geographical locations and timestamps**.**	Passage->story->event->topic	Introduced *passage threading*, which processes news at the passage level**.**	More accurate modeling of semantic information represented in the short piece is required, the current implementation used single main-subject and possible types of relations are restricted.

Topic: Germanwings plane crash.

Event1: Germamwings, carrying 144 passengers and six crew, crashed over the French alps.
Event 2: Investigators reached the crash site.
Event 3: Search crews trying to recover crash victims' DNA from the site.
Event 4: One of the Black-Boxes from the plane found.
Event 5: New road built to improve access to the crash site.
Event 6: Second black-box found after nine day hunt.
Event 7: Investigations at co-pilot's home.
Event 8: Two helicopters hover overhead the site for more widespread debris.
Event 9: Documents seized from co-pilot's home.
Event 10: Reactions from leaders around the world..
Event 11: Search revealed: 75 Germans were on board.
Event 12: Memory card recovered from the crash site.
Event 13: Investigators revealed the cause: co-pilot locked the captain and deliberately crashed the plane.
Event 14: Search crews had recovered all 150 crash victims' DNA from the site.

Fig. 5 List of component events for the topic "Germanwings plane crash".

6. REFERENCES

[1] A. Artikis, O. Etzion, Z. Feldman and F. Fournier. "Event Processing under Uncertainity". In Proc. of DEBS'12, ACM, New York, USA, pp. 32-43, 2012.

[2] C. C. Aggarawal, J. Han, J. Wang and P. S. Yu. "A Framework for On-Demand Classification of Evolving Streams". IEEE Trans. on Knowledge and Data Engineering, Vol: 18(5), pp. 577-589, 2006.

[3] A. Artikis, R. Marterer, J. Pottebaum and G. Paliouras. "Event Processing for Intelligent Resource Management". In Proc. of ECAI'12, Corum, France, pp. 943-948, 2012.

[4] A. Feng and J. Allan. "Finding and Linking Incidents in News". In Proc. of CIKM'07, ACM, New York, USA, pp. 821-830, 2007.

[5] A. Feng and J. Allan. "Incident Threading for News Passages". In Proc. of CIKM'09, ACM, New York, USA, pp. 1307-1316, 2009.

[6] C. C. Aggarawal, J. Han, J. Wang and P. S. Yu. "A Framework for On-Demand Classification of Evolving Streams". IEEE Trans. on Knowledge and Data Engineering, Vol: 18(5), pp. 577-589, 2006.

[7] C. C. Chen, Y. Chen and M. C. Chen. "An Aging Theory for Event Life-Cycle Modeling". IEEE Trans. Systems, Man and Cybernetics, Vol: 37(2), pp. 237-248, 2007.

[8] C. C. Yang and X. Shi. "Discovering Event Evolution Graph from Newswires". In Proc. of WWW'06, ACM, New York, USA, pp. 945-946, 2006.

[9] C. C. Yang, X. Shi and C. P. Wei. "Discovering Event Evolution Graphs from News Corpora". IEEE Trans. Systems, Man and Cybernetics, Vol: 39(4), pp. 850-863, 2009.

[10] C. C. Yang, X. Shi and C. P. Wei. "Tracing the Event Evolution of Terror Attacks from On-Line News". Intelligence and Security Informatics, Vol: 3975, pp. 343-354, Springer-Verlag, Hiedelberg, 2006.

[11] C. M. Kelly and G. D. Moulin. The web cannibalizes media., Tech. report, The Forrester Grouop, 2002.

[12] C. P. Wei, Y. H. Lee, Y. S. Shiang, J. D. Chen and C. C. Yang. "Discovering Event Episodes from News Corpora: A Temporal-Based Approach". In Proc. of ICEC'09, ACM, New York, USA, pp. 72-80, 2009.

[13] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu and Q. Huang. "Using Webcast text for semantic event detection in broadcast sports video". IEEE Trans. Multimedia, Vol: 10(7), pp. 1342-1355, 2008.

[14] D. C. Luckhman. "The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems". Addison-Wesly Longman Publishing Co., Inc., Boston, MA, USA, ISBN: 0201727897, 2001.

[15] D. Radev, J. Otterbacher, A. Winkel and S. Blair-Goldensohn. "NewsinEssence: Summarizing Online News Topics". Communications of the ACM-The digital society, Vol: 48(10), pp. 95-98, 2005.

[16] D. Shahaf and C. Guestrin. "Connecting the Dots Between News Articles". In Proc. of KDD'10, ACM, New York, USA, pp. 623-632, 2010.

[17] G. P. C. Fung, J. X. Yu, H. Liu and P. S. Yu. "Time Dependent Event Hierarchy Construction". In Proc. of KDD'07, ACM, New York, USA, pp. 300-309, 2007.

[18] G. Sharon and O. Etzion. Event Processing Network Model and Implementation. IBM System Journal, Vol: 47, pp. 321-334, 2008.

[19] H. Ishii, Q. Ma and M. Yoshikawa. "Casual Network Construction to Support Understanding of News". In Proc. of HICSS'10, IEEE Computer Society, Washington DC, USA, pp. 1-10, 2010.

[20] H. L. Chieu and Y. K. Lee. "Query Based Event Extraction along a Timeline. In Proc. of SIGIR'04, Sheffield, UK, pp. 425-432, 2004.

[21] I. Subasics and B. Berendt. "Discovery of Interactive Graphs for Understanding and Searching Time-Indexed

Corpora". Knowledge and Information Systems, Vol: 23, pp. 293-319, 2010.

[22] I. Subasics and B. Berendt. "Web Mining for Understanding Stories Through Graph Visualization". In Proc. of ISDM'08, IEEE Computer Society, Washington DC, USA, pp. 570-579, 2008.

[23] J. Allan, R. Papka and V. Lavrenko. "On-Line New Event Detection and Tracking". In Proc. of 21st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, Melbourne, Australia, pp. 37-45, 1998.

[24] J. Allan. "Topic Detection and Tracking: Event-Based Information Organization". The Information Retrieval Series, Springer, US, 2002.

[25] J. Carthy. "Lexical Chains for Topic Detection". Tech. report, Dept. of Comp. Sc., Dublin National Univ., Ireland, 2002.

[26] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, C. Sable, B. Schiffman and S. Sigelman. "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster". In Proc. of the Human Language Technology Conference, Mrgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 280-285, 2002.

[27] Q. Mei and C. Zhai. "Discovering Evolutionary Theme Patterns fron Text: An Exploration of Temporal Text Mining". In Proc. of KDD'05, ACM, New York, USA, pp. 198-207, 2005.

[28] Q. Mei, C. Liu, H. Su and C. Zhai. "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs". In Proc. of WWW'06, ACM, New York, USA, pp. 533-542, 2006.

[29] R. Choudhary, S. Mehta, A. Bagchi and R. Balakrishnan." Towards Characterization of Actor Evolution and Interactions in News Corpora". In Proc. of ECIR'08, 30th European Conference on Advances in Information Retrieval, Springer-Verlag, Hiedelberg, pp. 422-429, 2008.

[30] R. Nallapati, A. Feng, F. Peng and J. Allan. "Event Threading Within News Topics". In Proc. of CIKM'04, ACM, New York, USA, pp. 446-453, 2004.

[31] R. Swan and J. Allan. "Automatic Generation of Timelines". In Proc. of SIGIR'00, Athens, Greece, pp. 49-56, 2000.

[32] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li and Y. Zhang. "Evolutionary Timeline Summarization: A Balanced Optimization Framework via Iterative Susbstitution". In Proc. of SIGIR'11, Beijing, China, pp. 745-754, 2011.

[33] S. Wasserkrug, A. Gal, O. Etzion and Y. Turchin. Efficient Processing of Uncertain Events in Rule-Based Systems". IEEE Trans. on Knowledge and Data Engineering, Vol:24, pp. 45-58, 2012.

[34] Y. Yang, J. Carbonell, R. Brown, J. Lafferty, T. Pierce and T. Ault. "Multi-Strategy Learning for Topic Detection and Tracking". Topic Detection and Tracking: Event-Based Information Organization, Vol: 12, pp. 85-114, Norwell, MA: Kluwer, 2002.

[35] Y. Yang, J. Zhang, J. Carbonell and C. Jin. "Topic-Conditioned Novelty Detection". In Proc. of ACM SIGKDD'02, Edmonton, AB, Canada, pp. 688-693, 2002.

[36] Y. Yang, T. Ault, T. Pierce and C. W. Lattimer. "Improving Text Categorization Methods for Event Tracking". In Proc. of 21st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, , Greece, Athens, pp. 65-72, 2000.

[37] Y. Yang, T. Pierce and J. Carbonell. "A Study on Retrospective and On-Line Event Detection". In Proc. of 21st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, Melbourne, Australia, pp. 28-36, 1998.