ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

Evolutionary Timeline Summarization

S. N Deshmukh, S. S. Nandagaonkar
M.E.(computer engg-ii), Professor, computer department
VPCOE, Baramati, Maharashtra,

*Abstract— **Faced with thousands of news articles, people usually try to ask the general aspects such as the beginning, the evolutionary pattern and the end. General search engines simply return the top ranking articles according to query relevance and fail to trace how a specific event goes. General search engines simply return the top ranking articles according to query relevance and fail to trace how a specific event goes. Infomediaries such as Google and Yahoo are available to search across multiple newswires to retrieve news stories of any ongoing incidents. Techniques that are capable of extracting the underlying structure of the news events are desired. They are helpful for users to understand the evolution of events on the same topic. Timeline summarization, which aims at generating a series of concise summaries for news collection published at different epochs, can give readers a faster and better access to understand the evolution of news. The key of timeline summarization is how to select sentences which can tell readers the evolutionary pattern of topics in the event. It is very common that the themes of a corpus evolve over time, and topics of adjacent epochs usually exhibit strong correlations. Thus, it is important to model topics across different documents and over different time periods to detect how the events evolve. We develop a novel model called Evolutionary Hierarchical Dirichlet Process (EHDP) to capture the topic evolution pattern in timeline summarization. In EHDP, time varying information is formulated as a series of HDPs by considering time-dependent information. Then we construct an event evolution graph. An Event evolution graph is constructed to present the underlying structure of events for efficient browsing and extracting of information. Case study and experiments are presented to illustrate and show the performance of our proposed technique. It is found that our proposed technique outperforms the baseline technique and other comparable techniques in previous work.***

I. INTRODUCTION

Due to the popularity of the Internet, most news stories have electronic versions published on newswires such as CNN, BBC, CBS, etc. Infomediaries such as Google and Yahoo are available to search across multiple newswires to retrieve news stories of any ongoing incidents. Retrieving news of the same topic from multiple sources and keeping information updates becomes more convenient and easier. However, it also generates tremendous volume of news streams. Managing, interpreting and analyzing such a huge volume of information is a difficult task. General search engines simply return the top ranking articles according to query relevance and fail to trace how a specific event goes. Timeline summarization, which aims at generating a series of concise summaries for news collection published at different epochs can give readers a faster and better access to understand the evolution of news. In this paper, we propose EHDP: evolutionary hierarchical Dirichlet process (HDP) model for timeline summarization. In EHDP, each HDP is built for multiple corpora at each time epoch, and the time dependencies are incorporated into epochs under the Monrovian assumptions. Then we construct the event evolution graph. The objective of this paper is to develop a technique that to capture the topic evolution pattern in timeline summarization and then construct the event evolution graph that identify the relationship between the news of the event. These relationships denote how events evolve or develop from the beginning to the end of the specific news affair. The events together with their relationships are then used to build a well-defined structure, *event evolution graph* which presents the blueprint of a news topic. Such event evolution graphs show the sophisticated event interrelationships in a graphical structure for easy navigation and browsing. Accordingly, users are able to capture the major events and understand the flow of the stories within the incident.

II. RELATED WORK

A. TIMELINE SUMMARIZATION

1. Allan et al. (2001)

The task of time line summarization is firstly proposed by Allan et al.(2001) by extracting clusters of noun phases and name entities.

Drawbacks:

Noun phases and name entity not sufficient to summarize data

2. Chieu et

Chieu et al.(2004) built a similar system in unit of sentences with interest and burstiness.

Drawbacks:

However, these methods seldom explored the evolutionary characteristics of news.

3. Yan et al.(2011)

Yan et al.(2011) extended the graph based sentence ranking algorithm used in traditional multi-document summarization (MDS) to timeline generation by projecting sentences from different time into one plane. They further explored the timeline task from the optimization of a function considering the combination of different respects such as relevance, coverage, coherence and diversity (Yan et al., 2011b). However, their approaches just treat timeline generation as a sentence ranking or optimization problem and seldom explore the topic information lied in the corpus.

B. CONSTRUCTION OF EVENT EVOLUTION GRAPH

I. Topic Detection and Tracking

For event tracking mostly TDT technique is used. The aim of TDT is to organize news documents given a stream new stories coming from different news channels. There have been several techniques on detecting news topics and tracking new stories for a news topic.

1. James Allan et. Al [1] in on line News event detection, Clustering and Trackin. Considered each incoming news document as a query that was made on the previous clustered documents to determine if the incoming news document is similar to any of the clustered documents.

Drawbacks:

This technique only concentrated on the clustering of related document, it was failed to define the temporal relation between the events.

2. Yang et al. [2] employed the group average clustering and the single-pass cluster for topic detection. Yang in automated tracking of events from chronologically ordered document streams is a new challenge for statistical text classification.

Drawbacks:

For event detection and event tracking they used clustering of similar events, but did not defined event evolution.

3. Carthy [3], Allan [1], and Yang [2] used the natural language processing approach by combining lexical chains with keywords for topic tracking and extracted seven types of name entities. For finding similar keywords is not sufficient to decide it's similarity between documents.

Drawbacks:

This approach gives approximate result. To find event content similarity between two events and decaying factors required temporal proximity and document proximity.

II. Event Evolution Graph

Makkonen[4] and Nallapati[5] conduct investigation on event evolution as a subtopic of TDT.

4. Makkonen [4] was the first investigate event evolution. The news documents within a topic are temporally linearly ordered. A narrative begins when the first story of the topic is detected. A seminal event may lead to several other events. The events at the beginning may have more influence on the events coming immediately after than the events at the later time. The events and the event evolution relationship can be represented as a graph structure.

Drawbacks:

This approach did not define the concept of event evolution clearly and elaborate the structure of event evolution. It is lack of details in the event evolution model and fails to present any experimental evaluation results.

5. Nallapati [5] (Event threading within news topics) was defined the concept of event threading, given a small number of documents and events in a news topic. Their definition of event threading is close to event evolution except that they consider event threading as a tree structure rather than a graph.

Drawbacks:

Their dependence modeling methods only consider the average document similarity between events. Such methods are not effective and sufficient to identifying the event evolution relationships.

6. Wei and Y. Chang [6] proposed an event evolution pattern discovery technique that identifies event episodes together with their temporal relationships that occur frequently in a collection of events of the same type. Their study focuses on segmenting a sequence of news stories of specific event episode and their relationship. However, relationships among event episodes discussed only consist of temporal ones.

Drawbacks:

This approach define relationship among event episode only consists of temporal ones, which gives approximate results.

7. Wei and Chang [7] proposed an event evolution pattern discovery technique that identifies event episodes together with their temporal relationships that occur frequently in a collection of events of the same type. Their work differs from prior studies in that they focus on segmenting a sequence of news stories of a specific event into event episodes and generalizing event episodes across different events of similar topics. However, relationships among event episodes discussed only consist of temporal ones. In this work, formally define the event evolution by three logical rules. Besides, introduce the temporal relationship, event similarity, temporal proximity and document distributional proximity to identify the event evolution relationships to construct the event evolution graphs. Given such graphical representation of the underlying structure of events in a terror any incident, users can easily navigate the development of the incident and extract specific information for their needs.

Drawbacks:

This approach use classic clustering methods, but there is some remarkable characteristics when the news document are divided into diversified cluster along the time**.**

III. PROBLEM FORMULATION

Given a general query Q, we firstly obtain a set of query related documents news from news channel. We notate different corpus as C according to their published time where C_t denotes the document collection published at epoch t. Document D_{ti} is formulated as a collection of sentences. Each sentence is presented with a series of words w_{tj} and associated with a topic Θ^{t}_{ij} .V denotes the vocabulary size. The output of the algorithm is a series of timelines summarization I.· Then we construct the event evolution graph. In our system first we cluster news according to user query. Timeline summarization, which aims at generating a series of concise summaries for news collection published at different epochs can give readers a faster and better access to understand the evolution of news. Then we represent each news in event evolution graph which presents the blueprint of a news topic. Such event evolution graphs show the sophisticated event interrelationships in a graphical structure for easy navigation and browsing. Accordingly, users are able to capture the major events and understand the flow of the stories within the incident.

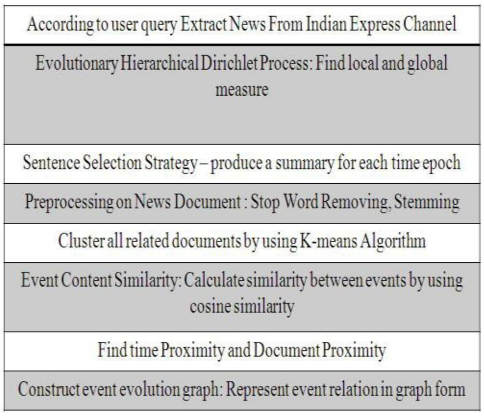
System Architecture

System works in four modules:

- Evolutionary Hierarchical Dirichlet Process (EHDP)
- Sentence selection strategy Module
- Preprocessing Module
- Event Evolution graph module.

As shown in figure, input to the system is user query in the form. The system retrieves relevant news from any news channel here we use Indian Express of search engine. Preprocessing of user query. Then according to the query it extracts related news documents. From this document we select particular sentence by using topic scoring algorithm. It gives summarized news. Then we find relation between news. This relation we represent into graph format. Event evolution graph represent relationship between two news.

Flow of System



According to user query Extract News From Indian Express Channel

Evolutionary Hierarchical Dirichlet Process: Find local and global measure

Sentence Selection Strategy – produce a summary for each time epoch

Preprocessing on News Document : Stop Word Removing, Stemming

Cluster all related documents by using K-means Algorithm

Event Content Similarity: Calculate similarity between events by using cosine similarity

Find time Proximity and Document Proximity

Construct event evolution graph: Represent event relation in graph form

Fig. 1. Flow of System

A. Evolutionary Hierarchical Dirichlet Process (EHDP) [9]

We focus solely on the news summarization and construction of event evolution graph using manually generated and annotated news events. It should also be noted that there are some online sources of well-generated news events (either manually or automatically by undisclosed techniques), Google News aggregates news from multiple sources such as CNN, ABC, BBC, etc. It also traces related news for each news article. These sources, after necessary preprocessing, can serve as inputs to our proposed event evolution identification technique .We only used the related news extracted from Indian Express to avoid duplicated news from multiple sources. Extracted news we stored in temporary database. EHDP to capture the topic evolution pattern in timeline summarization. In EHDP module each corpus Ct is modeled as a HDP. These HDP shares an identical base measure G_0, which serves as an overall bookkeeping of overall measures. We use G^t_0 to denote the base measure at each epoch and draw the local measure G^t_i for each document at time t from G^t_0. In EHDP, each sentence is assigned to an aspect Θ^t_{ij} with the consideration of words within current sentence. To consider time dependency information in EHDP, we link all time specific base measures G^t_0 with a temporal Dirichlet mixture model as follows:

$$G_0^t \sim DP(\alpha, \tfrac{1}{K}G_0 + \textstyle\sum_{\delta=0}^{\Delta} F(v,\delta)G_0^{t-\delta})$$

where $F(\gamma, \delta) = \exp(-\delta/\gamma)$ denotes the exponential kernel function that controls the influence of neighboring corpus. K denotes the normalization factor where $K = 1 + \sum_{\delta=0}^{\Delta} F(\gamma, \delta)$ Δ is the time width and λ is the decay factor.

B. Sentence Selection Strategy

The task of timeline summarization aims to produce a summary for each time and the generated summary should meet criteria such as relevance, coverage and coherence. To care for these three criteria, we propose a topic scoring algorithm.

C. Preprocessing

Event evolution graph represent relation between news .After summarization we have construct graph for that we preprocessed all news documents. In Preprocessing remove stop words, stemming is used . These sources, after necessary preprocessing, can serve as inputs to our proposed event evolution identification technique. We only used the related news extracted from Indian Express to avoid duplicated news from multiple sources. Extracted news we stored in temporary database.

D. Constructing Event Evolution Graph

In this section, we present our proposed technique for constructing an event evolution graph for a given news corpus of the same topic. Our technique can be decomposed into three steps:

- Generating representations of events from news stories.
- Modeling event evolution relationships between these events.
- Pruning edges that correspond to invalid or weak even evolution relationships.

1. Generation of Events

Manually generated events can eliminate the biases that may be created by different event detection and tracking techniques and, hence, provide a best platform on which we can fairly compare different event evolution identification techniques. It should also be noted that there are some online sources of well-generated news events.

a. Feature Extraction:

In feature extraction technique parses the documents in each document sequence to produce a list of nouns and noun phrases that exclude a set of pre-specified stop words.

b. Feature Selection:

For each document sequence, the feature selection phase select the top k-feature with the highest feature selection metric score to represent the documents in the document sequence. S

c. Document Representation:

In the document representation phase, each document is then representing using the representative feature selected for the document sequence to which that document belongs.

d. Document Clustering:

The documents clustering phase generates episode as cluster of documents. Specifically, we obtained the K-means algorithm.

2. Modeling Event Evolution Relationships

If an event evolution relationship exists from one event to another, the two events should share some common information in their content. Accordingly, we adopt the vector space model to measure the relatedness of events. A k-term vector for S is denoted as $\omega_i = \omega_{i1}, \omega_{i2}, \ldots, \omega_{ik}$. Let a story i be represented as a weighted term vector $\omega_i = \omega_{i1}, \omega_{i2}, \ldots, \omega_{ik}$..On the basis of the traditional TF-IDF function ω_{ik} is defined as

$$\boxed{\omega ik = \frac{\mathrm{tf}_{i\,k}}{\max \mathrm{tf}_{i\,l}} \log \frac{N}{dfk}}$$

where tfik is the frequency of term k in the news document i, N is the total number of news documents in that topic, dfk is the number of news documents in that topic containing term k, and max tfil is the maximum term frequency for all terms in document i.

The content similarity between two events is the cosine similarity of their event term vectors.

$$\boxed{\mathrm{cos\_sim}(e_i\,,\,e_j\,) = \frac{\sum_{x=1}^{k} \omega'_{ix}\,\omega'_{jx}}{\sqrt{\sum^{k}(\omega'_{ix})^2 + \sum^{k}(\omega'_{jx})^2}}}$$

3. *Temporal proximity*:[7] *and* *Document Distributional Proximity*

Temporal proximity measure their relative distance between two events. It defines as follows:

Where T is the event horizon defined as the temporal distance between the start time of the earliest event timestamp

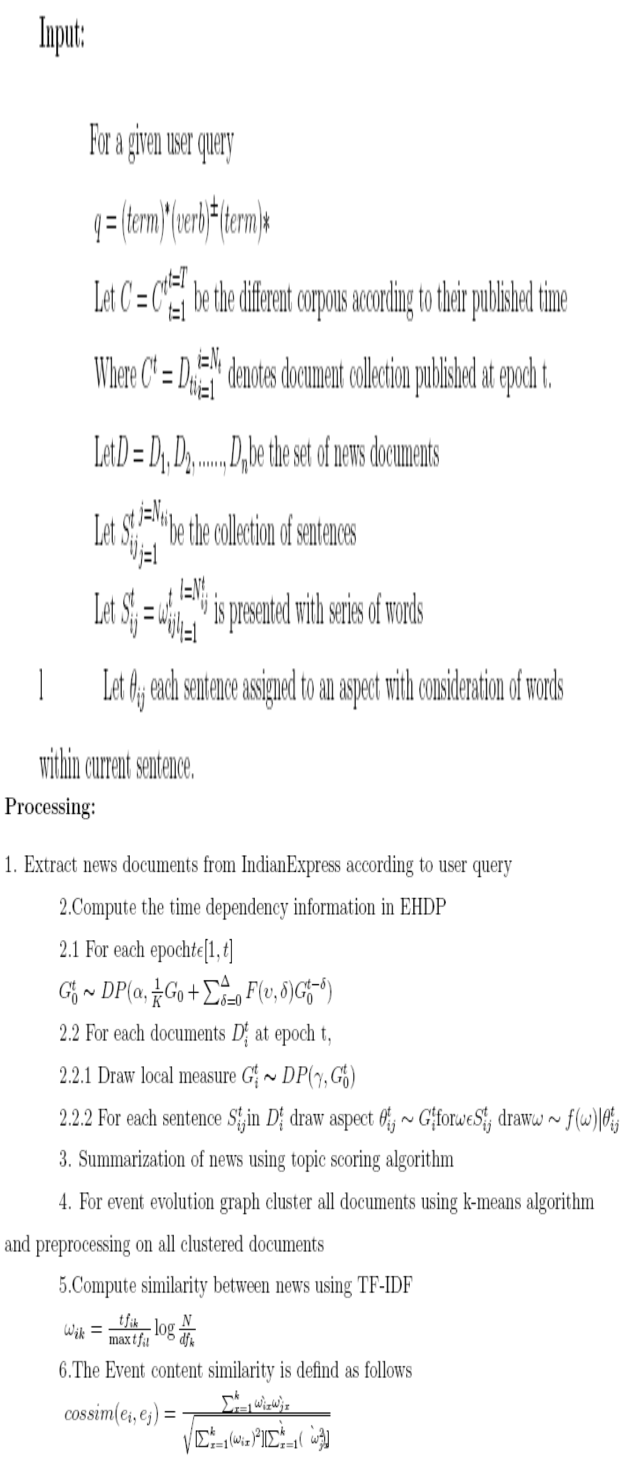
$$\boxed{tp(\mathrm{e}_1,\ \mathrm{e}_2) = e^{-\alpha[\frac{d(\tau(\mathrm{e}_1),\tau(\mathrm{e}_2))}{T}]}}$$

Document Distributional Proximity:[7]

$$\boxed{df\,(\mathrm{e}_1,\ \mathrm{e}_2) = e^{-\beta\frac{m}{n}}}$$

This algorithm represents total flow of our project.

Algorithm:



Input:

For a given user query

$q = (term)^{*}(verb)^{\pm}(term)*$

Let $C = C^{t}{}_{t=1}^{t=T}$ be the different corpous according to their published time

Where $C^{t} = D_{ti}{}_{i=1}^{i=N_t}$ denotes document collection published at epoch t.

Let$D = D_1, D_2, \ldots\ldots, D_n$be the set of news documents

Let $S_{ij}^{t}{}_{j=1}^{j=N_{ti}}$be the collection of sentences

Let $S_{ij}^{t} = \omega_{ijl}^{t}{}_{l=1}^{l=N_{ij}^{t}}$ is presented with series of words

l Let θ_{ij} each sentence assigned to an aspect with consideration of words within current sentence.

Processing:

1. Extract news documents from IndianExpress according to user query

2.Compute the time dependency information in EHDP

2.1 For each epoch$t\epsilon[1,t]$

$G_0^t \sim DP(\alpha, \frac{1}{K}G_0 + \sum_{\delta=0}^{\Delta} F(v,\delta)G_0^{t-\delta})$

2.2 For each documents D_i^t at epoch t,

2.2.1 Draw local measure $G_i^t \sim DP(\gamma, G_0^t)$

2.2.2 For each sentence S_{ij}^tin D_i^t draw aspect $\theta_{ij}^t \sim G_i^t$for$\omega\epsilon S_{ij}^t$ draw$\omega \sim f(\omega)|\theta_{ij}^t$

3. Summarization of news using topic scoring algorithm

4. For event evolution graph cluster all documents using k-means algorithm and preprocessing on all clustered documents

5.Compute similarity between news using TF-IDF

$\omega_{ik} = \frac{tf_{ik}}{\max tf_{il}} \log \frac{N}{df_k}$

6.The Event content similarity is defind as follows

$cossim(e_i, e_j) = \frac{\sum_{x=1}^{k} \grave{\omega}_{ix}\grave{\omega}_{jx}}{\sqrt{[\sum_{x=1}^{k}(\omega_{ix})^2][\sum_{x=1}^{k}(\ \grave{\omega}_{jx}^2)]}}$

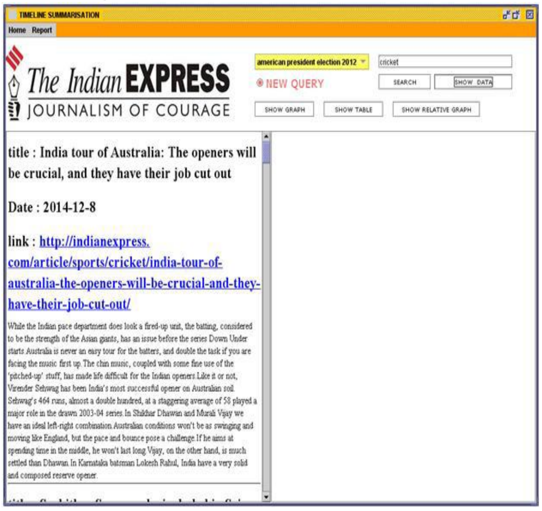
7.Compute temporal proximity and document proximity.

8. Represent vent relation in graph format.

Output: Summarized news and Evolution graph.//

IV. RESULT

In our system user enter query and system extract news from Indian Express news channel. System summarizes all extracted news from news channel by using EHDP technique. Figure 2. Shows summarized news.



title : India tour of Australia: The openers will be crucial, and they have their job cut out

Date : 2014-12-8

link : http://indianexpress.com/article/sports/cricket/india-tour-of-australia-the-openers-will-be-crucial-and-they-have-their-job-cut-out/

While the Indian pace department does look a fired-up unit, the batting, considered to be the strength of the Asian giants, has an issue before the series Down Under starts.Australia is never an easy tour for the batters, and double the task if you are facing the music first up.The chin music, coupled with some fine use of the 'pitched-up' stuff, has made life difficult for the Indian openers.Like it or not, Virender Sehwag has been India's most successful opener on Australian soil. Sehwag's 464 runs, almost a double hundred, at a staggering average of 58 played a major role in the drawn 2003-04 series.In Shikhar Dhawan and Murali Vijay we have an ideal left-right combination.Australian conditions won't be as swinging and moving like England, but the pace and bounce pose a challenge.If he aims at spending time in the middle, he won't last long.Vijay, on the other hand, is much settled than Dhawan.In Karnataka batsman Lokesh Rahul, India have a very solid and composed reserve opener.

Fig. 2. Summarized News

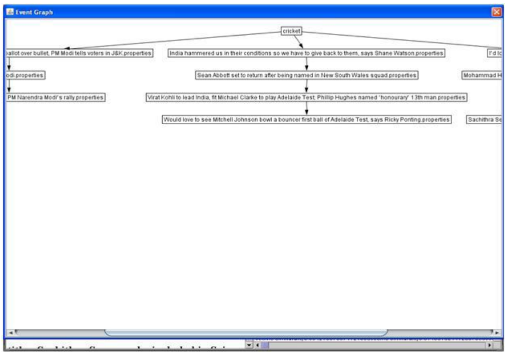

Event Graph
cricket
allot over bullet, PM Modi tells voters in J&K.properties
India hammered us in their conditions so we have to give back to them, says Shane Watson.properties
I'd lo
odi.properties
Sean Abbott set to return after being named in New South Wales squad.properties
Mohammad H
PM Narendra Modi's rally.properties
Virat Kohli to lead India, fit Michael Clarke to play Adelaide Test; Phillip Hughes named "honourary" 13th man.properties
Would love to see Mitchell Johnson bowl a bouncer first ball of Adelaide Test, says Ricky Ponting.properties
Sachithra Se

Fig. 3. Event Evolution Graph

When user clicks on Show Relative Graph button, system page get displayed event evolution graph as shown in figure 2.

A. Performance Analysis

Performance of system is analyzed by comparing the solution generated by manually annotated event evolution graph and system generated event evolution graph results. Performance of system is measured using standard IR measures: precision and recall. Assuming the manually annotated set of event evolution relationships as the truth set O and the system generated by certain algorithms as A.

Precision (P): It is the ratio of the number of true and valid event evolution relationships retrieved by the automatic system to the total number of event evolution relationships retrieved by the automatic system.

$$\textbf{Precision (P)} = \frac{|c|}{|o|}$$

Recall (R): It is the ratio of the number of true and valid event evolution relationships retrieved by the automatic system to the total number of true and valid event evolution relationships annotated manually.

$$\textbf{Recall (R)} = \frac{|c|}{|o|}$$

Table 1 shows precision and recall values for sample user queries. Figure 4 and 5 show the average precision and recall of the two techniques. Graphically, recall and precision can be shown as in figure 4 and 5 for different user queries.

Query	Performance Measure	
	Precision	Recall
Nanredra Modi	0.5	0.67
Anna Hazare	0.48	0.69
Cricket	0.44	0.63
Hockey	0.52	0.65

Table 1: Performance Analysis

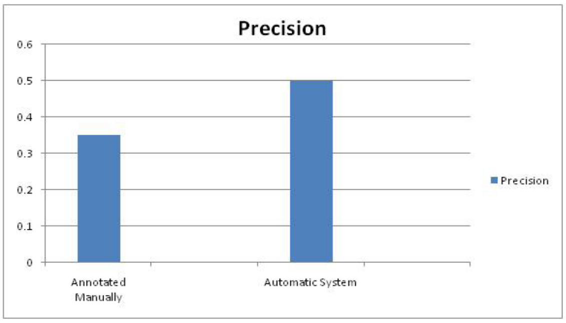


Fig 4: Precision of Proposed System

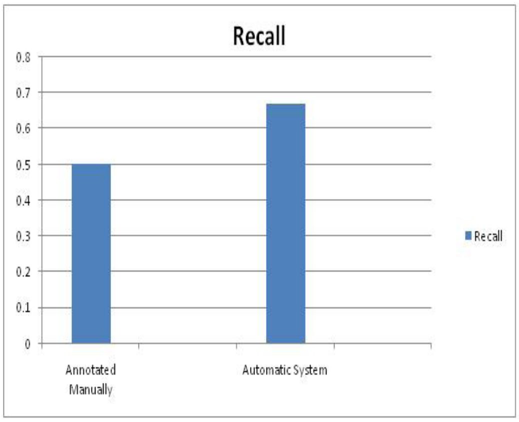


Fig 5: Recall of Proposed System

From the values of recall and precision obtained for sample scenarios, we conclude our system gives better result than the proposed technique is promising in producing an event evolution graph satisfactory precision and recall to support user navigation and understanding of the development of events in a given topic.

V. CONCLUSION

In this system we present an evolutionary HDP model for timeline summarization. Our EHDP extends original HDP by incorporating time dependencies and background information. We also develop an effective sentence selection strategy for candidate in the summaries. In this work, to capture the development of the events in these incidents efficiently and effectively, we develop the event evolution identification technique to automatically identify event evolution relationships and represent the underlying structure as an event evolution graph. We developed a novel model to capture the topic evolution pattern in timeline summarization. Users are able to capture the major events and understand the flow of the stories within the incident.

VI. FUTURE SCOPE

In our future work, we plan to develop information visualization tools that support users in conducting interactive browsing, extracting the main story line from the event evolution graph and extracting summary automatically for a specific path in the evolution graph.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko, ―On-line new event detection and tracking in Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, Melbourne, Australia, 1998, pp. 37–45.

[2] Y. Yang, T. Pierce, and J. Carbonell, ―A study on retrospective and online event detection, in Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, Melbourne, Australia, 1998, pp. 28–36.

[3] J. Carthy J. Carthy, ―Lexical chains for topic detection Dept computer Sci, Univ. Collage Dublin-National Univ. Ireland, Dublin, Ireland, 2002. Tech. Rep.

[4] J. Makkonen, ―Investigations on event evolution in TDT, in Proc. Conf. North Amer. Chapter Assoc. Linguistics Human Language Technol., HLT-NAACL Student Research Workshop, Edmonton, AB, Canada, 2003, pp. 43–48.

[5] R. Nallapati, A. Feng, F. Peng, and J. Allan, ―Event threading within news topics, in Proc. 13th ACM Int. Conf. Inf. Knowl. Management,

[6] Wei and Y. Chang, ―Discovering event evolution patterns from document sequences IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 37, no. 2, pp. 273–283, Mar. 2007

[7] Wei and Y. Chang, ―Discovering event evolution Graph from corpora IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 39, no. 4, July. 2009.

[8] S.S.Nandagaonkar, D.B.Hanchate, S, N.Deshmukh, "Survey on Event tracking and Event Evolution, ijcta Int.J.Comp.Tech.Appl, Vol 3 (1).

[9] Jiwei Li, Sujian Li "Evolutionary Hierarchical Dirichlet Process for Timeline Summarization" 2013 Association for Computational Linguistics

[10] James Allan, Rahul Gupta and Vikas Khandelwal. Temporal summaries of new topics. 2001. In Proceedings of the 24th annual international ACM SIGIR conference on R

[11] Hai-Leong Chieu and Yoong-Keok Lee. Query based event extraction along a timeline. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval

[12] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li and Yan Zhang. 2011a. Evolutionary Timeline Summarization: a Balanced Optimization Framework via Iterative Substitution. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.