Discovering event episodes from sequences of online news articles: A time-adjoining frequent itemset-based clustering method

Submitted by:-
Mayur Bhat(181CO132)
Sukruth N Bhat(181CO154)

Review 1-Paper

Why do organizations perform environmental surveillance?

- Organizations need to perform environmental surveillance to identify the important events and their developments so that they can incorporate the experience gained by doing that in decision making,strategy formulation and business action. In this paper focus is online news articles.
- Online news articles have become an integral part of environmental surveillance because of unprecedented growth of the internet.

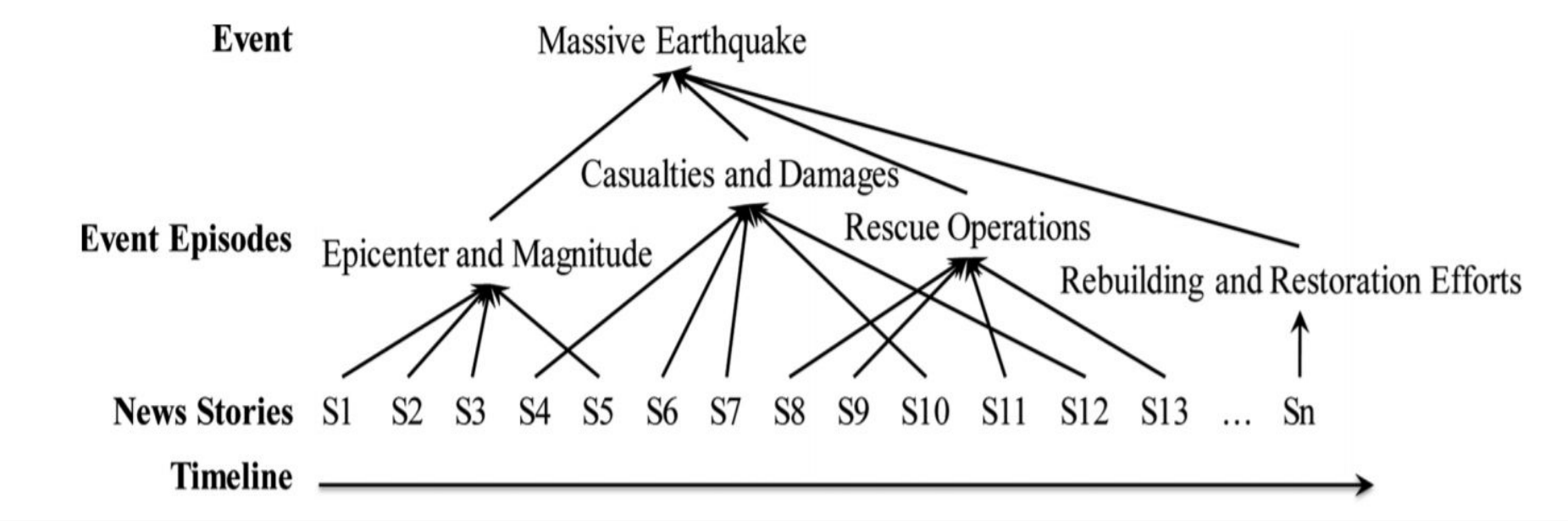
How do the companies do it?

- Companies use a technique known as Event evolution patterns (EEPs) in order to perform their environmental surveillance.
- Companies perform this because EEPs depict the evolution of a particular event type over time which inturn helps them to get prepared for similar kinds of events in the near future.

Structure of an event

There are several different event structures or taxonomies. Some of them are Story→Event→Topic, Story→Simple Event→Complex Event, Story→Component Event→Event, Story→Episode→Event. In this paper, Story→Episode→Event structure has been followed. The next slide explains the concept in detail with an example.

Example explaining an event's structure



Retrospective event detection and Event episode discovery

- A critical precursor to EEP discovery is identifying and grouping articles representing distinct episodes of an event from a sequence of news articles (documents) that pertain to that event.
- Retrospective event detection and event episode discovery are two approaches mentioned in this paper for doing the above process.

Retrospective event detection

- Retrospective event detection techniques generally discover events from a stream of news articles.
- They target an event-based classification by clustering a sequence of chronologically ordered news articles, available in different sources or languages, to identify a set of coherent topics (events) inherent to the news articles.

Event episode discovery

Event episode discovery explicitly aims at discovering an event as it evolves through different development stages and identifying news articles that pertain to each stage.

Which is better?

- Event episode discovery identifies distinct episodes of an event from a sequence of news articles related to that event
- Retrospective event detection identifies events from a stream of articles by segmenting the different events described by these articles.
- As a result, event episode discovery tends to perform analyses at a deeper level than retrospective event detection.

Different techniques used in Event episode discovery

- Two techniques used for event episode discovery have been discussed in this paper. They are Frequent itemset-based hierarchical clustering (FIHC) and Time-adjoining frequent itemset-based event-episode discovery (TAFIED).
- The next few slides explain these two techniques and also gives reason as to which among these two techniques is the best for solving our problem.

FIHC Method

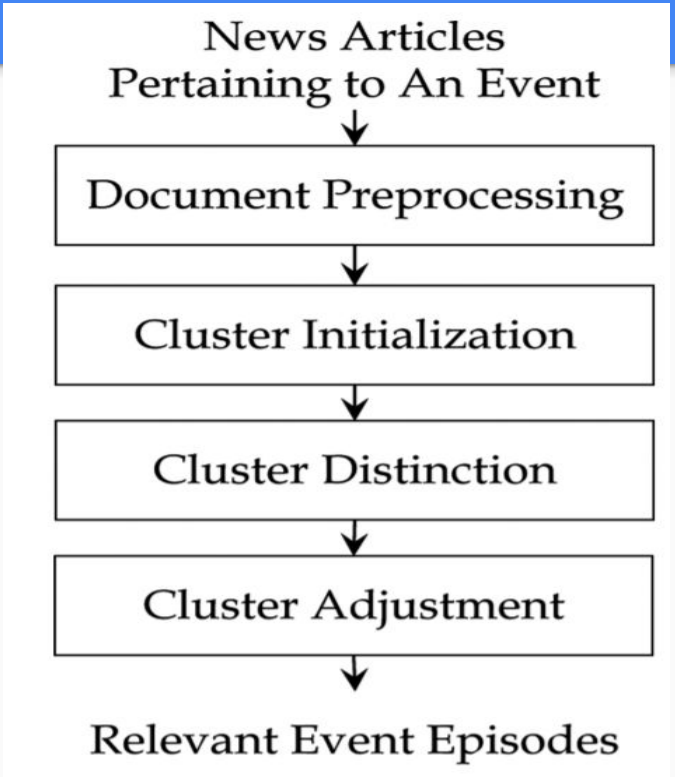
- In FIHC, the documents are called transactions and the features of a document are called items.
- FIHC selects features (items) with a document frequency greater than the prespecified minimum (gf) threshold and uses the identified frequent features (items) as cluster centroid.
- A document dj is initially assigned a set of candidate clusters, on the basis of its own frequent items.

$$Score(c_x \longleftarrow d_j) = [\sum_i (n(t_i) \times Cluster\_Support(t_i))] - [\sum_i (n(t_i') \times Global\_Support(t_i'))]$$

TAFIED Method

- TAFIED creates clusters in which documents are temporally adjacent and share features that frequently appear in a stream of news articles.
- The overall processing of this method consists of document preprocessing, cluster initialization, cluster distinction, and cluster adjustment.

TAFIED Method contd



Proposed TAFIED Model for our problem statement

TAFIED Method contd

- In **document preprocessing**, TAFIED extracts meaningful terms like nouns, noun phrases, and verbs from each news article, applies a rule based part-of-speech tagger to tag each word in the article.
- Stop words such as non semantic-bearing words get removed, and the remaining words are stemmed into their respective original forms.

TAFIED Method contd

- In **cluster initialization**, TAFIED constructs a set of initial clusters and assigns each news article to candidate clusters according to its own frequent items.
- Term ti is a frequent item if the ratio between its document frequency (number of news articles with term ti) and the total number of news articles exceeds the prespecified minimum global support gt.
- By viewing each frequent item as a class label, our method can create a set of initial clusters.

TAFIED Method contd

- After cluster initialization, each news document has been assigned to at least one candidate cluster. Each news article pertains to one and only one event episode so **cluster distinction** has to done.
- During this process, TAFIED assesses the fit between a document and each candidate cluster, selects the most appropriate cluster, and generates a final set of clusters. A fitness function is needed for this purpose.

TAFIED Method contd

$$Fitness(c_x \longleftarrow d_j) = \sum_{i=1}^{|T|} (\alpha \times CS(t_i, c_x) \times TFIDF(t_i, d_j) \times TP(c_x))$$

TAFIED Method contd

- In cluster adjustment, TAFIED merges the clusters that contain highly similar or relevant documents.
- A combined cohesion measure evaluates the appropriateness of merging two clusters.

Which is better suited for our requirement?

- Event episode discovery should properly consider two issues: news articles describing different episodes of a particular event have similar content, and different episodes could emerge concurrently within a time window.
- TAFIED satisfies both the above requirements as it basically extends FHIC by incorporating temporal locality, according to the fit between a cluster and a document.

Review 2-Implementation

Dataset used

- We generated our own dataset from NewsApi.
- The dataset contains articles representing all the major events published in India from 27-03-2021 to 20-04-2021.

Dataset description

- The dataset has 9900 news articles.
- The attributes of the dataset are source, author, title, description ,url , urltoImage, PublieshedAt, description.
- Preprocessing of this dataset has been done before running the algorithm.

Dataset used

	source	author	title	description	url	urlToImage	publishedA	content
0	{'id': 'the-v	Jon Porter	Twitter's v	Twitter ha	https://wv	https://cdi	2021-02-1	For when theres just way too much to typeIllustration by Alex Castro / The VergeTwitter has rolled out support for voice direct messages on iOS and Android in India sta
1	{'id': 'enga	Daniel Coc	Amazon fc	Amazon Pr	https://wv	https://o.a	2021-01-1	Amazon Prime Video and Bharti Airtel, India's second-largest mobile carrier, are teaming up to launch a mobile-only video service. Variety reports that Prime Video Mob
2	{'id': 'techc	Manish Sin	India bans	India has b	http://tecl	https://tec	2020-09-0	India has banned more than 100 additional apps with linkage to China including popular mobile game PUBG citing cybersecurity concerns as geopolitical tension betwee
3	{'id': 'enga	Steve Dent	Samsung b	With the C	https://wv	https://o.a	2020-07-0	With the COVID-19 crisis continuing unabated in India, more folks than ever are relying on their smartphone. At the same time, the pandemic means it's not easy to get
4	{'id': 'enga	Mariella M	Sony is lau	PlayStatio	https://wv	https://o.a	2021-01-0	PlayStation gamers in India will finally have the chance to get their hands on a PS5 within a few weeks' time. The official PlayStation India Twitter account has announce
5	{'id': 'techc	Manish Sin	Facebook	Facebook,	http://tecl	https://tec	2020-07-0	Facebook, which reaches more users than any other international firm in India, has identified a new area of opportunity to further spread its tentacles in the worlds sec
6	{'id': 'techc	Manish Sin	Facebook	As scores (	http://tecl	https://tec	2020-07-0	As scores of startups look to cash in on the content void that ban on TikTok and other Chinese apps has created in India, a big challenger is ready to try its own hand.Inst
7	{'id': 'techc	Manish Sin	Apple begi	Apple's co	http://tecl	https://tec	2020-07-2	Apples contract manufacturing partner Foxconn has started to assemble the current generation of iPhone units — the iPhone 11 lineup — in its plant near southern city
8	{'id': 'techc	Manish Sin	Uber cuts	Uber is cut	http://tecl	https://tec	2020-05-2	Uber is cutting 600 jobs in India, or 25% of its workforce in the country, it said on Tuesday as it looks to cut costs to steer through the coronavirus pandemic.The job cut
9	{'id': 'techc	Manish Sin	Amazon n	Amazon's	http://tecl	https://tec	2020-07-2	Amazons India business said on Thursday it has begun offering auto insurance to cover two and four-wheeler in the country, marking American giants first foray into thi
10	{'id': 'the-v	Adi Robert	India will r	India's legi	https://wv	https://cdi	2021-03-1	One of the strictest crackdowns worldwidePhoto by Michele Doying / The VergeIndia is reportedly moving forward with a sweeping ban on cryptocurrencies. According
11	{'id': 'techc	Manish Sin	PayPal is s	PayPal is s	http://tecl	https://tec	2021-02-0	PayPal is shutting down its domestic business in India, less than four years after the American giant kickstarted local operations in the worlds second largest internet ma
12	{'id': 'techc	Manish Sin	India plans	India plans	http://tecl	https://tec	2021-01-3	India plans to introduce a law to ban private cryptocurrencies such as bitcoin in the country and provide a framework for the creation of an official digital currency duri
13	{'id': 'techc	Manish Sin	Leverage E	Each year,	http://tecl	https://tec	2021-02-1	Each year, millions of students in India rush to get an admission in universities abroad. Often they dont know which program they should focus on, or the college that is
14	{'id': 'techc	Manish Sin	Indian trac	An India tr	http://tecl	https://tec	2021-02-1	An India trader group that represents tens of millions of brick-and-mortar retailers called New Delhi to ban Amazon in the country after a report claimed that the Ameri
15	{'id': 'techc	Manish Sin	Amazon is	Amazon o	http://tecl	https://tec	2021-03-0	Amazon on Tuesday issued a rare apology to users in India for an original political drama series over allegations that a few scenes in the nine-part mini series hurt religic
16	{'id': 'techc	Manish Sin	YouTube a	WhatsApp	http://tecl	https://tec	2021-01-1	WhatsApp has enjoyed unrivaled reach in India for years. By mid-2019, the Facebook-owned app had amassed over 400 million users in the country. Its closest app rival
17	{'id': 'techc	Rita Liao	Xiaomi fur	China's Xia	http://tecl	https://tec	2021-02-2	China's Xiaomi had dominated the Indian smartphone market for three consecutive years until recently losing the top spot to Samsung. It has played by the Indian gover
18	{'id': 'techc	Manish Sin	Top Faceb	Ankhi Das,	http://tecl	https://tec	2020-08-1	Ankhi Das, a top Facebook executive in India, has filed a criminal complaint against a journalist who she alleges attempted to defame her in a public Facebook post and
19	{'id': 'techc	Manish Sin	Indian star	Google, wl	https://tec	https://tec	2020-10-0	Google, which reaches more internet users than any other firm in India and commands 99% of the nations smartphone market, has stumbled upon an odd challenge in t
20	{'id': 'techc	Manish Sin	Uber is hir	Uber said (	http://tecl	https://tec	2020-10-1	Uber said on Thursday it is working to hire 225 engineers in India, strengthening its tech team in the key overseas market months after it eliminated thousands of jobs glo
21	{'id': 'techc	Manish Sin	WhatsApp	WhatsApp	http://tecl	https://tec	2020-11-0	WhatsApp, which began testing its payments service in India with 1 million users in early 2018, can finally start to expand the feature to more users in the world's secon
22	{'id': 'techc	Manish Sin	Reliance's	Reliance's	http://tecl	https://tec	2020-12-0	Reliances Jio Platforms, the largest telecom operator in India, plans to roll out a 5G network in the country in the second half of 2021, top executive Mukesh Ambani an
23	{'id': 'techc	Manish Sin	Indian tele	Vodafone	http://tecl	https://tec	2020-09-0	Vodafone Idea, one of the largest telecom operators in India, has rebranded to 'Vi' as it looks to better leverage the unified venture between British telecom giant Voda
24	{'id': 'techc	Manish Sin	Smartphor	Smartphor	http://tecl	https://tec	2020-10-2	Smartphone shipments reached an all-time high in India in the quarter that ended in September this year as the worlds second largest handset market remained fully op
25	{'id': 'the-v	Sam Byfor	Twitter's v	Twitter ha	https://wv	https://cdi	2020-06-1	"Fleets," Twitter's take on Snapchat/Instagram-style stories, just became available in India. The feature is gradually rolling out around the world; after initially launching

Numerical Analysis

Table 1
Example of Term Frequency of Frequent Items in Each Document.

	t_1	t_2	t_3	t_4	t_5
d_1	3	-	2	-	-
d_2	5	-	2	2	-
d_3	7	-	-	3	4
d_4	1	-	1	4	-
d_5	2	-	-	5	-
d_6	-	2	-	-	2
d_7	-	3	-	-	-
d_8	-	1	16	-	2
d_9	-	1	-	-	1
d_{10}	2	-	12	-	-

Table 2
Example Initial Clusters and Their Respective Member Documents.

Initial Set of Clusters	Member Documents
c_{t1}	$d_1, d_2, d_3, d_4, d_5, d_{10}$
c_{t2}	d_6, d_7, d_8, d_9
c_{t3}	$d_1, d_2, d_4, d_8, d_{10}$
c_{t4}	d_2, d_3, d_4, d_5
c_{t5}	d_3, d_6, d_8, d_9

Formula

$$Fitness(c_x \longleftarrow d_j) = \sum_{i=1}^{|T|} (\alpha \times CS(t_i, c_x) \times TFIDF(t_i, d_j) \times TP(c_x))$$

where T is a set of frequent items,

ti denotes a frequent item,

CS(ti, cx) is the cluster support,

TFIDF(ti, dj) represents the within-document term frequency f,

TP(cx) is a temporal proximity (TP) function,

$$Fitness(c_{t1} \longleftarrow d_3) = ((1 \times \frac{6}{6} \times 7 \times \log_2 \frac{10}{6}) + (1 \times \frac{4}{6} \times 3 \times \log_2 \frac{10}{4}) + (-1 \times \frac{1}{6} \times 4 \times \log_2 \frac{10}{4})) \times \frac{e^{20-29}}{1+e^{20-29}} = 0.001,$$

where $\lambda_{t1} = |6-1| \times 2^2 = 20$ and $\theta_{t1} = |2-1|^2 + |3-2|^2 + |4-3|^2 + |5-4|^2 + |10-5|^2 = 29.$

Performance Measures

- The cluster recall (CR) and cluster precision (CP) of the target event are the two major performance measures.
- CR=|CA| / |TA | and CP = |CA | / |GA |, where TA refers to the set of associations of documents in the true event episodes, GA denotes the set of associations of documents in the event episodes generated by a technique under evaluation, and CA is the set of associations of documents that exists in both the true and generated event episodes.

Results of our implementation(TAFIED)

```
#recall and precision
rec=ca/ta
pre=ca/ga
print(rec)
print(pre)
```

```
0.6666666666666666
0.42106618593870715
```

```
[108] f1=2*pre*rec/(pre+rec)
      print(f1)
```

```
0.5161392155315286
```

Results of our implementation(FIHC)

```
[234] #recall and precision
      rec=ca/ta
      pre=ca/ga
      print(rec)
      print(pre)
```

```
0.6055045871559632
0.285097192224622
```

```
[235] f1=2*pre*rec/(pre+rec)
      print(f1)
```

```
0.3876651982378854
```

Results comparison

	Cluster Recall	Cluster Precision	F-measure
TAFIED	0.667	0.421	0.516
FIHC	0.606	0.285	0.388

Improvements Done

- Due to the unavailability of the dataset given in the paper we created our own dataset using an Api. This enabled the dataset to become much more tailor made for our requirement as we could create it based on the events we specified.
- All the functions used in TAFIED method were implemented in python without using libraries.
- HAC implementation for the same dataset has also been done.

Future Scope

- Using API's, web scraping the TAFIED algorithm can be run in real time making the process of maintaining the episodes of a event less cumbersome.
- The fitness function used in cluster distinction step has 4 more sub functions which are mathematically complex and take a lot of computational time. Thus this method may fail if the dataset is too large.
- A common efficient implementation of the algorithm which can overcome it's current drawbacks.

Work distribution

FIHC code- Mayur Bhat.

TAFIED code-Sukruth N Bhat.

HAC code and presentation- Both of us.