International Journal of Computer Science and
Applications, ©Technomathematics Research Foundation
Vol. 14, No. 2, pp. 107 – 114, 2017

SEXUAL HARASSMENT AND NATURAL CALAMITY: NEWS EVENT SEQUENCE MINING

GAJENDRA WANI

Department of Computer Science, Bhusawal Arts, Science and P.O.Nahata Commerce College, Bhusawal, INDIA

gajuwani03@gmail.com

MANISH JOSHI

School of Computer Sciences, North Maharashtra University, Jalgaon ,INDIA

joshmanish@gmail.com

Newspapers are authentic and important source of news in India. Newspapers report significant events that impact society. Analysis of major news events that occurred throughout India in last five years to obtain genuine and innovative news sequences is the objective of this study. The news events of the same category reported on different dates corresponds to a sequence of events. Mining sequence data can unfold some innovative sequence of the events. The early information of events that can occur in future (after certain duration) in sequence with some already happened events shall prove very significant in many areas. Sequential pattern mining therefore is one of the popular knowledge discovery and pattern extraction techniques. Several researchers have obtained successfully 2-sequences from sequence data sets. The bold title statement of the paper emerges as our observation of 2-sequences obtained in our analysis.

In this paper, we present our approach to discover genuine 2-sequences and 3-sequences among news events. We did not come across any such type of work, where sequences are extracted from events that occur in certain regions of any country. We created a news events dataset by collecting and categorizing large number of news from appropriate and authentic source. We have a collection of more than 3000 news that are divided into 38 categories from the period of 2012 to 2016. We obtained a list of genuine 2-sequnce and 3-sequence events within certain time period. The algorithms and the results are presented in the paper.

Keywords: sequence; sequence pattern mining; time interval; news event; new category

1. Introduction:

FICCI (Federation of Indian Chambers of Commerce and Industry) and Pinkerton Pvt. Ltd. has been conducting "India Risk Survey" since 2012 to capture the risk that affects business sector in India. This survey quantifies 12 prominent risks and one of the major risk factor is workplace violence & sexual harassment. Indiatogether has published an article on 09 June 2017 regarding "Women/Gendered Tragedies" expressing the stories of number of rape cases increase after every big calamity. Mining sequence data can explore new sequence of events. Early event information can predict event occurring in future. Researchers are using sequential pattern mining to obtain sequences from dataset.

Sequence pattern mining algorithms such as [Srikant and Agrawal, (1996)], [Masseglia *et al.* (1998)], [Han *et al.* (2000)], [Pei *et al.*(2001)], [Zaki, (2001)], [Ayres *et al.* (2002)], [Giudici and Passerone, (2002)] discovers subsequences from large database. Mostly, time constraint based sequential pattern mining algorithms [Pei *et al.*

(2003)], [Sun *et al.* (2003)], [Sun *et al.* (2004)], [Jin *et al.* (2008)], are used to discover sequences. Discovering sequential patterns has several applications including customer buying pattern analysis, web log analysis, medical record analysis, stock market data analysis etc. Such patterns help in making predictions. In this paper, we discuss the use of sequential data mining to predict forthcoming news event from the previous one.

Every day we read about events reported in daily news papers. These events are related to natural disaster, earthquake, election, scam, woman welfare, sexual harassment, politics, viral infection, business etc. Our study is to obtain the sequence in which these events occurred with estimation of time interval among them.

A number of news events in a sequence are called as length of a sequence. For example, a sequence<A,B> such that news event A followed by news event B is called as 2-sequence. A sequence <A,B,C> such that news event A followed by news event B and event B follows by an event C. A sequence <A,B,C> contains three different news. A, B and C corresponds to three different events known as 3-Sequences. For example, swine-flew rising after flood-hit is a 2-sequence. Cholera rising after a drought followed by storms is a 3-sequence.

In our proposed approach, we proposed one pass algorithm to discover 2-sequence and 3-sequence of news events with estimation of time interval among them by keeping track of valid sequences. We count the number of sequences generated to determine whether sequence is frequent or not. We also calculate average value of estimated time interval between 2-sequences and 3-sequences of news events. We obtained very interesting sequential pattern among many as Natural calamities news events followed by Sexual harassment. This sequence occurred many times so the title of paper is named accordingly.

This paper is organized as follows: Section-2 described related work. Section-3 presents problem formulation, dataset and algorithms used for experiment, Section-4 present results and observations thereof; Section-5 concludes the paper.

2. Related Work:

[Jin *et al.* (2008)], proposed an algorithm for mining unexpected temporal associations (UTAR) to find expected temporal associations from data directly. The work is related to medical domain. An algorithm MUTARC is developed to find infrequent pair wise UTARS and to generate adverse drug reaction (ADR) signals from healthcare administrative database. [Huang and Huang, (2009)] uses sequential patterns and proposed collaborative recommender system. This system predicts the customer's time-variant purchase behavior in e-commerce environment. To discover sequential patterns time weight is introduced. [Radinsky and Horvitz, (2013)] describe methods to forecast forthcoming events from a corpus containing 22 years of news stories. System predict when and where disease outbreaks, depths and riots in advance of the occurrence of these events in the world. In this paper, the automated extraction and sequences of events is generated from news corpora and multiple web resources. [Lee *et al.* (2009)], proposed a new temporal data mining technique to extract temporal interval relational rules from temporal interval data by using Allen's theory. [Joshi *et al.* (2009)], proposed the one pass algorithm to estimate time period between sequential events. This generated a list of sensible pattern without using pre-specified time window.

3. Experimental Details:

In this section, we present all experimental details in 3 subsections namely problem formulation, algorithm and dataset.

3.1 Problem Formulation:

Let E = {e_1, e_2, e_3 ……., e_n} be a set of news event categories. Let S = {n_1, n_2, n_3, ……., n_p} be a set of news events. In set S, each n_i is con stituted with four fields namely n_i.category, n_i.time and n_i.description, where category Є E, time represent date on which event occur and description contains details of event. For experiment, we considered two fields namely category and time as shown in Fig (1). Our objective is to estimate an intermediate time interval for a 2-sequence and 3-sequence. An intermediate time interval is the time elapsed between successive occurrence of news events of the 2-sequence and 3-sequence. Fig (1) shows news events occurring in sequence order using timeline as day.

Time	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
News Categories	e_1	e_5	e_3	e_8	e_5	e_1	e_1	e_2	e_7	e_4
	e_3	e_6	e_7	e_9	e_{10}	e_2	e_6	e_9	e_8	e_8
	e_4	e_{14}	e_{11}	e_{19}	e_{16}	e_{13}	e_{16}	e_{10}	e_{13}	e_{10}
	e_{19}	e_{21}		e_{21}	e_{18}			e_{17}	e_{17}	e_{15}

Fig (1) Example of news events sequence database.

In Fig (1), sample of dataset is given for 10 days. On date1 (d_1), events e_1, e_3, e_4 and e_{19} occur. Similarly on date2 (d_2), events e_5, e_6, e_{14} and e_{21} occur and so on. For example, in this dataset, a 2-sequence $e_7 \rightarrow e_{10}$ is considered as valid because e_7 occurs on d_3 and e_{10} occurs on d_5. Time interval among e_7 and e_{10} is calculated as (d_5-d_3). Also e_7 and e_{10} occurs on d_9 and d_{10} respectively. A time interval is calculated as (d_{10}-d_9). The sequence $e_7 \rightarrow e_{10}$ occurs 2 times in a dataset. A 3-sequence $e_6 \rightarrow e_9 \rightarrow e_{13}$ is a valid sequence and appears 2 times in a dataset. In the first occurrence, time interval among $e_6 \rightarrow e_9$ is (d_4-d_2) and $e_9 \rightarrow e_{13}$ is ($d_6 - d_4$). In the second occurrence, time interval among $e_6 \rightarrow e_9$ is (d_8-d_7) and $e_9 \rightarrow e_{13}$ is ($d_9 - d_8$). The sequence $e_1 \rightarrow e_5 \rightarrow e_7$ is a valid and it appears only 1 time in dataset i.e. on d_1, d_2 and d_3. However this sequence is not valid for d_5, d_7 and d_9 because event e_5 occurs on d_5 and $e_1$on d_7 and ($d_5 - d_7$) < 0.

Using our proposed algorithms, we found the temporal relationship between all news categories. We obtained 2-sequences and 3-sequences of news categories with estimation of time intervals among them. In Fig (1), d_1 represent date1, d_2 represents date2 and so on. Similarly, we assign code for each news categories using java program. Table-1 represents news categories and its code used in our research.

Table-1: List of News categories

Code	News Category	Number of News	Code	News Category	Number of News
e_1	Jammu & Kashmir	50	e_{20}	Business-corporate-activity	33
e_2	Business-Government-achievement	53	e_{21}	Scam-bribery	88
e_3	Civil-disorder	87	e_{22}	Business-foreign-activity	04
e_4	Business-career-objective	10	e_{23}	Business-development	39
e_5	Business-Government-Decision	119	e_{24}	Scam-in- Government-grant	161
e_6	Terrorism&Naxal	33	e_{25}	Business-International-affair	51
e_7	Business-sensex-Nifty	170	e_{26}	Epidemic	20
e_8	Business-corporate-investment	70	e_{27}	Fire	22
e_9	Business-Market-update	105	e_{28}	Business-taxation	16
e_{10}	Business-corporate-achievement	50	e_{29}	Election	104
e_{11}	Sexual-harassment	165	e_{30}	Internal-security	108
e_{12}	Natural-Calamities	143	e_{31}	Business-corporate-affair	11
e_{13}	Government-activity	276	e_{32}	Scam-buy-and-sale	16
e_{14}	Business-corporate-decision	39	e_{33}	Business-commodity-market	04
e_{15}	Business-foreign-investment	30	e_{34}	Business-Government-Policy	14
e_{16}	Business-Government-Inverstment	30	e_{35}	Women-Welfare	40
e_{17}	International-terrorism	06	e_{36}	Religion	13
e_{18}	Military-teritary	09	e_{37}	Business-merger&aqusition	05
e_{19}	Business-Government-activity	266	e_{38}	Cow-Vigilant	15

We measured count of valid sequences and found out the ‘average of time intervals’ among 2 and 3 sequences. These statistical measures help to find out the prediction of forthcoming events within appropriate time.

3.2 Dataset:

For experiment, we have created a database of news retrieved from website http://www.mapsofindia.com. We recorded everyday news from this site and maintained a database from the period of 12 September 2012 to 21 August 2016. Around 3000 records are maintained in the database. Database is organized with fields namely – category, date of publication (time) and news description. Out of these fields, we only considered two main fields, category and date of publication. We categorized news category field using news description. Table-2 shows snapshot of original dataset.

Table-2: Snapshot of original dataset

News Category	Date of News	News Description
Women-Welfare	26/8/2016	Bombay HC Lifts Ban of Women Inside the Haji Ali Dargah
Sexual-harassment	25/8/2016	Dec 16 Delhi Gangrape Convict Attempts Suicide in Jail
Natural-Calamities	24/8/2016	Eastern Indian States Felt Tremors of Myanmar Earthquake
Natural-Calamities	23/8/2016	Earthquake Hits India-Myanmar Border Region
Natural-Calamities	23/8/2016	Centre Assures Help to Flood-Hit Areas
Bus-corporate-achievement	29/7/2015	Net Profit of Maruti Suzuki during April-June Quarter increases by 56 Per Cent
Business-Government-achievement	28/7/2015	Government to Raise Rs. 1,600 crore from 5 Per Cent Stake Sale of Power Finance Corporation
Business-corporate-achievement	29/7/2015	Net Profit of Maruti Suzuki during April-June Quarter increases by 56 Per Cent
Scam-bribery	09/06/2015	Cash for Vote Controversy Escalates as Arrested Telangana TDP MLA Revanth Reddy's House Raided by ACB
Scam-in-Government-grant	11/11/2014	Coal Scam: CBI Changes Finds many Evidences against the Accused
Natural-Calamities	27/10/2014	After Hudhud, it's Nilofer Cyclone
Government-activity	23/9/2014	Grain Supply increased by 40 per cent under Food Security Law
Epidemic	02/09/2014	6 Ebola suspect cases reported
Civil-disorder	25/8/2014	10 die in stampede in Madhya Pradesh
Jammu & Kashmir	24/3/2013	Kashmir militant attack leads to death of one civilian
Sexual-harassment	16/3/2013	Swiss woman gangraped by 8 persons in Madhya Pradesh
Fire	27/2/2013	Major fire broke at Sealdah area in Kolkata
Business-International-affair	15/10/2012	India, Russia to resolve liability for Kudankulam III and IV
Business-corporate-investment	24/8/2012	Maruti to invest Rs.1,700 crore in Manesar's new plant: Haryana minister

3.3 Algorithms:

In this section, we develop two algorithms for discovers temporal relationship between news events. Algorithm-1 discovers 2-sequence of news events and Algorithm-2 discovers 3-sequence of news events. We also estimate time interval among them and find average value of intermediate time intervals to predict appropriate time of fourth coming event. Both algorithms are based on algorithm [Joshi *et al.* (2009)]. The dataset S is preprocessed to sort on news category and date field. At first stage, compute support of each news category in a dataset S to obtain frequent 1-sequence dataset S1 which satisfy minimum support threshold value. (Using GSP algorithm).

Algorithm 1. Mining 2-sequences of news events.
Input : Frequent 1-sequence dataset S1 and S.
Output : Valid 2-sequence of news events.
Step 1. Scan dataset S1 to process various combinations of two news category (e_Xe_Y).
// Where e_X is called first news event and e_Y is the second news event of sequence (e_Xe_Y).
Validate 2-sequence (e_Xe_Y) from dataset S using Step 2.
Step 2. For each combination of 2 events (e_X, e_Y)
(i) For each event in a 2-sequence, initialize file pointer p1 and p2 :
p1 ← first occurrence of e_X Є S.
p2 ← first occurrence of e_Y Є S.
count ← 0
(ii) To determine whether <e_Xe_Y> sequence is acceptable or not.
For acceptable sequence:
If ((p1.newscategory = e_X) .and. (p2.newscategory = e_Y) .and. (p1.time≤ p2.time))
i) Estimate T as
T = p2.time – p1.time
ii) Increment count for sequence.
iii) Write sequence (e_X→ e_Y) with time interval T to result file.
EndIf
Advance p1, p2 pointers to next category in respective event list.
(iii) Calculate average of time intervals among each 2-sequence.

Algorithm 2. Mining 3-sequences of news events.
Input : Frequent 1-sequence dataset S1 and S.
Output : Valid 3-sequence of news events.
Step 1. Scan dataset S1 to process various combinations of three news category ($e_Xe_Ye_Z$).
// Where e_X is called first news event, e_Y is the second news event and e_Z is the third news event of sequence ($e_Xe_Ye_Z$).
Validate 3-sequence ($e_Xe_Ye_Z$) from dataset S using Step 2.
Step 2. For each combination of 3 events (e_X, e_Y, e_Z)
(i) For each event in a 3-sequence, initialize file pointer p1, p2 and p3 :
p1 ← first occurrence of e_X Є S.

p2 ← first occurrence of e_Y Є S.

p3 ← first occurrence of e_Z Є S.

count ← 0

(ii) To determine whether <$e_X e_Y e_Z$> sequence is acceptable or not.

For acceptable sequence:

If (($p1.newscategory = e_X$) .and. ($p2.newscategory = e_Y$) .

($p3.newscategory = e_Z$) .and. ($p1.time \leq p2.time \leq p3.time$))

i) Estimate T1 and T2 as

T1 = p2.time – p1.time

T2 = p3.time – p2.time

ii) Increment count for sequence

iii) Write sequence ($e_X \rightarrow e_Y \rightarrow e_Z$) with time interval T1and T2 to result file.

EndIf

Advance p1, p2, p3 pointers to next category in respective event list.

(iii) Calculate average of time intervals among each 3-sequence.

4. Results and Observations:

In this section, we present results of our approach. All implementation are done through java programming. Table-3 shows the example of 2-sequence obtained through experiment.

Table-3: Sample of 2-sequence

Association between 2 news categories : Sexual harassment (e_{11}) →Natural Calamities (e_{12})					
Date of e_{11}	Date of e_{12}	Date Difference	Date of e_{11}	Date of e_{12}	Date Difference
13/07/2012	15/07/2012	2	29/07/2014	01/08/2014	3
23/07/2012	20/08/2012	28	04/08/2014	04/08/2014	0
20/10/2012	27/10/2012	7	05/08/2014	05/08/2014	0
02/01/2013	02/01/2013	0	19/08/2014	22/08/2014	3
05/02/2013	17/02/2013	12	05/09/2014	05/09/2014	0
11/03/2013	11/3/2013	0	08/09/2014	08/09/2014	0
28/03/2013	30/03/2013	2	09/09/2014	09/09/2014	0
24/04/2013	01/05/2013	7	12/09/2014	14/09/2014	2
17/05/2013	19/05/2013	2	22/09/2014	24/09/2014	2
11/06/2013	16/06/2013	5	25/09/2014	16/10/2014	21
27/06/2013	27/06/2013	0	31/01/2015	31/01/2015	0
12/07/2013	20/07/2013	8	24/03/2015	30/03/2015	6
01/09/2013	01/09/2013	0	01/04/2015	01/04/2015	0
23/09/2013	30/09/2013	7	24/04/2015	25/04/2015	1
11/10/2013	12/10/2013	1	08/05/2015	12/05/2015	4
26/10/2013	26/10/2013	0	18/05/2015	27/05/2015	9
21/11/2013	21/11/2013	0	24/06/2015	25/06/2015	1
22/11/2013	22/11/2013	0	24/07/2015	03/08/2015	10
26/11/2013	26/11/2013	0	10/8/2015	11/08/2015	1
04/03/2014	11/03/2014	7	04/09/2015	15/09/2015	11
20/03/2014	20/03/2014	0	23/10/2015	26/10/2015	3
15/04/2014	18/04/2014	3	20/12/2015	26/12/2015	6
03/05/2014	13/05/2014	10	29/12/2015	04/01/2016	6
31/05/2014	18/06/2014	18	01/03/2016	04/05/2016	64
01/07/2014	16/07/2014	15	22/06/2016	22/06/2016	0
23/07/2014	23/07/2014	0			

In table-3 sequence e_{11} (Sexual harassment) → e_{12} (Natural calamity) occurs 51 times and average value of time is 5.6. Some more examples of 2-sequence we have obtained as shown in table-4.

Table-4: Samples of 2-sequence with the count and average of time intervals among them

Newscategory1	Newscategory2	Count	Average value of time intervals
Civil-disorder(e_3)	Internal-security(e_{30})	29	7.6
Election(e_{29})	Scam-in- Government-grant (e_{24})	31	6.6
Sexual-harassment(e_{11})	Election(e_{29})	30	8.5
Sexual-harassment(e_{11})	Civil-disorder(e_3)	42	9.2
Scam-bribery(e_{21})	Business-Government-Decision(e_5)	22	8.6
Epidemic(e_{26})	Natural-Calamities(e_{12})	10	10.6

We also found 3-sequences of the news events. Table-5 shows samples of 3-sequences.

Table-5: Samples of 3-sequence with the count and average of time intervals among them.

Newscategory1	Newscategory2	Newscategory3	Count	Average value of time interval among newscategory1 and newscategory2	Average value of time interval among newscategory2 and newscategory3
Civil-disorder(e_3)	Sexual-harassment(e_{11})	Natural-Calamities(e_{12})	18	14.6	2.0
Scam-in-Government-grant (e_{24})	Business-Government-activity(e_{19})	Government-activity(e_{13})	40	4.2	2.2
Election(e_{29})	Civil-disorder(e_3)	Government-activity(e_{13})	10	12	1.1
Epidemic(e_{26})	Natural-Calamities(e_{12})	Sexual-harassment(e_{11})	8	20.5	2.3

5. Conclusions:

This section contains two subsections. In section 5.1, we state findings and usage and in section 5.2, we state future research work.

5.1 Findings and Usage:

In this paper, we discuss new ideas of discovering 2-sequences and 3-sequences from news database. The sequence mining of news events extracted an innovative 2-sequences like (Sexual-harassment) → (Natural Calamity), (Civil-disorder) → (Internal security) and 3-sequences like (Civil-disorder) → (Sexual harassment) → (Natural calamity), (Election) → (Civil-disorder) → (Government-activity).

We obtained 2-sequences and 3-sequences of news events with time intervals among them. This is useful to predict forthcoming events with appropriate time. These results help government-aid organizations or other social organizations to be ready in advance for tackling any future mishap. At the same time, it will be useful to the survey agencies like FICCI and Pinkerton and others to broaden their area.

5.2 Future Research work:

We had characterized news categories manually which can be automated in future. Categories can further be subdivided to obtain results at detailed level. We can also calculate standard deviation to measure spread out time intervals. Using the standard deviation we have a way of knowing what are normal, what are extra large or extra small time intervals.

References:

Ayres, J., Gehrke, J. E., Yiu, T. et al. (2002). Sequential Pattern Mining using Bitmap. International Conference on Knowledge Discovery and Data Mining.

Cheng-Lung Huang, Wei-Liang Huang. (2009). Handling sequential pattern decay: Developing a two-stage collaborative recommender system. Electronic Commerce Research and Applications (ELSEVIER) , vol.8 issue 3, pp. 117-129.

Giudici, P., & Passerone, G. (2002). Data Mining of Association Structures to Model Consumer Behaviour. Computational Statistics & Data Analysis, 38, (pp. 533-541).

Han, J., Pei, J., Mortazavi-asl, B. et al. (2000). FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining . Sixth International Conference on Knowledge Discovery and Data Mining, (pp. 355-359).

Jin, H., Chen, J., He, H. et al. (2008). Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions. . Information Technology in Biomedicine,12, (pp. 488-500).

Kira Radinsky, Eric Horvitz. (2013). Mining the web to predict future events . sixth ACM international conference on Web search and data mining(WSDM'13) , (pp. 255-264).

Manish Joshi, Pawan Lingras, Virendra Bhavsar. (2009). An algorithm for the Estimation of a Time Period of 2-Sequences . 4th Indian International Conference on Artificial Intelligence (IICAI-09), Tumkur, India, (pp. 71-88).

Masseglia, F., Cathala, F., Poncelet, P. (1998). The PSP Approach for Mining Sequential Patterns . Second European Symposium on Principles of Data Mining and Knowledge Discovery, LNAI, 1510 , (pp. 176-184.).

Paliwal, R., Chakraborty, S., Vig, N., Gupta, S., Sarita, R., & Rahman, O. (2015). India Risk Survey 2015. New Delhi: FICCI and Pinkerton.

Pei, J., Han, J., Mortazavi-Asl, B. et al. (2001). PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Growth. International Conference on Data Engineering, (pp. 215-224).

Pei, J., Han, J., Wang, W.,. (2002). Constraint-Based Sequential Pattern Mining in Large Databases. International Conference on Information and Knowledge Management, (pp. 18-25).

Srikant, R., & Agrawal, R. (1996). Mining Sequential Patterns: Generalization and Performance Improvement . International Conference on Information and Knowledge Management, (pp. 3-17).

Sun, X., Orlowska, M. E., Li, X. (2004). Finding Negative Event-Oriented Patterns in Long Temporal Sequences. Lecture Notes on AI, 3056, (pp. 212-221).

Sun, X., Orlowska, M. E., Zhou, X. (2003). Finding Event-Oriented Patterns in Long Temporal Sequences. Lecture Notes on AI, 2637, (pp. 12-26).

Yong Joon Lee, Jun Wook Lee, Duck Jin Chai, Bu Hyun Hwang, Keun Ho Ryu . (2009). Mining temporal interval relational rules from temporal data. JSS (The Journal of Systems and Software).

Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. in Machine Learning Journal, 42,, (pp. 31-60).

Web Site http://www.mapsofindia.com

Web Site http://www.indiatogether.org