Sindh Univ. Res. Jour. (Sci. Ser.) Vol.43 (1-A)103-108 (2011)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Contextual Event Information Extractor for Emails

S. Wasi, Z. A. Shaikh and J. Shamsi

FAST-National University of Computer and Emerging Sciences, Karachi, Pakistan
zubair.shaikh@nu.edu.pk , zubair.shaikh@nu.edu.pk

Corresponding author: S. Wasi, email: shauakat.wasi@nu.edu.pk,
Revised

Abstract: A powerful tool for planning and announcement of Events is Email. Automatic detection of the Occurrence (Title) and its contextual information (Location, Temporal information, Participants) associated with the email is surely desirable to help the users manage and plan important Events. A lot of work has been done in the area of Event detection but it has various limitations from different perspectives. Firstly, the existing work mainly targets text streams like news stories, scientific documents, articles etc that are somewhat structured documents with sufficient event description as compare to the Emails that have structured, semi-structured and unstructured short descriptions with a plenty of description styles. Secondly the objective in most of the research is to detect new or hot events. Thirdly, much of the existing work aims on reporting events and our objective is to support Event Planning and Management. Another lacking thing is the use of publication time as the temporal information instead of actual temporal information contained within text that is indeed required for Event planning and management task. We have used Finite State Automata (FSA) to extract phrases revealing the Places, temporal information and the actual occurrence. The results are evaluated using different measures. Experiments show that the proposed approach performed well on the Email data Corpus.

Keywords: Events email, finite state automata (FSA).

INTRODUCTION

Emails are commonly used for broadcasting upcoming events to target groups and planning events among interested ones. One receives a lot of emails in his inbox on daily basis and it's difficult to manage all mails with due attention. Event emails are one of the most important ones as no one wants to miss an important event. But it happens many a times that very important events are missed not because the email was neglected but because of the lack of attention. Hence it is surely desirable that the actual event with its contextual information is automatically extracted and placed in one's personal or shared calendar as it helps significantly in planning and management of events.

The event detection task was part of the TDT research initiated in 1996 by DARPA, the University of Massachusetts, Carnegie Mellon, and Dragon Systems. The goal was to detect new events online (Papka, *et al.*, 1998) and to identify the news stories containing a specific event (Pierce, *et al.*,1998). The domain of TDT research was broadcast news stories in multiple languages. People have proposed various approaches for event detection in text streams but most of the research has mainly focused general text streams (news stories, scientific documents, official or corporate documents and articles etc). A few have worked on social text streams (Emails, blogs etc). Emails usually have short descriptions (may be one or two sentences or simply Labeled Points) and have a rich set of styles for event description. Due to these distinguishing features the existing approaches do not perform well in detecting event info in emails.

A lacking feature of the most of the existing work is that they have used the date of publication as temporal information instead of employing the actual temporal information available within the event description. In our case, the actual temporal information is vital for our objective of Planning and managing Events.

We have developed an automatic event info extractor. Our application may be attached to mail client as a plug in. We have used a simple technique to cater with our needs and goals. FSA are used for extracting the event information from Email

descriptions. For this work we consider the sender and receiver(s) of the email as participants. We evaluated our approach using miss rate, false alarm rate, precision, recall and F-Score measures.

MATERIAL AND METHOD

Literature Review

The TDT research work (Carbonell, *et al.,* 1998) included the Event detection and tracking task as part of the Topic detection and tracking task. After on a lot of research took place in this area with different objectives like new event detection (Papka *et al.,* 1998; Carbonell *et al.,* 1999; Fukumoto and Suzuki 2000; Makkonen, *et al.,*2004; Kumaran and Allan 2004; Zi *et al.,* 2007; Liao *et al.,* 2008; Li 2009), retrospective event detection (Pierce *et al.,*1998; Lei, and Liao 2008) and hot event detection (He, and Qu, 2006; Luesukprasert, *et al.,* 2007; Kotsakos *et al.,* 2008; Chen *et al.,* 2009; Bai, and Guo *et al.,* 2010). Most of these works targeted news stories or documents.

Event detection in social text streams like blogs, emails etc has also been addressed in recent years. X. Wan *et al.,* (Milios *et al.,* 2009), (Zhao and Mitra *et al.,* 2007) have taken an event definition different from ours. They have defined Event as a set of conversations on a specific topic spanning over a specified time period. Naive Bayes Classifier operating on words tagged with NEs was used to extract specific information from general email announcements (Pekar 2005). They fragmented the document at three levels; sections, sentences and lines and found that the evaluation results were best for lines method, yet it was also the hardest one. PET (Zhao *et al.,* 2010) proposed a statistical model for identifying the popular or hot event in a social community. They performed experiments on DBLP and Twitter. Similarly a framework based on relevance models for tracking news events discussed in Weblogs was proposed in (Mejova *et al.,* 2009). The framework also tracks the event intensities changing over time. A keyword based approach is also proposed for detecting and tracking new events (Hurst *et al.,*2009). A key word graph is constructed from the specific set of articles and a community detection method is used for identifying the events. Event identification from the data shared at social media sites is addressed in (Naaman *et al.,*2009; Naaman *et al.,* 2010) incorporating the contextual information available in form of tags and labels etc.

All of the above cited work except (Pekar, 2005) has used Time of publication as the temporal information and none have extracted the actual temporal information available within the text. Mayeng, 2004) highlighted and proved the significance of the actual temporal information in detection of an Event. They used Finite state Automata (FSA) to extract temporal information form Korean News Articles. The simple FSA method was also used by an information extraction system (Kim, *et al.,* 2008) developed for mobile devices to extract temporal information from Korean text messages. The system employed modified HMM based on syllable n-grams to extract information like topics and locations. FASTUS system (Hobbs, and Appelt, *et al.,* 1997) used FSA for extracting variety of information along with the temporal information as a part of the Message Understanding Conference (MUC) activities.

We have employed simple FSA method to extract event and its contextual information from emails and we have extracted the actual temporal information present within the email text.

The rest of the paper is organized as follows. Section 4 explains in detail our approach to event detection in emails. Section 5 provides the results of evaluation and finally Section 6 has the Conclusion and future directions.

Detection of Event and the Contextual information

In order to identify the place, time and the actual occurrence of an event, we have used a simple NLP approach. As a preprocessing phase, the signatures, the initials and all the quoted text is removed from the email text. The resulting email text is first tokenized and the tokens are tagged with Part of speech (POS) tags. We have defined the tag set for each information component separately for the convenience of understanding. Each tag set is an extended one as we have introduced special tags for our purpose. We have constructed a FSA set for each component. Each FSA determines a distinct phrase revealing place, time or occurrence. The FSA are constructed keeping in mind the different styles of Event expressions. Event expressions usually vary with the event categories (Official, Personal, Social etc). Another cause of variation is the type of email, whether it's for announcement or planning. We have tried to cover most of the expressions. As a training phase the FSA were operated manually and improvements were made where ever required. By using FSA for detecting the Event information, we get some time efficiency.

It should be noted that we have used POS tags as labels for the transitions. It means that actual phrases or words are not provided to FSA.

Table I: Extended POS tag set for Places

Tag	Category	Example Phrases
NN	Noun	School, Park
SYM	Symbols	-, : , /
NNA	Abstract noun	Meeting, Seminar, Concert….
NNP	proper noun	FAST-NU, …
PA	adverbial particle	at , in, on
AT	Articles	a,an,the
KWP	Keyword for Place	Venue, Location, Place….

Detection of Places

For the identification of Place(s) where the event occurs, we have constructed six FSA and an extended POS tag set. **(Table I)** shows the extended POS tag set for Place revealing phrases with the meaning of the tags and sample phrases. Some special tags are introduced to cater distinct place revealing phrases. For example the "KWP" tag is introduced for phrases like "Venue: CRUC Room". Some tags are extended to distinguish them for Place and time. A 't' included in a tag distinguishes the tag for time. Similarly to differentiate between date point and duration, 'd' is added to a tag for duration. The FSA for identifying the place are shown in **(Fig. 1)**. Our FSA are also able to discover the number of places where the event occurs. We consider an event taking place at two different locations as two different events.

Detection of Temporal information

We have used the same approach as employed by Pyung and Sung (Kim 2003) but we have modified and extended the things for our purpose. Firstly we have captured the exact time of an event with the date(s). Secondly we have extended the Tag set, FSA and the rules of canonicalization based on our data set. The tag set that helps in detecting phrases containing temporal information is shown in **(Table 2)**. Some POS tags are extended to have a't' or'd' in the suffix where't' indicates time and 'd' indicates date/day. To get the actual time value of an event, the time revealing phrases must be converted to some canonicalized form (King and Mayeng 2004). We have constructed three sets of FSA for the detection of temporal information, one for some specific point in time with respect to Date, Second for duration (days) and third for detecting the time (Specific time or duration).

Table 2: Extended POS tag set for Date/Time

Tag	Category	Example Phrases
NUM	Numeral	1994, 11, …, 3:00,..
SYM	Symbols	-, : , /
NNBU	Bound Noun for Units	Today, Evening, Year, Week, Hour ….
NNPD	demonstrative pronoun	This
NNP	proper noun	Labor Day, …
NNPB	Bound Proper Noun	Monday, January,…
XSNN	suffix	end of , beginning of…
QT	quantifier	one, two, first, …
PA	adverbial particle	at , in, on
NNAD	Adnoun	Next, last
PX	Auxiliary	From, to, for
DTT	Determiner for time	AM, PM
KWT	Time Keyword	Time, Date,….

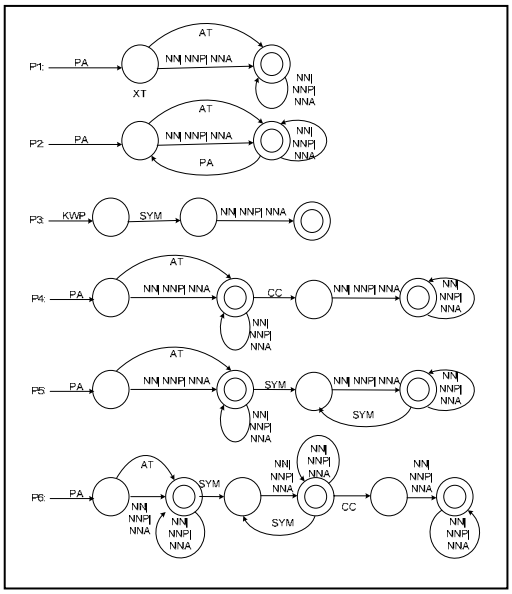

Fig. 1. Place revealing FSA

Detection of the Actual Occurrence

We have tried to capture the actual occurrence by determining the verb phrases that are strong candidates for describing the occurrence. Since Emails are used for just announcements or planning of events therefore these do not have discussion about the events. Generally the initial lines of the email contain the information about the actual occurrence either in form of title or description. We have constructed FSA for accepting phrases that are most likely to be used to describe the occurrence in an email. We have tried to cater with both types of expressions, the active and passive ones. When a phrase is recognized by some FSA as an occurrence revealing one, the whole

expression representing event is captured. **(Table 3)** contains the tag set for detecting actual occurrence.

Detection of the Actual Occurrence

We have tried to capture the actual occurrence by determining the verb phrases that are strong candidates for describing the occurrence. Since Emails are used for just announcements or planning of events therefore these do not have discussion about the events. Generally the initial lines of the email contain the information about the actual occurrence either in form of title or description. We have constructed FSA for accepting phrases that are most likely to be used to describe the occurrence in an email. We have tried to cater with both types of expressions, the active and passive ones. When a phrase is recognized by some FSA as an occurrence revealing one, the whole expression representing the event is captured. **(Table 4)** contains the tag set for detecting actual occurrence. The six FSA shown in **(Fig. 2 and 3)** are constructed to determine the actual occurrence. It must be noted that since our area is information extraction and not NLP, therefore we do not claim to address each and every type of occurrence revealing phrases.

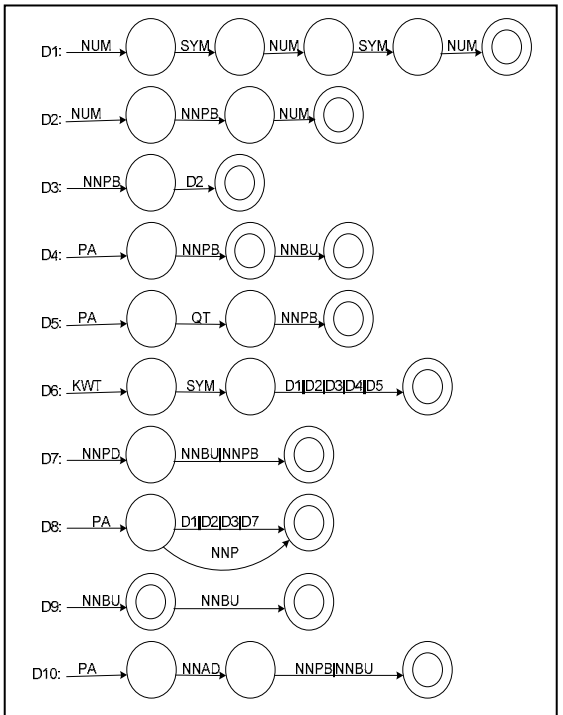

Fig. 2. (a) FSA for detecting Date/Day

Table 3: Extended POS tag set used for detecting Occurrence (What)

Tag	Category	Example Phrases
VB	Verb	Meet, visit,…
NN	Noun	Meeting, Seminar
NNP	proper noun	Web Seminar…
VBX	Auxiliary Verbs	Is, are, have…
VBPP	Verb, Past Participle	Has Invited, have scheduled,…
XT	Auxiliary Tokens	Invites all of you to…
VBFN	Finite Verbs	To attend, to play,…

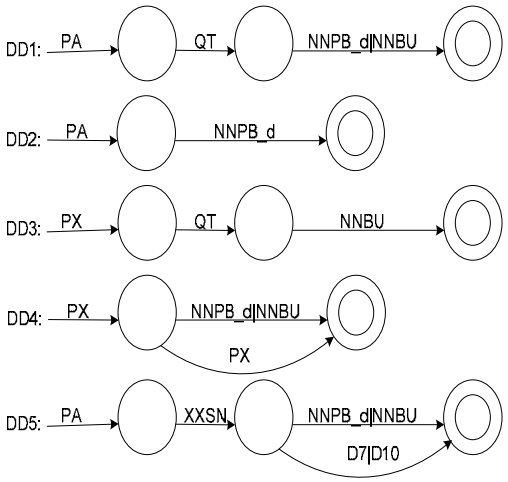

Fig. 2. (b) FSA for detecting Duration (Date/Day)

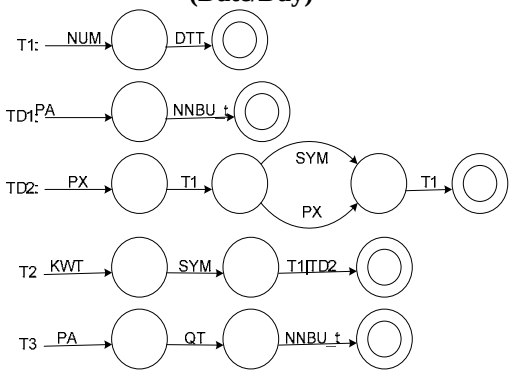

Fig. 2. (c) FSA for detecting Time point/Duration

RESULTS AND DISCUSSION

1000 emails from our inbox were selected as a data corpus. 230 of the emails were non-event emails and the rest 770 were related to different events like Conferences, Seminars, Lectures, Personal Meetings, Weddings, and Sport Events etc. We used Precision, Recall and FSCORE for evaluating our work. More

over the accuracy of Canonicalization process is also evaluated. The answers for What and Where were evaluated using R-measure and P-measure because exact string match is obviously not possible in all cases. R-measure is the fraction of words in the gold standard that are present in the extracted value and P-measure is defined as the fraction of words in the extracted value that are present in the gold standard. The threshold values used for location were (0.8, 0.8) and for actual occurrence were (0.5, 0.5). A value is marked correct if both R and P threshold values for the value are met. Results are shown in **(Table 4, 5 and 6)**. Table 6 shows that social events are the more difficult to identify. High precision in almost all cells proves that the rate of wrong detection is very low which is surely desirable in our system. A relatively low recall for occurrence means that the identification of the event title was not properly handled by FSA or we need some other technique to detect the actual occurrence. We have proposed another approach in the concluding section.

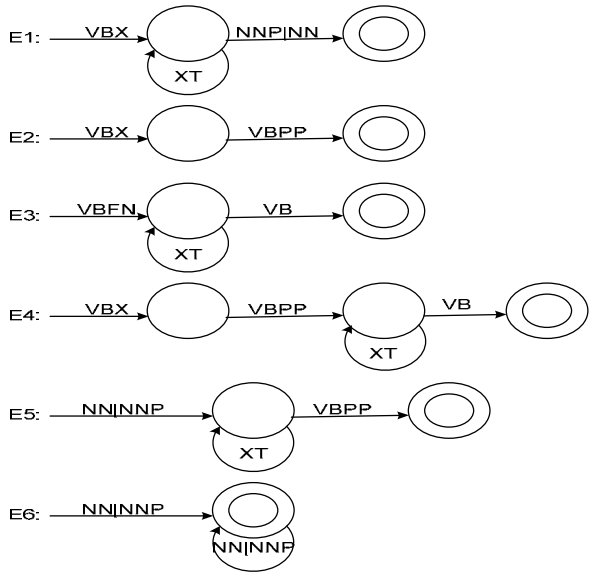

Fig. 3. FSA for detecting the actual occurrence

CONCLUSION

Automatic detection of Event and its contextual information significantly supports in Event Planning and Management tasks. It relieves the burden of manual identification of event emails and manually extracting the event info from emails to record for reminders. Results show that the system proposed in this paper performs the task well.

We aim to further improve the system with different features. Firstly the identification of actual participants is certainly desirable. Similarly, extracted values for venue and occurrence must be more accurate. Techniques like Hidden Markov Models or n-grams may help to achieve the above targets. More over there is a need to cater with emails containing more than one event and emails containing more than one value for any of the ubiquitous information component. We will improve the system with these extensions to make it much appropriate for the targeted objectives.

This is the extended version of our own paper presented and published as Conference proceedings in "International Conference on Computers & Emerging Technologies" (ICCET 2011) held on 22-23 April 2011 at Shah Abdul Latif University, Khairpur, Sindh, Pakistan. We pay thanks to Centre for Research in Ubiquitous Computing (CRUC) team at FAST-NU, Karachi who really supported us in sense of providing a research environment.

Table 4: Recall, Precision, Miss rate, F/A rate and F1 for individual information components

Information Component	Precision	Recall	F1
Actual Occurrence	0.96	0.76	0.84
Location	0.97	0.96	0.96
Date	0.99	0.97	0.98
Time	0.99	0.98	0.98

Table 5: Recall, Precision, Miss Rate, F/A Rate and F1 for Event and Non-Event Emails

Number of Emails	Precision	Recall	F1
1000 (770 Event, 230 Non-Event)	0.91	0.83	0.86

Table 6: Recall, Precision, Miss Rate, F/A Rate and F1 for Different Event Types

Event Type	Precision	Recall	F1
Personal(280)	0.92	0.88	0.90
Social(140)	0.57	0.36	0.44
Official(350)	0.97	0.97	0.97

REFERENCES

Allan, J., J. Carbonell, (1998) Topic Detection and Tracking Pilot Study Final Report. DARPA Broadcast News Transcription and Understanding Workshop.

Allan, J., and R. Papka, (1998) On-line New Event Detection and tracking. SIGIR'98, Melbourne, Australia, ACM.

Becker, H., and M. Naaman, (2009) Event identfication in social media. Twelfth International Workshop on the Web Databases (WebDB 2009), Providence, USA.

Becker, H., and M. Naaman, (2010) Learning Similarity Metrics for Event Identification in Social Media. WSDM, New York, USA, ACM.

Bai, J., and J. Guo, (2010) An Efficient Algorithm of Hot Events Detection in Text Streams. Cyber-Enabled Distributed Computing and Knowledge Discovery (Cyber C), 2010, Huangshan

Chen, C. C., and M. C. Chen, (2009) "An adaptive threshold framework for event detection using HMM-Based Life Profiles." ACM Transactions on Information Systems 27 (**2**) 22-30.

Chen, K., and L. Luesukprasert, (2007) "Hot Topic Extraction based on timeline analysis and multidimensional sentence modeling." IEEE Transactions on Knowledge and Data Engineering 19 (**8**) 567-575.

Fukumoto F. and Y. Suzuki (2000) Event Tracking based on Domain Dependency. Event Tracking based on Domain Dependency, Athens, Greece, ACM.

Ha-Thuc, V., and Y. Mejova, (2009) Event Intensity Tracking in Weblog Collections. ICWSM-DCW' 09, California, USA, AAAI.

Hobbs, J., and D. Appelt, (1997) FASTUS: ACascaded Finite-State Transducer for Extracting Information from Natural-Language Text. MUC, Cambridge, MA, MIT Press.

King, P. and S. H. Mayeng (2004) "Usefulness of Temporal Information Automatically Extracted from News Articles for Topic Tracking." ACM Transactions on Asian Language Information Processing 3 (**4**): 227-242.

Kumaran, G. and J. Allan (2004) Text Classification and Named Entities for New Event Detection. SIGIR'04, Sheffield, South Yorkshire, UK, ACM.

Kuo, Z., and L. J. Zi, (2007) New Event Detection Based on Indexing-Tree and Named Entity. SIGIR'07, Amsterdam, The Netherlands, ACM.

Kim, P., S.H. Myaeng, and J. C. Ryou (2003) Extracting Temporal Information from Korean News Articles for Event Detection and Tracking. 20th International Conference on Computer Processing of Oriental Languages.

Lin, C. X., and B. Zhao, (2010) PET: a statistical model for popular events tracking in social communities. SIGKDD, New York, USA, ACM.

Lei, Z., and J. Liao, (2008) Event Detection and Tracking Based on Improved Incremental K-Means and Transductive SVM ICIC 2008, Shanghai, China, Springer.

Makkonen, J., and H. Anonen-Myka, (2004) "Simple Semantics in TopicDetection and Tracking." Information Retreival Journal **7** (3-4): 347-368.

Platakis, M., and Dimitrios Kotsakos, (2008) Discovering Hot Topics in the Blogosphere. 2nd Panhellenic Scientific Student Conference, Samos, Greece.

Pekar, V. (2005) Information Extraction from Email Announcements. LNCS. Berlin Heidelberg, Springer Verlag**:** 372-375.

Sayyadi, H., and M. Hurst, (2009) Event detection and tracking in social streams. Association for Advancement of Artificial Intelligence (AAAI'09).

Seon, C.N., and H. Kim, (2008) Information extraction using finite state automata and syllable n-gramsin a mobile environment. ACL-08: HLT Workshop on Mobile Language Processing, Ohio, USA.

SAC'09, Honolulu, Hawaii, U.S.A, ACM. Xiaoming Zhang and Z. Li (2009) Online New Event Detection Based on IPLSA ADMA, Beijing, China, Springer.

Tingting He, Guozhong Qu, (2006) Semi-automatic Hot Event Detection. ADMA, Xian, China, Springer.

Wan, X., E. Milios, (2009) Link-based Event Detection in Email Communication Networks.

Yang, Y., and J. Carbonell, (1999) "Learning Approaches for detecting and tracking News Events." IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval **4**(14): 32-43.

Yang, Y., and T. Pierce, (1998). A Study on retrospective and online Event Detection. SIGIR'98, Melbourne, Australia, ACM.

Zhao, Q. and P. Mitra (2007) Event Detection and Visualization for Social Text Streams. ICWSM, Coloroda, USA.

Zhao, Q., and P. Mitra, (2007) Temporal and Information Flow Based Event Detection From Social Text Streams. American Association for Artificial Intelligence (AAAI 2007), Vancouver, British Columbia, Canada.