EXTRACTING INTELLIGENCE FROM ONLINE NEWS SOURCES

Smriti Sharma[1], Rajesh Singh[2], Pawan Bhadana[3]

[1]Student, [2]Lecturer, [3]HOD, Computer science & engineering, BSAITM, Haryana, India,
smritisharma3627@gmail.com,rajeshsingh22@gmail.com,pawanbhadana79@gmail.com

Abstract

This paper summarizes initiative for news extraction when we are investigating a simple approach for visualization of a range of content. To find specific information easily a novel approach of 5W1H is easiest & best suitable. Here we are "Extracting Intelligence from various online news sources". Intelligence here means "detecting &tracking, visualization". So our objective is not only extracting the news events occurred but to visualize it as well. This paper presents relatively lightweight approach of mapping the extracted news events. We present results of our work in news event extraction ,relevancy visualization, news visualization of extracted events, to enhance user interaction in information access and exploitation tasks. Here our news event extraction is done by 5W1H approach for detecting & tracking news events & then using its output to visualizing those events by personalizing maps.

Index Terms: *Event extraction, Visualization, Detecting & tracking, NER, NEXUS*

----------***----------

1. INTRODUCTION

Information overload is a main challenge in our world today. Applying techniques from text mining, automated machine learning and statistical analysis canhelp to reduce this overload of information Online system works in three steps: 1) Data Scraping from online sources, pre-processing and Named Entity Recognition (NER).2) Detecting new violent events and then clustering of related items to a thread.3) Mapping these related events on a timeline and location visualizer.For detecting & tracking news event extraction there are many more approaches but as above defined problems & To relieve "News Information Overload", we describe & using approach of 5W1H (who, what, whom, when, where, how) event semantic elements extraction for Chinese news event knowledge base construction. .This approach comprises a key event identification step, an event semantic elements extraction step and an event ontology population step. We first use a machine learning method to identify the key events from Chinese news stories. Then we extract event 5W1H elements by employing the combination of SRL, NER technique and rule-based methodVisualization means mapping the articles on a time-line and location map. The online system utilizes MIT's SIMILE library for timeline visualization. It is difficult to follow the entirity of the global of news sources , and the events happening every day. If an analyst in her area has to follow and map all these according to the timeline they happen, the task quickly becomes overwhelming. As we see in cluster centric approach where NEXUX system is used news can be mapped by live event tracking as shown in fig-1

Fig-1: Event Visualization in Google earth

We present a tool which attempts to ease the task of finding all news articles about an event, and mapping them without using Google earth.

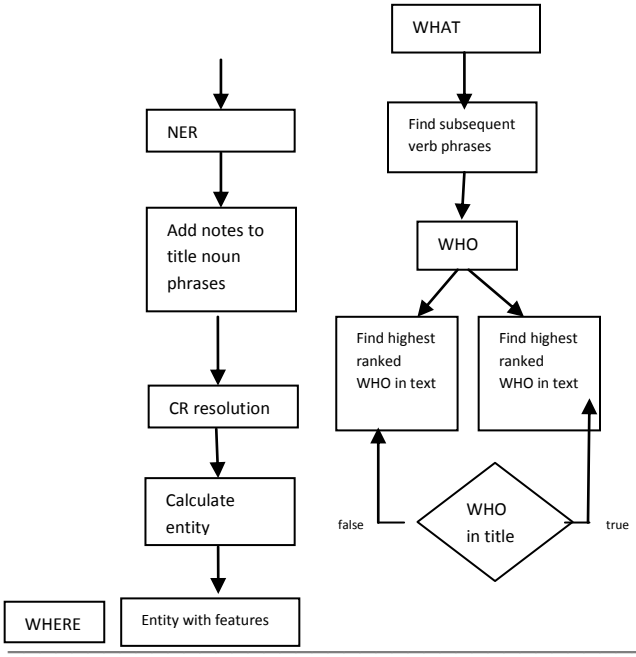
2. RELATED WORK

Wider area of the work presented in this paper fits in theareas of data visualization [3] and in particular in the areas oftext visualization [6] and more recently, developments of semantic web and visualization of ontologies and other knowledge structures [1].approach used for detecting & tracking of news events we used News Event Extraction Using 5W1H Approach & Its Analysis [12] . Most prominent is the overview publication from MITRE team [7] giving goodoverview over the approaches for visualization of different document types, including news stories. Their goals

aresimilar to the work presented here, but the actual approach is quite different. Their publication appeared also at [8]together with some other interesting approaches fordocument visualization.Another approach for visualizing trends in news documentsis the system ThemeRiver [9] developed at Pacific NorthwestNational Laboratory together with many other interestingapproaches for information text visualization [10].ThemeRiver in particular is specialized for analyzing andvisualizing trends in news stories over time, enablingefficient detection of trends in the vocabulary used in thetexts. Among others, we would also like to mention work ofthe authors on visualization of large text corpora [11]presented at "TELRI -Information in Corpora" workshopwhich directly precedes this work.

3. DETECTING & TRACKING

Formally, the task of event extraction is to automatically identify events infree text and to derive detailed information about them, ideally identifying Whodid what to whom, when, with what methods (instruments), where and eventually why?. Automatically extracting events is a higher-level information extraction(IE) task which is not trivial due to the complexity of natural language anddue to the fact that a full event description is usually scattered over severalsentences and documents. Further, event extraction relies on identifying namedentities and relations holding among them. Here detection & tracking is done using 5W1H approach [12].from A) key identification step we can get answers about specific information needed (Perpetrators , Event type, Victims, Date, Place, Weapons). Here We divide our approach into six sub-tasks and group them in three steps: (1) Title classification and topic sentences extraction for key event identification; (2) Semantic role labeling and 5W1H elements identification for event semantic elements extraction; (3) Collect extracted event facts automatically according to nature of existence.[12]



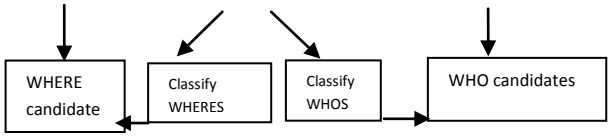


Fig-2: Processing chain of WHERE WHO & WHAT

Elements found from these processing chains we will get "News Event 5Ws": The {Time, Location, Subject, Predicate, Object} information which describe the when, where, who, what, whom of an event are called news event 5W elements.

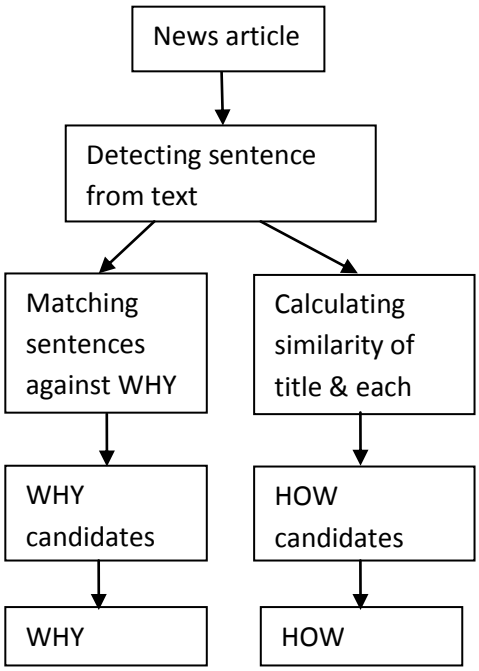


Fig-3: Processing chain of WHO & WHY

After this key identification step done above second B) Event Semantic Element Extraction step is there.We first label semantic roles in the headline and topic sentences and then we improve the results.[12]These steps also need POS-tagging, Sentence Detection , phrase chunking , Named entity recognition etc. Next step is C) Collecting Event facts, Event Extraction Process. After the text of the article is preprocessed the gradual extraction of the 5W1H starts. After extraction of 5W1H is done. Here we have a general problem that the subsequent verb phrase in long sentences contains a lot of information that we cannot ignore because it is semantically relevant. Finally it is a non-trivial task to filter out the minimal necessary information. We decided to solve this problem by limiting the verb phrase to length.[12].

4. VISUALIZATION

Multilingual crisis-related event tracking poses a number of practical issues, mainly related to the correct geo-spatial

visualization of the event together with its principal characteristics. Another concern is to minimize constraints on end users to rely on expensiveand proprietary desktop applications. We fulfill these issues by publishing the event datausing current internet standard formats, namely, KML and GeoRSS. In particular, the results of the event extraction are accessible in two ways: (a) via Google Earth applicationwhich is passed event descriptions in KML format; and (b) via a publicly accessible web client that exploits the Google Maps technologies and connects to our KML server.Paper presents a system for visualization of large amounts of new stories. In the first phase, the new stories are preprocessed for the purpose of name - entity extraction. Next, a graph of relationships between the extracted name entities is created, where each name entity represents one vertex in the graph and two name entities are connected if they appear in the same document. Text visualization is an area having the main goal to present textual contents of one or many documents in a visual form. The intention of producing visualization of the textual contents is mainly to create graphical form of content summary on different levels of abstraction. the documents are preprocessed in two different ways. First, the text is cleaned and bag- of- words representation is created, and next, the name - entities are extracted. All the documents are stored in three different representations (as alreadydescribed: plain text, bag -of -words and name - entities) in the database which is then used by the client software using efficient graphical user interface described in the following sections.

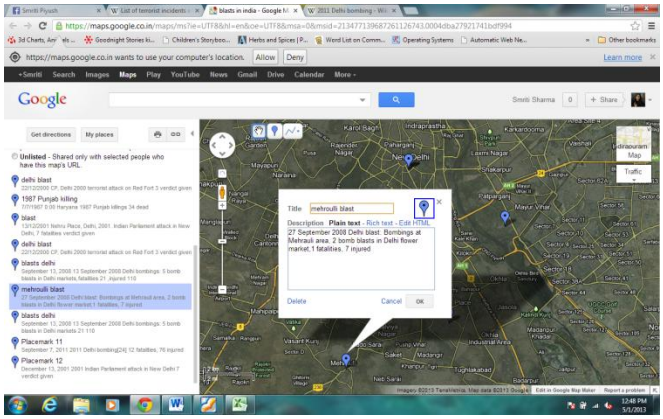


Fig-4: Visualization of news events on Google map

Here in this paper we have output of 5W1H approach. We use a very simple approach of mapping events.Here we d'nt need of designing the systemto get an efficient and quick understanding of largecorpus of general news stories at different levels ofabstraction.Here for extraction we already took 5W1H approach which gives us specific elements of events. We have to just connect output of extraction to input of map .Input to map is database obtained by extraction approach Used.

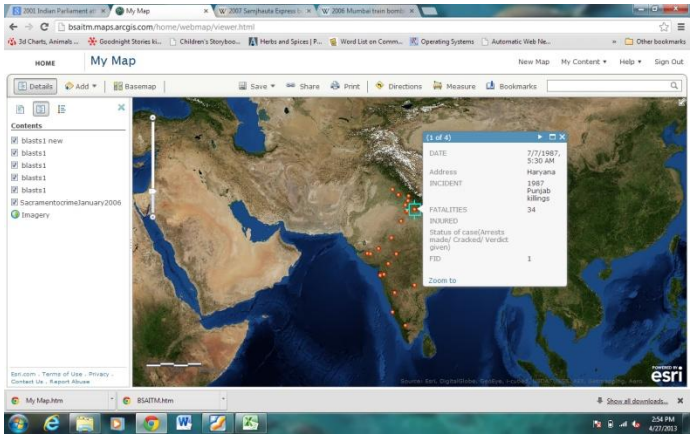


Fig-5: output map showing event with information

This implementation is done this by using ArcGIS map. There can be maps of different types .this is a simple approach of showing information.

5. EXPERIMENTS AND EVALUATION

Main work done to be evaluated here is its visualization. So for evaluating this we take a database table after news event extraction using 5W1H approach which should be in Microsoft Excel CSV format, which is comma separated values file.

Table-1 Extracted events with specific information

Date	Address	Incident	Fatalities	Injured	Status of case
7/7/1987	haryana	1987 Punjab killings	34		
15/6/1991	Ludhiana district, Punjab	1991 Punjab killings	88		
12/3/1993	Mumbai, Maharashtra	1993 Bombay bombings	257	713	verdict given
30/12/1996	Western Assam	Brahmaputra Mail train bombing	33		
14/2/1998	Coimbatore, Tamil Nadu	1998 Coimbatore bombings	58	200+	verdict given

22/12/2000	CP,Delhi	2000 terrorist attack on Red Fort			verdict given
24/9/2002	Gandhinagar, Gujarat	Terrorists attack the Akshardham temple in Gujarat	31		

Geographical Information System is among today's fast developing technologies and is being integrated with various other computer applications. This table is obtained after extraction process .Now we use ArcGIS maps. ArcGIS Explorer Online is an online application that lets you explore and present maps within an efficient and well-structured environment. Maps show you where things are, they tell you what they are and help you understand why they are that way. ArcGIS Explorer Online lets you open a map, add other content to it, navigate around it, ask questions the map can answer, and present and share the map with others. The visualization of news events extracted in table 1 is shown in fig-6.

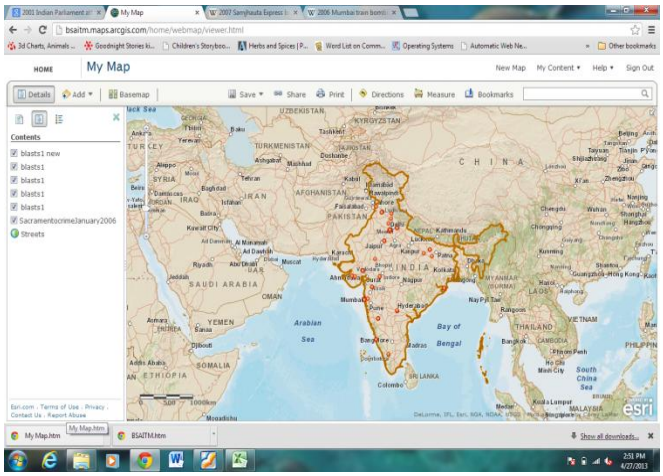


Fig-6: Blasts information extracted in table-1 on map

Here we can specifically see the information on map of nay event. There is no need to see output table of extraction. All information will be visualized.

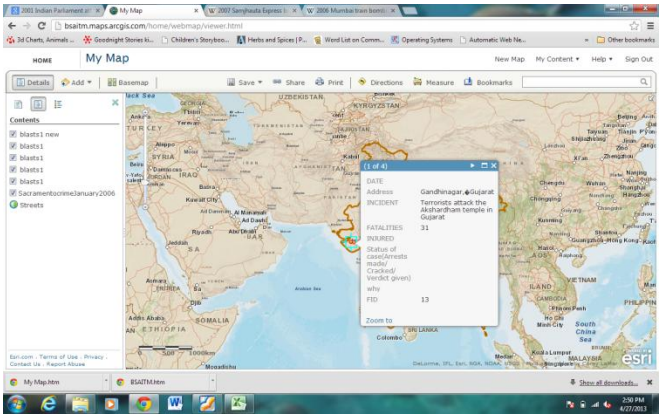


Fig-7: Gandhi nagar, Gujarat blast information on map

CONCLUSIONS

In this paper we presented an easy approach of detecting, tracking & visualization of news events. This paper presents an approach of event tracking n visualization. Research based in the fields of geographic information retrieval (GIR) and natural language processing (NLP) use methods to extract place-names and other spatial references from web documents that can be used to display event locations in a GIS. Vagueness common in spatial descriptions, such as north of the city or near the border, is a problem that must be addressed in geographic information systems in order to successfully represent text-based events. future work this vagueness should be investigated.

REFERENCES

[1] Geroimenko, V., Chen, C. (ed): Visualizing the SemanticWeb. Springer Verlag, (2003).
[2] Manning , C., Schütze, H.: Foundations of StatisticalNatural Language Processing. MIT Press (1999).
[3] Fayyad, U., Grinstein, G., Wierse, A.: InformationVisualization in Data Mining and KnowledgeDiscovery. Morgan Kaufmann (2001).
[4] Chakrabarti, S.: Mining the Web: Analysis of Hypertextand Semi Structured Data. Morgan Kaufman (2002).
[5] Jurafsky, D., Martin. J.H.: Speech and LanguageProcessing: An Introduction to Natural LanguageProcessing, Computational Linguistics and SpeechRecognition. Prentice Hall (2000).
[6] C hen, C.: Visualization of Knowledge Structures, In"Handbook of Software Engineering and Knowledgeengineering", World Scientific Publishing (2002)
[7] Chase, P., D'Amore, R., Gershon, N., Holland, R.,Hyland, R., Mani, I., Maybury, M., Merlino, A., RaysonJ. :Semantic Visualization. ACL- COLING Workshop onContent Visualization and IntermediaRepresentation.
[8] Content Visualization and Intermedia Representations(CVIR'98), http://acl.ldc.upenn.edu/W/W98/
[9] Havre, S., Hetzler, E. , Whitney, P., Nowell, L.:ThemeRiver: Visualizing Thematic Changes in

LargeDocument Collections. IEEE Transactions onVisualization and Comp. Graphics, V8, No.1, 2002.
[10] Pacific Northwest National Laboratory, InformationVisualization, http://www.pnl.gov/infoviz/
[11] Grobelnik, M., Mladenic, D.: Efficient visualization oflarge text corpora. 7thTELRI, Info.in Corpora (2002)
[12] SmritiSharma,RajeshKumar,Pawan Bhadana, News Event Extraction Using 5W1H Approach & Its Analysis,IJRET.

BIOGRAPHIES:

Smriti Sharma is B.Tech. , MBA (HRM), &M.Tech(pursuing). She has 5 yrs of experience in field of education. 8 research papers have already been published.

Rajesh Singh is B.Tech ,M.Tech .He has 6 yrs teaching experience .8 research papers have been published till now.

Pawanbhadana is HOD in BSAITM. He is B.E., M Tech ,PhD(pursuing).he has total 10 yrs of experience in the field of education .total research papers published till now are 17.