Topic Detection and Tracking Interface with Named Entities Approach

Omar Mabrook A. Bashaddadh and Masnizah Mohd
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 Selangor, Malaysia

Abstract—**Topic Detection and Tracking (TDT) refer to the technologies and techniques in analysing and handling the vast amount of information arriving continuously from the news stream. Some issues in news monitoring regarding interface design are considered. In this paper we present the techniques and works done in TDT domain. We discuss the named entities approach used in document representation and in user interface. Finally, we present the suggested approach to improve users' performance and to facilitate them to perform an effective TDT task.**

Keywords-Topic Detection and Tracking; Bag of Words; Named Entity; Interface.

I. Introduction

A vast amount of information arriving every day through a variety of news media such as, newswires, TV, broadcasts, newspapers, radio and website resources. In addition to that, the significant growth and dynamic environment of digital information has captured human's attention causing some challenges in Information Retrieval (IR) technology.

This amount of constant information, which readers can effectively use, is still limited compared to the continuous and rapid growth of online information. It has become significantly important to enhance and provide methods to find and present textual information effectively and efficiently. There are many technologies that have been proposed to solve the information overload issues including information customization, search engines IR and information agency [6]. Ranking mechanisms have been used to find related document and sort them by rank in IR systems using keyword search or query expansion [9].

Topic Detection and Tracking (TDT) technology investigates the IR methods in organizing a constantly arriving stream of news articles by the events (Topic- Event- Activity-story) that they discuss. The key terms in TDT are:

c. Topic: is a method specifically used for TDT research. Topics refer to specific events or activities, such as the crash of Yemen Airlines airplane (Yemenia) in Indian Ocean near the Comoros islands. On Jun 30, 2009, and involved all facts, events and activities that are directly associated and related to them.

d. Event: is something that happens at a specific time and place, and the unavoidable consequences. Crimes, specific elections, ceremony, accidents and natural disasters are examples of events.

e. Activity: is a combined set of actions that have a common focus or purpose. Specific campaigns, investigations, and disaster relief efforts are examples of activities.

f. Story: is a newswire article or a segment of a news broadcast with a coherent news focus.

TDT is an automatic technique which is applied in finding and organising topically related material in a stream of news stories. Moreover, it can assist people to interpret and analyse news stories quickly. This could be valuable in a wide variety of applications where efficient and timely information access is important such as CNN[1], Yahoo News[2] and Aljazeera News[3].

The key question here is how can we track and detect constant and dynamic information, such as news, using IR techniques? And how we can present the news to the user? Teevan, in her study [19], investigated how people re-find information on the Web. Mohd [14] has designed a TDT interface that uses effective features in helping professionals,, such as journalists, to perform the TDT tasks. There are a lot of IR systems available publicly which aim to increase the capability of tracking and detecting the news. For example, CNN and Aljazeera Mobile News have been developed to alert users on any new products. GoogleNews and NewsInEssence from the University of Michigan have provided a tracking service by email to the users when new articles about their interested subject become available [14].

Journalists often rely on the Rich Site Summary (RSS) news feeds to keep track of the most current information and events. Thus, there is an increasing need for automatic techniques to analyse, present and visualize news to users in a meaningful and efficient manner.

Dynamic information is the main topic dealt with in the area of research known as TDT. Recent Research in TDT aims to effectively retrieve and organize broadcast news (speech) and newswire stories (text) into groups of events. The majority of TDT research and evaluation has been on the system performance without any user involvement. TDT is part of Text Retrieval Conference (TREC). In recent years, much work has been done in TREC and the TDT domain to investigate methods for automatically organising news stories.

[1] http: //www.cnn.com
[2] http://www. news.yahoo.com
[3] http: //www. english.aljazeera.net

Very few TDT researchers have started working on user interfaces and user interaction.

There are two well-known approaches used to represent a document in TDT; the bag-of-words and the named entities approach. A common way of converting the document to such a form is to use the words in a document as attributes and the number of times the word occurs in the document, or some function of it, as the value of the attribute. Using such a method to convert documents to an actionable form results forgoing information contained in the grammatical structure in the document. Despite this drawback, the bag-of-words approach is one of the most successful and widely used methods of converting text documents into actionable form [15]. Named Entities (NEs) approach is used in Information Extraction (IE), Question Answering (QA) or other Natural Language Processing (NLP) applications. It is important to recognize significant information units such as names, including person, organization, location, and numeric expressions including time, date, money and percentage expressions. This is because the user tends to remember more of this information rather than the common terms [17].

The rest of the paper is organized as follows. In Section II, we discuss the related works that uses NEs in document representation and in user interface. Section III presents our proposed approach for document representation and to enhance the user interface. Finally in Section IV we conclude the paper with a summary.

II. NAMED ENTITIES

In this section we will discuss the related works that used Named Entities (NEs) in the document representation and user interface. The bag-of-words and NEs approaches are widely used in automatic keywords indexing on the basis of properties such as frequency and length or to controlled vocabulary of terms, and classifies the documents according to their content [21]. Named entity was first used in the Message Understanding Conferences (MUC) which influenced Information Extraction (IE) research in the U.S. in the 1990's [8]. At the sixth conference (MUC-6) the task of Named Entity Recognition and co-reference were added. At that time, MUC focused on IE tasks where structured information of company activities and military related activities was extracted from unstructured text, such as newspaper articles. Outside the U.S., there have been several evaluation-based projects for named entities, as one of the tasks of Information Retrieval and Extraction Exercise (IREX) in Japan [16]. There has also been evaluation for named entities in the shared task in the Conference on Computational Natural Language Learning (CoNLL) in 2002 and 2003 for four languages, English, German, Dutch and Spanish. In the IREX project, a new category known as artefact, an example of which might be Odyssey as a book title or Windows as a product name, was added to the original MUC categories. The named entities task in MUC was inherited by the Automatic Content Extraction (ACE) project in the U.S., where two new categories were added; Geographical and Political Entities (GPE), such as France or New York; and Facility, such as The Empire State Building. There are new fields where the named entities related task becomes an important component technology. For example, in bioinformatics, recognizing names of proteins or genes is crucial. As a result, there are on-going efforts to make extended named entities [17].

The importance of Named Entity Recognition (NER) was highlighted by the OKKAM European Project. This study investigates some aspects of the use of keywords in Web searching. Named entity is discussed from a journalism perspective to improve TDT interface [14].

A. Document Representation

Document representation is one of the most common and crucial stages of an information organization and access system. Several methods and models of document representation have been proposed based on the target application. Some of them are general enough to be applicable to almost any IR-based application. However, some tasks demand a different approach to document representation. Topic Detection and Tracking (TDT) is one of these domains. TDT began as a technology development and evaluation program [3]. TDT evaluation provides a standard set of news documents with a number of topics to be tracked and a list of relevant documents for each topic [7]. Researchers in this area claim that technology evaluation is the main focus of TDT and does not investigate user interface issues [1]. Similarly, TDT evaluation has traditionally been carried out in a laboratory setting, which does not involve real users and real tasks. Hence, researchers in this area have focused on developing techniques and algorithms for a better TDT performance; this is also the main activity in TREC evaluation.

Recently, NER has been receiving more attention in TDT where several efforts have been made to exploit it for document representation, in order to improve TDT systems. Yang [20] investigated and focused on location as a named entity for document representation. The DOREMI research group also looked at people and location NEs to obtain a final confidence score for each story [13]. Meanwhile Kumaran and Allan [11] split document representation into two parts: named entities and non-named entities. It was found that some classes of news could achieve better performance using NER such as Elections, Accidents, Violence and War, New Laws, Sports News, and Political and Diplomatic Meetings. For instance, the names of election candidates (Person name) are very important for stories of election class; the locations (Location name) where accidents happened are important for stories of accident class. While some other classes of news such as Natural Disasters, Criminal cases, Scandals/Hearings and Science could achieve better performance using non-named entities representation. [22] Investigating the average correlation between Part-of-Speech (POS) and news genre to model New Event Detection (NED) model revealed that terms of different types (Noun, Verb or Person name) have different effects for different genre of stories in determining whether two stories are on the same topic. For example, the names of election candidates (Person name) are very important for stories of election class; the locations (Location name) where accidents happened are important for stories of accident class.

B. User Interface (UI)

The previous scholarly studies in TDT tried to build better document models, enhancing and developing similarity metrics or better document representations. A lot of research efforts focused on enhancing and improving document representations by applying Natural Language Processing (NLP) technology such as Named Entity Recognition (NER) [20, 13, 11, 22]. Whereas, a few researches concentrating on Graphical User Interface (GUI) such as:

a. Event Organizer [2] to organize a constantly updating stream of news articles by the events that are discussed in the stories.

b. TDT Lighthouse [12] was designed to present results of a search session to the user. It provides a visualization of inter-document similarities in two or three dimensions, in addition to a typical ranked list search result.

c. TimeMine [18] is a prototype system (TDT system) to detect, rank and group semantic feature based on their statistical properties to generate an interactive timeline displaying the major events and uses it as a browsing interface to a document collection.

d. Topic Tracking Visualization tool [10] is a graphical tool used in TDT algorithm development. Whereas, the system uses colours to show the results of the TDT in relation to some ground truth e.g. on-topic stories, misses and false alarm are shown in green, red and blue respectively.

e. Interactive Event Tracking System (*i*Event) interface [14] to facilitate professionals, such as journalists or information analysts, to perform TDT tasks. Interestingly, this is the first TDT work which involves ‘the journalist’ as a user of an *i*TDT interface. It consists of three components: Cluster View (CV), Document View (DV) and Term View (TV) and two settings: keywords and named entities.

f. Stories in time [5] are methods and tools for learning an abstracted story representation system, and graphical support for story understanding and story search.

III. PROPOSED APPROACH

Previous research in TDT has concentrated primarily on the design and evaluation of algorithms or interface to carry out TDT tasks. This has motivated us to propose a new structure to improve the time vector and to enhance the user interface to support the tasks related to TDT. In this review paper, we will propose a new approach depending on [14] dissertation future work to carry out the following:

a. Built a new setup which will merge the Bag-of-words and Named Entities approach. We will be using all the previous interface views and features in the previous system specially Setup 1 and Setup2 in *i*Event interface [14] and new approaches to build Setup 3 as a new setup as depicted in Fig. 1

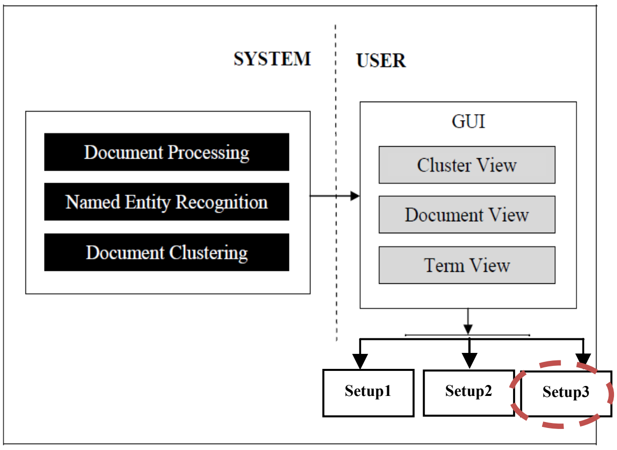

Fig. 1 *i*Event architecture with new setup

b. Build a new structure to model and store the clusters using data structure technique. We will use an array of vector to store the clusters (c). Each cluster (c_1, c_2… c_n) represent an object (n object) and each vector in array store file (f_1, f_2, … f_m). We will create a new array vector as an object which contains all index links, the keywords (K_w) and Named Entities (NEs) for each cluster as shown in Fig 2 or use a tree structure (Fig. 3) to facilitate the new setup (Setup 3). Thus, the new structure will affect the labelling of the clusters as shown in Fig. 4

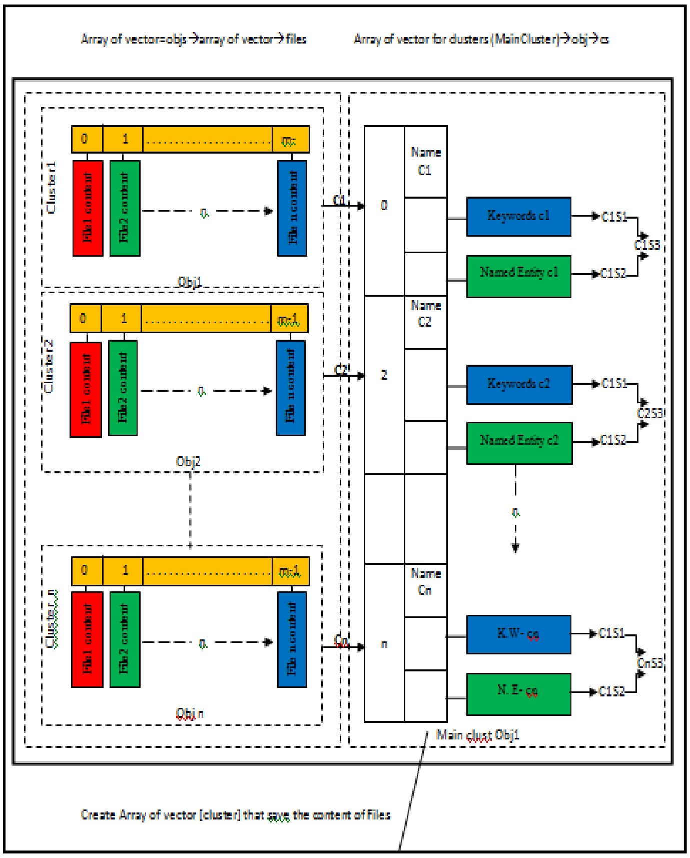

Fig. 2 New proposed structure

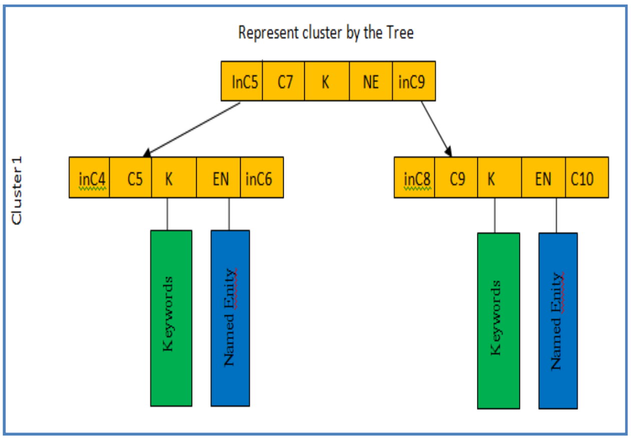

Fig. 3. Tree structure

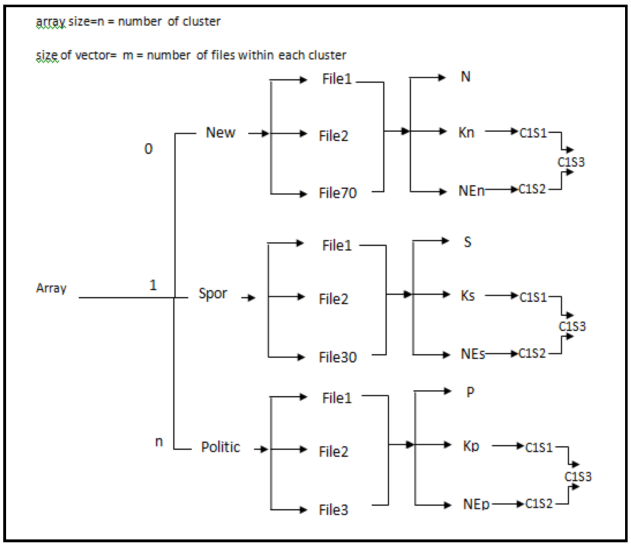

Fig. 4. The new proposed structure in labeling the clusters

The proposed works aims to improve the user performance when they perform the TDT tasks. We plan to conduct a user experiment with journalists to evaluate the new setup (Setup 3). They will perform the Tracking and the Detection task.

IV. CONCLUSIONS

This paper has investigated the related TDT works using named entities approach where the focus is on the user interfaces aspect. This has contributed to the proposed approach and towards the design of an interface to carry out the TDT tasks. We are hoping the new data structure and the new setup will increase the users' performance and facilitate them to perform effective TDT tasks.

REFERENCES

[1] J. Allan, "Topic Detection and Tracking: Event-based Information Organization." vol. 12: Kluwer Academic Publishers, 2002.

[2] J. Allan, et al., "Taking Topic Detection From Evaluation to Practice," presented at the Proceedings of the 38th Hawaii International Conference on System Sciences(HICSS'05), Washington DC, 2005.

[3] J. Allan, et al., "On-line new event detection and tracking," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 37-45, 1998.

[4] H. Becker, et al., "Event Identification in Social Media," presented at the TTwelfth InternationalWorkshop on theWeb and Databases (WebDB 2009),nJune 28, 2009, Rhode Island, USA, 2009.

[5] B. Berendt and I. Subasic, "STORIES in time: a graph-based interface for news tracking and discovery," in WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology vol. 03, 534 ed. Washington, DC, USA ©2009: IEEE Computer Society 2009, pp. 531- 534.

[6] H. Berghel, "Cyberspace 2000: dealing with information overload," Communications of the ACM New York, NY, USA, vol. 40, 1997.

[7] J. G. Fiscus and G. R. Doddington, "Topic detection and tracking evaluation overview," Topic Detection and Tracking: Event-Based information Organization, pp. 17- 31, 2002.

[8] R. Grishman and B. Sundheim, "Message Understanding Conference - 6: A Brief History.," in Proceedings of the 16th conference on Computational linguistics (COLING- 96). Copenhagen, Denmark: Association for Computational Linguistics, 1996, pp. 466- 471.

[9] J.-F. Hsueh, "Learning ontology from Web documents for supporting Web query," Master, Information Management, National Sun Yat-Sen University, 2002.

[10] G. J. F. Jones and S. M. Gabb, "A visualisation tool for topic tracking analysis and development," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, (August 11 - 15, 2002), 2002, pp. 389- 390

[11] G. Kumaran and J. Allan, "Text Classification and Named Entities for New Event Detection," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004.

[12] A. Leuski and J. Allan, "Lighthouse: Showing the Way to Relevant Information," Proceedings of the IEEE Symposium on information Visualization, pp. 125- 129, 2000.

[13] J. Makkonen, et al., Simple Semantics in Topic Detection and Tracking vol. 7: Information Retrieval, Springer, 2004.

[14] M. Mohd, "Design and Evaluation of an Interactive Topic Detection and Tracking Interface," PhD, Computer and Information Science, Strathclyde, Glasgow, 2010.

[15] N. Sahoo, et al., "Incremental Hierarchical Clustering of Text Documents," in Proceedings of the 15th ACM international conference on Information and knowledge management, ACM New York, NY, USA, 2006.

[16] S. Sekine and H. Isahara, "IREX: IR and IE Evaluation project in Japanese," Proceedings of International Conference on Language Resources & Evaluation (LREC 2000), 2000.

[17] S. Sekine, et al., "Extended Named entities Hierarchy," in Proceedings of International Conference on Language Resources & Evaluation (LREC 2000), Athens, Greece, 2002.

[18] R. Swan and J. Allan, "Automatic Generation of Overview Timelines," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (Athens, Greece, July 24 - 28, 2000), SIGIR '00. ACM, New York, NY, 2002, pp. 49- 56.

[19] J. Teevan, "The Re-Search Engine: Helping People Return to Information in Dynamic Information Environments," PhD, Massachusetts Institute of Technology, 2007.

[20] Y. Yang, et al., "Learning approaches for detecting and tracking news events," IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, IEEE Educational Activities Department, vol. 14(4), pp. 32 -43, 1999.

[21] C. ZHANG, et al., "Automatic Keyword Extraction from Documents Using Conditional Random Fields," Journal of Computational Information Systems, vol. 4, pp. 1169-1180, 2008.

[22] K. Zhang, et al., "New Event Detection Based on Indexing-tree and Named entities," presented at the SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, The Netherlands, 2007.

[23] W. Zheng, et al., "Topic Tracking Based on Keywords Dependency Profile," Lecture Notes in Computer Science, vol. 4993, pp. 129-140, 2008.