

# Evaluation of an Interactive Topic Detection and Tracking Interface

**Masnizah Mohd**

*Universiti Kebangsaan Malaysia, Malaysia*

**Fabio Crestani**

*University of Lugano, Switzerland*

**Ian Ruthven**

*University of Strathclyde, United Kingdom*

**Abstract.**

**Interactive Topic Detection and Tracking (*i*TDT) refers to the TDT works which focused on user interaction, user evaluation and user interfaces aspects. This paper investigates and identifies the elements of the design of an interface that aims to facilitate journalists to perform TDT tasks. Therefore we present an (*i*TDT) interface called Interactive Event Tracking (*i*Event) and evaluate the usability of the features introduced. We conduct a user experiment to investigate the effectiveness and usefulness of the features introduced and report what TDT tasks these features are facilitating. We show that the interface enables journalists to perform well in TDT tasks and we have identified potential guidelines for future designs of the *i*TDT interfaces for TDT task.**

**Keywords:** Interactive Topic Detection and Tracking (*i*TDT); TDT; user interface

## **1. Introduction**

Research in Topic Detection and Tracking (TDT) served as a core technology for a system that would monitor broadcast news. Thus they help alert the information professionals such as journalists to new and interesting events happening in the world. They are keen to track and in particular to know the latest news about a story from a large amount of information that arrives daily. This has created an opportunity and attempt in TDT research to focus on user interaction, user evaluation and user interfaces as a way to visualise and represent news in a meaningful way.

TDT tasks such as stream segmentation; link detection; story detection; and story tracking mainly focusing on the evaluation of the algorithms in the context of TREC and without user involvement [10]. We believed that TDT

is very much an interactive task, inter-related and complemented with the user interaction. Therefore research in TDT should also be tackled from the user perspective which enables us to view the complete TDT [2, 14]. For example user can validate and define a new story in new event detection task.

There are efforts been done on user interfaces to improve TDT system by investigating not just the interaction aspect but also user and task oriented evaluation [11]. We believe that interfaces play a vital role in interactive Topic Detection and Tracking (*i*TDT) and we set up to design a new interface for *i*TDT that is meant to support the user in all the tasks related to TDT. The importance of user interaction in the real world is the reason why *i*TDT is receiving more attention. It was clear from the literature that a well designed *i*TDT interfaces is important to guide user in performing the TDT tasks. Designing such interface should incorporate the best and successful component or features.

The term interactive TDT (*i*TDT) is used in this paper, to refer to the TDT works which focused on these aspects. It is important to provide a means for people such as journalists to understand and interpret what is happening in the news. TDT research is still active where the researchers in this area have focused on developing algorithm for better TDT performance and the evaluation of these algorithms is the main activity in TREC evaluation [17]. Fewer TDT researchers investigated techniques such as information visualisation and automatic timelines to support users with a dynamic and interactive use. Very few researchers worked on interfaces and user interaction for TDT. According to [4] and [13], an effective interface should be well designed and generate a positive feelings of success, competence, mastery, pleasure and clarity in the user community.

## 2. Related work

TDT research is still continuing where one of the focuses now is on *i*TDT and this work is on this direction. TDT researchers have attempted to build better document models, developing similarity metrics or better document representations [7]. This led to a series of research efforts that concentrated on improving document representation by applying Named Entity Recognition (NER) [6, 12, 15, 20, 21]. Then a few researchers have started to move from the laboratory style of experiment to the interactive TDT mainly focusing on graphical user interface (GUI). Event Organizer [9], TDTLighthouse [3], TimeMine [16] and Topic Tracking Visualisation tool [5] are an example of TDT works that investigate certain approach to improve TDT system performance using GUI. We review these works by discussing the features and the approaches used and how it motivates this work. The reviewed works on interactive TDT enabled us to identify the similarities and differences of the components and features used, as shown in Table 1.

Table 1  
Comparison of *i*TDT features

	Event Organizer	TDT Lighthouse	Timemine	Topic Tracking Visualisation
Document View (DV)	<ul style="list-style-type: none"> <li>• Story profile (important terms and document list)</li> <li>• Document timeline</li> </ul>	<ul style="list-style-type: none"> <li>• List of topics (important terms)</li> </ul>	<ul style="list-style-type: none"> <li>• Topic timeline</li> </ul>	<ul style="list-style-type: none"> <li>• List of topics</li> <li>• Document content</li> </ul>
Cluster View (CV)	-	<ul style="list-style-type: none"> <li>• Cluster visualisation</li> </ul>	-	<ul style="list-style-type: none"> <li>• Boxes visualisation</li> </ul>
Term View (TV)	-	-	<ul style="list-style-type: none"> <li>• List of important terms</li> </ul>	<ul style="list-style-type: none"> <li>• List of important terms</li> </ul>

Most of *i*TDT interfaces reviewed have the Document View (DV) as the important component that display information such as the document timeline, document content and the list of topics or documents. Meanwhile Cluster View (CV) is also an important component which presents the stories or documents by visualising it in a cluster or in a box form. Unfortunately Term View (TV) is not as popular compared to the rest. The exploration and combination of these three views (DV, CV and TV) with features such as cluster visualisation and the timeline on the user interface could be effectively used together to perform the TDT tasks. Based on the works reviewed, none of them measured the effectiveness of their approach and features, applied on the interfaces, from a formal user aspect. Most of it reported on the effectiveness of the technique to the system performance using the Information Retrieval (IR) and TDT style evaluation. Past research has proven that user interfaces can significantly improve the effectiveness of the TDT task [9]. Therefore, the challenging questions are how to effectively analyse and present news in a meaningful and efficient manner, and what kinds of additional and critical information will contribute to an interactive TDT interface design. This will be described thoroughly in the next section.

### 3. Interactive Event Tracking (*i*Event) interface

The reviewed work of *i*TDT interface discussed in Section 2 has affected the design of *i*Event. *i*Event is composed of three components which are the Cluster View (CV), Document View (DV) and Term View (TV). In this section we describe the design of *i*Event and discuss its components and features.

The layout and the order of the component displayed on the interface begins from the Cluster View followed by the Document View and finally the Term View. Cluster View is displayed on top of the interface as the main component since this is the starting point where users are presented with a large amount of information for rapid interpretation. Visualising the cluster based on the size and the density of the documents might help them to identify the important and related cluster based on the task given. Cluster View allows the users to browse the whole collection before they narrow their search to a specific cluster. That is the reason why the order of the Document View is after the Cluster View. The Document View allows the users to view the whole document in a cluster with the specific timeline. It provides an effective form of presentation and a very fast graphical overview of the information that a cluster contains. Document View generates an interactive timeline displaying the major events and uses it as a browsing interface to a document collection contained in a cluster. Finally the Term View is displayed at the bottom of the interface to be more specific on the named entities contained in the cluster. Named entities are information units like names, including person, organisation, location names, and numeric expressions including time, date, money and percent expressions. Users get the whole view of the corpus before they received specific information of the documents and the named entities occurred in a cluster. The sequence or the ordering of the components on *i*Event helps users to narrow down their browsing and to be focussed in their searching. Thus it helps them to perform the TDT tasks.

*i*Event has two settings. Setup 1 is the baseline setup that uses keywords and Setup 2 is the experimental setup that uses NER. We extract the named entities using ANNIE which is an information extraction component of GATE (General Architecture for Text Engineering). We use it for its accurate entity, pronoun and nominal co-references extraction [8]. In this paper, we evaluated the usability of *i*Event interface without comparing the setups. Therefore in Section 5 (Results), we presented the results for both settings. We associated the findings with the setup if there is a significant finding when comparing result between setups.

#### 3.1. Cluster View (CV)

The Cluster View displays information related to the size and the density of a cluster; and the ten most frequently keywords (Setup 1) or named entities (Setup 2) in a cluster. The clusters are visualised based on the size and the density. Clusters with a large size and high density contain lots of documents which have appeared over a short period of time, therefore they are supposed to represent very important events. On the other hand, clusters with a small size and low density contain a small number of documents which have appeared over a long

period of time, thus presenting recurring but relatively unimportant events. Cluster visualisation is intended to help the user to make a rapid interpretation of a topic. It should be noted that given the difficulty in story segmentation, sometimes a cluster with a large size and low density might indicate the presence of more than one topic in the cluster. Clusters are labelled using the three most frequently named entities. When user clicks on the cluster, additional information on the ten most frequently named entities in that cluster is presented. These features are useful in TDT tasks since it provides information on the most frequent named entities that occur in a specific cluster.

### 3.2. Document View (DV)

The Document View displays information about the document timeline and the documents contained in a cluster. The document timeline is displayed in a histogram form to show the occurrence and the document frequency for a specific date. The height of histogram indicates the number of documents that occurred on that specific date in a cluster. This feature is an attempt to support the user in analysing the discourse or the information flow in a press article. Discourse analysis is a general term that includes many approaches to analysing the use of languages, and one important application of it is on news [18].

Timelines are a useful way to present information that has a temporal dimension. Journalists often generate timelines to describe the course of events. This will be evaluated to prove that automatically generated timelines could prove invaluable for navigating the results of a TDT system and for interactive TDT. The timeline feature is offered both in the Document View and in the Term View. Users would be able to see the occurrence of the document and named entities within the timeline in a histogram form for each cluster. Users would also be able to see the document content, with named entities highlighted where only Setup 2 highlighted named entities.

### 3.3. Term View (TV)

The Term View displays information related to named entities timeline in a cluster. The timeline is displayed in a histogram form to show the named entities occurrence and its frequency for a specific date. The histogram with the timeline shows the relevance score of named entities using term frequency ( $tf$ ). The timeline feature provides journalists with the whole view of named entities occurrences in the cluster. This is helpful in providing information about when the event occurred and supporting the new event detection task. This feature also helps the user in the Topic Detection task by presenting information about the latest occurrence of a named entity from the timeline.

## 4. User experiment

We conduct a user experiment to evaluate the *iEvent* interface. The participants were a combination of journalists and postgraduate journalism students from the Scottish Centre for Journalism Studies (SCJS), University of Strathclyde. 20 participants were recruited, half were journalists and half were students. The average participant age was 30-40 years. In terms of education background, 70% of participants had or were pursuing a postgraduate degree, 25% had undergraduate degree and 5% had a Higher National Diploma<sup>1</sup> (HND). 85% of participants had working experience in journalism with 30% having more than 10 years experience and the average type of journalist was the daily news reporter.

Participants had a chance to perform the tasks using the interface. A Latin square is used to construct the experimental design. It allows us to evaluate the same topic using different setups. The order of topics assigned in

---

<sup>1</sup> A Higher National Diploma (HND) is a higher education qualification in the United Kingdom. This qualification can be used to gain entry into universities, and is considered equivalent to the first two years of a university course.

the Tracking tasks and the order of clusters given in the Detection task were rotated to avoid any learning and fatigue factor. This principle also applied during the Training Session. Topic 1 (Oprah Lawsuit) for example has a chance to be the first, second, third and fourth in order, during the Tracking task. The clusters assigned in the Detection task were invisible when participants performed the Tracking task to avoid any intersection of clusters. This is important because the intersection will affect the participants' performance because they might have come across the clusters used in the Detection task during the Tracking task. Therefore it will make the tasks challenging to the participant. The selection of topics and clusters given in the user experiment has a combination of good and poor clustering performance based on the F1-measure. This is important to justify whether the *iEvent* interface helps the participant to perform the TDT tasks even though they were given a bad cluster to track or a bad topic to detect.

#### 4.1. User tasks with *iEvent*

The participants were given two types of tasks: Tracking and Detection. In the Tracking task, the participant has to track the cluster that contains the given topic and show that the system provides a sufficient amount of information on the event. This is in line with the journalist's task of reporting news. There are two sub activities in this task which are Reporting and Profiling. Reporting requires the participant to write an article of a topic by drafting the important facts while in Profiling, the participant has to make a profile of a story by providing the important keywords. Meanwhile in the Detection task, the participant has to identify the topic dealt with by a specific cluster. This is in line with the journalist's task of identifying some important event that happened on a specific day.

There were eight topics for the Tracking task and four clusters for the Detection task in two sessions. After completion of the tasks, participants completed a questionnaire about using the interface. They were given two hours to attempt the entire Tracking task and were given 15 minutes for each topic. While in the Detection task, they had 40 minutes to complete it and were given 10 minutes for each cluster. The whole user experiment took about 2 hours 40 minutes to 3 hours excluding a short training session. The time assigned to each task was sufficient based on the feedback received from the Pilot Test conducted. Each participant session lasted between one and one-and-a-half hours, depending on the time taken to complete the assigned tasks and the time taken by the participant to complete the questionnaires. Participants were offered a short break (5 to 15 minutes) after the first session. Each session consisted of the following steps:

- a. Participants were welcomed and asked to read the introduction to the experiment provided on an Information Sheet. This set of instructions was developed to ensure that each participant received precisely the same information. Participants could retain the information sheet after the experiment.
- b. The participants were given a short overview of what the experiment would entail. We also explained our role in this experiment i.e. to observe participant interaction with the systems, to provide participants with technical support and to remind participants on the time taken in performing the tasks.
- c. Participants were then asked to complete an *Entry* questionnaire. This provided background information on the participant's education, work experience and previous experience of news network tools used.
- d. Participants were given a demonstration of the *iEvent* interface with both setups by following the experimental design as shown in Figure 9. This includes the features available on the interface, followed by a training session. The training session was the same for all participants using both setups. It gave participants a chance to familiarize themselves with the interface. Participants could ask questions or ask for general assistance at any time during the session.
- e. Tracking task
  - i. Once comfortable with *iEvent*, participants were asked to perform the Tracking task. There are two sub-activities in this task which are Reporting and Profiling. Reporting requires the participant to write an article on a topic by drafting the important facts while in Profiling; the participant had to make a profile of a story by providing important keywords. They were given 15 minutes to search and could stop early if they were unable to find any more relevant information. Searching in this experiment refers to identifying the cluster related to a given topic.

- ii. After completing the search (successfully or otherwise), the participant was asked to complete the questionnaire.
- iii. The remaining tasks were given to the participant in the second session using a different setup, following steps 5a-b. Participants were offered a short break after the first session.
- f. Detection task
  - i. Participants were given 10 minutes to search and could stop early if they were unable to find any more relevant information. Searching in this experiment refers to detecting the topic for a given cluster.
  - ii. After completing the search (successfully or otherwise), the participant was asked to complete the questionnaire.
  - iii. The remaining tasks were given to the participant in the second session using a different setup, following steps 6a-b.
- g. At the end of the experiment, participants were asked to complete the post-evaluation questionnaire and an informal post-experiment interview was conducted. The post-evaluation questionnaire compares participants' performance between setups.

The start time to perform the task was defined as the moment where the participant started using *iEvent* and 15 minutes later is defined as the end time to perform the Tracking task for a topic. This principle also applies to the Detection task with 10 minutes allowed for each cluster. Since *iEvent* was new and unfamiliar to participants, they received a training session on how to use it. A short time, around 30 minutes, was allocated for training at the start of the experiment. In all cases this appeared sufficient for participants to familiarize themselves with *iEvent*. The training session was broken down into a series of stages:

- a. The purpose of *iEvent* were explained i.e., to cluster news stories into the same group of events or topics by visualizing the clusters.
- b. Participants were introduced to the interface components and features that appeared in *iEvent* interface. We also printed a screenshot of the interface to describe the components and features of *iEvent* which we believe would help the participants to understand how *iEvent* works and perform the task better.
- c. Participants were given a live demonstration of each setup using the same topic *General Motors Strike* for the Tracking task and *Cluster 24 (Pope visits Cuba)* for the Detection task.
- d. A training session was issued and participants were given the chance to attempt the Tracking and the Detection tasks. It gave participants an opportunity to use *iEvent* in a realistic news tracking and detection context and become accustomed to the interface features.
- e. The training session stopped once participants felt comfortable using *iEvent*.
- f. Participants were allowed to comment or ask questions at any point during the session.

## 5. Result

Participant performance was analysed to identify the effectiveness of *iEvent* interface in facilitating them to perform the Tracking and the Detection tasks. During the experiment 240 tasks were performed. 160 (66.67%) of these tasks were Tracking and 80 (33.33%) of tasks were Detection.

Findings revealed that 70% of the participants liked *iEvent* and 50% of the participants like to use *iEvent* in both tasks. A possible explanation for these results might be the participants' success in performing both tasks (as reported in Section 5.1.1 and 5.2). 20% of participants disliked *iEvent* and 10% of participants were not sure. Those who disliked *iEvent* were all journalists that have an average age of 30-40 years and average working experience of more than 10 years. From the interview session, these participants had previously used news network tools such as [PressDisplay.com](http://www.pressdisplay.com)<sup>2</sup>, [PaidContent.org](http://www.paidcontent.org)<sup>3</sup> and [Google Fast Flip](http://www.google.com/fastflip)<sup>4</sup>. Thus they have a high

---

<sup>2</sup> <http://www.pressdisplay.com/pressdisplay/viewer.aspx>

expectation when using *iEvent*. 10% were not sure, although they mentioned some interesting features of *iEvent*. However they disliked the fact that they have to scroll and mouse over the Cluster View to find the topic in the Tracking task.

Participants were asked about their topic familiarity and topic interest before they started using *iEvent*. Each degree of agreement is given a numerical value from one to five where a higher value corresponds to more familiarity. Findings revealed that there were no statistical significance difference between the topics and participants topic familiarity (Mann-Whitney Test,  $p=0.483$ ). This indicated that the topics given during the experiment did not influence the participants' topic familiarity. In addition, the participants were not familiar with the topics given in the Tracking task (mean=2.01 sd=1.03). There were also no statistical significance difference between the participants and their topic interest (Mann-Whitney Test,  $p=0.842$ ). Their topic interest was average (mean=3.27 sd=1.09). This is a good indication to the experiment since the participants are not affected by external factors such as their topic familiarity and topic interest because we compare their results after using *iEvent*. Wilcoxon Signed Ranks test proved that there was a statistical significance difference in both topic familiarity and topic interest before and after using *iEvent*. The mean for topic familiarity (before=2.01, after=3.26) and topic interest (before=3.27, after=3.63) was increased after using *iEvent*.

There was an increasing percentage (5 times higher) for participants who were familiar with the topic before (8%) and after (46%) using *iEvent*. The percentage decreased for participants who were not familiar with the topic before (69%) and after (27%) using *iEvent*. 69% of participants were not familiar with the topic because the collection used was in year 1998 (TDT2 and TDT3 corpus). Thus it supports the evaluation that *iEvent* influenced their topic familiarity and topic interest. If participants were given more recent topics, they might be familiar with it and probably have better knowledge of the topics that will influence their performance in the Tracking task.

For the topic interest, there was an increasing percentage for participants who were interested with the topic before (50%) and after (59%) using *iEvent*. Meanwhile the percentage of participants who were not interested in the topic before decreased to 10% after using *iEvent*. These indicate that the participants were more familiar and more interested with the topics in the Tracking task after using *iEvent*.

A Mann-Whitney Test confirmed that there was no statistical significance difference ( $p=0.492$ ) in topic interest before using *iEvent* across setups. However there was a statistical significance difference in topic interest after using *iEvent* across setups (Mann-Whitney Test,  $p=0.003$ ). The participants were more interested with a topic in the Tracking task after using Setup 2 (mean=3.81 sd=1.032). Participants found using Setup 2 had enhanced their topic interest. There is a ratio of 7 participants to 1 who found that they were more interested in a topic after using Setup 2. They found that using Setup 2 of *iEvent* had significantly enhanced their topic interest with 46.3% of participants agreeing that they were interested (scale 4) with a topic. These results indicate that the participants were more familiar with the topics in the Tracking task after using *iEvent*. They were also more interested in the topics in the Tracking task after using Setup 2 of *iEvent*.

The participants were given an entry questionnaire before they performed the Tracking and Detection task. They were asked to list out the news network, tools or search engines used. They mostly used Google (95%) and BBC news (90%) as their main news networks tools. Participants were also asked to rate their experience in using the news network tools. 45% of participants found the news network tools that they used was easy (scale 4) (mean=4.05 sd=0.759). 50% of the participants found that the news network tools was relaxing (scale 4) (mean=3.45 sd=0.887). 35% of participants agreed that the news network tools was neither simple nor complex (scale 3) (mean=3.10 sd=0.912). Based on participants' satisfaction, 35% of them were dissatisfied (scale 2) and found that the news network tools was average (scale 3) (mean=3.00 sd=0.918). Finally based on participants' interest, 45% of them found that the news network tools was averagely interesting (scale 3) and interesting (scale 4) (mean=3.35 sd=0.617). Participants interviewed mentioned the Google style of searching contribute to the ease of use of the news network tools, thus making the searching process more relaxed and interesting.

---

<sup>3</sup> <http://paidcontent.org/>

<sup>4</sup> <http://fastflip.googlelabs.com/>

## 5.1. Tracking task

Several analyses were performed on the captured data. The following sections present the findings. First, the overall participants' opinions of *iEvent* are examined. Next, we investigate the participants' performance using *iEvent* in the Reporting task i.e. the amount of news written. Then, we investigate the features of *iEvent* that participants perceived as useful, effective, helpful and interesting.

### 5.1.1 Overall opinions

The *iEvent* interface that participants perceived as easy, relaxing, simple, satisfying and interesting during the Tracking task were analysed.

#### *Easy*

A ratio of 3 participants to 1 found that *iEvent* was easy to use (mean=3.57 sd=1.079) with 38.8% of participants agreeing that it was easy (scale 4). During the interview session, the participants informed us that *iEvent* was easy because it has a structured and clear components; Cluster, Document and Term Views, thus making it easy to use. There was a statistical significance difference in participants' opinions (easy) across setups (Mann-Whitney Test,  $p=0.004$ ). 45.0% of participants agreed that Setup 2 (mean=3.85 sd=0.828) was easy (scale 4). Interestingly there were 67.5% of participants who found that Setup 2 was easier compared to 5% who found it difficult. This indicates that 14 participants found that using Setup 2 of *iEvent* made the Tracking task easier. The participants interviewed agreed that Setup 2 provides significant information on important named entities which makes the Tracking task easier.

#### *Relaxing*

4 participants to 1 found that *iEvent* was relaxing (mean=3.50 sd=0.883). 38.8% of participants agreed that it was relaxing (scale 4). Participants interviewed again associated the relaxing factor with the structured and clear components of *iEvent* which also support the perceived easiness of using *iEvent* to perform the Tracking task. There was a statistical significance difference in participants' opinions (relaxing) across setups (Mann-Whitney Test,  $p=0.003$ ). 41.3% of participants agreed that Setup 2 (mean=3.71 sd=0.860) was relaxing (scale 4). 60% of participants who found that Setup 2 was more relaxing compared to 7.5% who found it stressful. This indicates that 8 participants found using Setup 2 of *iEvent* makes the Tracking task more relaxing.

#### *Simple*

There were 2 participants to 1 who found that *iEvent* was simple (mean=3.30 sd=1.039) with 37.5% of participants indicating that it was simple (scale 4). Mann-Whitney Test confirmed that there was no statistical significance difference on participants' opinion in simple ( $p=0.840$ ) in conjunction with the setups. The participants interviewed also relate this opinion with the clear and structured components of *iEvent* but there were also suggestions to revise the layout of *iEvent* especially the Cluster View to be vertical instead of horizontal. Thus the layout issue would be interesting for the future work on *iEvent*.

#### *Satisfying*

There were 5 participants to 1 who found *iEvent* to be satisfying (mean=3.50 sd=0.854) with 40% of participants agreeing that it was satisfying (scale 4). A Mann-Whitney Test confirmed that there was no statistical significance difference on participants' opinion in satisfying ( $p=0.500$ ) in conjunction with the setups.

We measure participants' satisfaction by analysing their agreement on enough information gathered during the Tracking task and the Reporting task results. We believe participants were satisfied with *iEvent* if they managed to perform the Tracking task by receiving enough information for a topic and they managed to report the story assigned by tracking the correct cluster. Participants were deemed to be satisfied if they found the information that they needed. Further analysis showed that 39.4% of participants agreed that they have gathered enough

information using *iEvent* (mean=3.50 sd=1.082) during the Tracking task. There were 3 participants to 1 agreed that they had gathered enough information using *iEvent*.

Interestingly, the satisfying factor was also related to the high percentage of correct clusters to be tracked. This gives strong evidence that *iEvent* mostly helped to facilitate the participants in tracking the correct cluster (mean=3.87 sd=0.49). We classify the correctness of cluster as being tracked into four categories:

- i. none- where participants did not provide any information or they did not complete the task
- ii. wrong- where participants tracked the wrong cluster.
- iii. partially correct- where participant list out the minor cluster as their main finding.
- iv. correct- where participant list out the major cluster as their main finding.

The entire Tracking task was successful with 91.9% of task being correct and 4.4% being partially correct. There were 2 participants (1.3%) who did not complete the task on topic National Tobacco Settlement. Participants were using Setup 2 (experimental setup) which displays information on named entities (e.g. Congress, Clinton) however they were looking for the term tobacco. This is the reason participants spent the full 15 minutes allocated and were still not able to find the correct cluster. There were 4 participants (2.5%) who were wrong about the topic Mobil-Exxon Merger. Participants were confused with this topic when they were using Setup 1 (baseline setup) from the term merge which also highlights the cluster on topic Microsoft Merger. These uncompleted and wrong tasks represented just a small percentage compared to the successful tasks. This proved that *iEvent* managed to facilitate the participants in performing well in the Tracking task.

### *Interesting*

This opinion of *iEvent* received the highest ratio with 9 participants to 1 finding that *iEvent* was interesting (mean=3.89 sd=0.956). 34.4% of participant agreed that it was very interesting (scale 5). There was a statistical significance difference in participants' opinion (interesting) across setups (Mann-Whitney Test,  $p < 0.05$ ). Setup 2 (mean=4.23 sd=0.779) received the highest percentage with 43.8% of participants agreeing that it was very interesting (scale 5). Surprisingly none of the participants found that Setup 2 was boring and 78.8% of participants found that it was more interesting. This indicates that the participants found using Setup 2 of *iEvent* makes the Tracking task more interesting than Setup 1.

The participants interviewed agreed that the histogram with the timeline was one of the most interesting features and from our observation participants mostly used it as their main strategy when performing the Tracking task. One of the participants quoted that "This is a new paradigm of monitoring news in journalism and absolutely interesting".

### 5.1.2 Reporting task

This section reports the findings of participants' performance during the Reporting task as one of the sub activities of the Tracking task. We analyse the number of lines that participants wrote. This is an important measure of how effective *iEvent* is in providing information to the participants. The more they wrote indicated that the participant received enough information and was able to deliver it in a written form. We also analysed the number of lines that participants wrote across setups. There was no statistical significance difference on the amount of news written in conjunction with the setups (Mann-Whitney Test,  $p = 0.434$ ) and no statistical significance difference on the amount of news written for different topics (Mann-Whitney Test,  $p = 0.202$ ). These indicate that the participants managed to write the amount of news equally using both setups and they managed to write the amount of news equally for every topic given in this experiment.

Findings revealed that the participants wrote on average nine lines using *iEvent* (mean=9.44 sd=6.455). There was a statistical significance difference in the amount of news written in conjunction with the type of participants (Mann-Whitney Test,  $p < 0.05$ ). The journalists (mean=7.09 sd=5.45) wrote less than the students (mean=11.79 sd=6.56). The reason was the journalists were more selective and critical when writing news.

*iEvent* also facilitated the participants to report the correct news (mean=3.80 sd=0.708). 91.3% of participants manage to report the correct news and interestingly none of the participants provided the wrong

information. There was no statistical significance difference on the amount of correct news written in conjunction with the setups (Mann-Whitney Test,  $p=0.651$ ) and this indicated that the participants managed to write the amount of correct news equally using both setups. We classify the correctness of news written into four categories:

- i. none- where participants did not provide any information or they did not complete the task.
- ii. wrong- the news written did not match the topic.
- iii. partially correct- part of the news written match the topic.
- iv. correct- the news written match the topic.

### 5.1.3 Features

In this section, we analyse each feature of *iEvent* and assessed which setup participants perceived as useful during the Tracking task as shown in Table 2.

Table 2

Percentage of participants who perceived the features of *iEvent* as useful in the Tracking task

FEATURES	Scale (%)					(-) ive		(+) ive	Ratio
	1	2	3	4	5				
<b>CLUSTER VIEW (CV)</b>									
cluster labelling	5.6	15.6	20.0	<b>35.6</b>	23.1	21.2	58.7	3:1	
top terms	3.8	18.1	27.5	<b>38.1</b>	12.5	21.9	50.6	2:1	
cluster visualisation	3.8	3.1	26.9	<b>36.3</b>	30.0	6.9	66.3	<b>10:1</b>	
<b>DOCUMENT VIEW (DV)</b>									
histogram with the timeline	8.1	3.8	21.3	31.9	<b>35.0</b>	11.9	66.9	<b>6:1</b>	
document content	1.3	11.9	23.8	<b>38.8</b>	24.4	13.2	63.2	5:1	
<b>TERM VIEW (TV)</b>									
keyword approach	5.6	12.5	24.4	<b>28.8</b>	<b>28.8</b>	18.1	57.6	3:1	
histogram with the timeline	6.3	6.3	16.3	<b>44.4</b>	26.9	12.6	71.3	<b>6:1</b>	

(-)ive=scale 1, 2; (+)ive=scale 4,5

(scale from 1 to 5, higher=better; highest value shown in bold)

### Useful

The highest ratio for this opinion was for the ‘CV: cluster visualisation’ feature. 10 participants to 1 found that this feature was useful in the Tracking task (mean=3.86 sd=1.008). 36.3% thought the ‘CV: cluster visualisation’ feature was useful (scale 4) as shown in Table 2. The size and the density of the clusters contained in this feature allow the participant to identify how many topics are in each cluster, such that clusters with large size and high density indicate a high number of documents where the distribution of the documents are over a long period of time.

There were also two features perceived as useful by the participants with ratio 6:1 which were the ‘DV: histogram with the timeline’ (mean=3.82 sd=1.192) and the ‘TV: histogram with the timeline’ (mean=3.79 sd=1.099). 44.4% thought the ‘TV: histogram with the time line’ feature was useful (scale 4) and 35% found the ‘DV: histogram with the timeline’ feature to be very useful (scale 5). These features allow the participant to see the document and the term occurrence for a specific date. Topic such as Jonesboro Shooting did mentioned the

date 29<sup>th</sup> of April as the hearing case and using these features was an advantage in reporting the outcome of the trial.

Moreover, 60% of the participants agreed that the 'DV: document histogram with the timeline' feature is a way to analyse discourse analysis. Discourse analysis is important in journalism as it studies the information flow in a press article. These findings support the reason why participants gave a high score (scale 5) on the usefulness of this feature in Tracking task. In addition 40% of participants agreed that the document histogram with the timelines was the best feature of *iEvent*.

There was a statistical significance difference on the 'CV: cluster visualisation' feature (Mann-Whitney Test,  $p=0.002$ ) and the 'DV: histogram with the timeline' feature (Mann-Whitney Test,  $p<0.05$ ) between students and journalists. These 2 features were significantly more popular among students compared to journalists. Students found the 'CV: cluster visualisation' feature was useful (scale 4) and the 'DV: histogram with the timeline' was very useful (scale 5).

Findings also revealed that there was a statistical significance difference on the 'CV: cluster labelling' feature across setups (Mann-Whitney Test,  $p<0.05$ ). The 'CV: cluster labelling' feature in Setup 2 of *iEvent* was more useful (mean=3.94 sd=0.919) compared to Setup 1 (mean=3.16 sd=1.267). 9 participants to 1 found that 'CV: cluster labelling' feature in Setup 2 of *iEvent* was perceived as significantly useful with 38.8% of participants agreeing that it was useful (scale 4). This indicates that the participants found 'CV: cluster labelling' feature in Setup 2 of *iEvent* more useful than Setup 1.

### *Effective*

11 participants to 1 found that the 'DV: document content' feature was effective in the Tracking task (mean=4.41 sd=0.968). It can be seen from the data in Table 3 that the most striking results were 45% found the 'DV: document content' feature was very effective (scale 5). Further analyses on the interaction logs among the successful Tracking tasks proved that there was high activity using the 'DV: document content' feature with 71.4% of participants using it. This indicates that this feature was effective in facilitating the participant in tracking the correct cluster.

There were also two further features perceived as effective by the participants. 7 to 1 participant found that the 'DV: histogram with the timeline' (mean=3.97 sd=1.096) was effective. Moreover 39.4% found this feature to be very effective (scale 5). 6 participants to 1 found that the 'CV: cluster visualisation' feature (mean=3.73 sd=0.951) was effective too with 36.9% rating it as effective (scale 4).

Table 3

Percentage of participants who perceived the features of iEvent as effective in the Tracking task

FEATURES	Scale (%)					(-) ive (%)		Ratio
	1	2	3	4	5	(-) ive	(+) ive	
<b>CLUSTER VIEW (CV)</b>								
cluster labelling	1.9	19.4	28.8	<b>33.1</b>	16.9	21.3	50.0	2:1
top terms	0.6	14.4	25.0	<b>37.5</b>	22.5	15.0	60.0	4:1
cluster visualisation	1.3	8.1	30.6	<b>36.9</b>	23.1	9.4	60.0	6:1
<b>DOCUMENT VIEW (DV)</b>								
histogram with the timeline	4.4	5.6	18.1	32.5	<b>39.4</b>	10.0	71.9	7:1
document content	1.3	5.6	16.3	31.9	<b>45.0</b>	6.9	76.9	<b>11:1</b>
<b>TERM VIEW (TV)</b>								
keyword approach	3.8	8.1	31.9	<b>35.6</b>	20.6	11.9	56.2	5:1
histogram with the timeline	6.9	13.8	13.8	<b>41.3</b>	24.4	20.7	65.7	3:1

(-)ive=scale 1, 2; (+)ive=scale 4,5

(scale from 1 to 5, higher=better; highest value shown in bold)

There was a statistical significance difference in perception of the ‘CV: cluster visualisation’ feature (Mann-Whitney Test,  $p=0.001$ ) and ‘DV: histogram with the timeline’ feature (Mann-Whitney Test,  $p<0.05$ ) between students and journalists. The ‘CV: cluster visualisation’ feature was popular among the students. The students interviewed mentioned that it was effective since it gave them quick information on the number of documents and the density. They mentioned that clusters with large size and high density had more than 1 topic so they preferred to investigate on the clusters with medium size with medium or high density. The ‘DV: histogram with the timeline’ was popular among the journalists. The journalists interviewed claimed that the ‘DV: histogram with the timeline’ was effective since there were critical when looking for a very specific information. This feature allows them to answer the question on ‘*when was the event?*’.

Findings also revealed that there was a statistical significance difference on the ‘CV: cluster labelling’ feature across setups (Mann-Whitney Test,  $p=0.008$ ). The ‘CV: cluster labelling’ feature in Setup 2 of iEvent was more effective (mean=3.66 sd=0.927) compared to Setup 1 (mean=3.21 sd=1.110). 5 participants to 1 found that ‘CV: cluster labelling’ feature in Setup 2 of iEvent was perceived as significantly effective with 41.3% of participants agreed that it was effective (scale 4). This indicates that the participants found ‘CV: cluster labelling’ feature in Setup 2 of iEvent more effective than Setup 1.

### Helpful

There were 12 participants to 1 who found that the ‘TV: histogram with the time line’ feature was helpful in the Tracking task. The participants interviewed mentioned that they could see the specific occurrence for a specific term. The topic *Jonesboro Shooting* for example, allows them to scan the timeline for significant terms such as *Mitchell Johnson* and *Andrew Golden*. Thus 38.1% of the ‘TV: histogram with the time line’ feature was perceived to be very helpful (scale 5) as shown in Table 4.

Table 4

Percentage of participants who perceived the features of *iEvent* as helpful in the Tracking task

FEATURES	Scale (%)					(-) ive (%)		Ratio
	1	2	3	4	5	(-) ive	(+) ive	
<b>CLUSTER VIEW (CV)</b>								
cluster labelling	1.3	15.0	27.5	<b>41.3</b>	15.0	16.3	56.3	3:1
top terms	0.6	9.4	<b>31.3</b>	30.0	28.8	10.0	58.8	6:1
cluster visualisation	0.6	9.4	31.3	<b>35.6</b>	23.1	10.0	58.7	6:1
<b>DOCUMENT VIEW (DV)</b>								
histogram with the timeline	1.3	11.3	26.9	<b>30.6</b>	30.0	12.6	60.6	5:1
document content	2.5	8.8	22.5	28.8	<b>37.5</b>	11.3	66.3	6:1
<b>TERM VIEW (TV)</b>								
keyword approach	0.6	15.0	20.6	<b>38.1</b>	25.6	15.6	63.7	4:1
histogram with the timeline	0.6	5.0	25.0	31.3	<b>38.1</b>	5.6	69.4	<b>12:1</b>

(-)ive=scale 1, 2; (+)ive=scale 4,5

(scale from 1 to 5, higher=better; highest value shown in bold)

There were three features that were perceived to be helpful by 6 participants to 1. They were the ‘CV: top terms’ feature (mean=3.77 sd=0.992), the ‘CV: cluster visualisation’ feature (mean=3.71 sd=0.948) and the ‘DV: document content’ feature (mean=3.90 sd=1.083).

There was a statistical significance difference on the ‘TV: histogram with the time line’ feature (Mann-Whitney Test,  $p=0.029$ ) between the topics. This feature was particularly popular for Topic 7 (German Train Derails) because it requires the participant to report the accident where timeline is an important feature to track the story of the accident, investigation and the consequences from the accident. Further analysis of the interaction logs for Topic 7 proved that the participants were using this feature more frequently for this topic, with 10.2% of activity compared to an average usage of 9%. There was a statistical significance difference on the ‘CV: cluster visualisation’ feature (Mann-Whitney Test,  $p<0.05$ ) and the ‘DV: histogram with the timeline’ feature (Mann-Whitney Test,  $p<0.05$ ) between students and journalists. The two features were more popular among the students compared to the journalists. They found the features were helpful (scale 4).

Findings also revealed that there was a statistical significance difference on two features across setups. The ‘CV: top terms’ feature (Mann-Whitney Test,  $p=0.033$ ) and ‘TV: keyword approach’ features (Mann-Whitney Test,  $p=0.011$ ) in Setup 2 of *iEvent* were more helpful than Setup 1. 17 participants to 1 found that the ‘CV: top terms’ feature in Setup 2 of *iEvent* was perceived as significantly helpful with 35.0% of participants agreeing that it was very helpful (scale 5). 9 participants to 1 found that the ‘TV: keyword approach’ feature in Setup 2 was perceived as significantly helpful with 35.0% of participants agreeing that it was helpful (scale 4). This indicates that the participants found the ‘CV: top terms’ and ‘TV: keyword approach’ features in Setup 2 of *iEvent* more helpful than Setup 1.

### *Interesting*

It was apparent from Table 5, that there were three features that participants perceived as interesting which have a high ratio (more than 10:1) compared to other features.

Table 5

Percentage of participants who perceived the features of iEvent as interesting in the Tracking task

FEATURES	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-) ive	(+) ive	
<b>CLUSTER VIEW (CV)</b>								
cluster labelling	.6	5.6	25.6	25.6	<b>42.5</b>	6.2	68.1	<b>11:1</b>
top terms	1.3	5.6	20.6	33.8	<b>38.8</b>	6.9	72.6	<b>11:1</b>
cluster visualisation	4.4	6.9	<b>36.3</b>	30.0	22.5	11.3	52.5	5:1
<b>DOCUMENT VIEW (DV)</b>								
histogram with the timeline	.6	3.1	<b>43.1</b>	25.0	28.1	3.7	53.1	<b>14:1</b>
document content	4.4	8.8	27.5	26.9	<b>32.5</b>	13.2	59.4	5:1
<b>TERM VIEW (TV)</b>								
keyword approach	4.4	16.3	20.0	29.4	<b>30.0</b>	20.7	59.4	3:1
histogram with the timeline	1.3	12.5	16.3	<b>38.8</b>	31.3	13.8	70.1	5:1

(-)ive=scale 1, 2; (+)ive=scale 4,5

(scale from 1 to 5, higher=better; highest value shown in bold)

14 participants to 1 found that the ‘DV: histogram with the timeline’ feature was interesting (mean=4.04 sd=0.983). The participants found that the ‘CV: cluster labelling’ feature (mean=4.04 sd=0.983) was interesting with 42.5% finding it very interesting (scale 5). During the informal interview session, the participants found this feature was very interesting because they received quick information on the topic using the 3 most frequent terms for the cluster. 11 to 1 participants found that the ‘CV: top terms’ feature (mean=4.03 sd=0.968) was interesting too, with 38.8% indicating they felt this feature was very interesting (scale 5).

There was a statistical significance difference on the ‘CV: cluster visualisation’ feature (Mann-Whitney Test,  $p=0.049$ ) between students and journalists. This feature was popular among students since they not only found it effective and helpful but also interesting (scale 4).

Findings also revealed that there was a statistically significant difference on four features across setups. The ‘CV: cluster labelling’ (Mann-Whitney Test,  $p=0.033$ ), ‘CV: top terms’ (Mann-Whitney Test,  $p=0.026$ ), ‘DV: document content’ (Mann-Whitney Test,  $p=0.013$ ) and ‘TV: keyword approach’ features (Mann-Whitney Test,  $p=0.035$ ) in Setup 2 of iEvent was more interesting than Setup 1. There were two features in Setup 2 perceived as very interesting (scale 5) by the participants. They were the ‘CV: cluster labelling’ feature (48.8%) and the ‘CV: top terms’ feature (43.8%). There were also two further features in Setup 2 perceived as interesting (scale 4) by the participants; the ‘DV: document content’ feature (36.3%) and the ‘TV: keyword approach’ feature (37.5%). Surprisingly these four features received a high percentage (70%-82.5%) of participants who found it interesting. This indicates that the participants found the ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features in Setup 2 of iEvent were more interesting than Setup 1.

## 5.2. Detection task

The entire detection task was successful with 85% of task results being correct and 15% of being partially correct. Surprisingly there was no unsuccessful detection task or participants who wrongly detected the topics which proved that iEvent managed to facilitate the participants to perform well in the Detection task. We also classify the correctness of topic detected into four categories:

- i. none- where participants did not provide any information or they did not complete the task
- ii. wrong- where participants detected the wrong topic.

- iii. partially correct-where participant list out the minor topic as their main finding
- iv. correct-where participant list out the major topic as their main finding

Interestingly, there were 11 participants to 1 who found that it was easy to detect the topic in this task. 51.3% found that it was easy to detect the topic (scale 4) and 20% of found it was very easy (scale 5) using *iEvent*. There was no statistical significance difference on the ease of detecting a topic in conjunction with the cluster given (Mann-Whitney Test,  $p=0.735$ ). This proved that although participants were given a combination of good and poor cluster performance, they manage to complete the Detection task and perform well using *iEvent*. Further results from the interaction logs among the successful tasks showed that participants took 4 minutes and 49 seconds (mean of click=39) to perform this task on average, much less that the 10 minutes given to complete the task.

There was a statistical significance difference between participants opinion on the ease to detect a topic and the setups (Mann-Whitney Test,  $p<0.05$ ). 60% of participants agreed that it was easy (scale 4) to detect a topic using Setup 1. Surprisingly none of the participants found that it was hard to detect a topic using Setup 1 and 92.5% found that Setup 1 makes the Detection task easier than Setup 2. Results also show that there was no statistical significance difference on the ease to detect a topic in conjunction with the type of participant (Mann-Whitney Test,  $p=0.477$ ). Both students and journalists found out that *iEvent* assisted them in detecting the topic easily (mean=3.85 sd=0.813).

The highest percentage of features used was the 'CV: top terms' (83.8%) while the lowest was the 'CV: cluster visualisation' (53.8%). The participants were using the 'CV: top terms' feature to get more information when detecting the topics and less on the 'CV: cluster visualisation' feature. A possible explanation for this might be that participants only deal with a specific assigned cluster without having to compare them with other cluster which makes this feature less useful in the Detection task. Further analyses on the interaction log proved that there was a low activity (2.7%) using the 'CV: cluster visualisation' feature.

There were three features which also received a high percentage of use; namely the 'DV: document content' (81.3%), 'TV: keyword approach' (81.3%) and 'DV: histogram with the timeline' (80%) features. Further analyses on the interaction logs proved that there was high activity using the 'DV: document content' feature with 77.6% of participants using it. There was also 3.9% of activity using the 'TV: keyword approach' and 7.2% of activity using the 'DV: histogram with the timeline'.

A possible explanation for this might be that the participants received more information from the 'DV: document content' feature to detect the topic and the 'DV: histogram with the timeline' feature gives an overall view on the distribution of the topics for the specific cluster. Participants could identify how many topics the cluster contains. While the 'TV: keyword approach' feature gives good information on the most frequent terms appearing in the cluster, thus allowing the participant to relate to the topics easier.

## 6. Discussion

We investigate the effectiveness of *iEvent* (*iTDT* interface) in facilitating the journalists in performing the TDT tasks. Thus we set out to determine which features of *iEvent* facilitate the Tracking and the Detection tasks.

This experiment has shown that generally *iEvent* facilitates the participants to perform well with a high percentage of successful Tracking and Detection tasks. Surprisingly there was only 3.8% of unsuccessful tasks in Tracking and none in Detection task. Findings have revealed that the participants were more familiar with the topics in the Tracking task after using *iEvent*. They were also more interested in the topics in the Tracking task after using Setup 2 of *iEvent*.

Table 6

The ratio of each feature across participants' opinion in the Tracking task

	USEFUL	EFFECTIVE	HELPFUL	INTERESTING
<b>CLUSTER VIEW (CV)</b>				
cluster labelling	3:1	2:1	3:1	11:1
top terms	2:1	4:1	6:1	11:1
cluster visualisation	10:1	6:1	6:1	5:1
<b>DOCUMENT VIEW (DV)</b>				
histogram with the timeline	6:1	7:1	5:1	14:1
document content	5:1	11:1	6:1	5:1
<b>TERM VIEW (TV)</b>				
keyword approach	3:1	5:1	4:1	3:1
histogram with the timeline	6:1	3:1	12:1	5:1

(higher=better; highest value shown in bold)

These were the features with the highest ratio that participants perceived as useful, effective, helpful and interesting as shown in Table 6. The results revealed that generally Cluster View was useful and interesting Document View was effective and interesting, and Term View was helpful.

For Cluster View 11 participants to 1 agreed the 'CV: cluster labelling' and 'CV: top terms' features were interesting. The participants found the 'CV: cluster labelling' feature in Setup 2 of *iEvent* was more useful, more effective and more interesting than Setup 1. They also found that the 'CV: top terms' feature in Setup 2 of *iEvent* was more helpful and more interesting than Setup 1. Meanwhile 10 participants to 1 perceived the 'CV: cluster visualisation' feature as useful during the Tracking task, thus this feature received the highest ratio for usefulness. However this feature is the lowest in the Detection task because the participants only deal with one specific cluster to detect the related topics compared to the Tracking task where participants have to track several related clusters. For the Detection task, there was only one feature in Cluster View which received the highest percentage which is the 'CV: top terms'.

For Document View, 14 participants to 1 found that the 'DV: histogram with the timeline' feature was interesting and 11 participants to 1 agreed that the 'DV: document content' was effective. Interestingly these two features also received the highest ratio on the opinions mentioned. The participants also found the 'DV: document content' feature in Setup 2 of *iEvent* was more interesting than Setup 1. It also appears that the Document View was an important component since two features in it - 'DV: histogram with the timeline' and 'DV: document content' - received a high percentage in the Detection task. These indicate that the Document View with the features in it does facilitate the participants in performing both tasks. In addition the 'DV: document content' feature in Setup 2 of *iEvent* was used more frequently compared to Setup 1 during the Detection task.

For Term View, 5 participants to 1 agreed that the 'TV: keyword approach' feature was effective. They also found the 'TV: keyword approach' feature in Setup 2 of *iEvent* was more helpful and more interesting than Setup 1. Finally the 'TV: histogram with the time line' feature received the highest ratio with 12 participants to 1 agreeing it was helpful. Meanwhile in the Detection task, there was only one feature in Term View that received a high percentage which was the 'TV: keyword approach'. In addition the 'TV: keyword approach' feature in Setup 1 of *iEvent* was used more frequently compared to Setup 2 during the Detection task.

Table 7

The comparison of each feature in facilitating the TDT tasks

	TRACKING	DETECTION
<b>CLUSTER VIEW</b>		
cluster labelling	√	√
top terms	√	√
cluster visualisation	√	×
<b>DOCUMENT VIEW</b>		
histogram with the timeline	√	√
document content	√	√
<b>TERM VIEW</b>		
keyword approach	√	√
histogram with the timeline	√	×

As shown in Table 7, the Cluster View with the features in it such as the ‘CV: cluster labelling’ feature and the ‘CV: top terms’ feature facilitate the participants in performing both tasks. Meanwhile the ‘CV: cluster visualisation’ feature only facilitates the participant during the Tracking task but not for the Detection task due to the nature of the task itself. The Document View with the features in it such as the ‘DV: histogram with the timeline’ and the ‘DV: document content’ does facilitate the participants in performing both tasks. The participants found that the ‘TV: keyword approach’ feature was popular in both tasks. This is because participants need to detect the related topics and ‘TV: keyword approach’ feature allows them to see the most frequent terms in the specific cluster assigned. Meanwhile the ‘TV: histogram with the timeline’ feature was popular during the Tracking task. This probably has to do with the participants’ behaviour in trying to match the pattern of the ‘DV: histogram with the timeline’ feature with the ‘TV: histogram with the timeline’ feature. These results indicate that the ‘TV: keyword approach’ feature facilitate the participants in the both tasks while the ‘TV: histogram with the timeline’ feature only facilitates the participants in the Tracking task. We believe these results could be the guideline for the designs of *iTDT* interfaces.

## 7. Conclusion

Overall these findings reveal that the *iEvent* interface generally facilitated the journalists in performing well in the TDT tasks. There were few features in Setup 2 of *iEvent* that facilitated the journalists to perform well in the TDT tasks. This indicates that highlighting the named entities with different colours has affected the participants’ opinions of *iEvent*. Thus it would be interesting to merge Setup 1 and Setup 2 in one interface for the future work on *iEvent*. Therefore journalists have an option to enable the highlighting of named entities in the features of *iEvent*. Some comments were made suggesting revision of the *iEvent* layout which is interesting for the future work on *iEvent*. A key contribution of this work was the design of a novel *iTDT* interface. The findings of this work have fundamental implications for the design of *iTDT* interface and its evaluation. The set of guidelines reported in this paper is useful for future *iTDT* interface design which could enhance the effectiveness of users’ performance in performing the TDT tasks. Thus, the contributions made in this work would benefit the *iTDT* research community.

## 8. References

- [1] Leuski, and J. Allan, Interactive cluster visualization for information retrieval. In: *Proceedings of ECDL'98*, (Springer, Berlin, 1998).
- [2] Leuski, and J. Allan. Filtering: Improving realism of topic tracking evaluation. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 89-96. (ACM Press, New York, 2002).
- [3] Leuski, and J. Allan. Lighthouse: Showing the Way to Relevant Information. In: *Proceedings of the IEEE Symposium on information Visualization 2000 (INFOVIS)*, (IEEE Computer Society, Washington, 2000) 125-129.
- [4] Shneiderman, *Designing the user interface: strategies for effective human-computer interaction*. (Addison-Wesley, Reading, 1997).
- [5] G. J. Jones and S. M. Gabb. A visualisation tool for topic tracking analysis and development. In: *Proceedings of the 25th Annual International ACM SIGIR Conference* (ACM, New York, 2002).
- [6] G. Kumaran and J. Allan, Text Classification and Named Entities for New Event Detection. In: *Proceedings of the 27th Annual International ACM SIGIR Conference* (ACM, New York, 2004).
- [7] H. Chen and L.W. Ku, A NLP & IR approach to topic detection. In: J. Allan (ed.), *Topic Detection and Tracking - Event-based Information Organization*, (Kluwer Academic Publisher, Norwell, 2002) 243-264.
- [8] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. (ACL, New Jersey, 2002).
- [9] J. Allan, S. Harding, D. Fisher, A. Bolivar, Guzman-Lara, S., and Amstutz, P. (2005). Taking Topic Detection From Evaluation to Practice. In: *Proceedings of the 38th Annual Hawaii international Conference on System Sciences (HICSS'05)* (IEEE Computer Society, Washington, 2005).
- [10] J. Allan. *Topic Detection and Tracking: Event-based Information Organization* (Kluwer Academic Publishers, Norwell, 2002).
- [11] J. G. Fiscus and G. R. Doddington, Topic detection and tracking evaluation overview. In: J. Allan (ed.), *Topic Detection and Tracking: Event-Based information Organization*, (Kluwer Academic Publishers, Norwell, MA, 2002).
- [12] J. Makkonen, H. Ahonen-Myka and M. Salmenkivi. Applying Semantic Classes in Event Detection and Tracking. In: *Proceedings of International Conference on Natural Language Processing (ICON 2002)* (Allied Publisher, New Delhi, 2002) 175-183.
- [13] M. Hearst, *Search User Interfaces*, (Cambridge University Press, 2009).
- [14] M. Mohd, F. Crestani and I. Ruthven. Design of an Interface for Interactive Topic Detection and Tracking. In: *Proceedings of the 8th International Conference on Flexible Query Answering Systems (FQAS 2009)*. (Springer, Berlin, 2009) 227-238.
- [15] Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. *Simple Semantics in Topic Detection and Tracking. Information Retrieval*, 7(3-4): 347-368.
- [16] 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*. (ACM Press, New York, 2000) 49-56.
- [17] T. Brants and F. Chen, A system for new event detection. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, (ACM, New York, 2003) 330-337.

- [18] V. Dijk, Discourse Analysis: Its Development and Application to the Structure of News. The *Journal of Communication*, 33(2) (1983) 20-43.
- [19] Y. Yang, J. Zhang, J. Carbonell and C. Jin, Topic conditioned Novelty Detection. In: *Proceedings of the 8th ACM SIGKDD International Conference*, (ACM Press, New York, 2002) 688–693.
- [20] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B.T. Archibald and X. Liu, (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval*, 14(4) (1999) 32-43.
- [21] Z. Kuo, L.J. Zi and W. Gang, New Event Detection Based on Indexing-tree and Named entities. In: *Proceedings of SIGIR'07*, (ACM, New York, 2007).