Monitoring the News: a TDT demonstration system

David Frey, Rahul Gupta, Vikas Khandelwal,
Victor Lavrenko, Anton Leuski, and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

ABSTRACT

We describe a demonstration system built upon Topic Detection and Tracking (TDT) technology. The demonstration system monitors a stream of news stories, organizes them into clusters that represent topics, presents the clusters to a user, and visually describes the changes that occur in those clusters over time. A user may also mark certain clusters as interesting, so that they can be "tracked" more easily.

1. TDT BACKGROUND

The Topic Detection and Tracking (TDT) research program investigates methods for organizing an arriving stream of news stories by the topics the stories discuss.[1, 4, 7, 8] Topics are de£ned to be the set of stories that follow from some seminal event in the world—this is in contrast to a broader subject-based notion of topic. That is, stories about a particular airline crash fall into one topic, and stories from other airline crashes will be in their own topics.

All organization is done as stories arrive, though variations of the task allow £nal organizational decisions to be postponed for minutes, hours, or even days. The formal TDT evaluation program includes the following research tasks:

1. Segmentation is used to separate a television or radio program into distinct news stories. This process is not needed for newswire services, since those stories arrive pre-segmented.

2. Detection is the task of putting all arriving news stories into bins that represent broad news topics. If a new topic appears in the news, the system must create a new bin. Neither the set of bins nor the total number of them is known in advance. This task is carried out without any supervision—i.e., the system never knows whether or not the stories it is putting together actually belong together.

3. Tracking is the task of £nding all stories that follow are on the same topic as an initial small set. This task is different from detection in that the starting stories are *known* to be on the same topic. Typically tracking is evaluated with 2-4 on-topic stories.

.

The TDT research workshops also include a few other tasks (£rst story detection, and story link detection). TDT has also inspired other event-based organization methods, including automatic timeline generation to visualize the temporal locality of topics[10], and the identi£cation of new information within a topic's discussion[3].

This demonstration system illustrates event-based news organization by visualizing the creation of, changes within, and relationships between clusters created by the detection task. It leverages the segmentation results so that audio stories are distinct stories, but does not directly visualize the detection. Tracking is implicity presented by allowing clusters to be marked so that they receive special attention by the user.

2. ARCHITECTURE

The TDT demonstration system is based upon Lighthouse, an interactive information retrieval system developed by Leuski.[6] Lighthouse provides not only a typical ranked list search result, but a visualization of inter-document similarities in 2- or 3-dimensions. The user interface is a Java client that can run as an application or an applet. Lighthouse uses http protocols to send queries to a server and receive the ranked list, summary information about the documents, and the visualization data.

The TDTLighthouse system requires a TDT system running in the background. In this version of the demonstration, the TDT system is only running the segmentation and detection tasks described above. Stories arrive and are put into clusters (bins).

The TDTLighthouse client can query its server to receive up-to-date information about the clusters that the TDT system has found. The server in turn queries the TDT system to get that information and maintains state information so that changes (cluster growth, additional clusters, etc.) can be highlighted.

3. DEMONSTRATION DATA

The data for this demonstration was taken from the our TDT 2000 evaluation output on the TDT cluster detection task [8]. The sytem is running on the TDT-3 evaluation collection of news articles, approximately 40,000 news stories spanning October 1 through December 31, 1998.

We simulated incremental arrival of the data as follows. At the end of each day in the collection, we looked at the incremental output of the TDT detection system. At this point, every story has been classi£ed into a cluster. Every story seen to date is in one of the clusters for that day, even if the cluster has the same contents as it did yesterday.

The demonstration is designed to support text summarization tools that could help a user understand the content of the cluster. For our purposes, each cluster was analyzed to construct the following information:

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2001		**00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Monitoring the News: a TDT demonstration system	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts,Center for Intelligent Information Retrieval,Department of Computer Science,Amherst,MA,01003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**5**	

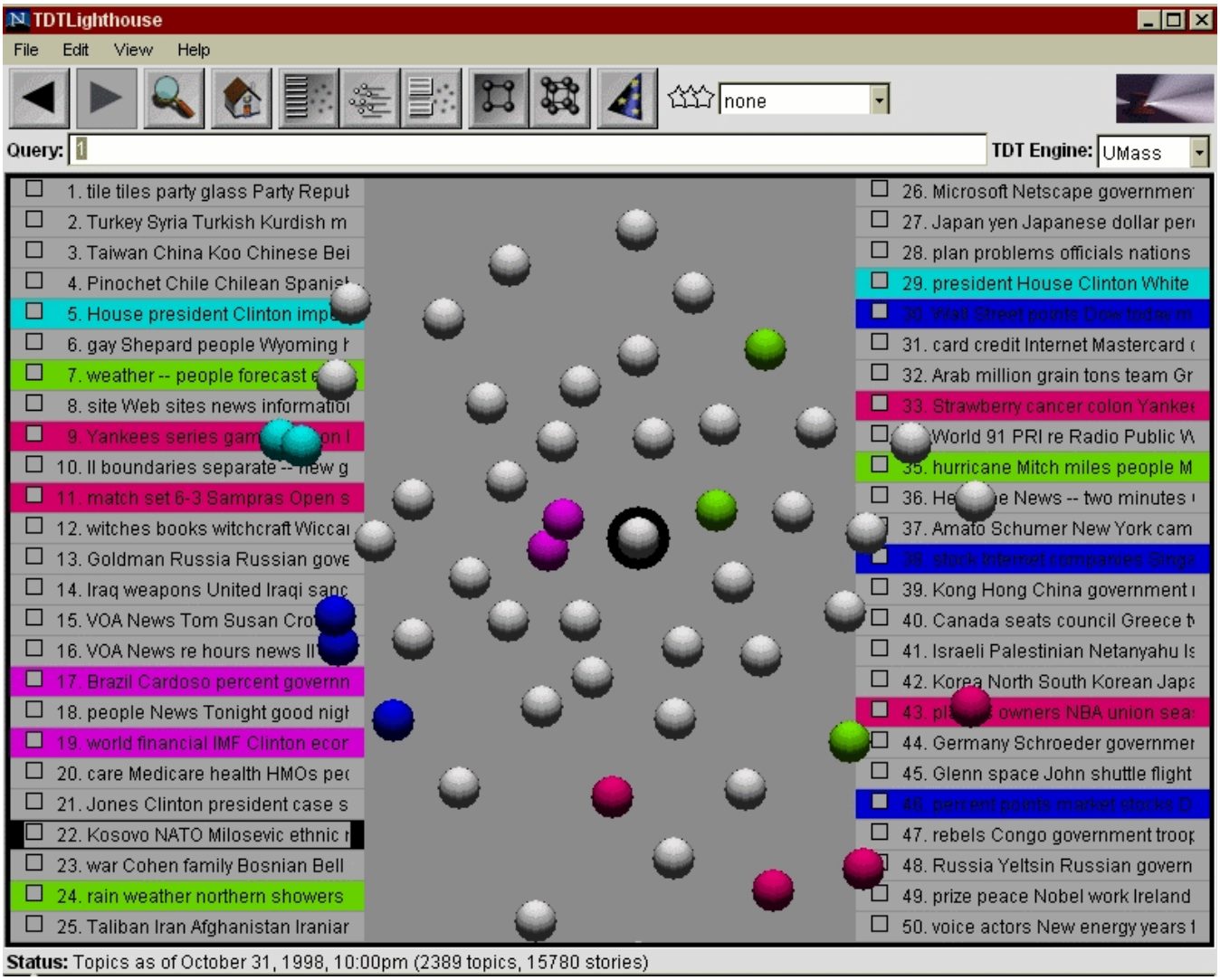

Figure 1: TDT demonstration system running on TDT-3 data, approximately four weeks into the collection.

1. The *title* was generated by selecting the 10 most commonly occurring non-stopwords throughout the cluster. A better title would probably be the headline of the most "representative" news story, though this is an open research question.

2. The *summary* was generated by selecting the £ve sentences that were most representative of the entire cluster. Better approaches might generate a summary from the multiple documents [9] or summarize the changes from the previous day [5, 2].

3. The *contents* of the cluster is just a list of every story in the cluster, presented in reverse chronological order. Various alternative presentations are possible, including leveraging the multimedia (radio and television) that is the basis for the TDT data.

The demonstration system was setup so that it could move from between the days. All of the input to the client was generated automatically, but we saved the information so that it could be shown more quickly. It typically takes a few minutes to generate all of the presentation information for a single day's clusters.

4. DEMONSTRATION SYSTEM

Figure 1 shows the client window. This snapshot shows the system on October 31 at 10:00pm, approximately four weeks into the data. The status line on the lower-left shows that at this point the system has already encountered almost 16,000 stories and has broken them into about 2400 topic clusters.

The system is showing the 50 topics with the largest number of stories. The ranked list (by size) starts on the upper-left, shows the £rst 25, and the continues in the upper-right. The "title" for each of those topics is generated in this case by the most common words within the cluster. Any system that does a better job of building a title for a large cluster of stories could be used to improve this capability.

In addition to the ranked list of topics, the system computes inter-topic similarities and depicts that using the spheres in the middle. If two topics are highly similar, their spheres will appear near each other in the visualization. This allows related topics to be detected quickly. Because the 50 largest topics are shown, the topics are more unalike than they would be with a wider range, but it is still possible to see, for example, that topics about the Clinton presidency are near each other (the cyan pair of spheres overlapping rank number 9, topic rank numbers 5 and 29). The spheres and the ranked list are tightly integrated, so selecting one causes the other to be highlighted.

Topics can be assigned colors to make them easier to pick out in future sessions. In this case, the user has chosen to use the same color for a range of related topics—e.g., red for sports topics, green for weather topics, etc. The color selection is in the control of the user and is not done automatically. However, once a color is assigned to a topic, the color is "sticky" for future sessions. A user might choose to color a critical topic bright red so that changes to it stand out in the future.

Figure 2 shows the same visualization, but here a summary of a selected topic is shown in a pop-up balloon. This summary was generated by selecting sentences that contained large numbers of key concepts from the topic. Any summarization of a cluster could be used here if it provided more useful information.

To illustrate how the demonstration system shows changes in TDT clusters over time, Figure 3 shows an updated visualization for two weeks later (November 14, 1998). The topic colors are

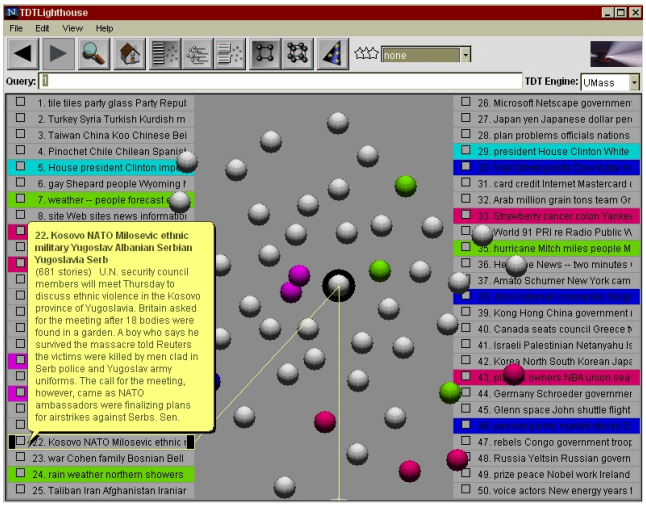

Figure 2: Similar to Figure 1, but showing a pop-up balloon.

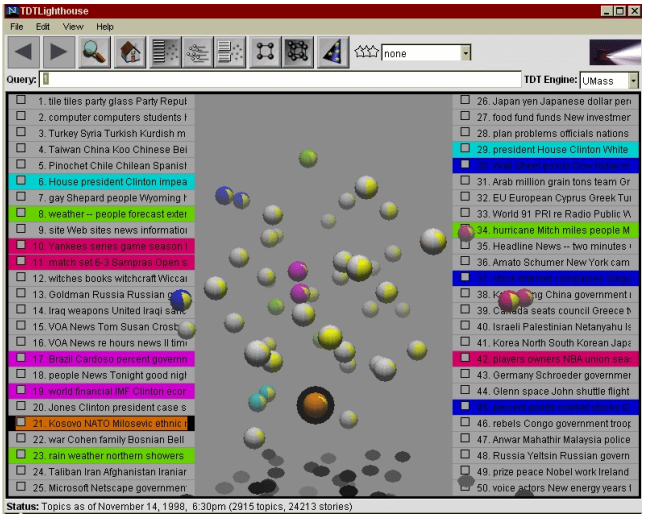

Figure 4: A 3-dimensional version of Figure 3.

persistent from Figure 1, though one of the marked topics ("Strawberry cancer colon Yankee") is no longer in the largest 50 so does not appear.

Most of the spheres include a small "wedge" of yellow in them. That indicates the proportion of the topic that is new stories (since Figure 1). Some topics have large numbers of new stories, so have a large yellow slice, whereas a few have a very small number of new stories, so have only a thin wedge. The yellow wedge can be as much as 50% of the sphere (which would represent an entirely new topic), and only covers the top of the sphere. This restriction ensures that the topic color is still visible.

The controls at the top of the screen are for moving between queries, issuing a query, and returning the visualization to a "home" point. The next £ve controls affect the layout of the display, including allowing a 3-D display: a 3-D version of Figure 3 is shown in Figure 4. The £nal control enables a browsing wizard that can be used to £nd additional topics that are very similar to a selected topic color (that set is chosen using the pull-down menu that has "none" in it).

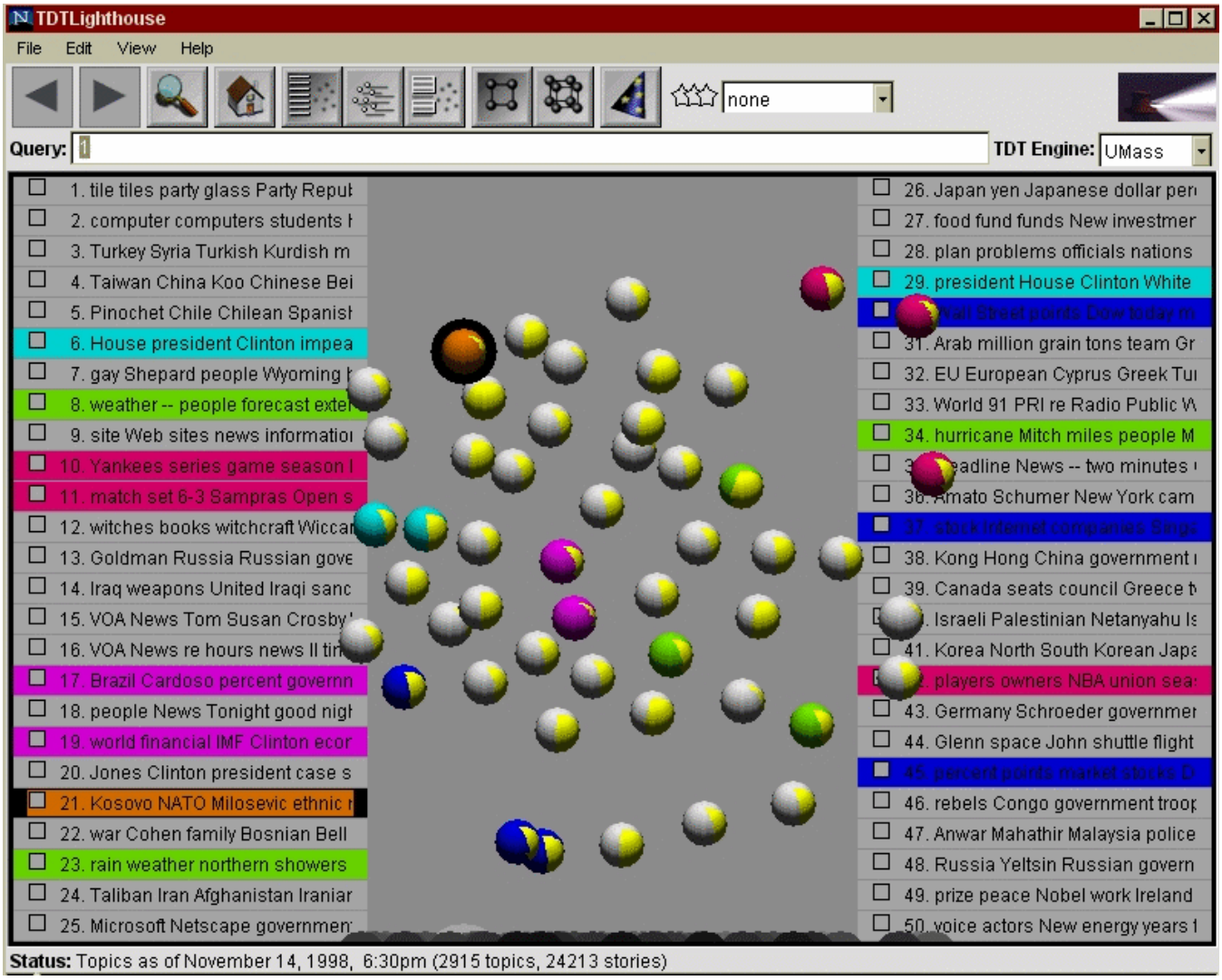

Figure 3: TDT demonstration system running on TDT-3 data, approximately six weeks into the collection.

5. CONCLUSION AND FUTURE WORK

The demonstration system described above illustrates the effect of TDT technology. It is also interesting in its own right, allowing a user to track news topics of interest and to see how changes occur over time. There is no reason that the same system could not be used for non-TDT environments: any setting that clusters documents might be appropriate for this system.

We are working to extend the demonstration system to include some additional features.

- Considering the large number of topics (almost 3,000 in Figure 3), it is unlikely that all "interesting" topics will be £ndable. The query box at the top of the display will be used to allow the user to £nd topics that match a request. The ranked list will display the top 50 topics that match the query.

- Related to querying, we hope to include an "alert" feature that will ¤ag newly-created topics that match a query. For example, an analyst interested in the Middle East might develop a query that would identify topics in that region. When such a topic appeared, it would be ¤agged for the user (probably with a "hot topic" color).

- We hope to allow user "correction" of the topic breakdown provided by the TDT system. The state-of-the-art in TDT still makes mistakes, sometimes pulling two similar topics together, and sometimes breaking a single topic into multiple clusters. We intend that a user who sees such a mistake be able to indicate it to the system. That information will, in turn, to be relayed back to the TDT system to affect future processing.

- We will be implementing an "explode this topic" feature that will show the stories within a topic analogously to the way the current system shows the topics within the news. If the topic is small enough, for example, the spheres would represent stories within the topic. If the topic is larger, the spheres might represnt sub-clusters within the topic.

Acknowledgments

This material is based on work supported in part by the Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and in part by SPAWARSYSCEN-SD contract number N66001-99-1-8912. Any opinions, £ndings and conclusions or recommendations expressed in this material are the authors' and do not necessarily re¤ect those of the sponsor.

6. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[2] J. Allan, R. Gupta, and K. Khandelwal. Temporal summaries of news topics. Technical Report IR-226, University of Massachusetts, CIIR, 2001.

[3] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, £nal report. Available at http://www.clsp.jhu.edu/ws99/tdt, 1999.

[4] DARPA, editor. *Proceedings of the DARPA Broadcast news Workshop*, Herndon, Virginia, February 1999.

[5] V. Khandelwal, R. Gupta, and J. Allan. An evaluation scheme for summarizing topic shifts in news streams. In *Notebook proceedings of HLT 2001*, 2001.

[6] A. Leuski and J. Allan. Lighthouse: Showing the way to relevant information. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pages 125–130, 2000.

[7] NIST. Proceedings of the TDT 1999 workshop. Notebook publication for participants only, March 2000.

[8] NIST. Proceedings of the TDT 2000 workshop. Notebook publication for participants only, November 2000.

[9] D. R. Radev, H. Jing, and M. Budzikowska. Summarization of multiple documents: clustering, sentence extraction, an d evaluation. *ANLP/NAACL Workshop on Summarization, Seattle, WA*, 2000.

[10] Russell Swan and James Allan. Automatic generation of overview timelines. In *Proceedings of SIGIR*, pages 49–56, Athens, Greece, 2000. ACM.