# Information Retrieval and Visualization using SENTINEL

Margaret M. Knepper[1], Robert Killam[1], Kevin L. Fox[1], and Ophir Frieder[2]

[1]Harris Corporation
Information Systems Division
PO Box 98000
Melbourne, FL 32902-9800

[2]Department of Computer Science
Illinois Institute of Technology
10 W. 31st Street
Chicago, IL 60616

## 1. INTRODUCTION

Harris Corporation focuses on information retrieval support for various Government agencies. Time constraints and interest-level limit our user to reviewing the top documents before determining if the results of a query are accurate and satisfactory. In such cases, retrieval times and precision accuracy are at a premium, with recall potentially being compromised. To meet user demands our system, called SENTINEL, was designed to yield efficient, high precision retrieval.

This is the second time Harris has participated in the Text Retrieval Conference (TREC). We learned a lot from our first TREC [Knepper-97]. This year, we enhanced several aspects of our retrieval system and improved our performance over last year's results.

## 2. SENTINEL OVERVIEW

SENTINEL is a fusion of multiple information retrieval technologies, integrating n-grams, a vector space model, and a neural network training rule. SENTINEL is a C++ implementation of an Object-Oriented design. The basic structure of SENTINEL includes the following components:

- A web browser-based user interface that provides users with a mechanism to build queries for a topic of interest, execute the queries, examine retrieved documents, and build additional or refine existing queries.
- Multiple retrieval technologies utilizing n-grams and a Vector Space Model (VSM) to query the document corpus.
- A fusion component that combines and ranks the results of each of the retrieval engines.
- A *3*-dimensional viewer that provides users with a mechanism to explore various aspects of retrieved documents, looking for additional relevant documents.

A user begins by defining a topic of interest, then proceeds to define one or more queries for that topic. User queries to SENTINEL can take the form of

- Keyword(s) or phrases
- Example document(s)

SENTINEL focuses on an interactive multi-pass approach. We do not assume that the information will be found immediately, and therefore the user needs to iteratively refine the query. SENTINEL allows the user to review the documents and select the documents most relevant to the topic. Relevant documents can be used as queries to further refine the topic. The user can then quickly query over the data with the additional queries.

## 3. SENTINEL'S ENHANCEMENTS

This year, an improved ranking algorithm and a 3-dimensional visualization capability were incorporated into SENTINEL.

### 3.1 Ranking Algorithm

Last year only the VSM was used for document scoring. This year the results from each of the retrieval engines were integrated into a final score. The retrieval engines maintain only the high level scores. The user adjusts the lowest acceptable score and retrieval engine weight to effect score results to favor/disfavor a particular engine. SENTINEL standardizes the scores from each retrieval engine to range from 0 to 1. A ranking algorithm fuses the results of the retrieval engines and ranks the

documents based on a variety of factors: the number of times the document was selected, highest score, lowest score, average score, location in the query list and number of retrieval engines locating the document. Irrelevant documents and queries are removed.

Each retrieval engine is assigned a specific percentage. The document scores for the retrieval engines are reduced by the specified percentage. Depending upon the query type, different retrieval engines can be emphasized.

- Potentially misspelled words may put more emphasis on the *n*-Gram retrieval
- Document example queries place more emphasis on the VSM retrieval engine

An algorithm was developed to rank the document scores from different retrieval engines. The algorithm rates the following items:

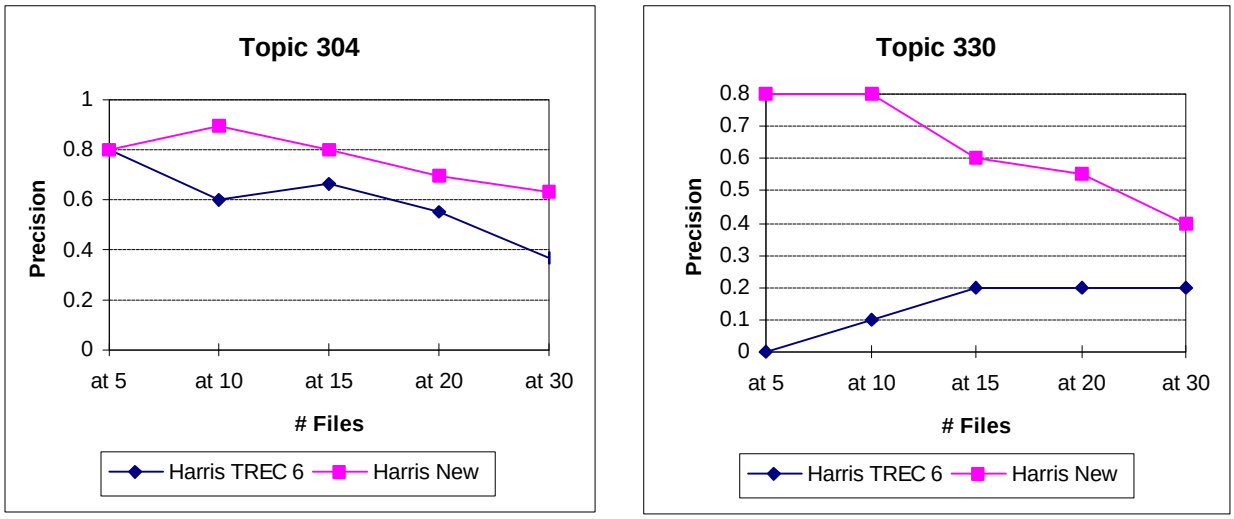
- Number of times document identified
    - Per query
    - Per retrieval engine
- Maximum score
- Minimum score
- Average score
- Penalty points

For all the items except the penalty points each item is ranked, the higher the number the lower the score. The individual items are totaled, and the lowest final score indicates the best document. Penalty points are assigned to a document based on the number of retrieval engines locating the document and the document location in the individual query list.

A penalty is added to a document for each retrieval engine not identifying it as relevant. Multiple engines retrieving a document is a strong indication of the document being relevant. This relevance correlation was also shown in [Lee-97]. Retrieval engines based on different search strategies and retrieving the same document are yet an even stronger indication of relevance [Alaoui-98]. The score must be above the minimum acceptable value after it is calculated for scaling and retrieval engine weight. During recent testing of the system, the team placed a lot of emphasis on the number of times the file was identified by multiple queries and multiple retrieval engines locating the same file. Setting high penalties on these values brought relevant documents to the top of the list. More experimentation with different types of data will help identify the values that should be assigned to the parameters.

Last year, the scoring algorithm took all query scores and put them into one final list. No consideration was given to the document's list location in the individual queries. The algorithm was modified so that each document receives a penalty point for its location in each individual query list. This is intended to reward the documents that are located close to the top of the list in the individual queries, by assigning fewer penalty points. Using this new technique we saw an improvement in the retrieval of relevant documents, Figure 1. We also added the ability to review individual results of the queries. This helped us quickly identify the usefulness of the queries and allowed us to focus on a query giving good results to help develop additional queries.

***Figure 1 Sample results from the improved ranking algorithm applied to TREC-6 topics.***



## 4. *3*-D VISUALIZATION

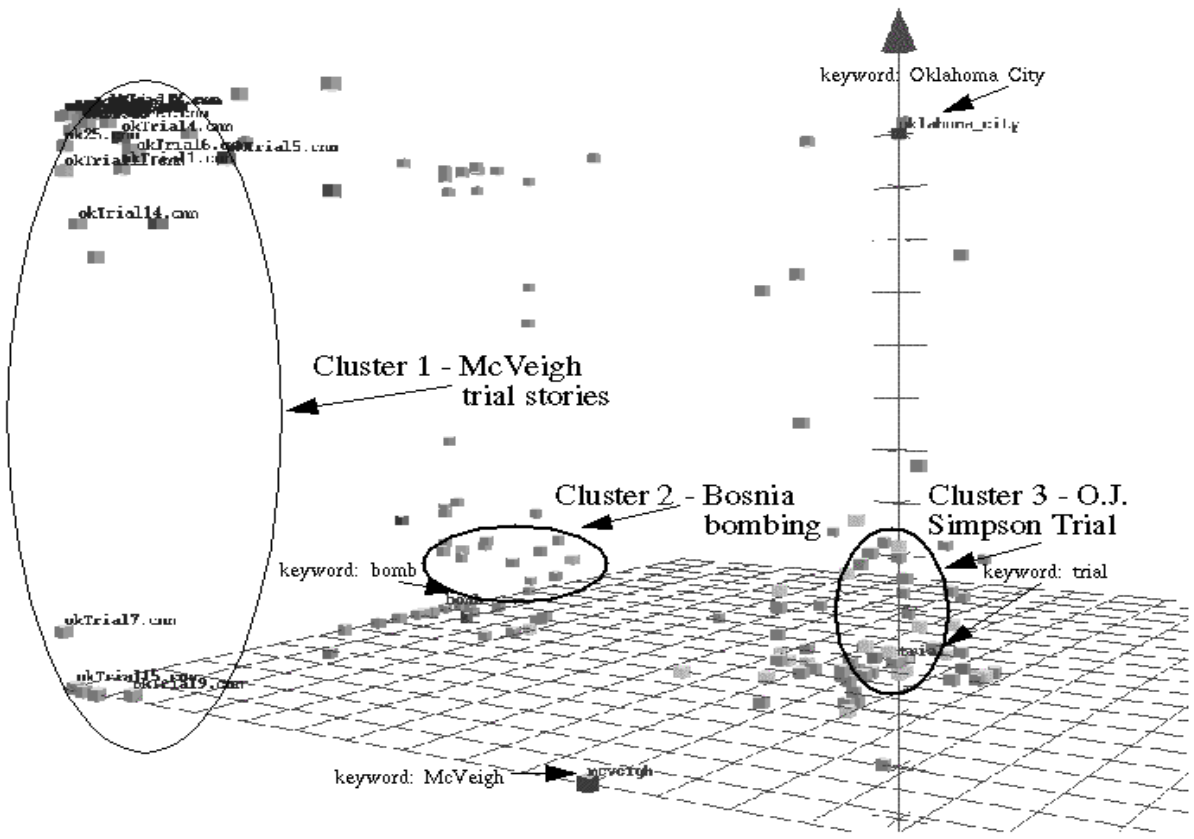
Primary user interaction with SENTINEL is through a web-browser-based user interface. Users can build and tailor queries as the topic of interest is further defined, moving from a generic search to specific topic areas through query inputs. Queries may consist of a single keyword, multiple keywords (or phrases), keyword clusters, an example document, and document clusters.

In SENTINEL, we enhance user understanding of the retrieved document set through the use of a Harris-developed *3*-dimensional visualization toolkit. This visualization tool supports multiple levels of data abstraction, clustered document presentation, data thresholding, and a variety of user interaction paradigms. The *3*-dimensional document visualization display enables the user to view different aspects of the document's topic, and provides an intuitive display of document relationships and similarity.

A set of documents retrieved for a particular topic may exhibit a variety of aspects. An example of different article aspects can be demonstrated by stories from the Oklahoma City bombing of 1996. There are articles about the bomb, damage from the bomb blast, rescue work, the victims, suspects, the Timothy McVeigh trial, the Terry Nichols trial, and the victims' memorial just to name a few. Each of these represents a different aspect of the Oklahoma City bombing.

Displaying the documents in a *3*-dimensional space enables a user to see document clusters, the relationships of documents to each other, and also aids in the location of additional documents that may be relevant to a query. Documents near identified relevant documents (identified through SENTINEL queries) can be easily reviewed for topic relevance. The user is able to manipulate the dimensional view to gain new views of document relationships. Changing the display axes allows the information to be viewed for different topic aspects to aid in further identification of relevant documents. SENTINEL is able to reduce the display down to the most important aspects of the document.

***Figure 2 Example of document clustering***



As illustrated in Figure 3, each document in the retrieved document corpus is represented mathematically in the *3*-D display space by a cube. The dark text indicates the relevant documents found through text queries submitted to SENTINEL, reviewed, and marked as relevant. The relevant stories appear to separate from the other stories, in the upper section of the view are surrounded by a dark lined box in Figure 3. These documents close to each other in the 3-D view, but are not close to each other on the list, Table 1. In the 3-D display, documents FT924-6336, LA050990-0002, and LA090389-0160 appear close to each other. They appear as on the list at the positions 2, 96, and 295 respectively. Using the 3-D tool gives the user a different view of document groupings that we don't ordinarily obtain from just a list.

***Figure 3 Example of a 3-D view for Query 372***

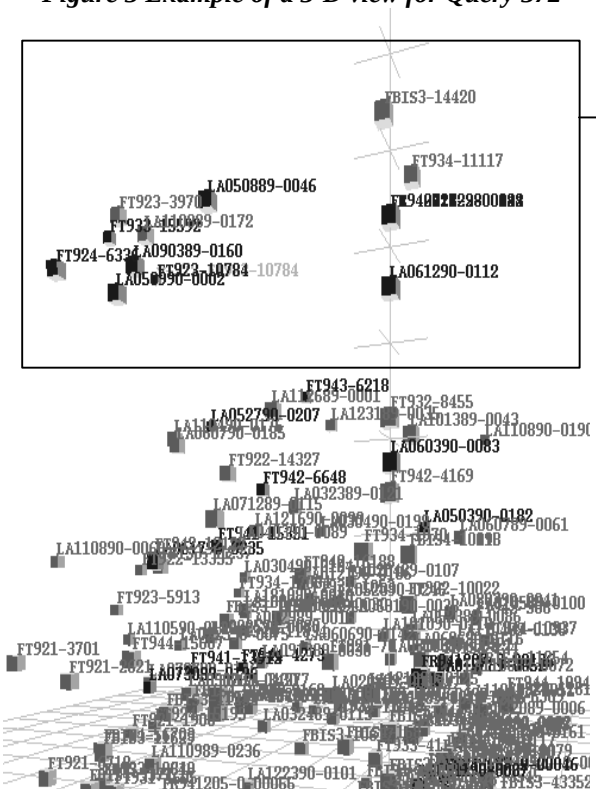



***Table 1 - The 3-D view shows a different set of relationships, not observed by just looking at the location of the documents on the list***

| File Name | List Ranking | Judged as a relevant document |
|---|---|---|
| FT924-6336 | 2 | yes |
| FT923-10784 | 3 | yes |
| LA052790-0207 | 5 | yes |
| LA050889-0046 | 6 | yes |
| FT942-12298 | 10 | yes |
| LA050990-0002 | 96 | yes |
| FT934-11117 | 101 | no |
| FBIS3-14420 | 150 | no |
| FT943-6218 | 172 | yes |
| LA110889-0172 | 198 | no |
| FR940617-2-00244 | 265 | yes |
| LA061290-0112 | 271 | yes |
| FT923-3970 | 284 | no |
| LA123189-0035 | 293 | no |
| LA090389-0160 | 295 | yes |

## 5. TREC 7 RESULTS

As part of our evaluation of SENTINEL, we participated in both TREC-6 and TREC-7. We saw a dramatic improvement in the average recall and precision results, illustrated in the graphs of Figure 4. Overall, we also reduced the number of queries this year, as shown in Table 2. Since each query is a pass through the database, fewer queries reduce the retrieval time.

*Figure 4 Average Recall and Precision From TREC-6 and TREC-7*

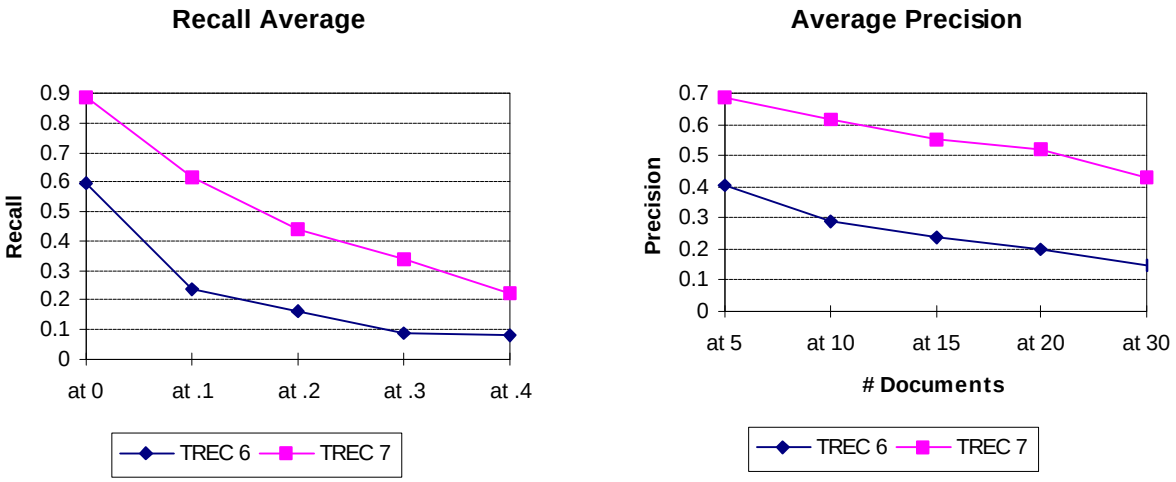



*Table 2 Number of Queries*

| | Number of n-gram queries | Number of VSM queries | Total Number of Queries |
|---|---|---|---|
| TREC-6 | 404 | 544 | 948 |
| TREC-7 | 449 | 443 | 812 |

## 6. CONCLUSION

The SENTINEL prototype is an efficient, high-level precision focused information retrieval and visualization system. It allows interactive formation of query refinement. It fuses results from multiple retrieval engines to leverage the strengths of the each. It has been designed for efficient maintenance, making it easy to add new documents. SENTINEL allows for multiple dictionaries and vocabularies – thus allowing a user to develop role-based dictionaries or vocabularies. Finally, SENTINEL provides a web-browser based interface for user interaction as well as a 3-D viewer for exploring the documents retrieved in response to a user's query. From a personal standpoint, we significantly gained a greater understanding of our query results using our visualization component. We see the importance of being able to respond real-time as the user rates the story and the need to filter the results shown to the user, i.e., show only the results from specific queries, or only show the first 30 documents.

## 7. REFERENCES


[Alaoui-98] S. Alaoui Mounir, N. Goharian, M. Mahoney, A. Salem, O. Frieder "Fusion of Information Retrieval Engines (FIRE)", 1998 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'98), Las Vegas, July 1998

[Knepper-97] Knepper, Margaret M., Cusick, Gregory J., Fox, Frieder, Ophir, Killam, Robert A. "Ad Hoc Retrieval with Harris SENTINEL", The Sixth Text Retrieval Conference (TREC-6), NIST Special Publication 500-240, E. M. Voorhess and D. K. Harman, Editors, August 1998, Pp. 503-509

[Lee-97] Joon Ho Lee, "Analysis of Mulitple Evidence Combination", in the proceeding of the 20th annual ACM SIR Conference, (Philadelphia, PA), 1997.