# IRIS at TREC-8

Kiduk Yang, and Kelly Maglaughlin
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599-3360 USA
{yangk, maglk}@ils.unc.edu

## 0 Submitted Runs

*unc8al32, unc8al42, unc8al52* – Category A, automatic ad-hoc task runs
*unc8iap* – interactive track run

## 1 Introduction

We tested two relevance feedback models, an adaptive linear model and a probabilistic model, using massive feedback query expansion in TREC-5 (Sumner & Shaw, 1997), experimented with a three-valued scale of relevance and reduced feedback query expansion in TREC-6 (Sumner, Yang, Akers & Shaw, 1998), and examined the effectiveness of relevance feedback using a subcollection and the effect of system features in an interactive retrieval system called IRIS (Information Retrieval Interactive System[1]) in TREC-7 (Yang, Maglaughlin, Mehol & Sumner, 1999).

In TREC-8, we continued our exploration of relevance feedback approaches. Based on the result of our TREC-7 interactive experiment, which suggested relevance feedback using user-selected passages to be an effective alternative to conventional document feedback, our TREC-8 interactive experiment compared a passage feedback system and a document feedback system that were identical in all aspects except for the feedback mechanism. For the TREC-8 ad-hoc task, we merged results of pseudo-relevance feedback to subcollections as in TREC-7.

Our results were consistent with that of TREC-7. The results of passage feedback, whose system log showed high level of searcher intervention, was superior to the document feedback results. As in TREC-7, our ad-hoc results showed high precision in top few documents, but performed poorly overall compared to results using the collection as a whole.

## 2 IRIS: Key Components

IRIS is an interactive retrieval system designed to provide users with ample opportunities to interact with the system throughout the search process. For example, users can supplement the initial query with two-word collocations suggested by the system, perform relevance feedback by selecting relevant documents or passages, or add and delete query terms. IRIS, first created in 1996 at the School of Information and Library Science at UNC-CH, has been under continuous development, evolving with each participation in TREC experiments. Below is a description of its key components used in TREC-8 experiments.

### 2.1 Text Processing

IRIS processes the text first by removing punctuation, and then excluding the 390 high-frequency terms listed in the WAIS default stopwords list as well as "IRIS stopwords,[2]" which were arrived at by examining the inverted index and identifying low frequency terms that appeared meaningless.

[1] IRIS was first developed by Kiduk Yang, Kristin Chaffin, Sean Semone, and Lisa Wilcox at the School of Information and Library Science (SILS) at the University of North Carolina. They worked under the supervision of William Shaw and Robert Losee.

[2] IRIS stopwords are defined as all numeric words, words that start with a special character, words consisting of more than 25 non-special characters, and words with embedded special characters other than a period, apostrophe, hyphen, underline, or forward or backward slash.

After the punctuation and stopword removal, IRIS conflates each word by applying one of the four stemmers implemented in the IRIS Nice Stemmer module,[3] which consists of a simple plural remover (Frakes & Baeza-Yates, 1992, chap. 8), the Porter stemmer (Porter, 1980), the modified Krovetz inflectional stemmer (Krovetz, 1993),[4] and the Combo stemmer that uses the shortest whole word (i.e., word that appears in a dictionary) returned by the three stemmers. We used the Krovetz stemmer in TREC-7 for its conservative conflation tendencies, but we opted for the simple plural remover in TREC-8 to speed up the indexing time. Simple stemmer was chosen over the porter stemmer to minimize the overstemming effect, with the hope that understemming effect would be compensated by the feedback query expansion process.

## 2.2 Phrase Construction

Phrase construction method employed in our TREC-7 interactive experiments was used to construct the phrase indexes in TREC-8. Using the online dictionary and the clause recognition algorithm built into the Nice Stemmer, we constructed a two-word noun-noun phrase index by first extracting adjacent word pairs of noun and proper noun combinations within a clause,[5] and then discarding the phrases occurring 20 or less times in the collection to reduce indexing time and to conserve computer resources. The phrase occurrence frequency threshold of 20 was arrived at by selecting the number that produced the phrase index whose size was most comparable to that of the collocation index. To augment the proper nouns in the online dictionary, all capitalized words not occurring at the beginning of a sentence were considered to be proper nouns. Hyphenated words were broken up and stemmed by the simple plural remover before the noun-noun phrase construction module was applied. Hyphenated words in their raw form (i.e. as they appear in documents sans punctuation) were added to the index as well.

## 2.3 Ranking Function and Term Weights

IRIS ranks the retrieved documents in decreasing order of the inner product of document and query vectors,

$$\mathbf{q}^{\mathrm{T}}\mathbf{d}_i = \sum_{k=1}^{t} q_k d_{ik} , \tag{1}$$

where $q_k$ is the weight of term $k$ in the query, $d_{ik}$ is the weight of term $k$ in document $i$, and $t$ is the number of terms in the index. We used SMART *Lnu* weights for document terms (Buckley, Singhal, Mitra, & Salton, 1996; Buckley, Singhal, & Mitra, 1997), and SMART *ltc* weights (Buckley, C., Salton, G., Allan, J., & Singhal, A., 1995) for query terms. *Lnu* weights attempt to match the probability of retrieval given a document length with the probability of relevance given that length (Singhal, Buckley, & Mitra, 1996). Our implementation of *Lnu* weights was the same as that of Buckley et al. (1996, 1997) except for the value of the *slope* in the formula, which is an adjustable parameter whose optimal value may depend, in part, on the properties of the document collection.

According to the pre-test experiments, an *Lnu* slope of 0.5 performed best with feedback, especially when using both single term and phrase indexes. Based on these findings, we used a slope of 0.5 to optimize performance with feedback.

## 2.4 Feedback Models

### 2.4.1 Adaptive Linear Model

We used the same implementation of the adaptive linear model (Wong & Yao, 1990; Wong, Yao, Salton, & Buckley, 1991) in TREC-8 as in TREC-7. The basic approach of the adaptive linear model, which is based on

[3] Nice stemmer was implemented by Kiduk Yang, Danqi Song, Woo-Seob Jeong, and Rong Tang at SILS at UNC. For an interactive demonstration, please visit http://ils.unc.edu/iris/nstem.htm.

[4] The modified Krovetz inflectional stemmer implements a modified version of Krovetz's inflectional stemmer algorithm and restores the root form of plural ("-s," "-es," "-ies"), past tense ("-ed"), and present participle ("-ing") words, provided this root form is in our online dictionary.

[5] IRIS identifies a clause boundary by the presence of appropriate punctuation marks such as a comma, period, semicolon, question mark, or exclamation mark.

the concept of the preference relation from decision theory (Fishburn, 1970), is to find a *solution vector* that, given any two documents in the collection, will rank a more-preferred document before a less-preferred one (Wong et al., 1988).

In the relevance feedback interface of IRIS, users can evaluate documents as "relevant," "marginally relevant," or "nonrelevant." By adapting the concept of the user preference relation to extend the relevance scale from a binary to a three-valued scale, we constructed the following formula for the starting vector. Note that this formula can be adjusted for any multivalued relevance scale:

$$\mathbf{q}_{(0)} = c_0 \mathbf{q}_{rk} + \frac{c_1}{N_{new\,rel}} \sum_{new\,rel} \mathbf{d} + \frac{c_2}{N_{new\,mrel}} \sum_{new\,mrel} \mathbf{d} - \frac{c_3}{N_{new\,nonrel}} \sum_{new\,nonrel} \mathbf{d}\,, \qquad (2)$$

where $\mathbf{q}_{rk}$ is the query vector that produced the current ranking of documents; $c_0$, $c_1$, $c_2$, and $c_3$ are constants; $N_{new\ rel}$, $N_{new\ mrel}$, and $N_{new\ nonrel}$ are the number of new relevant, new marginally relevant, and new nonrelevant documents respectively in the current iteration; and the summations are over the appropriate new documents. The detailed description of the adaptive linear model can be found in Yang et. al. (1999).

### 2.4.2 Passage Feedback Model

The conventional relevance feedback models assume the user's relevance judgement to be about an entire document and treat documents as the information units. The unit of a document, however, is sometimes determined by arbitrary reasons such as convenience or convention rather than content, which can produce a document containing subsections of various information content. In such instances, the user's determination of relevance is likely to be based on certain portions of a document rather than the entirety of it. Even in a document of consistent information content, the user may be interested in only a specific information described in certain passages of the document. Thus, we think that the conventional practice of using document as the unit of feedback may be less effective than using user-defined passages in relevance feedback.

To test this theory, we implemented the "passage feedback model" in TREC-7 with the following formula for feedback vector creation:

$$\mathbf{q}_{\mathbf{new}} = \mathbf{q}_{\mathbf{old}} + \sum_{rel} \mathbf{p} - \sum_{nonrel} \mathbf{p}\,, \qquad (4)$$

where **q** is the query vector and **p** is the passage vector determined by the user's selection of the relevant and nonrelevant portions of documents. Since the normalization factor of the *Lnu* weight is based on document length, an inverse document frequency weight was used for the passage vector **p**. The passage feedback approach differs fundamentally from the philosophy of the adaptive linear model in that it simply expands the query vector to make it more "similar" to relevant passages and "dissimilar" to nonrelevant passages rather than trying to rank a document collection in the preference order defined by a training set

Though the underlying implementations of the passage feedback model are the same in our TREC-7 and TREC-8 interactive experiments, we made a significant modification to the user interface in TREC-8. One of the prevalent user comments of our TREC-7 passage feedback system was about the clunkiness of its feedback interface. Users had to first select passages by highlighting with mouse, copy the highlighted portions, toggle to the passage feedback window, and then paste the copied selection into appropriate windows. System logs as well as user comments seemed to indicate the difficulty of these steps required for the passage feedback, which we thought kept users from fully enjoying the benefits of the passage feedback system. Consequently, we simplified the feedback interface by using an embedded java applet, which consolidated document display window and feedback window as well as simplified the overall passage feedback operation.

## 3 Pre-test Experiments

In our TREC-8 pre-test experiments, we compared two feedback query expansion size (300 terms, 30 terms) as well as several pseudo-feedback approaches using TREC-7 queries and relevance judgements on the full TREC-7 document collection. In keeping with the experimental design of both TREC-7 and TREC-8, queries were first submitted to each subcollection of FBIS, Federal Register, Financial Times, and LA Times, after which several pseudo-feedback runs were executed, and finally the results were merged by the similarity scores to produce the top 1000 ranked documents.

The pseudo-feedback runs consisted of using the full length queries with100th document as non-relevant and various number of top documents as relevant with additional parameter of 300-term feedback vector vs. 30-term feedback vector. The effect of changing the number of pseudo-relevant documents was negligible, but 300-term feedback vector outperformed 30-term vector, which is consistent with finding from previous experiments.

# 4 Ad-hoc Experiments

## 4.1 Research Question

We continued exploration of our approach of subcollection retrieval in TREC-8 ad-hoc experiments. Though we are aware that retrieval performance of using the whole collection is superior to that of using subcollections with initial retrieval, we wanted to see if we could minimize the performance loss with relevance feedback. If the subcollection retrieval with simple pseudo-relevance feedback can be shown to be competitive to that of using a whole collection, then subcollection retrieval may be a desirable strategy in real world situations, where the whole document collection statistics is unavailable or too costly to compute.

In this light, We posed the following question.

- Is the subcollection retrieval results using pseudo-relevance feedback competitive to the initial retrieval results using the whole collection?

## 4.2 Research Design

There are two main issues in the subcollection retrieval as we have defined it. First, there is the problem of "collection fusion", where various methods of merging the results of subcollection retrieval have been examined (Dumais, 1993; Voorhees, Gupta, & Johnson-Laird, 1995; Savoy, Calve, & Vrajitoru, 1997). Then there is the question of how best to implement the pseudo-feedback process, which has been one of the main concerns of TREC ad-hoc participants in the past.

Our research design in TREC-8 was strongly influenced by the desire to lay the groundwork for building a large scale system such as IRISWeb[6]. Consequently, our approach was to use the simpler methods over more complex ones in order to increase the system efficiency. We selected the top two performing subcollection retrieval approaches from the pre-test experiments as well as a run with a medium length query and a shorter feedback vector. The final results were produced by first retrieving 10% of documents in each collection and merging the results by their raw query-document similarity scores:

- *unc8al32*: Pseudo-relevance feedback with the top 5 retrieved documents as relevant and the 100th document as non-relevant using the top 250 positive-weighted terms and the lowest 50 negative-weighted terms.
- *unc8al42*: Pseudo-relevance feedback with the top 10 retrieved documents as relevant and the 100th document as non-relevant using the top 250 positive-weighted terms and the lowest 50 negative-weighted terms.
- *unc8al52*: Pseudo-relevance feedback with the top 3 retrieved documents as relevant and the 100th document as non-relevant using the top 25 positive-weighted terms and the lowest 5 negative-weighted terms.

[6] IRISWeb is an experimental Web search engine, which is a variation of IRIS system. Please see http://ils.unc.edu/iris for more information.

Both *unc8al32* and *unc8al42* used the full query (i.e. title, description, narrative), whereas *unc8al52* used only the title and description fields to construct the queries. The default system constructs for the ad-hoc experiment were:

- *Lnu* 0.5 weights for documents, and *ltc* weights for queries
- adaptive linear model for feedback
- use of single-term and noun-noun phrase index
- conflation by removal of simple plurals

### 4.3 Results

The TREC-8 ad-hoc collection consists of 130,471 FBIS, 55,630 Federal Register, 210,158 Financial Times, and 131,896 LA Times documents. Each document collection was first processed individually to generate single-word indexes of 244,458 terms and phrase index of 60,822 terms for FBIS, 118,178 single and 28,669 phrases terms for Federal Register, 290,880 single and 87,144 phrases terms for Financial Times, and 228,507 single and 62,995 phrase terms for LA Times collection.

TREC evaluation measures of the top 1000 documents (Table 2) showed consistent results with pre-test, where the variation in the number of relevant documents of the pseudo-relevance feedback made little difference (*unc8al32* vs. *unc8al42)*. The best performance of the three was achieved by using the medium length query and shorter feedback query (*unc8al52),* which was somewhat unexpected. To investigate this matter further, we ran a series of post analysis runs, where we tested the effect of the initial query length in combination with feedback vector length. The results consistently showed the runs with medium length query to perform better than ones with full length query regardless of the feedback vector length. It appears that better initial retrieval result achieved using the medium length query (Table 1) overpowers any advantage gained by the longer feedback vector.

It is unclear why our system performed better with the medium length query, whereas the usage of full length query has shown to be advantageous by the best performing ad-hoc systems in the past (Voorhees & Harman, 1999). It is possible that we need better query processing and expansion approaches to minimize the noise in the narrative portion of the query vector while maximizing its descriptiveness.

As expected, our performance using subcollection retrieval is somewhat worse off than the median performance of TREC-8 participants (Table 3). Table 1 and 2 show the subcollection retrieval results to be competitive to retrieval results using the whole collection. The gap in performance levels between subcollection and whole collection retrieval is narrowed by feedback, which suggests the possibility that simple pseudo-relevance feedback methods may be more effective in a subcollection retrieval setting.

**Table 1**. Initial Retrieval Results of top 1000 documents

| | *Subcollection Retrieval* | | *Whole Collection Retrieval* | |
|---|---|---|---|---|
| | *long query* | *medium query* | *full query (unc8alb1)* | *medium query (unc8alb2)* |
| Average Precision | 0.1031 | 0.1406 | 0.1687 | 0.1715 |
| Precision at 5 docs | 0.2200 | 0.3480 | 0.4320 | 0.4040 |
| Precision at 10 docs | 0.2080 | 0.2980 | 0.3820 | 0.3600 |
| Precision at 100 docs | 0.1460 | 0.1612 | 0.1824 | 0.1686 |
| Number of Relevant Documents | 2272 | 2272 | 2262 | 2287 |

**Table 2**. Pseudo-feedback Results of top 1000 documents

| | *unc8al32* | *unc8al42* | *unc8al52* | *unc8alb1** | *unc8alb2*** |
|---|---|---|---|---|---|
| Average Precision | 0.1347 | 0.1372 | 0.1669 | 0.1722 | 0.1746 |
| Precision at 5 docs | 0.3280 | 0.3280 | 0.4120 | 0.4160 | 0.4040 |
| Precision at 10 docs | 0.3080 | 0.3160 | 0.3580 | 0.3800 | 0.3700 |
| Precision at 100 docs | 0.1642 | 0.1664 | 0.1690 | 0.1846 | 0.1738 |
| Number of Relevant Documents | 2249 | 2261 | 2281 | 2285 | 2290 |

* Same as *unc8al42* except for using the whole collection
** Same as *unc8al52* except for using the whole collection

**Table 3**. Best, Median, Worst Results (top 1000 documents) of all TREC8 ad-hoc participants

| | *long query** | | | *medium query*** | | |
|---|---|---|---|---|---|---|
| | *Best* | *Median* | *Worst* | *Best* | *Median* | *Worst* |
| Average Precision | 0.4338 | 0.2570 | 0.0186 | 0.4339 | 0.2261 | 0.0001 |
| Number of Relevant Documents | 3854 | 2784 | 598 | 3807 | 2791 | 37 |

* Statistics computed over 37 automatic ad hoc runs that used the entire topic statement.
** Statistics computed over 59 automatic ad hoc runs that used the title and description sections of the topic statement (4 runs used description only).

# 5. Interactive Experiment

## 5.1 Research Question

In our TREC-7 interactive experiments, we examined the effects of user interface on retrieval performance by comparing a system with complex interface with one with a simpler interface. We also compared in TREC-7 the effectiveness of a "passage feedback" system, where user-defined passages were used to expand the feedback query, with a conventional "document feedback" system that used relevance judgement based on documents to perform the relevance feedback using the adaptive linear model.

In keeping with our hypothesis, passage feedback system results were better than that of the document feedback results in our TREC-7 experiments. However, the results of the simple interface system versus the complex interface system was rather unexpected in that it performed slightly worse than the complex interface system. After further examination, which revealed more searcher intervention steps in the system logs of the complex system, we concluded that complex system allowed users more opportunities to intervene, thereby positively affecting the retrieval performance.

Furthermore, we noticed that the passage feedback features in TREC-7 were somewhat underutilized. Though there were more retrieval iterations per query in the passage feedback system than in the document feedback system, there were more reformulation of the initial query cycles than the expansion of the feedback vector by using the passage feedback interface. One of the prevalent user comments of our TREC-7 passage feedback system was about the clunkiness of its feedback interface. Indeed, our TREC-7 passage feedback interface required several keystrokes or mouse clicks including the toggling between two windows.

Based on these observations, we hypothesized the following in our TREC-8 experiment:

- Passage feedback in an interactive system can perform better than the document feedback.
- Improving the usability of the passage feedback interface will invite more usage of it, thereby resulting in more positive user intervention.
- User intervention can positively affect the retrieval performance.

## 5.2 Methodology

To test our hypothesis, we constructed a passage feedback system with a streamlined feedback interface for our TREC-8 interactive experiment and compared its performance with that of a document feedback system. If our hypothesis were correct, our TREC-8 passage feedback system should show better results and more user intervention steps than both our document feedback system in TREC-8 and our “difficult” passage feedback system in TREC-7.

The passage feedback system and the document feedback system in TREC-8 are identical in all aspects except for how relevance feedback is implemented. Both systems have exactly the same features and interfaces for initial query formulation (Figure 1), initial query modification (Figure 2), and feedback query modification (Figure 4). The one and only difference between systems occurs in the relevance feedback interface. The document feedback system employ the conventional feedback mechanism of judging the relevance of a document as a whole (Figure 3.1) using the adaptive linear model, but the passage feedback system allows users to select relevant and nonrelevant portions of a document with which to expand the feedback query vector (Figures 3.2).

User intervention can occur in the initial query modification phase where the user can supplement the initial query with “suggested phrases” selected by the system, in the feedback query modification phase where the user can add new terms or delete existing terms from the feedback query, and in the relevance feedback phase.

The underlying system constructs for both systems were essentially the same as that of the ad-hoc system except for interactive feedback mechanism. Document term weights of *Lnu* 0.5 and *ltc* query term weights were used to maximize the relevance feedback influence, while the feedback query with 250 terms with highest positive weights and 50 terms with lowest negative weights was used in order to optimize the system for efficiency. A phrase index of adjacent noun-noun pairs were also used in suggesting potentially useful phrases for the initial query as well as in expanding the feedback vectors.

## 5.3 Searchers

Table 4 shows the information about each searcher's background and search experience gathered by pre-study questionnaires. All searchers were either working on or had received a graduate degree, most (17) of which were in Library and Information Science. The searchers had been searching between 1 and 14 years, with 4.5 being the average. Nine of the 24 searchers were male.

**Table 4.** Response Frequency of Searchers on Pre-Study Questionnaire

| | *No Experience* | | *Some Experience* | | *Great Deal of Experience* |
|---|---|---|---|---|---|
| 1.Using a point-and-click interface | | | 3 | 5 | 18 |
| 2.searching elec. library catalogs | | 2 | 3 | 9 | 10 |
| 3.searching on CD ROM systems | 1 | 3 | 12 | 8 | |
| 4.searching commercial systems | 5 | 11 | 2 | 5 | 1 |
| 5.using WWW search services | | | 5 | 8 | 11 |
| 6.searching other systems | 4 | | 2 | 1 | 2 |
| | *Never* | *Once or twice a year* | *Once or twice a month* | *Once or twice a week* | *Once or twice a day* |
| 7.Searching frequency | | | 2 | 9 | 13 |
| | *Strongly Disagree* | *Disagree* | *Neutral* | *Agree* | *Strongly Agree* |
| 8.Enjoys information searches | | 1 | 6 | 10 | 7 |

## 5.4 Results

The performance of IRIS measured by the mean instance precision (MIP) and mean instance recall (MIR) measures confounded both our hypothesis and the previous results of the passage feedback system. The results of the TREC-8 passage feedback system with an “improved” interface was the worst among all our systems tested in both TREC-7 and TREC-8, which is the exact opposite of what we expected (Table 5). However, the difference between systems ($H_a$: $\mu_{df} > \mu_{pf}$) was not statistically significant in either MIP ($p=0.34$) or MIR ($p=0.09$).

**Table 5**. Interactive Experiment Result Statistics

| | *TREC-8* | | *TREC-7* | |
|---|---|---|---|---|
| | *document feedback (df)* | *passage feedback (pf)* | *document feedback* | *passage feedback* |
| Mean Instance Precision | 0.643 | 0.625 | 0.673 | 0.778 |
| Mean Instance Recall | 0.277 | 0.228 | 0.281 | 0.314 |

A further examination of the results with system order consideration suggests that the passage feedback might have been more difficult to learn than the document feedback system (Table 6). In comparing the results of the first system shown to the searchers, the document feedback system outperforms the passage feedback system. This is not true when they were the second systems searched. Also, the passage feedback system's precision and recall scores improve when the system was the second system searched. There was actually a slight decrease in precision and recall in the document feedback system when it was the second system searched.

**Table 6**. Order Effects in TREC-8 Systems

| | *First System Searched* | | *Second System Searched* | |
|---|---|---|---|---|
| | *document feedback (df1)* | *passage feedback (pf1)* | *document feedback (df2)* | *passage feedback (pf2)* |
| Mean Instance Precision | 0.685 | 0.575 | 0.600 | 0.676 |
| Mean Instance Recall | 0.295 | 0.196 | 0.259 | 0.260 |

Table 7 shows the p-values of various system differences. The system difference with statistical significance ($\alpha=0.05$) occurs between the first systems searched, thus giving evidence to the hypothesis that the searchers would do better with the document feedback system as the first system. The improvement in performance of the passage feedback system ($H_a$: $\mu_{pf2} > \mu_{pf1}$) is statistically significant in MIP ($\alpha=0.05$), and not significant in MIR. Overall lack of significance in system learning effect suggests that either there was not sufficient time for any system learning to take place, or that there is no system learning to be gained in the first place.

**Table 7**. Statistical Significance (p-values) of System Differences

| | $H_a$: $\mu_{df1} > \mu_{pf1}$ | $H_a$: $\mu_{pf2} > \mu_{df2}$ | $H_a$: $\mu_{df1} > \mu_{df2}$ | $H_a$: $\mu_{pf2} > \mu_{pf1}$ |
|---|---|---|---|---|
| Mean Instance Precision | 0.03 | 0.11 | 0.08 | 0.05 |
| Mean Instance Recall | 0.02 | 0.49 | 0.26 | 0.09 |

The examination of the system logs show less user intervention in the passage feedback system than the document feedback system, as was the case in our TREC-7 interactive experiments. Even with the improved interface, the searchers seemed to have more difficulty using the passage feedback system. Given the time constraint of the experiment, the searchers might be more comfortable making document-level relevance judgement by quickly scanning documents than having to identify relevant passages, which would possibly require more time and mental effort.

In conclusion, we believe the poor results of the passage feedback system is largely due to our failure to make the passage feedback system more usable. Though we improved the passage selection interface, we did little to ease the cognitive burden of having to identify relevant passages rather than documents. Thus, the main challenge for the passage feedback system lies in helping searchers identify relevant passages quickly and easily.

## References

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. In D. K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Spec. Publ. 500-225, pp. 69-80). Washington, DC: U.S. Government Printing Office.

Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC 5. In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5)*.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART: TREC 4. In D. K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* (NIST Spec. Publ. 500-236, pp. 25-48). Washington, DC: U.S. Government Printing Office.

Dumais, S. T. (1993). LSI meets TREC. In D. K. Harman (Ed.), *Proceedings of the First Text REtrieval Conference (TREC-1), 137-152.*

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: John Wiley & Sons.

Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures & algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Krovetz, R. (1993). Viewing morphology as an inference process. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191-203.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14,* 130-137.

Savoy, J., Calve, A., & Vrajitoru, D. (1997). Report on the TREC-5 experiment: Data fusion and collection fusion. In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5).*

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

Sumner, R. G., Jr., & Shaw, W. M., Jr. (1997). An investigation of relevance feedback using adaptive linear and probabilistic models. In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5)*.

Sumner, R. G., Jr., Yang, K., Akers, R., & Shaw, W. M., Jr. (1998). Interactive retrieval using IRIS: TREC-6 experiments. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6)*.

Voorhees, E., Gupta, N. K., & Johnson-Laird, B. (1995). The Collection fusion problem. In E. M. Voorhees & D. K. Harman (Eds.), *Overview of the Third Text REtrieval Conference (TREC-3)*.

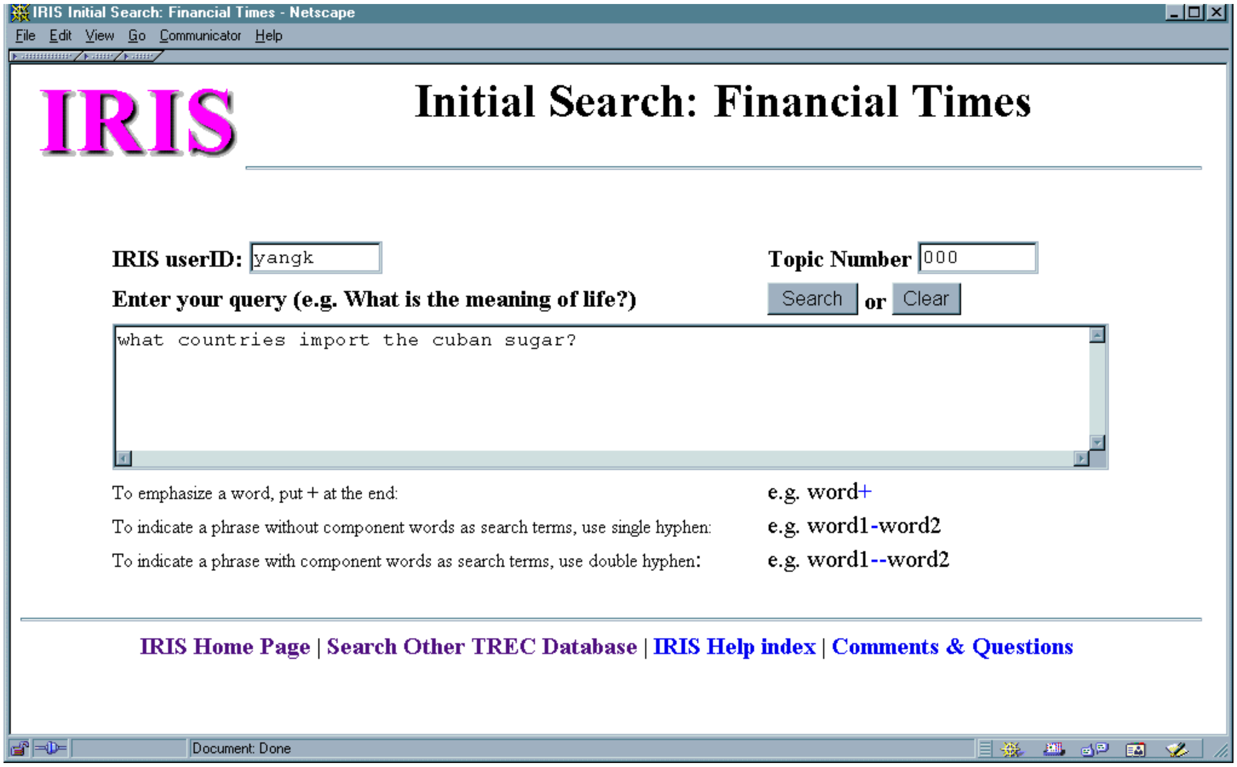
Voorhees, E., & Harman, D. K.. (1999). Overview of the Seventh Text REtrieval Conference (TREC-7). In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7)*.

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, *41*, 334-341.

Wong, S. K. M., Yao, Y. Y., Salton, G., & Buckley, C. (1991). Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, *42*, 723-730.

Yang, K., Maglaughlin, K., Meho, L., & Sumner, R. G., Jr. (1999). IRIS at TREC-7. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).*

**Figure 1** Initial Query Formulation Interface


IRIS Initial Search: Financial Times - Netscape
File Edit View Go Communicator Help
IRIS
Search
Clear
Document: Done

## Initial Search: Financial Times

**IRIS userID:** yangk **Topic Number** 000

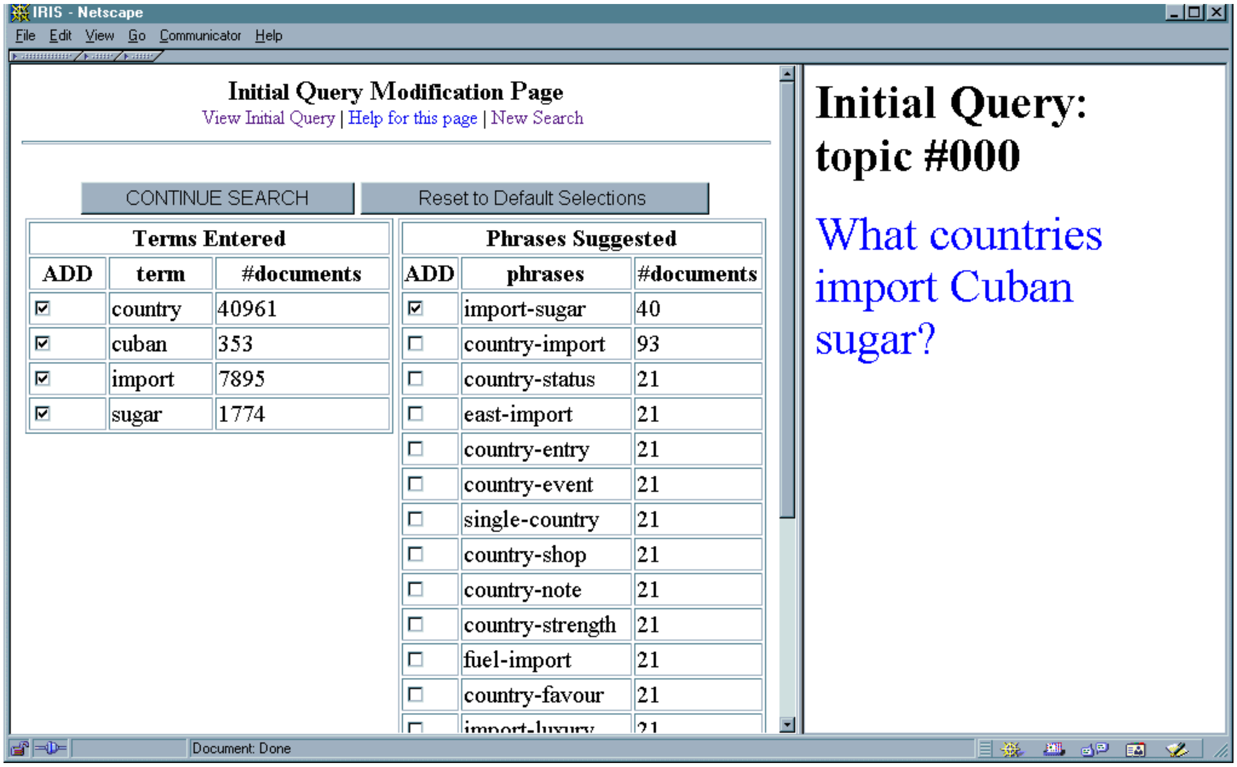
**Enter your query (e.g. What is the meaning of life?)** Search **or** Clear

```
what countries import the cuban sugar?
```

| | |
|---|---|
| To emphasize a word, put + at the end: | e.g. word+ |
| To indicate a phrase without component words as search terms, use single hyphen: | e.g. word1-word2 |
| To indicate a phrase with component words as search terms, use double hyphen: | e.g. word1--word2 |

**IRIS Home Page** | **Search Other TREC Database** | **IRIS Help index** | **Comments & Questions**

**Figure 2** Initial Query Modification Interface


IRIS - Netscape
File Edit View Go Communicator Help
CONTINUE SEARCH
Reset to Default Selections
Document: Done

**Initial Query Modification Page**

View Initial Query | Help for this page | New Search

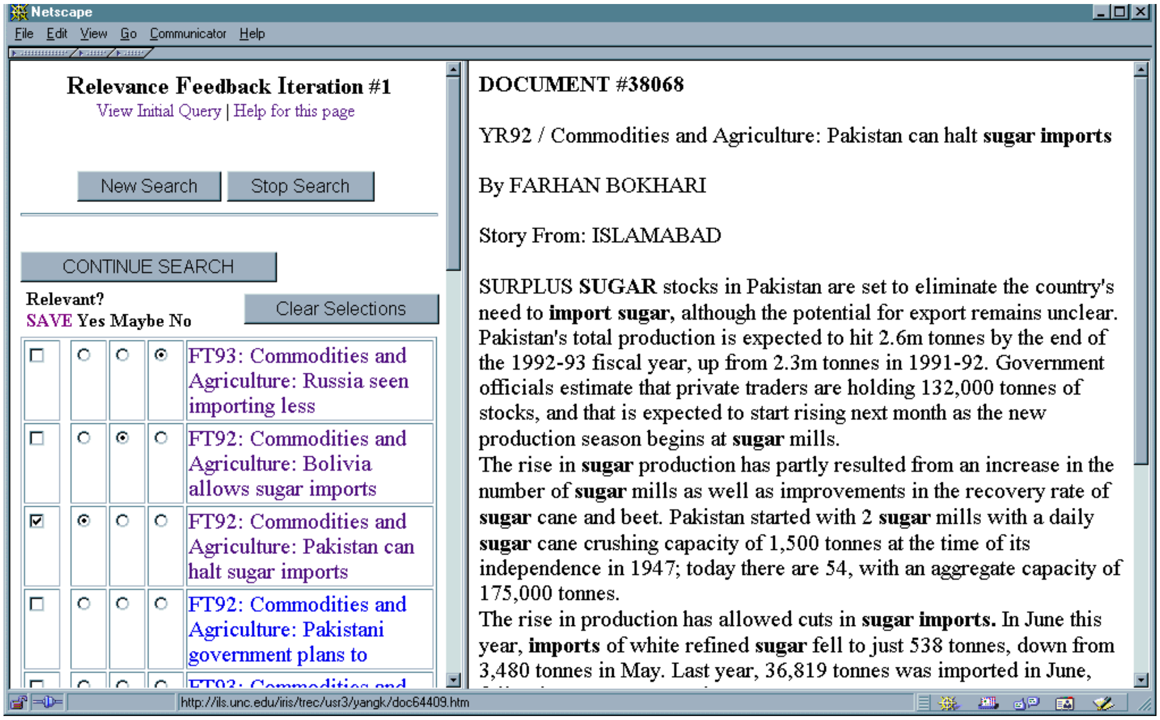
| Terms Entered | | |
|---|---|---|
| **ADD** | **term** | **#documents** |
| ☑ | country | 40961 |
| ☑ | cuban | 353 |
| ☑ | import | 7895 |
| ☑ | sugar | 1774 |

| Phrases Suggested | | |
|---|---|---|
| **ADD** | **phrases** | **#documents** |
| ☑ | import-sugar | 40 |
| □ | country-import | 93 |
| □ | country-status | 21 |
| □ | east-import | 21 |
| □ | country-entry | 21 |
| □ | country-event | 21 |
| □ | single-country | 21 |
| □ | country-shop | 21 |
| □ | country-note | 21 |
| □ | country-strength | 21 |
| □ | fuel-import | 21 |
| □ | country-favour | 21 |
| □ | import-luxury | 21 |

## Initial Query: topic #000

What countries import Cuban sugar?

**Figure 3.1** Document Feedback Interface


Netscape
File Edit View Go Communicator Help
New Search
Stop Search
CONTINUE SEARCH
Clear Selections
http://ils.unc.edu/iris/trec/usr3/yangk/doc64409.htm

**Relevance Feedback Iteration #1**

View Initial Query | Help for this page

Relevant?

| SAVE | Yes | Maybe | No | |
|---|---|---|---|---|
| ☐ | ○ | ○ | ● | FT93: Commodities and Agriculture: Russia seen importing less |
| ☐ | ○ | ● | ○ | FT92: Commodities and Agriculture: Bolivia allows sugar imports |
| ☑ | ● | ○ | ○ | FT92: Commodities and Agriculture: Pakistan can halt sugar imports |
| ☐ | ○ | ○ | ○ | FT92: Commodities and Agriculture: Pakistani government plans to |
| ☐ | ○ | ○ | ○ | FT93: Commodities and |

**DOCUMENT #38068**

YR92 / Commodities and Agriculture: Pakistan can halt **sugar imports**

By FARHAN BOKHARI

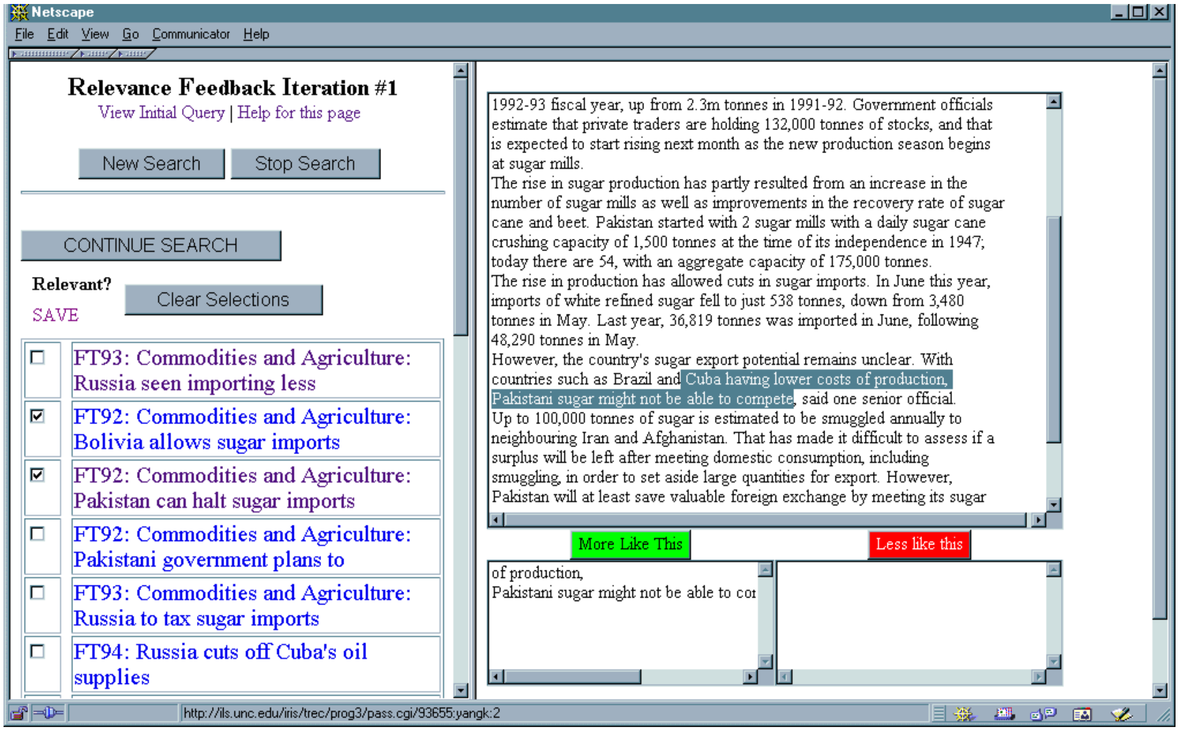
Story From: ISLAMABAD

SURPLUS **SUGAR** stocks in Pakistan are set to eliminate the country's need to **import sugar,** although the potential for export remains unclear. Pakistan's total production is expected to hit 2.6m tonnes by the end of the 1992-93 fiscal year, up from 2.3m tonnes in 1991-92. Government officials estimate that private traders are holding 132,000 tonnes of stocks, and that is expected to start rising next month as the new production season begins at **sugar** mills.

The rise in **sugar** production has partly resulted from an increase in the number of **sugar** mills as well as improvements in the recovery rate of **sugar** cane and beet. Pakistan started with 2 **sugar** mills with a daily **sugar** cane crushing capacity of 1,500 tonnes at the time of its independence in 1947; today there are 54, with an aggregate capacity of 175,000 tonnes.

The rise in production has allowed cuts in **sugar imports.** In June this year, **imports** of white refined **sugar** fell to just 538 tonnes, down from 3,480 tonnes in May. Last year, 36,819 tonnes was imported in June,

**Figure 3.2** Passage Feedback Interface


Netscape
File Edit View Go Communicator Help
New Search
Stop Search
CONTINUE SEARCH
Clear Selections
More Like This
Less like this
http://ils.unc.edu/iris/trec/prog3/pass.cgi/93655:yangk:2

**Relevance Feedback Iteration #1**

View Initial Query | Help for this page

Relevant?

SAVE

| | |
|---|---|
| ☐ | FT93: Commodities and Agriculture: Russia seen importing less |
| ☑ | FT92: Commodities and Agriculture: Bolivia allows sugar imports |
| ☑ | FT92: Commodities and Agriculture: Pakistan can halt sugar imports |
| ☐ | FT92: Commodities and Agriculture: Pakistani government plans to |
| ☐ | FT93: Commodities and Agriculture: Russia to tax sugar imports |
| ☐ | FT94: Russia cuts off Cuba's oil supplies |

1992-93 fiscal year, up from 2.3m tonnes in 1991-92. Government officials estimate that private traders are holding 132,000 tonnes of stocks, and that is expected to start rising next month as the new production season begins at sugar mills.

The rise in sugar production has partly resulted from an increase in the number of sugar mills as well as improvements in the recovery rate of sugar cane and beet. Pakistan started with 2 sugar mills with a daily sugar cane crushing capacity of 1,500 tonnes at the time of its independence in 1947; today there are 54, with an aggregate capacity of 175,000 tonnes.

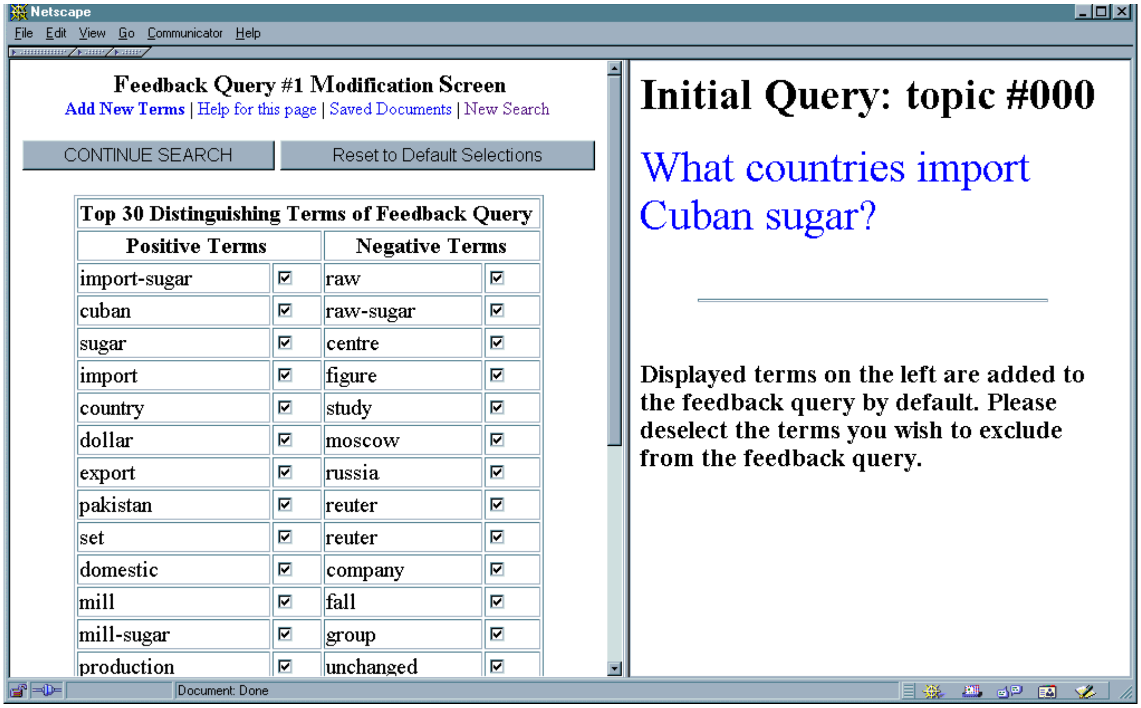
The rise in production has allowed cuts in sugar imports. In June this year, imports of white refined sugar fell to just 538 tonnes, down from 3,480 tonnes in May. Last year, 36,819 tonnes was imported in June, following 48,290 tonnes in May.

However, the country's sugar export potential remains unclear. With countries such as Brazil and Cuba having lower costs of production, Pakistani sugar might not be able to compete, said one senior official.

Up to 100,000 tonnes of sugar is estimated to be smuggled annually to neighbouring Iran and Afghanistan. That has made it difficult to assess if a surplus will be left after meeting domestic consumption, including smuggling, in order to set aside large quantities for export. However, Pakistan will at least save valuable foreign exchange by meeting its sugar

of production,
Pakistani sugar might not be able to cor

**Figure 4** Feedback Query Modification Interface


Netscape
File Edit View Go Communicator Help
CONTINUE SEARCH
Reset to Default Selections
Document: Done

**Feedback Query #1 Modification Screen**

**Add New Terms** | Help for this page | Saved Documents | New Search

| Top 30 Distinguishing Terms of Feedback Query | | | |
|---|---|---|---|
| **Positive Terms** | | **Negative Terms** | |
| import-sugar | ☑ | raw | ☑ |
| cuban | ☑ | raw-sugar | ☑ |
| sugar | ☑ | centre | ☑ |
| import | ☑ | figure | ☑ |
| country | ☑ | study | ☑ |
| dollar | ☑ | moscow | ☑ |
| export | ☑ | russia | ☑ |
| pakistan | ☑ | reuter | ☑ |
| set | ☑ | reuter | ☑ |
| domestic | ☑ | company | ☑ |
| mill | ☑ | fall | ☑ |
| mill-sugar | ☑ | group | ☑ |
| production | ☑ | unchanged | ☑ |

**Initial Query: topic #000**

What countries import Cuban sugar?

**Displayed terms on the left are added to the feedback query by default. Please deselect the terms you wish to exclude from the feedback query.**