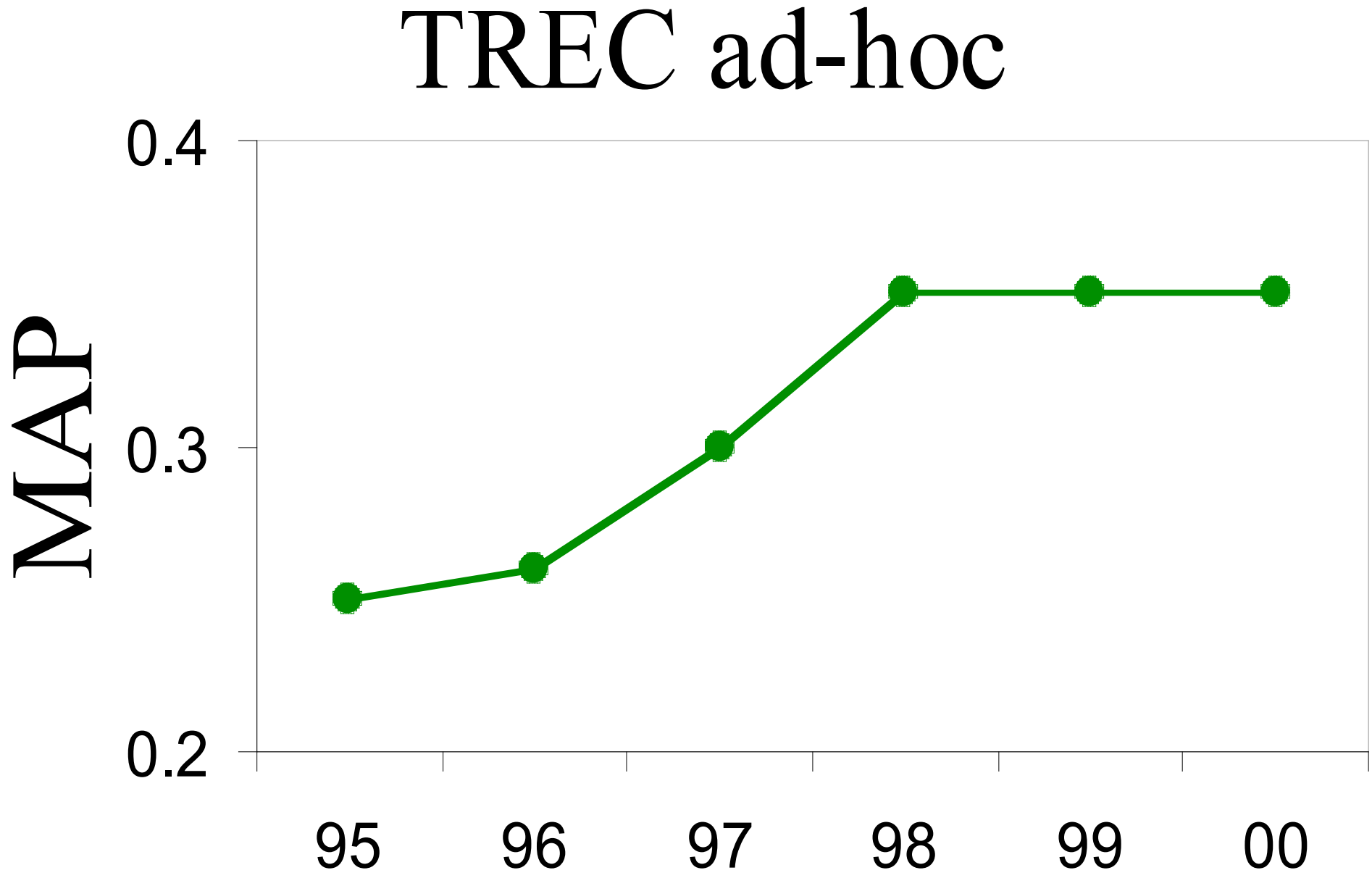
Why Batch and User Evaluations Do Not Give the Same Results

A. Turpin
Curtin University of Technology
Perth, Australia

W. Hersh
Oregon Health Sciences University
Portland, Oregon

Presented at SIGIR2001 New Orleans

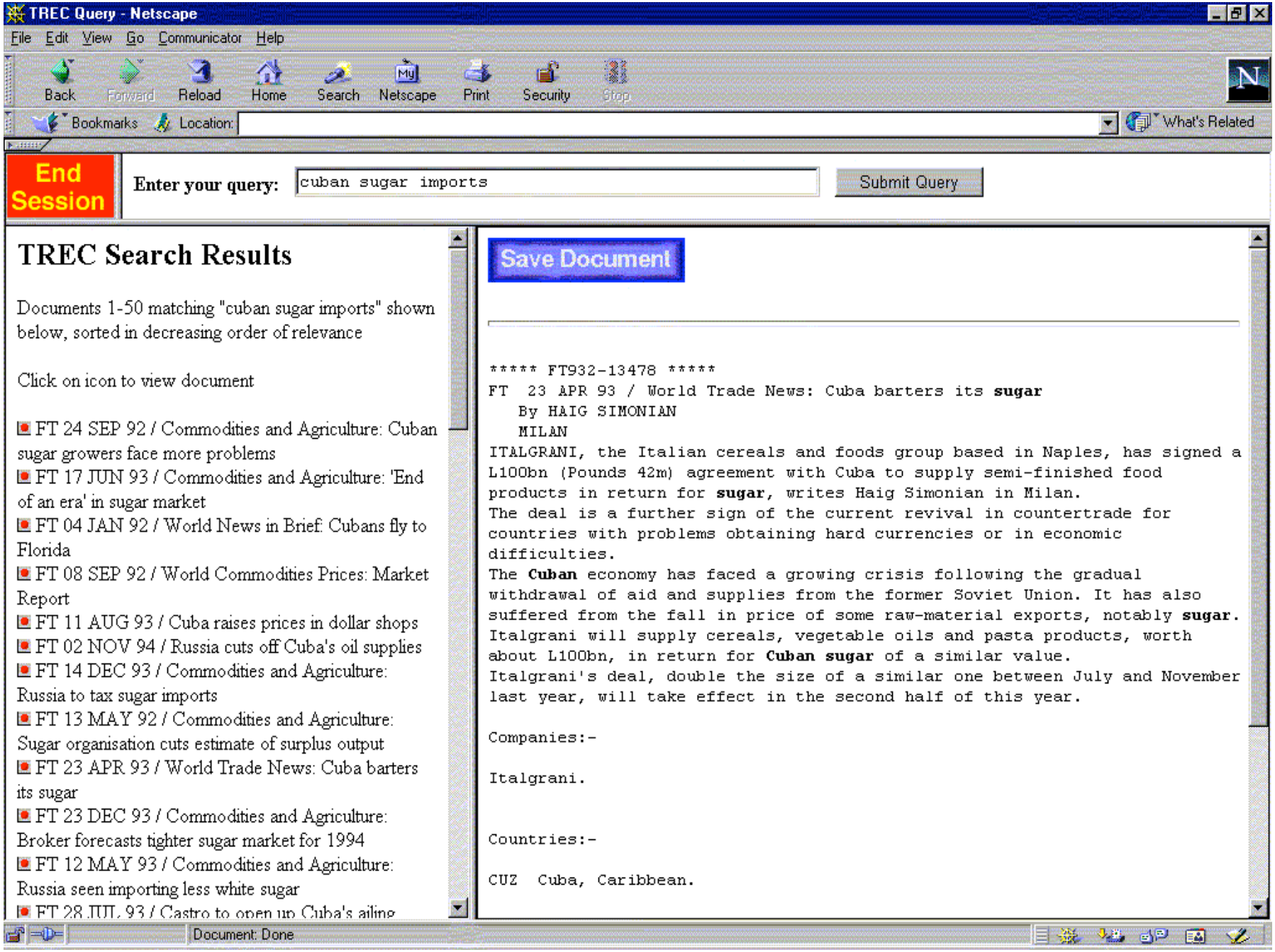

TREC ad-hoc
MAP
0.4
0.3
0.2
95
96
97
98
99
00

Experimental method

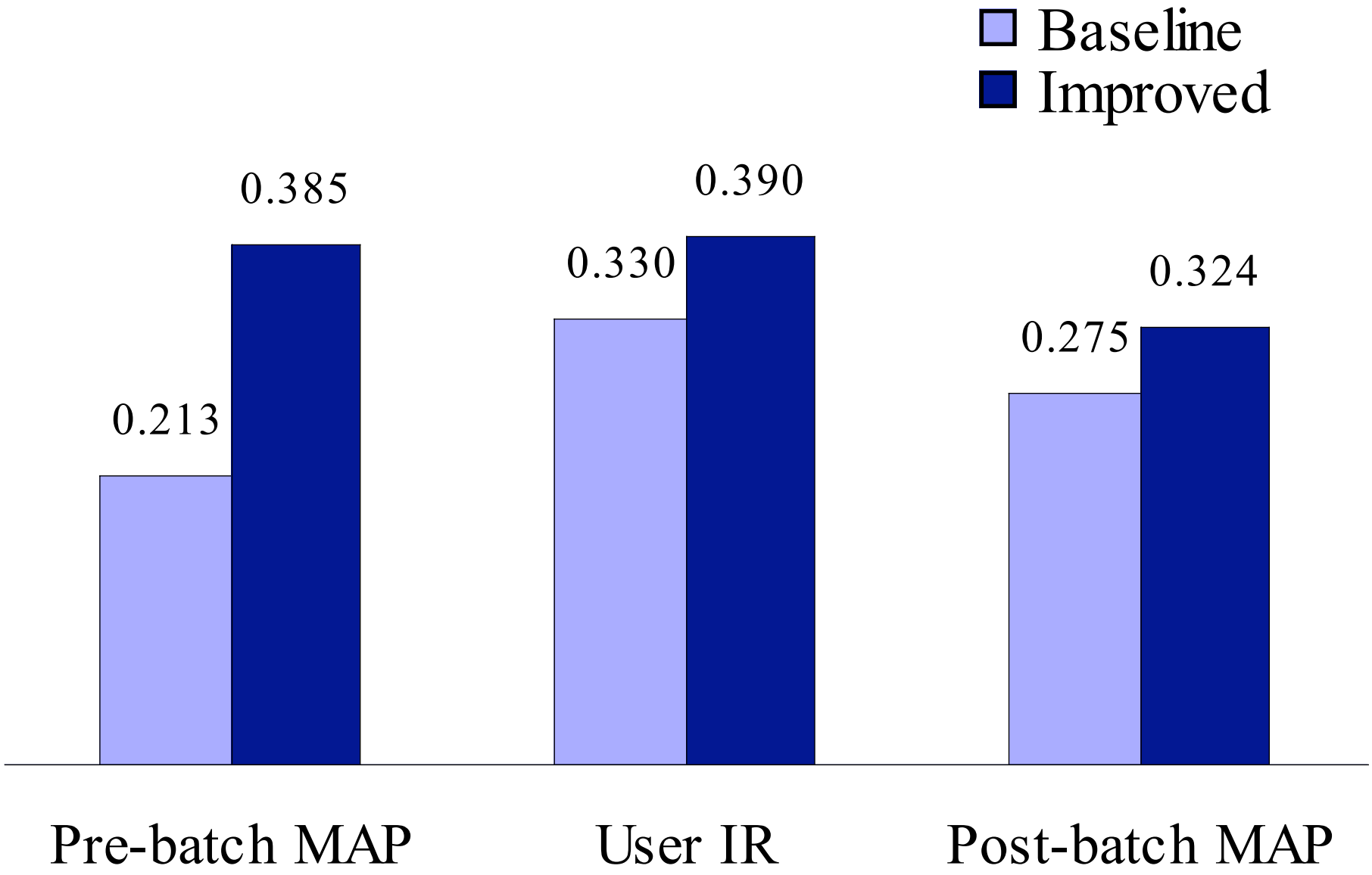
1. Set baseline system to basic Cosine Vector weights
2. Identify “super” system using batch experiments
3. Run 24 users on the 2 systems with same topics
4. Send results off to NIST
5. Get relevance judgments
6. Analyse user results
7. Check batch results

Example instance recall query

Number:
414i
Title:
Cuba, sugar, imports
Description:
What countries import Cuban sugar?
Instances:
In the time alloted, please find as many DIFFERENT countries of the sort described above as you can. Please save at least one document for EACH such DIFFERENT country.
If one document discusses several such countries, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT countries of the sort described above as possible.


TREC Query - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: What's Related
End Session
Enter your query: cuban sugar imports
Submit Query
TREC Search Results
Documents 1-50 matching "cuban sugar imports" shown below, sorted in decreasing order of relevance
Click on icon to view document
FT 24 SEP 92 / Commodities and Agriculture: Cuban sugar growers face more problems
FT 17 JUN 93 / Commodities and Agriculture: 'End of an era' in sugar market
FT 04 JAN 92 / World News in Brief: Cubans fly to Florida
FT 08 SEP 92 / World Commodities Prices: Market Report
FT 11 AUG 93 / Cuba raises prices in dollar shops
FT 02 NOV 94 / Russia cuts off Cuba's oil supplies
FT 14 DEC 93 / Commodities and Agriculture: Russia to tax sugar imports
FT 13 MAY 92 / Commodities and Agriculture: Sugar organisation cuts estimate of surplus output
FT 23 APR 93 / World Trade News: Cuba barters its sugar
FT 23 DEC 93 / Commodities and Agriculture: Broker forecasts tighter sugar market for 1994
FT 12 MAY 93 / Commodities and Agriculture: Russia seen importing less white sugar
Save Document
***** FT932-13478 *****
FT 23 APR 93 / World Trade News: Cuba barters its sugar
By HAIG SIMONIAN
MILAN
ITALGRANI, the Italian cereals and foods group based in Naples, has signed a L100bn (Pounds 42m) agreement with Cuba to supply semi-finished food products in return for sugar, writes Haig Simonian in Milan.
The deal is a further sign of the current revival in countertrade for countries with problems obtaining hard currencies or in economic difficulties.
The Cuban economy has faced a growing crisis following the gradual withdrawal of aid and supplies from the former Soviet Union. It has also suffered from the fall in price of some raw-material exports, notably sugar.
Italgrani will supply cereals, vegetable oils and pasta products, worth about L100bn, in return for Cuban sugar of a similar value.
Italgrani's deal, double the size of a similar one between July and November last year, will take effect in the second half of this year.
Companies:-
Italgrani.
Countries:-
CUZ Cuba, Caribbean.
Document: Done

Experiment 1 - Instance Recall



8 Q&A queries

1) What are the names of three US national parks where one can find redwoods?

2) Identify a site with Roman ruins in present day France

3) Name four films in which Orson Welles appeared

4) Name three countries that imported Cuban sugar during the period of time covered by the document collection

8 Q&A queries

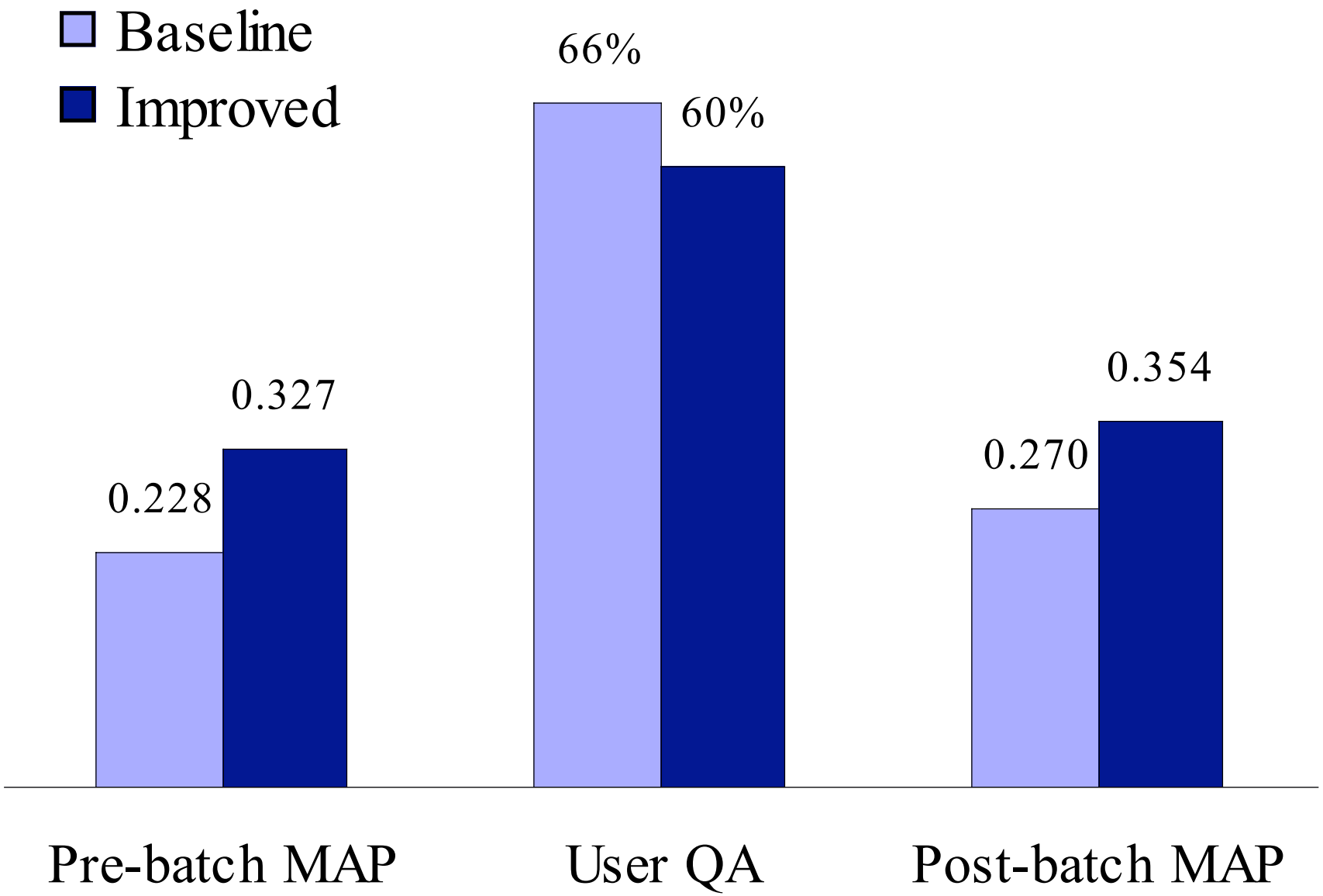
5) Which childrens TV program was on the air longer the original Mickey Mouse Club or the original Howdy Doody Show?

6) Which painting did Edvard Munch complete first Vampire or Puberty?

7) Which was the last dynasty of China Qing or Ming?

8) Is Denmark larger or smaller in population than Norway?

Experiment 2 - Question Answering



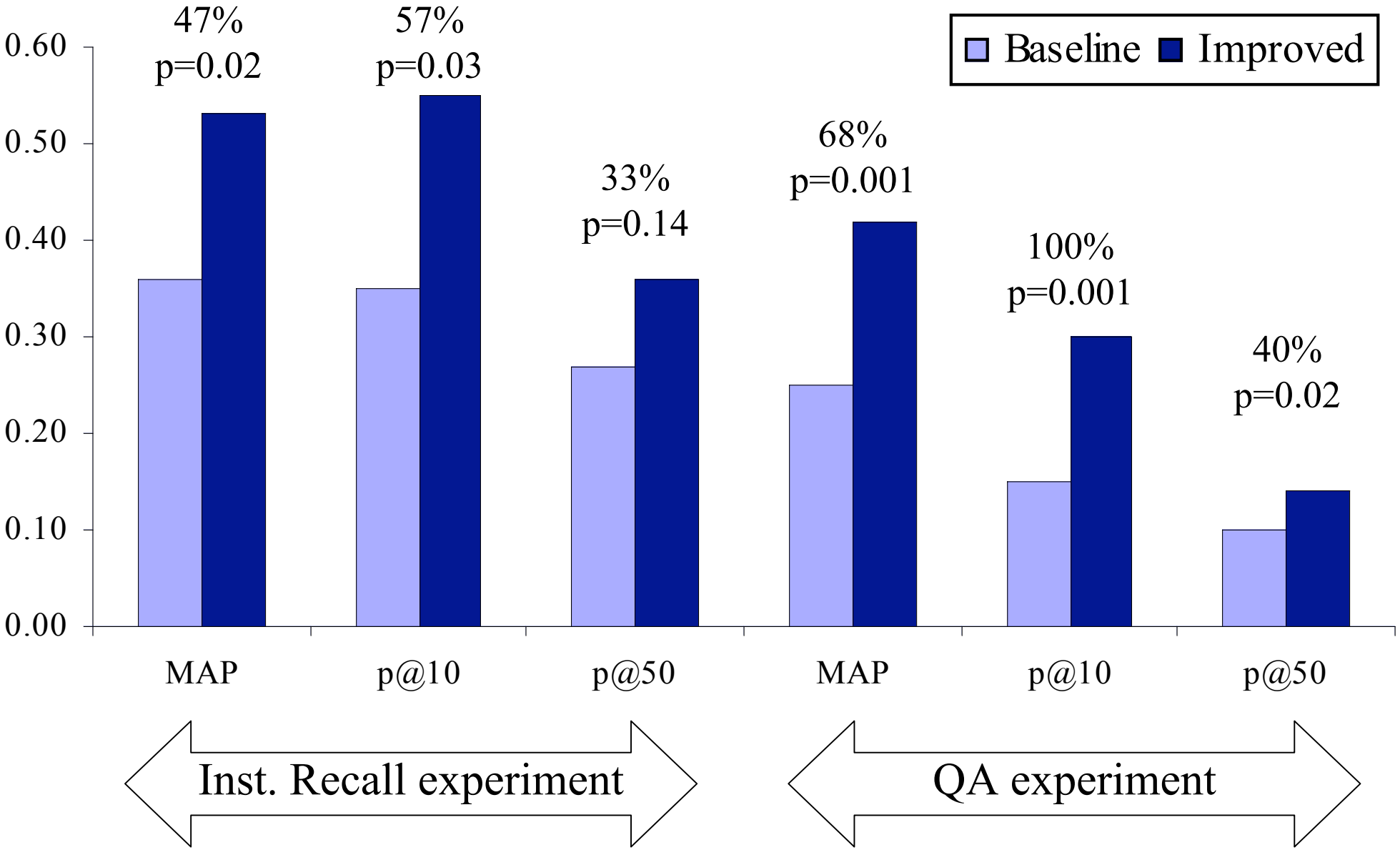
Results Summary

	Predicted	Actual
Instance recall	81%	15% (p = 0.27)
Question answering	58%	-6% (p = 0.41)

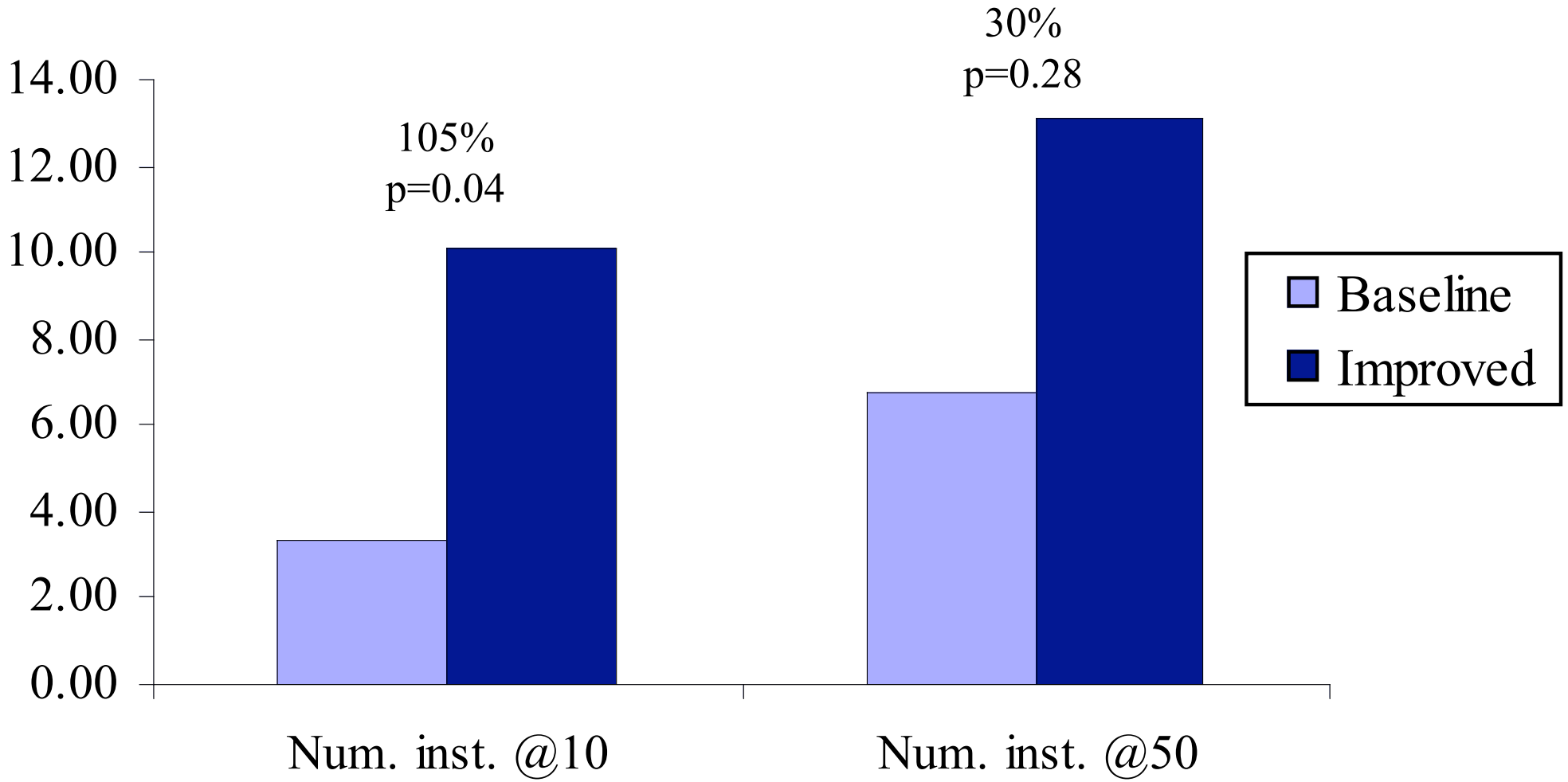
Why?

1. Systems no different on topics and collection used
2. There was a difference, but users ignored it

Precision metrics on user queries and collection



Number of instances on user queries and collection



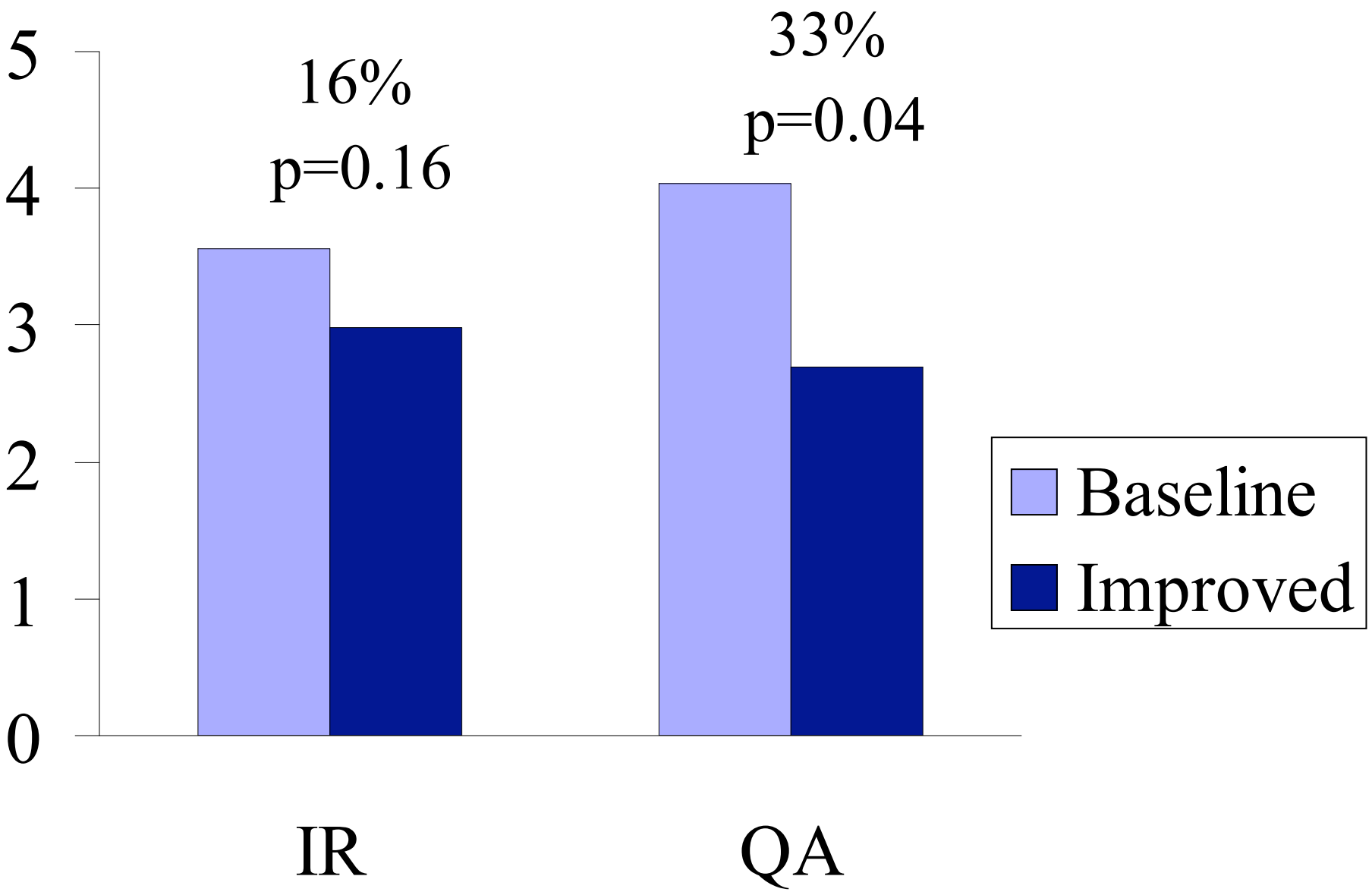
So what happens to the difference?

- Users compensate for the lack of relevant docs within time limit

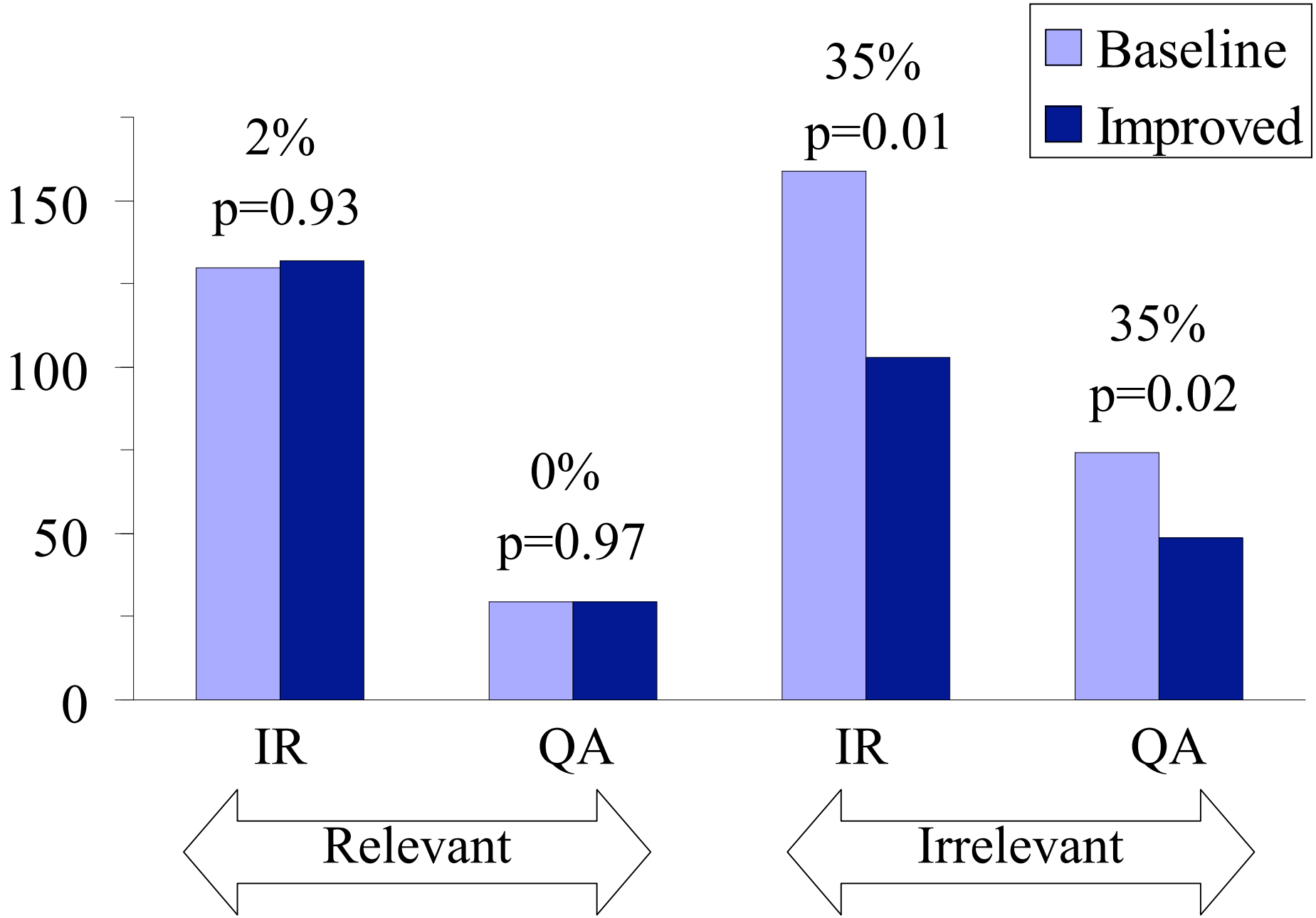
- Users ignore high ranked relevant documents
  - Maybe obscure document titles?
  - Don't read the list from the top?

- "Extra" relevant docs give no new information

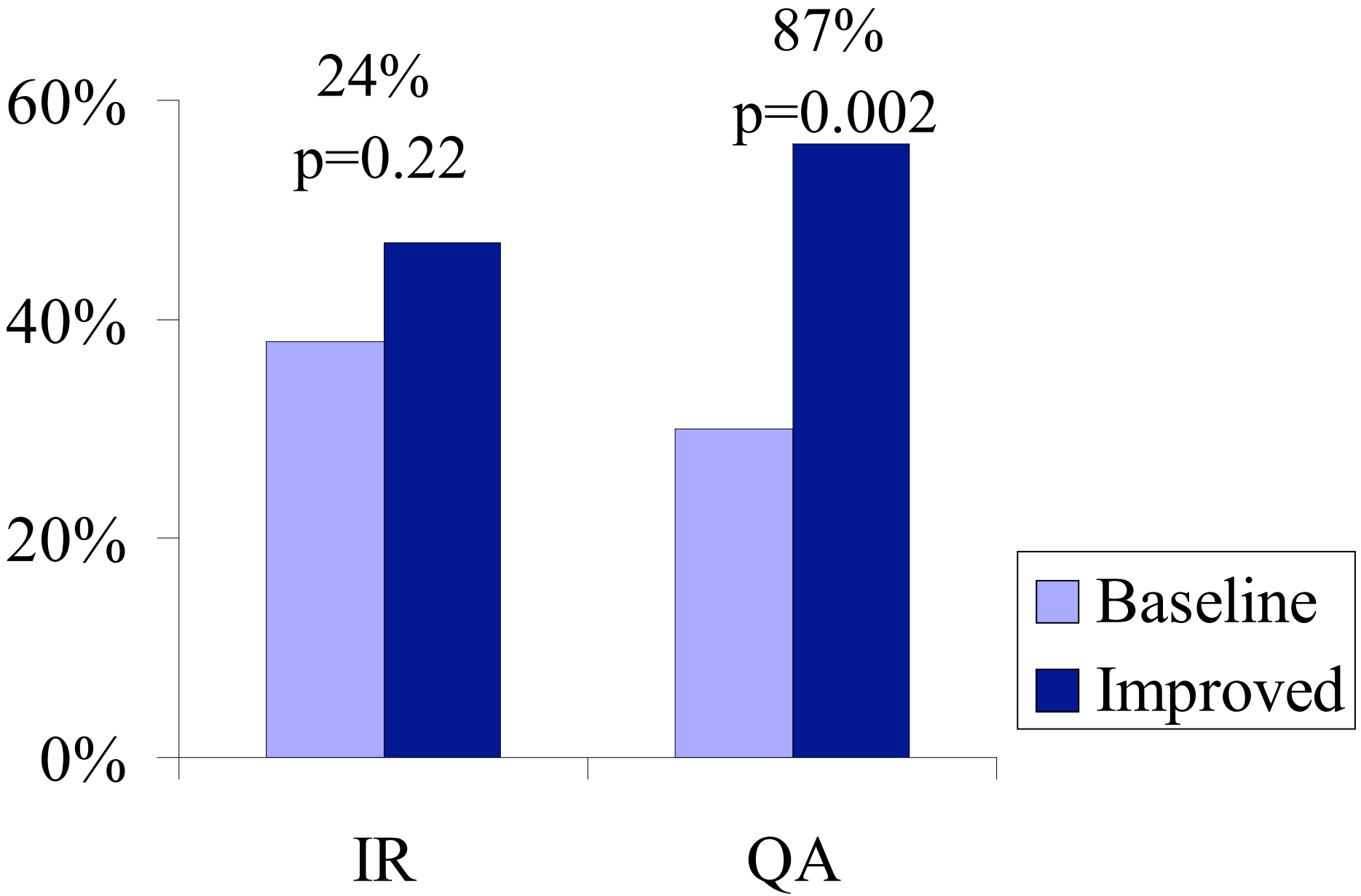
Number of queries per topic



Number of docs retrieved



Number top 10 relevant docs ignored



Conclusion

- In these two tasks there is no use providing users with a good weighting scheme because
  - They will ignore high ranking relevant docs
  - They will happily issue a few extra queries
- They find answers just as well with old technology
- User interface effects?
- Task effect?

Basic cosine

$$\sum_{t \in T_{q,d}} \frac{TF(t,d) \times IDF(t)}{\sqrt{\sum_{t \in T_d} TF(t,d)^2}}$$

Okapi

$$\sum_{t \in T_{q,d}} \frac{IDF(t)^2 \times f_{d,t}}{f_{d,t} + W_d}$$

Pivoted Okapi

$$\sum_{t \in T_{q,d}} f_{q,t} \times \ln\left(\frac{N - f_t}{f_t}\right) \times \frac{f_{d,t}}{f_{d,t} + W_d'}$$