

Evaluation Experiments and Experience from the Perspective of Interactive Information Retrieval

Ross Wilkinson
Mingfang Wu
ICT Centre
CSIRO, Australia

Outline

- A history of information retrieval (IR) evaluation
 - System-oriented evaluation
 - User-oriented evaluation
- Our experience with user-oriented evaluation
- Our observation
- Learnt lessons

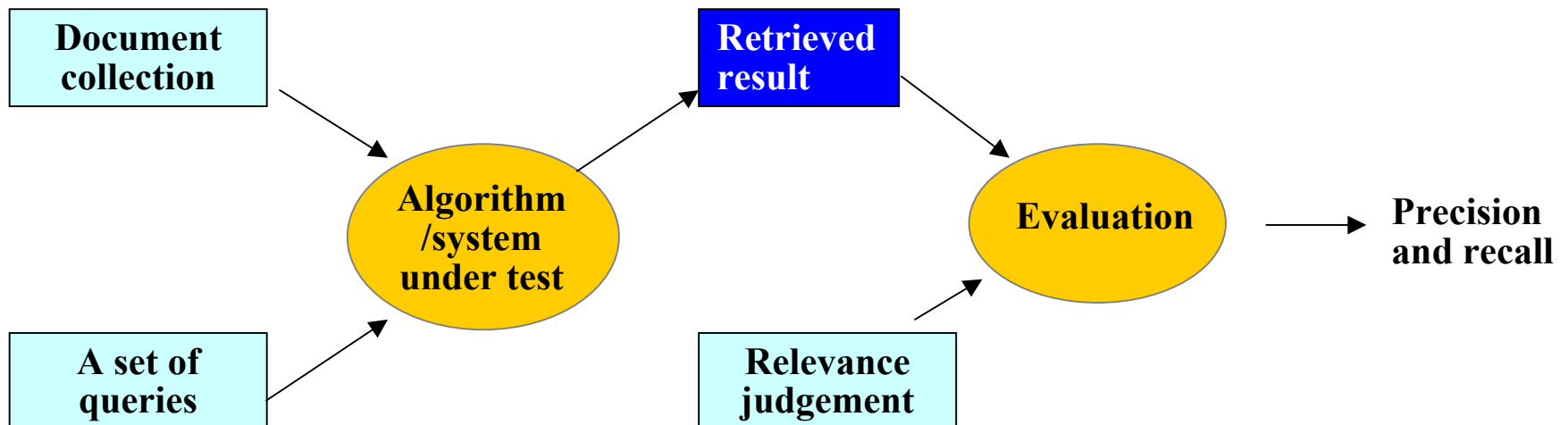
Information Retrieval

Why evaluate an IR system?

- To select between alternative systems/algorithms/models
- What is the best for:
 - Ranking function (dot-product, cosine, ...)
 - Term selection (stop word removal, stemming...)
 - Term weighting (TF, TF-IDF,...)

The traditional IR evaluation

- **Test collection:** a collection of documents, a set of queries, the relevance judgement
- **Process:** input the documents, put each query to the system, collect the output
- **Measurement:** usually precision and recall



Early Test Collections

Different research groups used different and small test collections:

- Hard to generalize the research outcomes
- Hard to compare systems/algorithms across sites

The TREC Benchmark

- Text Retrieval Conference - organized by NIST, started in 1992, about 93 groups from 22 countries participated in 2003.
- Purposes:
 - To encourage research in IR based on large text collections.
 - To provide a common ground/task evaluation that allows cross-site comparison.
 - To develop new evaluation techniques, particularly for new applications, e.g.
 - filtering, cross-language retrieval, web retrieval, high precision, question answering

Problems with the system-oriented experiment

- Pros:
 - Advanced the system development
- Cons:
 - System is an input-output device, while most real searches involve interaction.
 - Relevance is binary and judged independently of context, while relevance is
 - Subjective: Depends upon a specific user's judgment.
 - Situational: Relates to user's current needs.
 - Cognitive: Depends on human perception and behavior.
 - Dynamic: Changes over time.

TREC interactive track

- Goal: to investigate searching as an interactive task by examining the *process* as well as the *outcome*.

Interactive track tasks

- TREC3-4: finding relevant documents
- TREC5-9: finding any N short answers to a question, to which there are multiple answers of the same type.
- TREC10-11: finding any N short answers to a question and finding any N websites that meet the need specified in the task statement
- TREC12: topic distillation

Interactive track tasks

- TREC3-4: finding relevant documents
- TREC5-9: finding any N short answers to a question, to which there are multiple answers of the same type.
- TREC10-11: finding any N short answers to a question and finding any N websites that meet the need specified in the task statement
- TREC12: topic distillation

How to measure outcome?

- Aspectual precision
 - The proportion of the documents identified by a subject that were deemed to contain topic aspects.
- aspectual recall
 - The proportion of the know topic aspects contained in the documents identified by a subject.

How to measure process?

- Objective measures:
 - No. of query iterations
 - No. of document surrogates seen
 - No. of documents read
 - No. of documents saved
 - Actual time used
- Subjective measures:
 - searchers' **satisfaction** with the interaction
 - searchers' **self-perception** of their task completeness
 - searchers' **preference** of an search system/interface

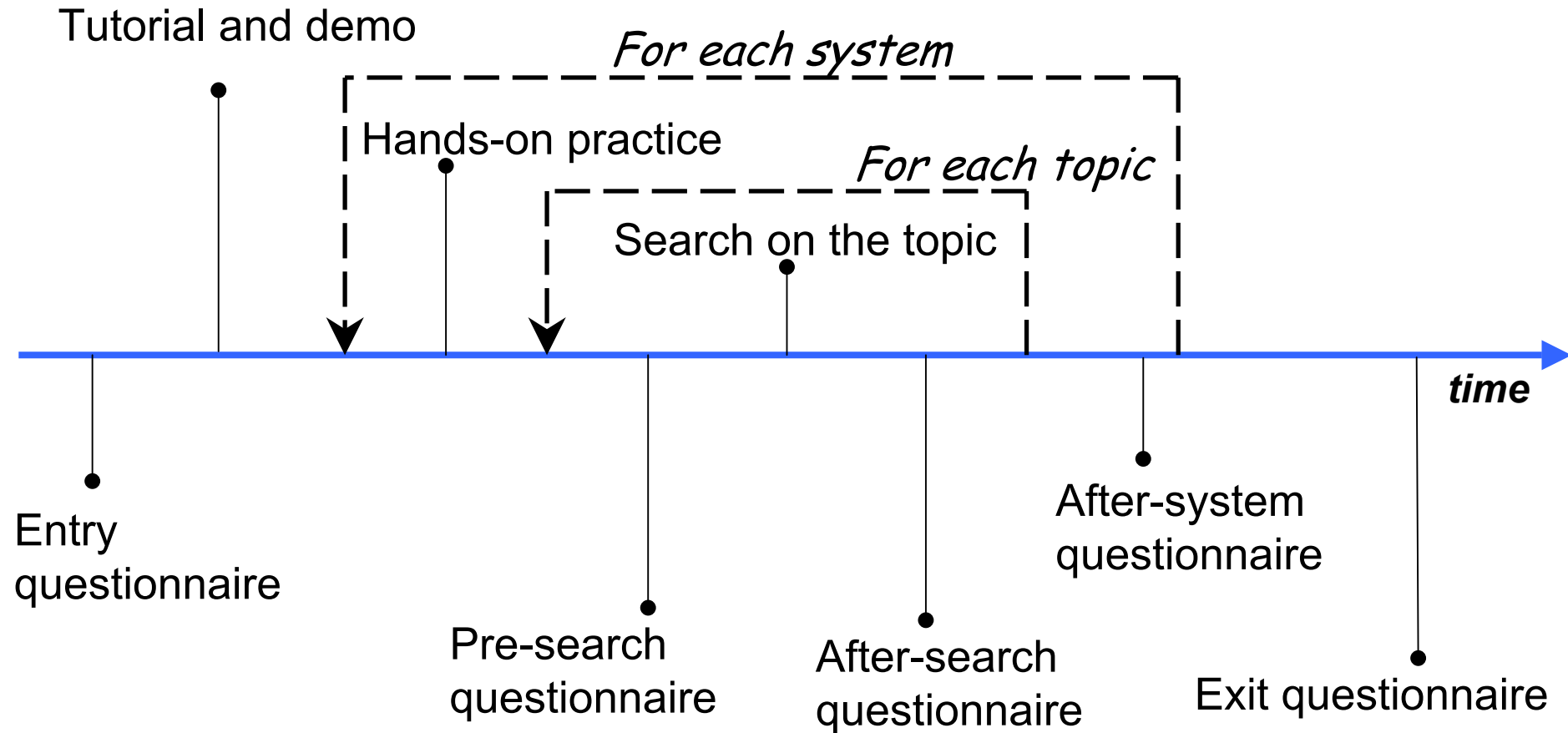
Experimental Design

- Factors: searchers, topic, and system
- Latin square experimental design

| Searchers | System, Topic | |
|-----------|---------------|-------|
| 1 | E, B1 | C, B2 |
| 2 | C, B2 | E, B1 |
| 3 | E, B2 | C, B1 |
| 4 | C, B1 | E, G2 |

E: Experimental System, C: Control System
B1 and B2 are two blocks of (4) topics

Experimental Procedure



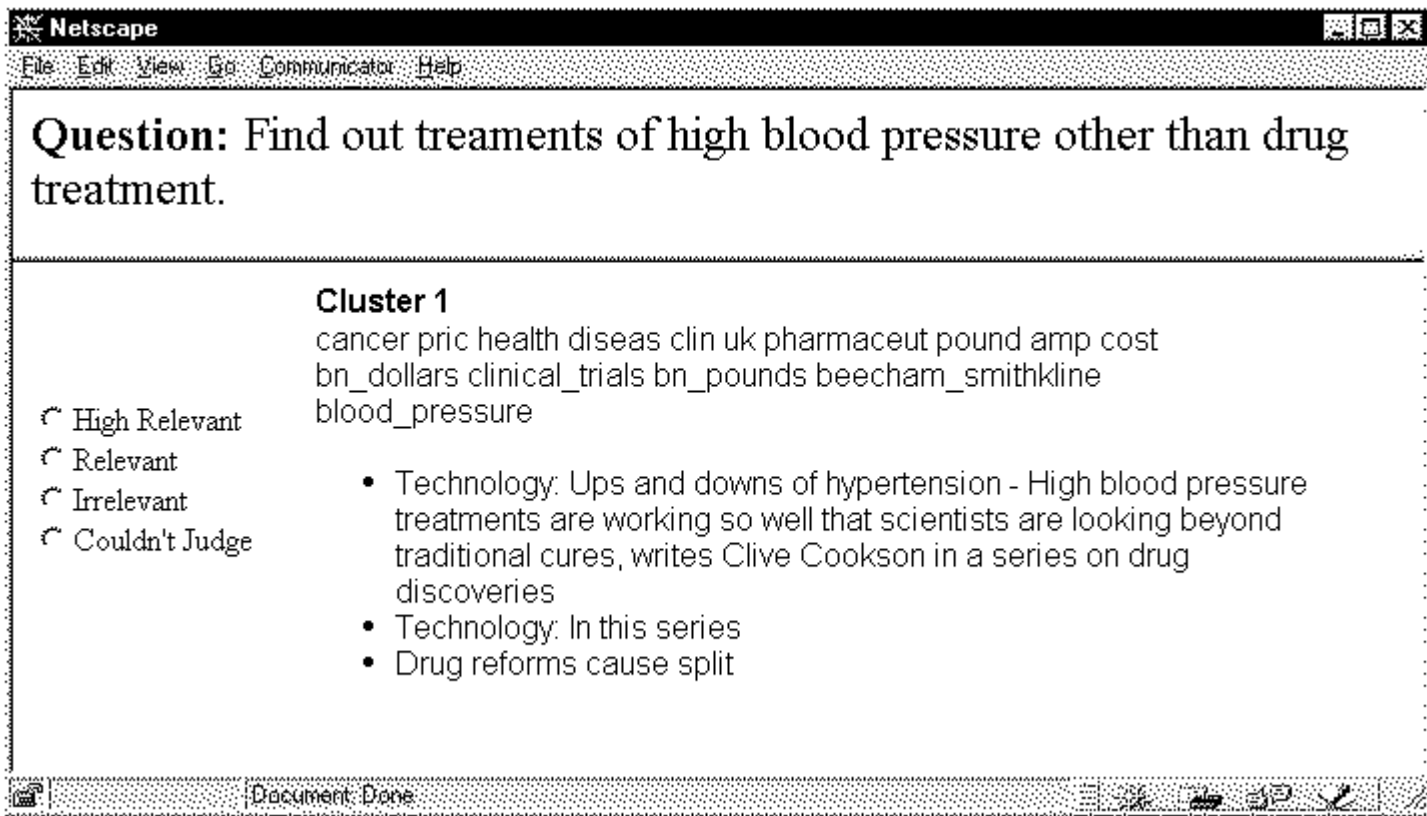
Experiment I – clustering vs ranked list (I)

- Hypothesis: clustering structure is more effective than a ranked list for the aspect finding task.

Experiment I – clustering vs ranked list (II)

- Stage I – Can subjects recognize good clusters?
- Experimental task: to judge the relevance of a cluster to the topic based only on the description of cluster
- Non-standard TREC experiment, four subjects are involved.

The interface for judging the relevance of clusters



Experiment I – clustering vs ranked list (III)

- Stage II: Can clusters be used effectively for aspect finding task?
- TREC experiment: 8 topics, 16 searchers

The list interface

MDS TREC7 Interactive Experiment - Netscape

File Edit View Go Communicator Help

Topic (example): Find out treatments of high blood pressure other than drug treatment. [Next Topic](#)

A list of retrieved documents:

- [FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries](#)
- [FT 29 NOV 94 / Technology: In this series](#)
- [FT 07 FEB 92 / Technology \(Worth Watching\): Relief in sight for asthma sufferers](#)
- [FT 14 NOV 92 / Drug reforms cause split](#)
- [FT 05 APR 93 / Leading Article: Drugs on trial](#)
- [FT 02 JAN 93 / UK Company News: Three groups' shares rise on drugs approval](#)
- [FT 29 JUL 94 / Technology: Brains on their minds - Drugs researchers are seeking a stroke treatment that could transform current therapy](#)
- [FT 02 SEP 94 / Technology: Towards a cure for blindness](#)
- [FT 08 FEB 94 / UK Company News: Glaxo asthma drug wins US approval](#)
- [FT 01 NOV 94 / UK Company News: British Biotech new cancer drug - Third promising treatment makes company one of best in sector](#)
- [FT 07 SEP 93 / Technology: A renaissance in treatment - New drugs to treat schizophrenia are finally becoming available](#)
- [FT 20 FEB 93 / New cancer drugs show promise](#)
- [FT 20 JUL 92 / Wellcome expects good news on Aids drug tests](#)
- [FT 31 MAR 94 / Technology: Deadly challenge proves costly - Daniel Green examines the continuing search for an effective sepsis treatment, in a series on drugs](#)
- [FT 03 NOV 93 / International Company News: Merck drug withdrawn](#)
- [FT 28 MAR 94 / Glaxo drug rival wins licence for UK](#)
- [FT 19 APR 91 / World News in Brief: Drug money fund](#)
- [FT 07 DEC 93 / Drug row flares as calls for tighter guidelines grow](#)
- [FT 18 MAR 94 / New cancer drug shows promise in first tests](#)
- [FT 10 FEB 94 / Clinton in new drugs push](#)
- [FT 14 FEB 92 / Technology \(Worth Watching\): Watching high blood pressure](#)
- [FT 28 MAR 92 / Trials and tribulations: Large-scale clinical drugs testing](#)
- [FT 08 APR 94 / Effective Aids drugs 'a long way off': Doctors divided over HIV treatment as study casts doubts on leading](#)

FT931-2736
_AN-DCRCHAGAFT
930318

FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries

By CLIVE COOKSON

Drugs to bring down high blood pressure are one of the great successes of pharmaceutical research. Over the past decade the industry has given doctors dozens of new drugs to treat hypertension - the medical name for the condition - by several different mechanisms. Their sales are worth more than Dollars 10bn (Pounds 7bn) a year, three times as much as the total market for cancer drugs. 'The treatment of hypertension is very good now and the side effects are minor,' says Desmond Julian, medical director of the British Heart Foundation, 'and because there is a range of drugs, you can normally find one to suit any particular patient.'

In industrialised countries, 15 to 20 per cent of the adult population has high blood pressure. Julian says patients with mild or moderate hypertension should not be put on drugs straightaway; their doctors should urge them to make changes in diet and lifestyle. But for the 5 per cent of people with severe hypertension, drugs are usually required to bring blood pressure down to a safe level. Clinical trials have shown that the greatest benefit of hypertension treatment is a 40 per cent reduction in the risk of suffering a stroke, which is caused by the rupture of blood vessels in the brain. The effects on other forms of cardiovascular disease are less clear-cut; indeed there is

[Add New Aspect](#)

Relevant aspects already identified for this topic:

1. calcium
2. regular exercise
3. biofeedback

The clustering interface

MDS TREC7 Interactive Experiment - Netscape
File Edit View Go Communicator Help

Topic (example): Find out treatments of high blood pressure other than drug treatment. Next Topic

Group 1
cancer pric health diseas clin uk pharmaceut pound amp cost
bn_dollars clinical_trials bn_pounds beecham_smithkline blood_pressure
• [FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries](#)
• [FT 29 NOV 94 / Technology: In this series](#)
• [FT 14 NOV 92 / Drug reforms cause split](#)
▶ (214 more documents)

Group 2
protein red pig arter bod cel worth human clot del
blood_human haemoglobin_human blood_substitutes blood_red blood_cells
• [FT 07 FEB 92 / Technology \(Worth Watching\): Relief in sight for asthma sufferers](#)
• [FT 14 FEB 92 / Technology \(Worth Watching\): Watching high blood pressure](#)
• [FT 09 DEC 94 / Technology \(Worth Watching\): New class of heart drugs for trial](#)
▶ (7 more documents)

Group 3
ulcer shar migrain produc imigran markes approv exchange fison pound
drug_food administration_drug bn_pounds pre_tax dr_sykes
• [FT 08 FEB 94 / UK Company News: Glaxo asthma drug wins US approval](#)
• [FT 30 DEC 92 / Glaxo wins US approval for migraine drug](#)
• [FT 17 JAN 92 / US drug regulator criticises Fisons product](#)
▶ (24 more documents)

Group 4
hous fund war law le seiz cor sen prison graham
control_drug core_hard enforcement_law drug_treatment drug_policy
• [FT 19 APR 91 / World News in Brief: Drug money fund](#)
• [FT 10 FEB 94 / Clinton in new drugs push](#)
• [FT 21 OCT 93 / US drug policy aims to reduce hard-core use](#)
▶ (12 more documents)

Group 5
intern administer rorer writ markes yesterda kar zagor recommends advisor
administration_drug drug_food dollars_dollars poulenc_rhone close_dollars
• [FT 02 JAN 93 / UK Company News: Three groups' shares rise on drugs approval](#)
• [FT 03 NOV 93 / International Company News: Merck drug withdrawn](#)
• [FT 10 SEP 93 / International Company News: US group's Alzheimer's drug wins approval](#)
▶ (15 more documents)

Group 6
test servic contamin don haemophiliac donor nat new par heat
hiv_virus blood_transfusion blood_transfusions blood_products service_transfusion
• [FT 22 APR 92 / World News In Brief: Blood donors by order](#)
• [FT 10 APR 92 / World News in Brief: Drug bust](#)
• [FT 28 APR 92 / World News In Brief: HIV payments](#)
▶ (18 more documents)

Group 7
account part program control program paperoidal uk health time

A list of documents in Group 1:

- [FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries](#)
- [FT 29 NOV 94 / Technology: In this series](#)
- [FT 14 NOV 92 / Drug reforms cause split](#)
- [FT 05 APR 93 / Leading Article: Drugs on trial](#)

FT931-2736
_AN-DCRCHAGAFT
930318

FT 18 MAR 93 / Technology: Ups and downs of hypertension - High blood pressure treatments are working so well that scientists are looking beyond traditional cures, writes Clive Cookson in a series on drug discoveries

By CLIVE COOKSON

Drugs to bring down high blood pressure are one of the great successes of pharmaceutical research. Over the past decade the industry has given doctors dozens of new drugs to treat hypertension - the medical name for the condition - by several different mechanisms. Their sales are worth more than Dollars 10bn (Pound 7bn) a year, three times as much as the total market for cancer drugs. 'The treatment of hypertension is very good now and the side effects are minor,' says Desmond Julian, medical director of the British Heart Foundation, 'and because there is a range of drugs, you can normally find one to suit any particular patient.' In industrialised countries, 15 to 20 per cent of the adult population has high blood pressure. Julian says patients with mild or moderate hypertension

Add New Aspect

Relevant aspects already identified for this topic:

1. calcium
2. snakeroot plant
3. regular exercise

Experiment I – findings

- Clustering structure works for some topics, but overall there is no significant difference between the clustering structure and the ranked list.
- Subjects preferred the clustering interface.

Experiment II - Document summary

- The relevant facts may exist within small chunks of a document, and these small chunks may not necessarily be related to the main theme of the document.
- These small chunks usually contain the keywords, and in the form of a complete sentence. We call this sentence the *answer indicative sentence* (AIS).
- When a user is scanning through a document to search for facts, s/he usually uses zoom-out strategy - keywords -> sentence -> document

Experiment II - hypothesis

- Hypothesis: The answer indicative sentences are better surrogate of a document than the first N words for the purpose of interactive fact finding.

The AIS

- An AIS should contain at least one query word and be at least ten words long.
- The AIS' are first ranked according to the number of unique query words contained in each AIS. If two AIS' have the same number of unique query words, they will be ranked according to their appearing sequence in the document.
- The top three AIS are then selected.

Control System (FIRST20)

The screenshot shows a Microsoft Internet Explorer browser window with the address bar displaying `http://ceres.mds.rmit.edu.au:8080/cgi-bin/TREC9.main.cgi?userid=2&qno=5`. The browser's menu bar includes File, Edit, View, Favorites, Tools, and Help. The toolbar contains icons for Back, Forward, Stop, Refresh, Home, Search, Favorites, History, Mail, Print, and Edit. The main content area displays the following information:

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

china dynasty qing ming

- [Claim Record Sale For Porcelain Ming Vase](#)
A 14th century **Ming dynasty Chinese** vase sold at auction Tuesday for nearly \$2.2 million, a record price for ...
- [Chinese Dish For Calligraphy Brushes Brings Record Price](#)
A 12th century **Chinese** ceramic dish used to wash calligraphy brushes was sold at auction Tuesday for \$2.82 million, ...
- [Hosokawa Meets With Chinese Radio-TV Minister](#)
BFN [From "News 7" program] [Text] Visiting **Chinese** Radio, Film, and Television Minister Ai Zhisheng met ...
- [FT 04 JUN 94 / Collecting: Rarities in blue and white - Susan Moore samples the numerous wares of quality Chinese porcelain](#)
Chinese porcelain takes pride of place at the seasonal Oriental art shows in London this month. Even the British ...
- [Classical Paintings Exhibit To Tour United States](#)
Eighty classical **Chinese** paintings dating back to the 14th century will be exhibited in five U.S. cities next year ...
- [Preserving Minority Cultures Said Urgent Task](#)
Language: English Article Type:BFN [By staff reporter Zhang Xia: "Successes Mixed With Anxiety"] ...
- [Jade Brightens Sotheby's Sale In Hong Kong ---- By Michael Duckworth Staff Reporter of The Wall Street Journal](#)
PARAGRAPH 2 ##### The three-day Sotheby's sale last week was somewhat smaller than the auction by rival Christie's ...

Experimental System (AIS3)

http://ceres.mds.rmit.edu.au:8080/cgi-bin/TREC9.main.cgi?userid=1&qno=4 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit

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

china dynasty qing ming Search Clear Query Next Topic

- [Claim Record Sale For Porcelain Ming Vase](#)
 - ☐ [A 14th century Ming dynasty Chinese vase sold at auction Tuesday for nearly \\$2.2 million, a record price for a porcelain piece, an auction house said.](#)
 - ☐ [The vase was part of a 118-piece private collection of porcelain from the Ming and Qing dynasties \(1368-1911\) that fetched more than \\$8.7 million Tuesday, Sotheby's said.](#)
 - ☐ [" Thompson said the record was set a year ago when a piece of Chinese porcelain sold for about \\$1.4 million.](#)
- [Chinese Dish For Calligraphy Brushes Brings Record Price](#)
 - ☐ [A Ming Dynasty \(1368-1644\) decorated basin sold for \\$2.61 million, auction officials said.](#)
 - ☐ [A 12th century Chinese ceramic dish used to wash calligraphy brushes was sold at auction Tuesday for \\$2.82 million, officials said.](#)
 - ☐ [A private collector, Sunrider International of Los Angeles, bought the rare mellow flower-shaped brushwasher from the imperial court of the Song Dynasty \(1127-1279\), according to Sotheby's.](#)
- [Hosokawa Meets With Chinese Radio-TV Minister](#)
 - ☐ [In the meeting, Radio, Film, and Television Minister Ai described recent progress in filming of NHK's special programs on China , and explained an agreement with NHK on filming of Gugong \[or Zijincheng\], which used to be the national palace of the Ming and Qing dynasties .](#)
 - ☐ [BFN \[From "News 7" program\] \[Text\] Visiting Chinese Radio, Film, and Television Minister Ai Zhisheng met with Prime Minister Morihiro Hosokawa today \[6 April\].](#)
 - ☐ [He noted the Chinese Government takes a positive stance toward cooperation with Japanese television stations' operations in China .](#)
- [FT 04 JUN 94 / Collecting: Rarities in blue and white - Susan Moore samples the numerous wares of quality](#)

Internet

Experiment II - findings

- Topic by topic, AIS3 has more successful sessions than the First20 in 7 topics (out of 8 topics).
- Subject by subject, 10 subjects are more successful with the AIS3 than the First20, 2 subjects are more successful with the First20 than the AIS3.
- Subjects thought the AIS3 is easier to use, preferred the AIS3, and takes less interactions with the AIS3.

Experience

- TREC interactive track evaluation platform
 - Pros:
 - leverage the effort to build the evaluation platform
 - Well developed experimental design and procedure
 - Cons:
 - small number of subjects and topics
 - hard to repeat experiment
- difficult to interpret results
 - E.g. performance vs. preference
- effective delivery works in the right context