

## Towards better contextual advertising

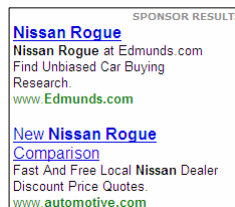
Evgeniy Gabrilovich  
[gabr@yahoo-inc.com](mailto:gabr@yahoo-inc.com)  
Yahoo! Research

*Joint work with* Aris Anagnostopoulos, Andrei Broder,  
Marcus Fontoura, Amruta Joshi, Vanja Josifovski,  
Lance Riedel, and Tong Zhang



## Background: Online Advertising

- A \$17B+ market in 2006, 20% annual growth
- Banner ads vs. textual ads



- Textual ads should be contextually relevant!



# Contextual advertising

The screenshot shows a Yahoo! search results page for the query "car insurance". The search bar at the top contains "car insurance" and the Yahoo! logo is on the right. Below the search bar, there are several sponsored search results for car insurance, including GEICO, Progressive, Esurance, and AAA. A speech bubble labeled "Sponsored Search" points to these results. Below the sponsored results, there is a search result for "Mesothelioma" from e-Mesothelioma. A speech bubble labeled "Content Match" points to this result. The "Mesothelioma" result includes a title, a brief description, and a "How To Sleep" link. The Yahoo! Research logo is visible in the bottom right corner of the screenshot.

## Part I

# Robust Classification of Rare Queries Using Web Knowledge (SIGIR 2007)

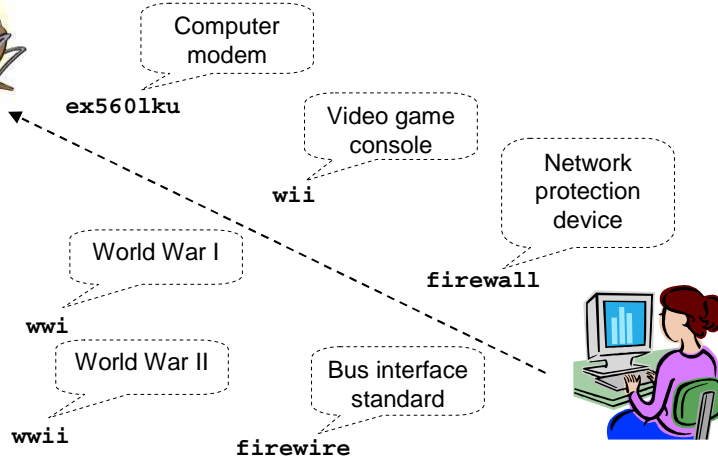
## Background: Web Search

- Web search: a factor in the life of billion+ people
  - Multi-billion dollar advertising industry
- Web search is difficult because queries are very short (2.4-2.7 words)
- How to infer the semantics of the queries?
  - Multiple uses in improving both search and advertising results



## Queries, queries, queries ...

Search engine



## Why care?

- If we knew more about queries, we could ...
  1. Return **better search** results
  2. Serve **more relevant ads**
    - Currently, some queries yield no ads at all
  3. Use selected databases to augment general search results
    - Focused metasearch
  4. Sensitivity filtering (adult, gambling, ...)
- One way to know more is to classify the queries w.r.t. to a large knowledge taxonomy
  - Using exogenous knowledge



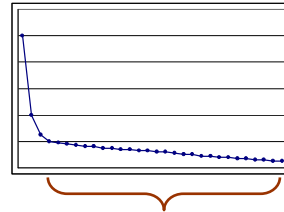
## Examples

- ex5601ku
  - Computing/Computer Hardware/Computer Peripherals/**Computer Modems**
- Heated intraoperative intraperitoneal chemotherapy, cachexia ascites
  - Health and Beauty/Medical Conditions/Respiratory Disorders/**Mesothelioma**
  - Professional Services/Legal Services/Class Action Legal Services/**Asbestos Class Actions**



## Problem statement: Query classification

- Query distribution follows a power law – a few common queries, **numerous** rare ones
- Rare queries are **difficult**
  - little training can be done
- Rare queries are **important**
  - when combined, they amount to a great mass



YAHOO!  
RESEARCH

## Our approach

- Ideally, we would like to consult a (human) domain expert
  - Too many queries to handle manually ...
- Use blind relevance feedback
  - Learn from top search results
    - Classify search results (either summaries or full pages)
    - Let them vote to determine the query class(es)

YAHOO!  
RESEARCH

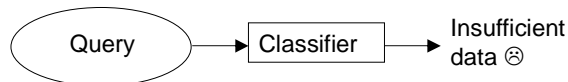


# Mesothelioma

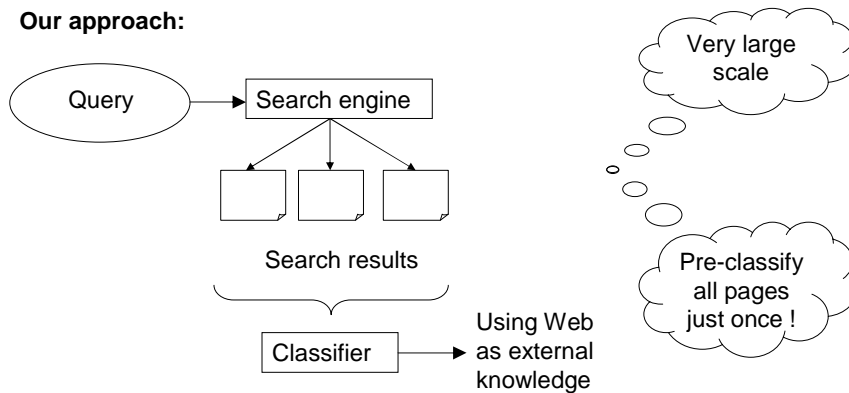
The screenshot shows a Yahoo! Search results page for the query 'Mesothelioma'. At the top, there are navigation links for 'Web', 'Images', 'Video', 'Local', and 'Shopping'. The search bar contains 'Mesothelioma' and a 'Search' button. Below the search bar, there are links for 'Answers', 'Search Services', 'Advanced Search', and 'Preferences'. The search results section shows '1 - 10 of about 15,000,000 for Mesothelioma - 0.37 sec. (About this page)'. A lightbulb icon suggests alternative terms: 'Also try: mesothelioma lawyer, mesothelioma asbestos, malignant mesothelioma More...'. The results are categorized into 'SPONSOR RESULTS' and 'What is Mesothelioma'. The 'SPONSOR RESULTS' section includes four items: 'Mesothelioma Collect Millions of Dollars' (www.mesothelioma-lawsuits-asbestos.com), 'Mesothelioma Help' (www.mesotheliomaweb.org), 'Mesothelioma Help and Legal Options' (www.mesotheliomaoptions.com), and 'Mesothelioma - Free Info Packet' (mesothelioma-lung-cancer.org). The 'What is Mesothelioma' section includes 'Mesothelioma Lawyers' (Millions of dollars for mesothelioma victims. 28 yrs of experience...), 'Mesothelioma Treatment Update' (Medical treatment & legal help for Mesothelioma victims and families.), and 'Mesothelioma Information' (Get legal help from experienced Mesothelioma attorneys today.).



## Traditional approach:



## Our approach:

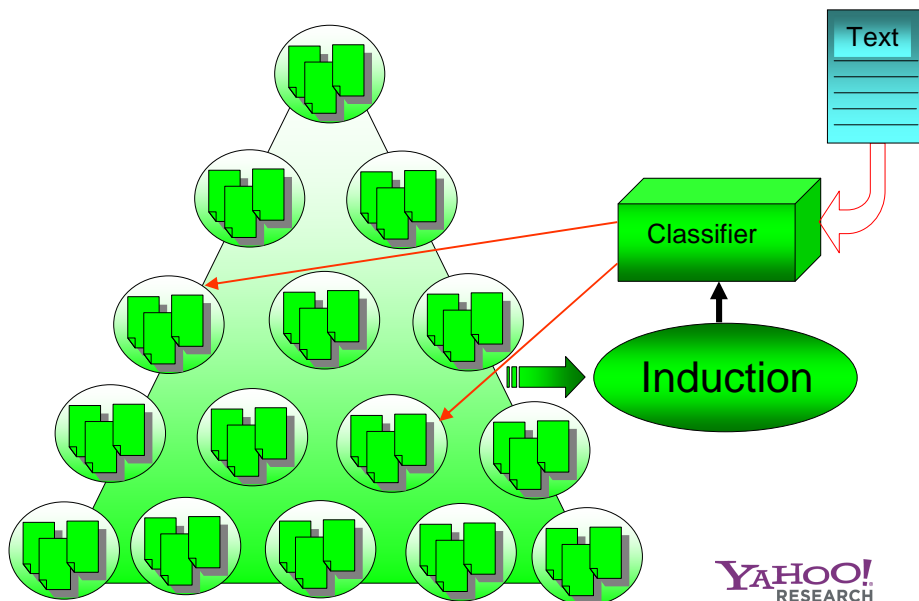


## Building the document classifier for the target taxonomy

- Classification taxonomy of 6,000+ nodes
  - Median depth 5, max depth 9
- Populated with ~150 queries per node
  - Bid phrases of ads
  - Editorially selected
  - Good coverage of commercial topics
- Centroid classifier

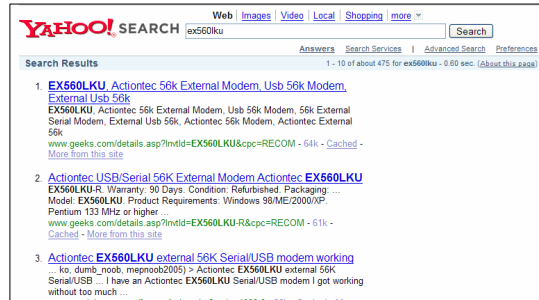
YAHOO!  
RESEARCH

## Text classification (cont'd)



# Research questions

Snippets or full pages?



Number of search results to obtain

Number of classes per search result

Aggregation:  
bundling or voting?

YAHOO!  
RESEARCH

## Data sets (distinctions are based on advertising properties)

Two sets of 1000 actual queries

- Set #1: Broad match
  - Mean length: 3.5 words
  - Fraction of queries with quotation marks: 8%
- Set #2: Uncovered
  - Mean length: 4.4 words
  - Fraction of queries with quotation marks: 14%

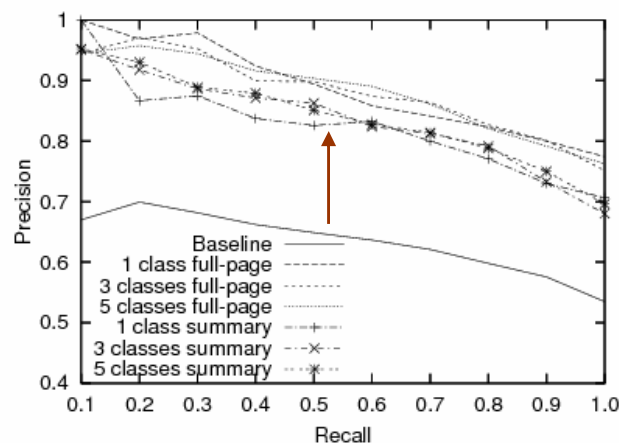
YAHOO!  
RESEARCH

## Evaluation procedure

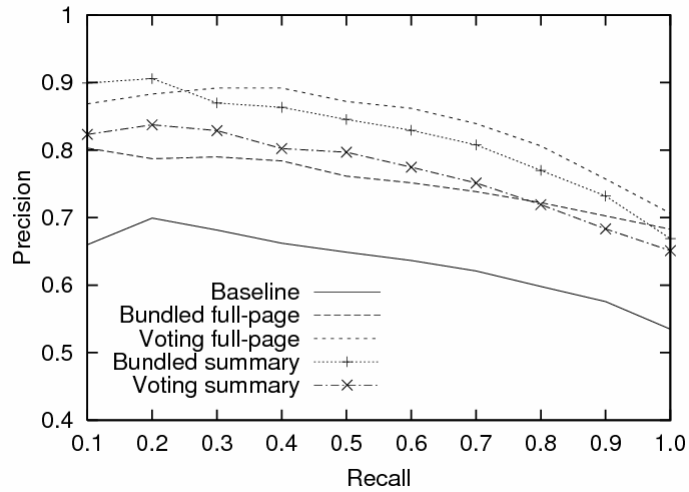
- Human judges reviewed top 3 classes assigned to each query
- Rating scale:
  - 1 (highly relevant) to 4 (irrelevant)
- Standard metrics: Precision, Recall, F1
- Baseline: standard query expansion, grouping of terms using a phrase recognizer, NN classifier



## The effect of external knowledge

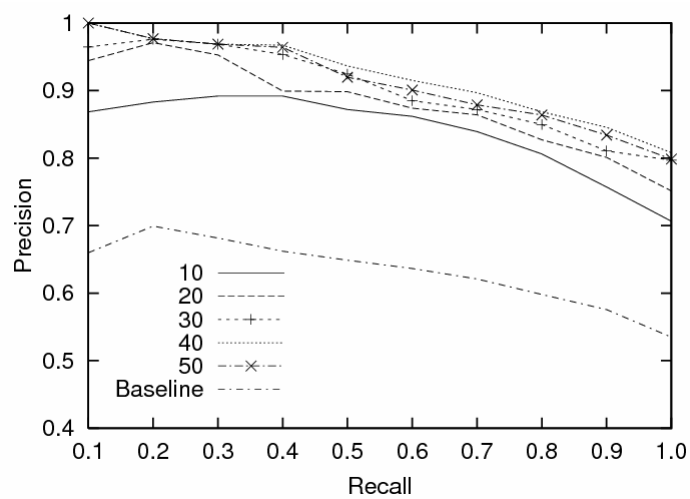


## Aggregation techniques



YAHOO!  
RESEARCH

## Varying the number of search results



YAHOO!  
RESEARCH

## Failure analysis

- Queries containing random string, e.g., phone numbers (~ 5% of queries)
  - incoherent search results
- No search results at all (~8-15% of queries)
- Queries corresponding to recent events (~ 5% of queries)
  - insufficient coverage



## Conclusions (Part I)

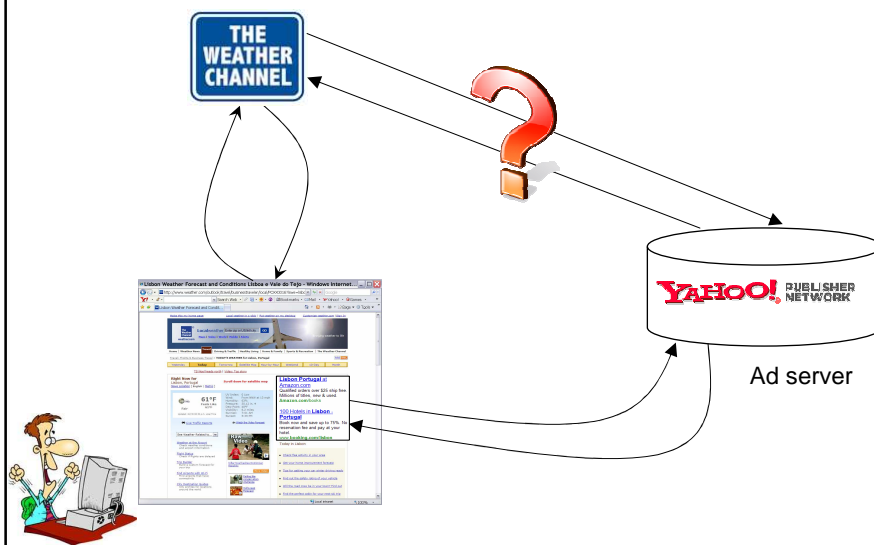
- Using search results substantially improves classification accuracy of rare queries
- If pages are pre-classified in the search index, the approach is very fast
- *Future work:* using query classes as features for better ad matching



Part II  
Just-in-Time  
Contextual Advertising  
(CIKM 2007)



Problem statement



## Problem statement (cont'd)

- Many pages cannot be analyzed in advance ☹
  - Dynamic / personalized / frequently updated
  - Invisible Web
  - Authorization or cookies needed for access
- Shortcomings of existing approaches
  - High latency
  - Low relevance ads
  - High load (frequent crawling)

YAHOO!  
RESEARCH

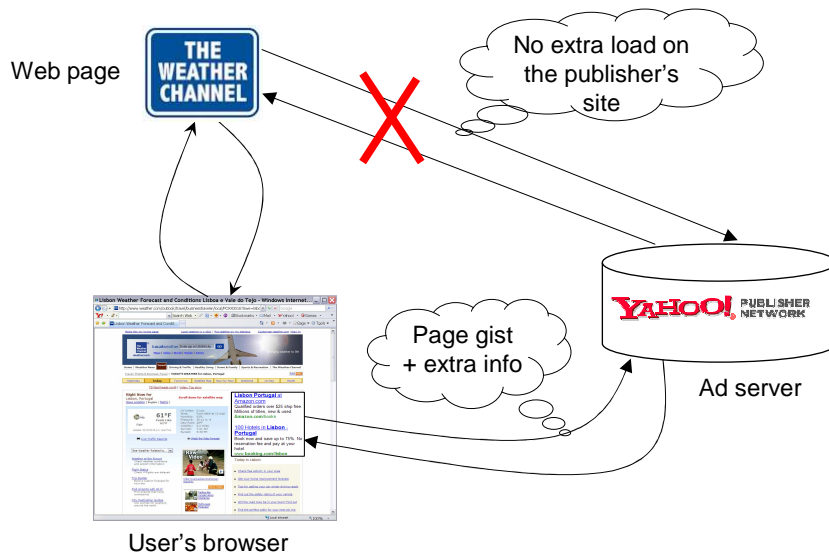
## Our goal

- Relevant ads
- No pre-crawling ahead of time
- Modest processing and communication resources



YAHOO!  
RESEARCH

## Just-in-time ad matching: an overview

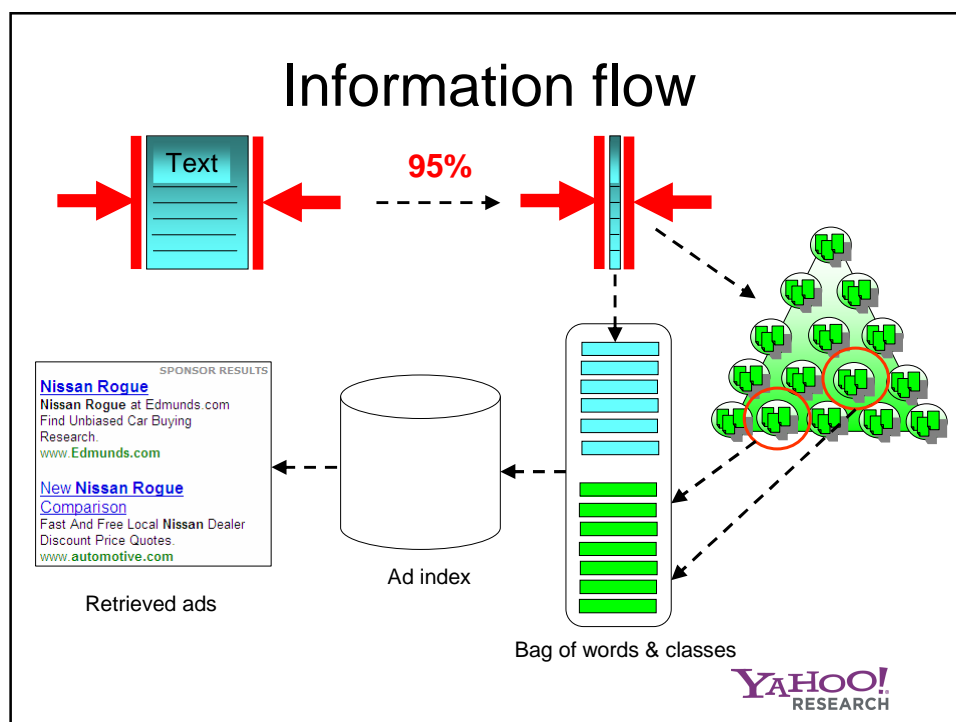


YAHOO!  
RESEARCH

## Our approach

- Text summarization
  - *less content to transfer and analyze*
- Using external knowledge
  - URL + Referrer URL
  - Page classification w.r.t. an external taxonomy of commercial topics

YAHOO!  
RESEARCH



- ## Text summarization
- Title
  - Meta keywords and description
  - Headings (<h1>, <h2> etc.)
  - Tokenized URL and referrer URL
  - First N bytes of the page text
  - Anchor text of the outgoing links
  - *Baseline: full text (+ URL + referrer URL)*
- YAHOO! RESEARCH**

## Text classification

- Classification taxonomy of 6,000+ nodes
  - Median depth 5, max depth 9
- Populated with ~150 queries per node
  - Bid phrases of ads
  - Editorially selected
  - Good coverage of commercial topics
- Centroid classifier
- $score(page, ad) = \alpha \cdot sim_{BOW}(page, ad) + \beta \cdot sim_{class}(page, ad)$



## Datasets

- Dataset 1
  - 105 pages of various content
  - 2,680 unique ads
  - 2,946 page-ad judgments
- Dataset 2
  - 827 pages of various content
  - 5,065 unique ads
  - 9,748 page-ad judgments



## Evaluation metrics

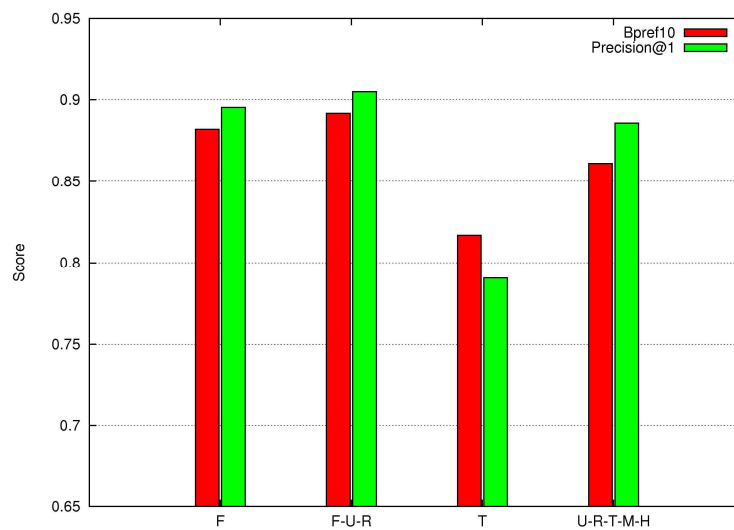
- Precision @ k
- Mean Average Precision (MAP)
- Bpref-10

$$\text{bpref-10} = \frac{1}{R} \sum_r 1 - \frac{|n \text{ ranked higher than } r|}{10 + R}$$

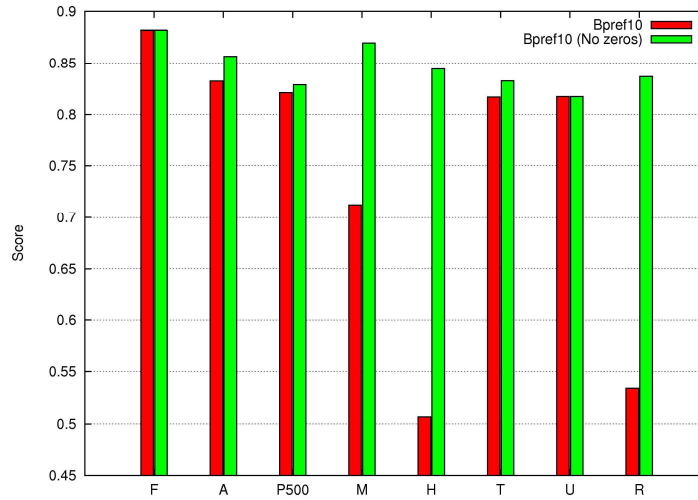
- Robust to incomplete judgment sets
- Only uses judged documents



## Empirical evaluation

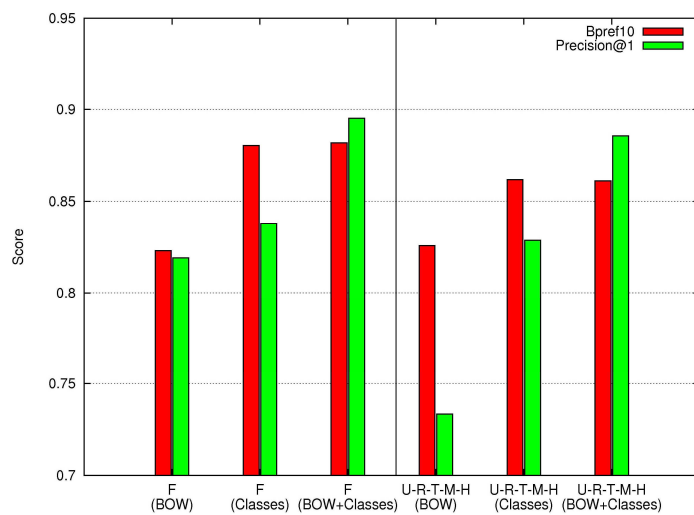


## Contribution of individual fragments



YAHOO!  
RESEARCH

## To classify or not to classify?



YAHOO!  
RESEARCH

## Conclusions (Part II)

- Web pages are often not available (or too expensive) to analyze ahead of time
- We proposed a methodology for contextual Web advertising in real time
- Carefully selected page excerpts make a decent summary
- Classification is important
- 5% of content yield 97-99% of relevance



Questions?

