



CS834 - Introduction to Information Retrieval

Presentation #3

Hussam Aldeen Hallak

The Two Papers:

- Fast generation of result snippets in web search

SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval
Pages 127-134

- Good Abandonment in Mobile and PC Internet Search

Jane Li, Scott B. Huffman, Akihito Tokuda, Google Inc-
Proc. SIGIR '09 , 2009

The First Paper: Seminal paper from the book

- Fast generation of result snippets in web search

SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval
Pages 127-134

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.4357&rep=rep1&type=pdf>

Snippets

- Sentences or fragments extracted from a document
- Gives a content preview
- Allows the user to get an idea of what's inside the document
- Sometimes, it contains the answer to the search query

Google Snippets

ymca suffolk

AllMapsNewsShoppingImagesMoreSettingsTools

About 445,000 results (0.90 seconds)

Suffolk Family YMCA | YMCA of South Hampton Roads

<https://www.ymcashr.org/locations/suffolk-family-ymca>

Suffolk Family YMCA. 2769 Godwin Boulevard Suffolk, VA 23434 A little Halloween fun for ALL AGES today at the Suffolk Y! We certainly do have some ...

Hours of Operation | Suffolk Family YMCA | YMCA of South Hampton ...

<https://www.ymcashr.org/hours-operation-suffolk-family-ymca>

Hours of Operation: Monday - Friday: 5am - 10pm. Saturday: 8am - 6pm. Sunday: 11am - 6pm. Outdoor Pool Hours June 19 - September 4. Monday - Friday: ...

Suffolk Family YMCA - Home | Facebook

<https://www.facebook.com> > Places > Suffolk, Virginia > Child Care Service



★★★★★ Rating: 4.6 - 144 votes

Suffolk Family YMCA, Suffolk, VA. 3K likes. For Youth Development For Healthy Living For Social Responsibility.

Suffolk Family YMCA - YMCA of the USA

www.ymca.net > ABOUT US > FIND YOUR Y

Visit this Y's website now. 2769 Godwin Boulevard Suffolk, VA 23434. Phone: 757-934-9622. CEO Information Not Available Ryan Harrell, Chief Volunteer ...



See photos

See outside

Suffolk Family YMCA ★

Community center in Suffolk, Virginia

Website

Directions

Community-focused nonprofit established in 1844 with recreational programs & services for all ages.

Address: 2769 Godwin Blvd, Suffolk, VA 23434

Hours: Open today · 5AM-10PM

Phone: (757) 934-9622

[Suggest an edit](#)

5

5 / 44

Results Page Construction

- Lexicon Engine: Maps query terms to integers
- Ranking Engine: Generates inverted lists and ranks documents
- Snippet Engine: Generates snippets from query term numbers and document numbers
- Meta Data Engine: Fetches other information to construct results page (e.g., URI, page title, document type, size, ...etc.)

Search Engine Architecture

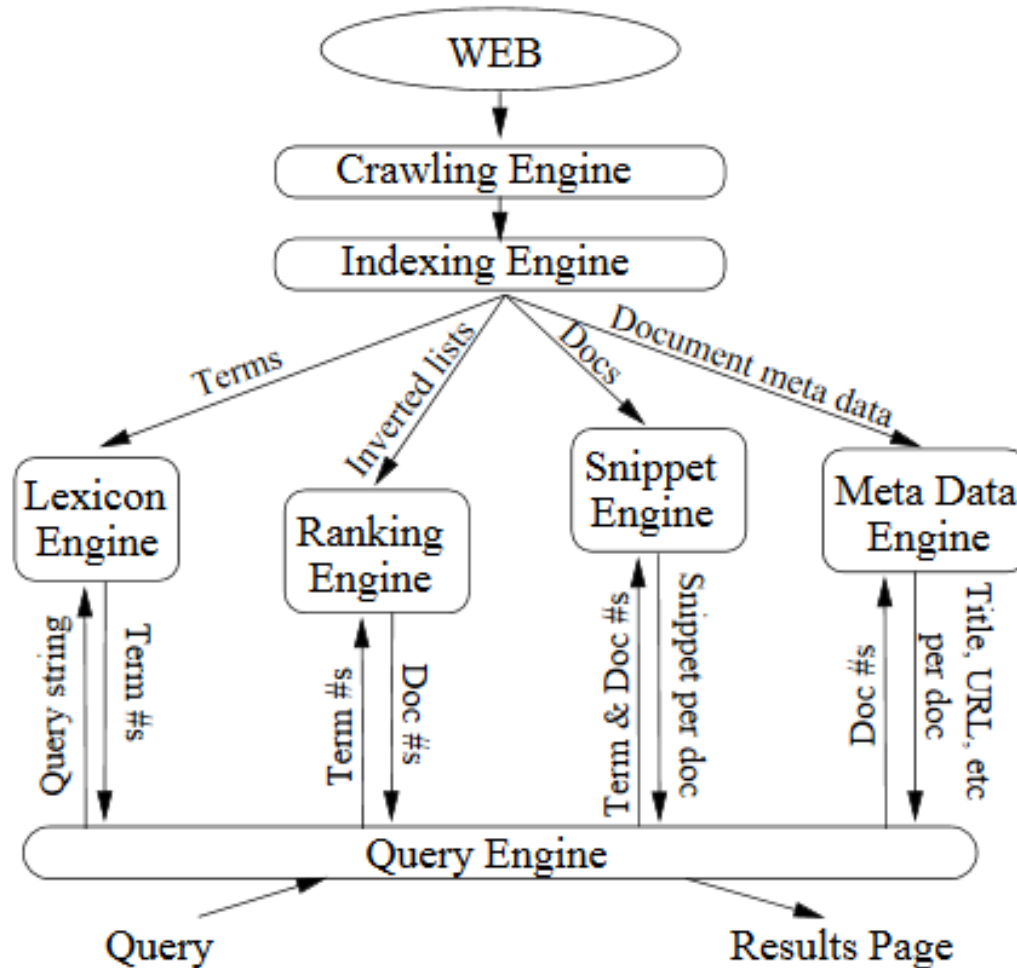


Figure 1: An abstraction of some of the sub-systems in a search engine. Depending on the number of documents indexed, each sub-system could reside on a single machine, be distributed across thousands of machines, or a combination of both.

Two types of snippets:

- Static:
 - Generated from a document and stored
 - Fixed for each document
 - Query independent
- Dynamic:
 - Generated from the query and the document
 - Query-biased
 - Different snippets are generated for the same document based on the query.

Simple Sentence Ranker

IN A document broken into one sentence per line, and a sequence of query terms.

- 1 For each line of the text, $\mathcal{L} = [w_1, w_2, \dots, w_m]$
- 2 Let h be 1 if \mathcal{L} is a heading, 0 otherwise.
- 3 Let ℓ be 2 if \mathcal{L} is the first line of a document, 1 if it is the second line, 0 otherwise.
- 4 Let c be the number of w_i that are query terms, counting repetitions.
- 5 Let d be the number of distinct query terms that match some w_i .
- 6 Identify the longest contiguous run of query terms in \mathcal{L} , say $w_j \dots w_{j+k}$.
- 7 Use a weighted combination of c, d, k, h and ℓ to derive a score s .
- 8 Insert \mathcal{L} into a max-heap using s as the key.

OUT Remove the number of sentences required from the heap to form the summary.

Figure 2: Simple sentence ranker that operates on raw text with one sentence per line.

Dynamic Snippets: Challenges & Solutions

- **Challenges:**

- Generating dynamic snippets takes time.
- Billions of web pages to store and access their files on the disk to generate dynamic snippets.
- Hundreds of millions of search queries.

- **Solutions:**

- Disk access reduction: **Cache:**

- 1- Frequently accessed documents

- 2- pre-computed result pages for popular queries

- Compression:

Smaller documents size → more can be cached.

- Compaction:

Only cache the important part of the document.

Baseline Snippet Engine:

- Uses a well-known adaptive compressor:
Compress all documents
Decompress as needed.
- Implemented using **zlib**
- Each document is stored in a single file
- Snippet Generation:
 - Documents are decompressed one at a time.
 - Linear search for provided query terms.

Compressed Token System (CTS) & CTS Snippet Engine

- Uses a semi-static compression:
 - 1- Fast decompression
 - 2- Minimal compression loss
- Mapping words and non-words to single integer tokens.
- Uses vbyte coding scheme.
- Words and non-words alternate in compressed files.
- Stores all documents contiguously in one file.
- Auxiliary table saves documents' start offsets.



Advantages of CTS over Baseline:

- Faster decompression
- Scoring sentences without document decompression
- Only decode the sentences returned as part of a snippet.

Experiments on CTS and Baseline:

- Data Collections:
WT10G, WT50G, WT100G
- Search Queries:
Excite Log files from 1997
<http://msxml.excite.com/>
- Search Engine:
Zettair
- Document Ranker:
Okapi BM25 (Simulating PageRank)

Storage: CTS vs Baseline

	WT10G	WT50G	WT100G
No. Docs. ($\times 10^6$)	1.7	10.1	18.5
Raw Text	10,522	56,684	102,833
Baseline(<i>zlib</i>)	2,568 (24%)	10,940 (19%)	19,252 (19%)
CTS	2,722 (26%)	12,010 (21%)	22,269 (22%)

Table 1: Total storage space (Mb) for documents for the three test collections both compressed, and uncompressed.

Zlib wins by about 2% in disk space saving

Reduction in time using CTS:

	WT10G	WT50G	WT100G
Baseline	75	157	183
CTS	38	70	77
Reduction in time	49%	56%	58%

Table 2: Average time (msec) for the final 7000 queries in the Excite logs using the baseline and CTS systems on the 3 test collections.

CTS wins by about 50% in snippet generation time

How CTS spends time?

% of doc processed	Seek	Read	Score & Decode
100%	45	4	21
50%	45	4	11

Table 3: Time to generate 10 snippets for a single query (msec) for the WT50G collection averaged over the final 7000 Excite queries when either all of each document is processed (100%) or just the first half of each document (50%).

- Majority of the time is spent locating the document on disk.
- Even when the collection size is cut in half, only 14% time reduction is observed!
- It seems logical to try and reduce its impact through caching.

Snippet Generation Time: CTS vs Baseline

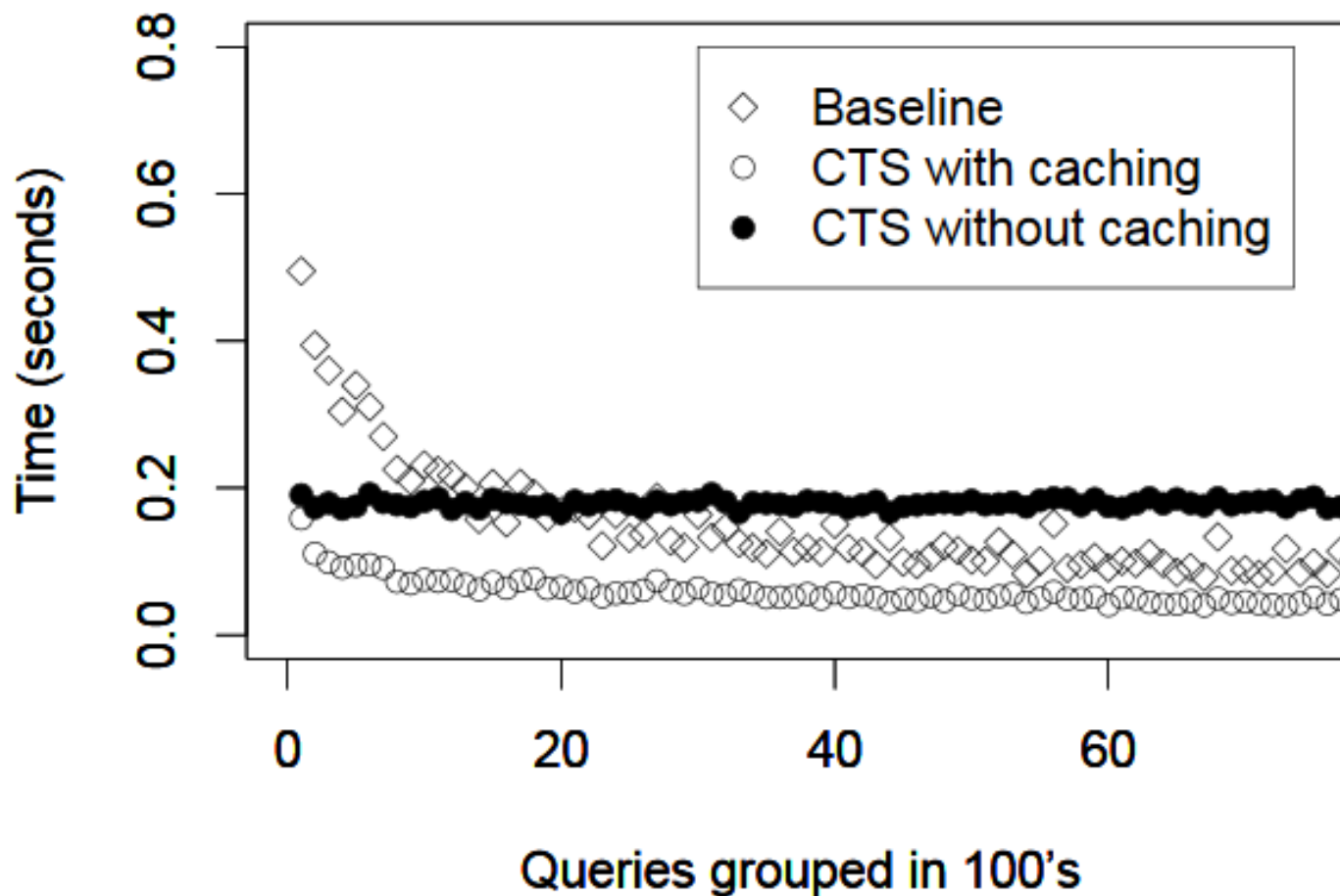


Figure 3: Time to generate snippets for 10 documents per query, averaged over buckets of 100 queries, for the first 7000 Excite queries on wt10g.

Document Compaction:

Only the important sentences count

- Natural order: 1st sentence is the most important. Does not work for poorly written documents.
- Query Log Based (QLT): Sentences containing past query terms are the most important.
- Query Log Based (QLU): Same as QLT, but only counts unique terms.
- Significant Terms (ST): Score based on term frequency (threshold)

$$f_{d,t} \geq \begin{cases} 7 - 0.1 \times (25 - s_d), & \text{if } s_d < 25 \\ 7, & \text{if } 25 \leq s_d \leq 40 \\ 7 + 0.1 \times (s_d - 40), & \text{otherwise,} \end{cases}$$

Natural Order vs (QLT vs QLU vs ST)

- **Significant Terms method wins since it leads to the smallest change in the sentence scoring components.**
- **The greatest change over all methods is in sentence position (h+l) component of the score.**

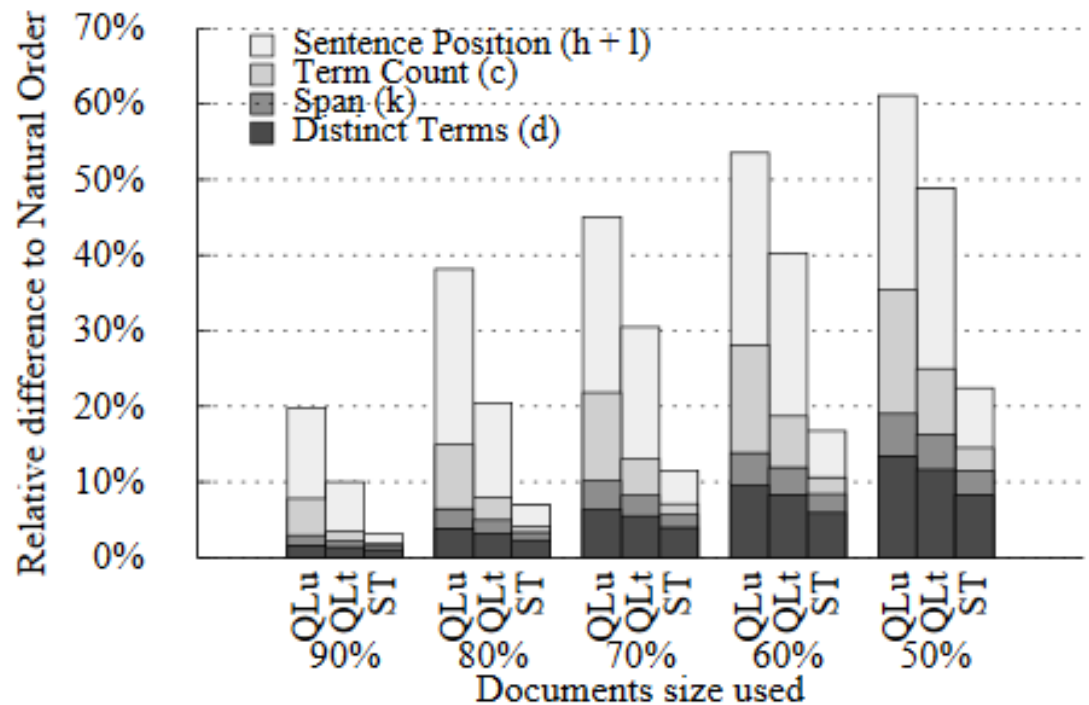


Figure 6: Relative difference in the snippet score components compared to Natural Ordered documents when the amount of documents processed is reduced, and the sentences in the document are reordered using Query Logs (QLt, QLu) or Significant Terms (ST).

The Second Paper:

- Good Abandonment in Mobile and PC Internet Search

Jane Li, Scott B. Huffman, Akihito Tokuda, Google Inc-
Proc. SIGIR '09 , 2009

<http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/35486.pdf>

What is a Good Abandonment?

- The user's information need was successfully addressed by search result page with no need to click on a result or refine the query.
- The snippets under the search results were good enough to provide what the user wanted.
- This user won't be clicking on any of the results, but that doesn't mean that the search has failed.

Example of Good Abandonment

Search results for "ymca suffolk" (About 445,000 results, 0.90 seconds)

Suffolk Family YMCA | YMCA of South Hampton Roads
<https://www.ymcashr.org/locations/suffolk-family-ymca>
Suffolk Family YMCA. 2769 Godwin Boulevard Suffolk, VA 23434 A little Halloween fun for ALL AGES today at the Suffolk Y! We certainly do have some ...

Hours of Operation | Suffolk Family YMCA | YMCA of South Hampton ...
<https://www.ymcashr.org/hours-operation-suffolk-family-ymca>
Hours of Operation: Monday - Friday: 5am - 10pm. Saturday: 8am - 6pm. Sunday: 11am - 6pm. Outdoor Pool Hours June 19 - September 4. Monday - Friday: ...

Suffolk Family YMCA - Home | Facebook
<https://www.facebook.com> > Places > Suffolk, Virginia > Child Care Service
★★★★★ Rating: 4.6 - 144 votes
Suffolk Family YMCA, Suffolk, VA. 3K likes. For Youth Development For Healthy Living For Social Responsibility.

Suffolk Family YMCA - YMCA of the USA
www.ymca.net > ABOUT US > FIND YOUR Y
Visit this Y's website now. 2769 Godwin Boulevard Suffolk, VA 23434. Phone: 757-934-9622. CEO Information Not Available Ryan Harrell, Chief Volunteer ...

Suffolk Family YMCA ★
Community center in Suffolk, Virginia
[Website](#) [Directions](#)

Community-focused nonprofit established in 1844 with recreational programs & services for all ages.
Address: 2769 Godwin Blvd, Suffolk, VA 23434
Hours: Open today · 5AM–10PM
Phone: (757) 934-9622
[Suggest an edit](#)



The Goal:

- To approximate the prevalence of good abandonment.
- Identify the types of information need that may lead to good abandonment.

Why is Good Abandonment so important?

- Abandoning usually indicates user dissatisfaction:
 - 1- The user did not click on any result
 - 2- The user did not refine the query
- No search engine wants that? **NOT TRUE** because:
 - 1- Mobile devices are clunky, slow, have formatting and content omission issues.
 - 2- Mobile users need quicker responses.
 - 3- Search engines use mobile user's location to find local addresses, phone numbers, ... etc.
 - 4- Search Engines want to accurately answer your query as quick as possible.

Definitions:

- Potential Good Abandonment:
 - User's query should mostly be answered by the result page.
 - It is judged by looking at the query to generate an upper bound of good abandonment.
- Likely Good Abandonment:
 - Subset of potentially good abandonment.
 - Information need was clearly met on the results page.
 - Determined by examining the actual results provided by Google.

PGA and LGA Query Examples:

Query	Information need	<i>Potential good abandonment?</i>	<i>Likely good abandonment based on Google results?</i>	Comment
[quote MRK]	Stock quote	Yes	Yes	Query was answered by search feature “Stock Quotes”.
[weather New York, NY]	Weather report	Yes	Yes	Query was answered by search feature “Weather”.
[1 USD in GBP]	Currency exchange	Yes	Yes	Exchange rates were returned in calculator and results snippets.
[taxi buffalo]	Local listings	Yes	Yes	A list of local businesses returned.
[who is the lead singer for tonic?]	A quick answer	Yes	Yes	Answer was provided in title and snippets
[taxi fare nyc from la guardia]	A quick answer	Yes	No	No answer provided on results page
[baby come back]	Lyrics	Maybe	Maybe	Partial lyric was returned in results snippets.
[yamaha psr-172]	Prices/vendors information	Maybe	Maybe	Some price and model information displayed, but may not satisfy the user’s need.
[myspace.com]	Homepage	No	No	Search results can provide a link to the desired site, but the “answer” is the site itself
[free ringtones]	Downloads	No	No	The “answer” is presumably actual downloads, which won’t surface directly in search results page.
[how to ace an amazon phone interview]	Detailed information on a topic	No	No	The “answer” involves reading more detailed information than would be reasonably surfaced in results snippets.

Experiment:

- Study was conducted on:
Two modalities: PC, Mobile devices
For
Three countries: USA, CHINA, JAPAN
- Google's PC and Mobile search log from a week in Sep-Oct 2008
- Sampled 400 abandoned queries from Japan, US, and 1000 from China for both mobile and PC.

Upper Bound Estimate of Good Abandonment:

- The proportion of abandoned queries classified as 'YES' or 'MAYBE' a potential good abandonment out of all abandoned queries.

Percentage of PGA Queries:

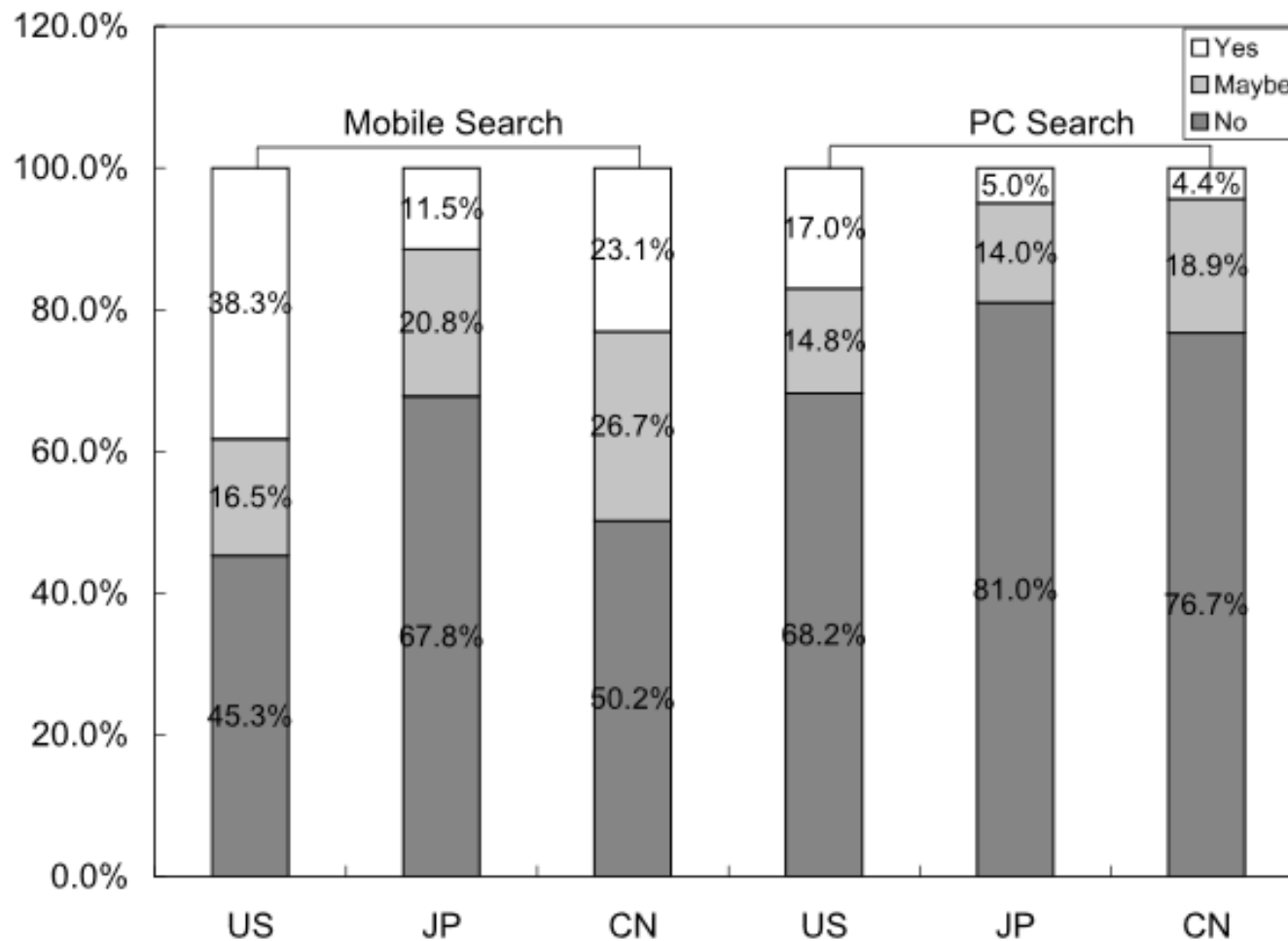


Figure 1: Percentages of queries classified as “Yes”, “Maybe”, “No” with respect to the *potential good abandonment* definition in six abandoned query samples.

Jane Li (2007)

Observations:

- Japan has lowest potential abandonment for mobile search.
- Higher potential good abandonment rate in mobile search because:
 - 1- Retrieving web pages on mobile devices is a clumsy experience.
 - 2- Mobile users perform simple searches
 - 3- Mobile users expect to find what they need without clicking on any result.
 - 4- Mobile searches are dependent on activity at the time, current location, ... etc.

Likely Abandonment Rates:

- It measures how often does the user actually get the desired result out of all queries that could potentially lead to good abandonment.
- YES: queries that turned out to be good abandonment queries.
- MAYBE: authors were not sure if the result page would have satisfied the user
- NO: authors thought it would be good abandonment but turns out it is not.

Percentage of LGA Queries:

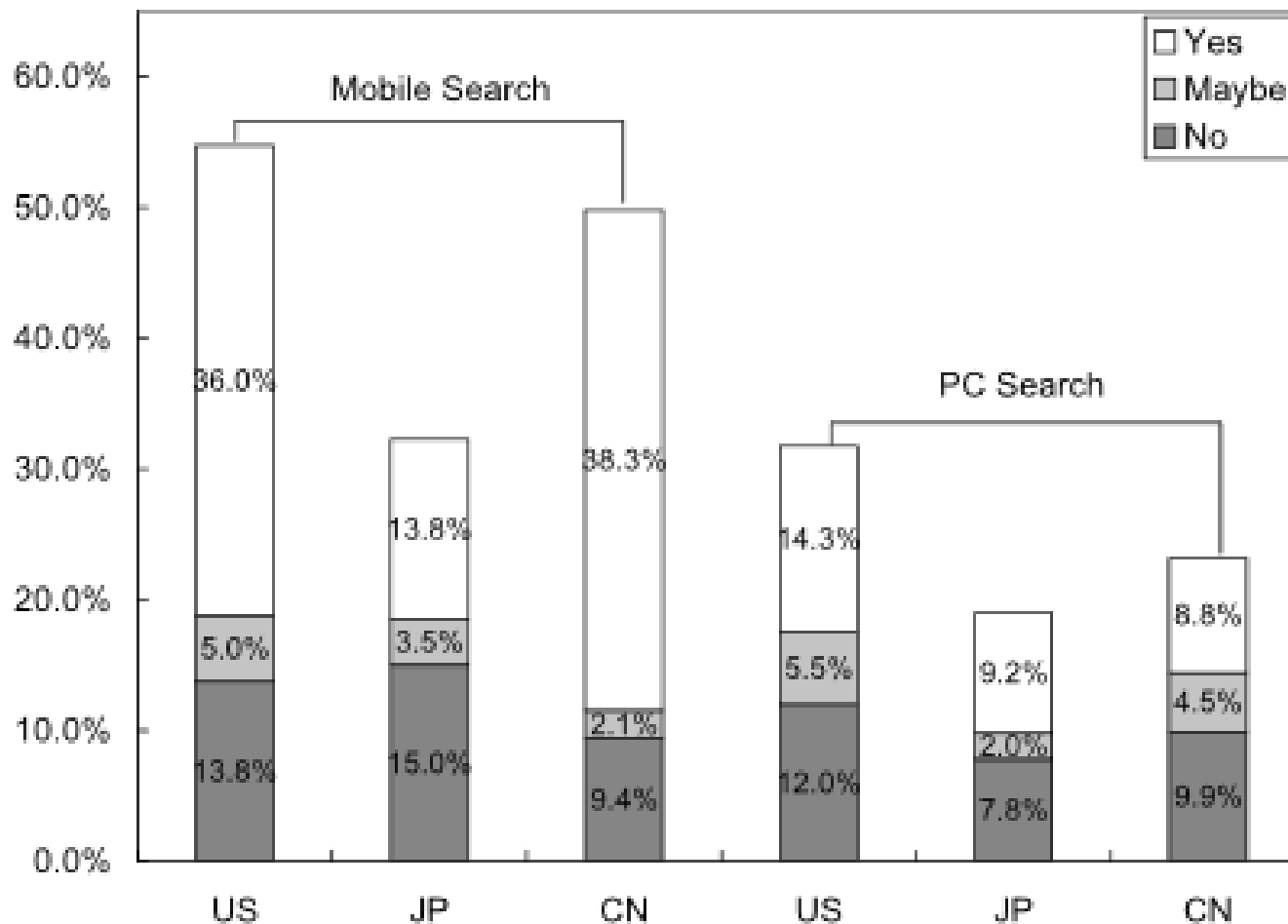


Figure 2: Percentage of potential good abandonment queries which are classified as Yes/Maybe/No with respect to the *likely good abandonment* definition



Observations:

- For mobile search 70% of the potential good abandonment queries were met on the result page.
- 56% for PC.
- Over 80% potential good abandonment queries are answered in Chinese mobile searches.

Classification by Information Need:

- The authors categorized potential or likely good abandonment queries based on information need.

Task Categories Example:

Category	Definition	Query Example
Answer	User seeks a short answer to a question.	[age of consent in PA]
Currency	User seeks currency conversion.	[1 USD in GBP]
Click-to-call	User typed a phone number into search, and either meant to call it, or seeks to learn whose number it is.	
Definition	User seeks the definition of a term.	[what is nvmd]
Images	User wants pictures of a person or thing.	[cubs logo]
Local	User seeks a local listing (address and/or phone number).	[at&t wireless, bartlett, tn]
Lyrics	User seeks song lyrics.	[abettes with time lyrics]
Map	User seeks for a map of a location (address or geographic locations).	[E 83rd St, Los Angeles, CA 90001]
Celebrities	User seeks news or images of a celebrity.	[john stamos]
News	User seeks current news on a topic.	[santa cruz wild fires]
Product	User seeks simple product information such as price range and typical vendors.	[2000 gsx 750f for sale]
Person	User is looking for contact information or vanity information on a (non-celebrity) person.	
Quotation	User seeks for the reference of a quoted sentence/phrase.	["(a) %20 aqueous phosphoric acid (H3PO4) or (b) an aqueous solution]
Stock	User seeks a current stock price.	[quote AKAM]
Sports	User seeks sports scores.	[chicago cubs]
Showtimes	User seeks movie showtimes.	[The Dark Knight]
Spelling	User seeks correct spelling of a word.	[unfortunatly]
SMS	User seeks short (greeting) messages to send.	[Short greeting message] (in Chinese)
Translation	User seeks for the translation of a word in foreign language.	[dounika translation]
Weather	User seeks a weather report.	[Weather New York]

Table 2: Task categories and their definitions

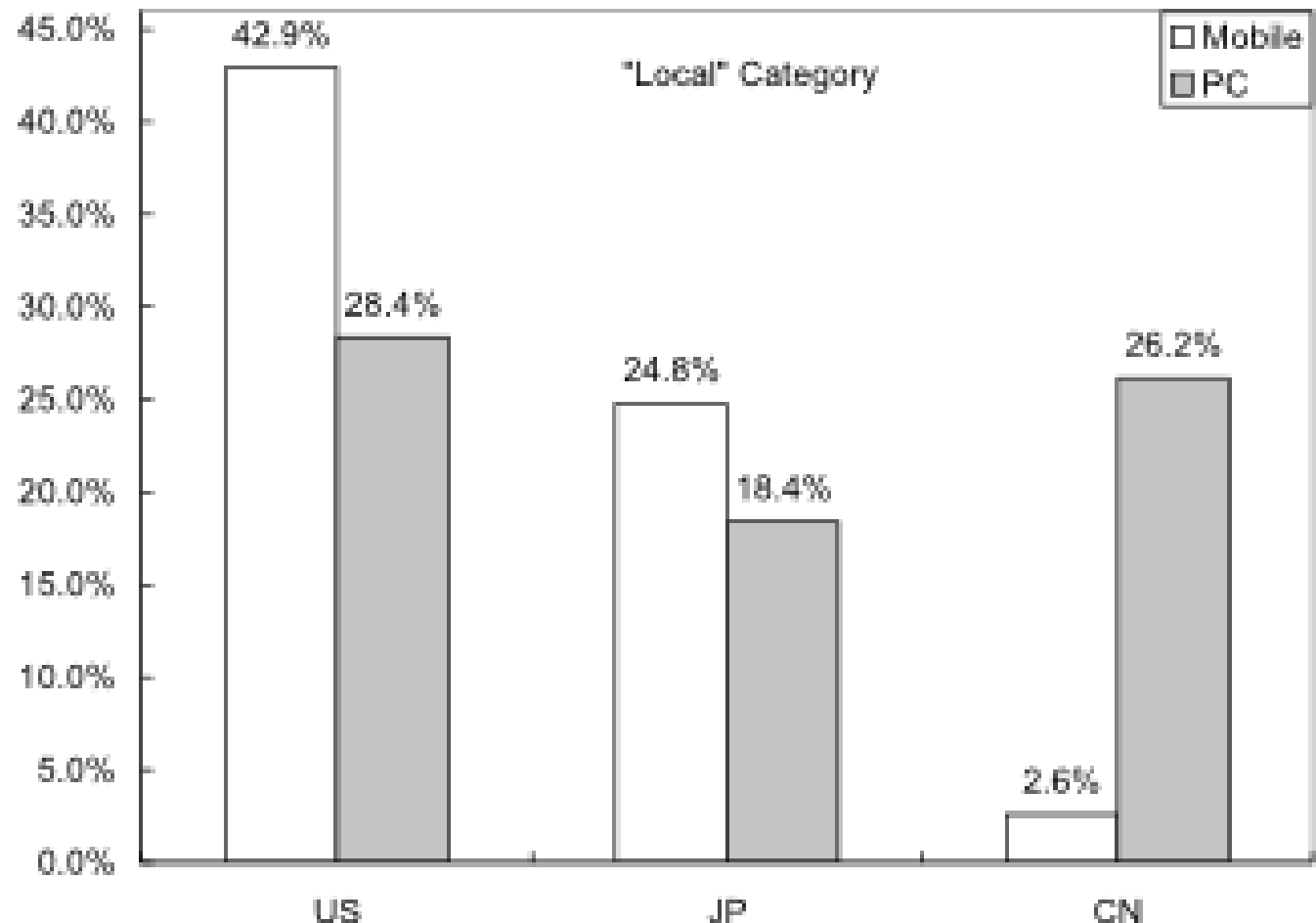


Findings:

- The distribution of good abandonment based on information needs varies enormously across search modalities and locales.
- Queries seeking local information or short answers are the top classes leading to good abandonment in PC search, consistently across locales.
- Significant portion of entertainment related searches in Japan and China, especially in mobile search.

Surprise!

Figure 4:
Jane Li (2009)



- **Less than 3% “Local” potential good abandonment queries for Chinese mobile search. By contrast, it is the top class for Chinese PC search.**
- **The actual search results were poorly addressed on the results page.**
- **On mobile devices, where text entry is difficult and loading takes longer, people may usually give up.**

Some Interesting Results:

- US users click more often when searching celebrities.
- Japanese mobile users browse images or news outlines (satisfied by the snippets).
- For Stocks, more good abandonment in mobile because:
 - 1- Mobile users look for stock quotes in the snippets.
 - 2- PC users may seek more information and graphs.
- China has most simple mobile searches for weather.

What to improve?

- Categories like **Celebrities, Local, Answer** have scope for improvement.
- search engines should address “Local”, “Image” information needs by adding features like “Shortcut Insertions”.
- The “Answer” category requires more intelligent snippets.
- Longer snippets is better for mobile users since loading a web page is a hassle.



Figure 5: Heat map depicting the headroom to drive additional good abandonment by information-need category. Darker squares indicate more headroom.



Findings:

- Queries leading to good abandonment are a significant portion of all abandoned queries.
- Good abandonment rate for Mobile is higher than PC across all locales.
- Good abandonment vary by both, locale and modality.



Take Away Message:

- Query abandonment is not a negative signal.
- Evaluation models derived from CLICKS should be improved by taking good abandonment into account.
- Search engines still have headroom for good abandonment especially for mobile users.



Good Citation?

- The first paper introduces a method for fast dynamic and smart snippet generation which leads to higher good abandonment rate, the subject of the second paper.

References:

[1] Fast generation of result snippets in web search

SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval

Pages 127-134

[2] Good Abandonment in Mobile and PC Internet Search

Jane Li, Scott B. Huffman, Akihito Tokuda, Google Inc- Proc. SIGIR '09 , 2009