Microblog Search

Yulu Wang
Dept. of Computer Science
University of Maryland
Outline

• Search
  – What’s search?
  – How to search?
  – Microblog Search
• Microblog Novelty Detection
• Search Evaluation
• Summary
Outline

• Search
  – What’s search?
  – How to search?
  – Microblog Search

• Microblog Novelty Detection

• Search Evaluation

• Summary
What’s search?
What is search?

Obtain information resources relevant to an information need from a collection of information resources.

Search is everywhere!
What is search?
What is search?

COMPUTER SCIENCE BOOKS?
What is search?
What is search?
How to search?
How to search?

COMPUTER?

Dictionary / Thesaurus

DATA: SORT ALPHABETICALLY
HOW: FIND C, O, M ....
How to search?

COMPUTER SCIENCE BOOKS?

DATA: SORT BY CATEGORY
HOW: FIND COMPUTER SCIENCE
How to search?

HASHTABLE?

DATA: ?
HOW: BRUTE FORCE?

NO! TAKE AGES
What do we learn from the first two examples?
How to search?

Data structure + Data access
To narrow down your search range if information resources are huge.
How to search – inverted index

• Inverted Index
  – Find documents by words, instead of finding words by documents
  – Can quickly locate the relevant documents for each query (search keywords)
How to search – inverted index

watch

doc1
doc2
doc3
doc4
doc5
doc6
doc7
doc8
doc9
doc10
doc11
doc12
doc13
doc14
doc15

......
How to search – inverted index
How to search – inverted index

term

apple

macbook

watch

.

.

.

.


docid

1  5  10  12  14

1  3  5  9

5  7  10  15

.

.

.

.


Dictionary

Postings
How to search – inverted index

• Collect the documents to be indexed
  Apple watch releases   Apple New Macbook

• Tokenize the text, turning each document into a list of tokens
  Apple  watch  releases  Apple  New  Macbook

• Do linguistic preprocessing, producing a list of normalized tokens, which are the indexing terms
  apple  watch  release  new  macbook

• Index the documents that each term occurs in by creating an inverted index, consisting of a dictionary and postings
How to search – inverted index

- Query: apple watch
- Postings Merge

<table>
<thead>
<tr>
<th>term</th>
<th>docid</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>1 5 10 12 14</td>
</tr>
<tr>
<td>watch</td>
<td>5 7 10 15</td>
</tr>
</tbody>
</table>

- Search results:
How to search – inverted index

- Query: apple watch
- Postings Merge

Search results: doc5
How to search – inverted index

• Query: apple watch
• Postings Merge

<table>
<thead>
<tr>
<th>term</th>
<th>docid</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>1 5 10 12 14</td>
</tr>
<tr>
<td>watch</td>
<td>5 7 10 15</td>
</tr>
</tbody>
</table>

• Search results: doc5
How to search – inverted index

• Query: apple watch
• Postings Merge

<table>
<thead>
<tr>
<th>term</th>
<th>docid</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>1 5 10 12 14</td>
</tr>
<tr>
<td>watch</td>
<td>5 7 10 15</td>
</tr>
</tbody>
</table>

• Search results: doc5 doc10
How to search – inverted index

• Query: apple watch

• Postings Merge

• Search results: doc5 doc10
How to search – inverted index

- Query: apple watch
- Postings Merge

<table>
<thead>
<tr>
<th>term</th>
<th>docid</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>1 5 10 12 14</td>
</tr>
<tr>
<td>watch</td>
<td>5 7 10 15</td>
</tr>
</tbody>
</table>

- Search results: doc5 doc10
How to rank the results?
How to rank?

• Assign score to each query and returned document, $S(Q, D)$
  – Term Weighting: weight the importance of a term in a document, based on the statistics of occurrence of the term
How to rank?

• Term Weighting
  – Tf-idf model
    • Tf (term frequency): frequency of a term in a document
    • Df (document frequency): the number of documents that contain a term
    • Idf (inverted document frequency) of term t:
      \[ idf_t = \log \frac{N}{df_t} \]  \( (N : \text{total number of documents}) \)
    • Tf-idf of term t in document d:  \( tf - idf_{t,d} = tf_{t,d} \times idf_t \)
    • Score(Q, D) = \( \sum_{t \in Q} tf - idf_{t,d} \)
How to rank?

• Term Weighting
  – Tf-idf model
    • Score(Q, D) = \sum_{t \in Q} tf - idf_{t,d}
    • Higher tf, higher score
    • Higher idf(lower df), higher score
Microblog Search

• Microblog: Twitter
  – Short: 140 characters limit (term weighting not working)
  – Informal: 2m, 2moro, tm, tmrw... (hard to match query)
  – Real-time (recent microblogs should be searchable)

• Web, News Media
  – Long
  – Formal
  – Relatively static
Improve effectiveness?
(To return more relevant documents)
Microblog Search Effectiveness

• Short tweet? Expand the tweet
  – Document Expansion

• Query doesn’t tell enough story? Expand the query
  – Query Expansion
Microblog Search Effectiveness

• Document Expansion (before query time)
  – Expand the document with any URL link page, which provides more relevant information

CM Christian Mail @ChristianMail · Apr 10
Apple Watch Release Date, Price, How to Order Yours? Smartwatch Available Via Online... goo.gl/fb/2axCiY

The sales, according to the site, will only be made possible "through a reservation system" and this will be true "for the indefinite future."

Apple Watch Release Date, Price, How to Order Yours? Smartwatch Available Via Online Reservations Only

Christian Post Report – With the release date of Apple’s highly anticipated smartwatch, the Apple Watch, already drawing nearer, a recent report indicated that for those who are planning to buy the device should first make their online reservations.

Reuters/Photo
A report from IGN cited CNET saying that would-be buyers of Apple’s first wearable technology should bear in mind that unlike the company’s iPhone and iPad products, the Apple Watch won’t be available for them if they plan to just walk into their store to buy it.
Microblog Search Effectiveness

• Query Expansion (At query time)
  – System returns an initial set of n documents, assuming they are relevant

Drudge Report News @Drudge_Report_ · 3h
#APPLE Quietly Removes #AppleWatch Release Date from Website... 9to5mac.com/2015/04/16/whe...

Squawk on the Street @SquawkStreet · 5h
Global #AppleWatch pre-orders over 2.3M: Report cnb.cx/1Hx7SiR

The Telegraph @Telegraph · 5h
Memo reveals #AppleWatch will not be available to buy in-store at launch telegraph.co.uk/technology/app...
Microblog Search Effectiveness

• Query Expansion (At query time)
  – System takes terms from these documents to reformulate the query
Microblog Search Effectiveness

• Query Expansion (At query time)
  – Search with the new query

Akihiro,mituhata @akihiromituhata · Apr 13
smeitsolutions: **Apple Watch available** for pre-order: pcpro.co.uk/wearables/2514…

PC Pro @pcpro · Apr 10
The **Apple Watch** is now **available** for pre-order (like you didn’t already know).
pcpro.link/1adJepV

Christian Mail @_ChristianMail · Apr 10
**App Watch Release** Date, Price, How to **Order** Yours? Smartwatch **Available**
Via Online: fb.gi/fb/2axCiY
Microblog Search Effectiveness

• Query Expansion (At query time)
  – Search twice
  – Return the results in the second search
Real-time search?
Microblog Real-time Search

• Real-time
  – Property
    • Large volume of new data are generated every second
    • Users are eager to get the response in short time
    • Users prefer results appearing recently
  – Difference from web search
    • New docs should be indexed as they become available so that they can be searchable
    • Web search can wait, because no preference for the recent docs
Microblog Real-time Search

• Twitter Earlybird
Microblog Real-time Search

- Each instance of Earlybird manages multiple index segments, and each segment holds a relatively small number of tweets.
- Ingested tweets first fill up the segment before proceeding to the next one.
  - At any given time, there is at most one index segment actively being modified, whereas the remaining segments are read-only and can be searched from.
- Once an index segment ceases to accept new tweets, we can convert it to read-only to make it searchable.
Outline

• Search
  – What’s search?
  – How to search?
  – Microblog Search

• Microblog Novelty Detection

• Search Evaluation

• Summary
Microblog Novelty Detection - Motivation

• Duplicate tweets

WSJD @WSJD · 2h
April 24 is still launch day for Apple Watch. But Apple is estimating a June delivery for Watch orders placed today on.wsj.com/1b70K0l

Geoffrey Fowler retweeted
WSJD @WSJD · 2h
April 24 is still launch day for Apple Watch. But Apple is estimating a June delivery for Watch orders placed today on.wsj.com/1b70K0l

Raza Jaffri retweeted
WSJD @WSJD · 2h
April 24 is still launch day for Apple Watch. But Apple is estimating a June delivery for Watch orders placed today on.wsj.com/1b70K0l
Microblog Novelty Detection-Motivation

• Near-duplicate tweets

  WSJD @WSJD · 2h
  April 24 is still launch day for Apple Watch. But Apple is estimating a June delivery for Watch orders placed today on.wsj.com/1b70K01

  Shining Star @PSSP92 · 4m
  Apple Watch website removes 'April 24' availability date, postponed launch to June: Apple may be rich in sales... binged.it/1CQJKSX

• And we want to remove those annoying tweets in the search results
Microblog Novelty Detection-Process

• Compute similarity between tweets
  – Jaccard similarity coefficient
  – Clustering

• Remove similar ones
Microblog Novelty Detection – Jaccard similarity coefficient

• Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets

• Jaccard similarity between A and B:

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]

\[ 0 \leq J(A, B) \leq 1 \]
Microblog Novelty Detection – Jaccard similarity coefficient

• For every pair of tweets $T_i, T_j$, treat them as bag of words

• Compute $J(T_i, T_j)$
  
  $T_i \cap T_j = \#$ of common words
  
  $T_i \cup T_j = \text{total} \ # \ \text{of unique words in} \ T_i \ \text{and} \ T_j$

• Remove $T_j$ if $J(T_i, T_j)$ is above some threshold
Microblog Novelty Detection – Jaccard similarity coefficient

- $T_i$: Apple watch releases
- $T_j$: Apple watch launches today

- Words in common: apple, watch
- Total unique words: apple, watch, releases, launches, today

$$J(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} = \frac{2}{5} = 0.4$$
Microblog Novelty Detection - Clustering

- Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups (clusters).
Microblog Novelty Detection - Clustering

• Clustering the documents, then choose cluster center to be the novel document (leave the clustering algorithm)
Outline

• Search
  – What’s search?
  – How to search?
  – Microblog Search
• Microblog Novelty Detection
• Search Evaluation
• Summary
What’s a better search engine?
Search Evaluation

• User Study
• Cranfield Paradigm
• Implicit Feedback
Search Evaluation – User Study

• How
  – Hire actual users
  – Have them use the systems to finish a task
  – Measure their performance
    • Number of relevant documents found in a given time
    • Time to find required answers
Search Evaluation – User Study

• Pros
  – Close to real

• Cons
  – Expensive
  – Time-consuming
Search Evaluation – Cranfield Paradigm

**Test collection:**
- Document Collection D
- Topic T
- Relevance Judgments R (gathered from assessors, called “”)

**System run:**
- Retrieve lists L from D for topic T

**Measure**

\[ \text{Score} = f(R, L) \]

Higher score, better system
Search Evaluation – Cranfield Paradigm

• Measures
  – Unranked measures: for unranked documents
  – Ranked measures: for ranked documents
Search Evaluation – Cranfield Paradigm

• Measures
  – Unranked measures: for unranked documents
    • Precision: of all the returned documents, how many of them are truly relevant
    • Recall: how many relevant documents I’ve found
    • F-score: combination of Precision and Recall
  – Ranked measures: for ranked documents
Search Evaluation – Cranfield Paradigm – Precision and Recall

• Precision(P) is the fraction of retrieved documents that are relevant

\[
Precision = \frac{\text{# of relevant documents retrieved}}{\text{# of retrieved documents}}
\]

• Recall(R) is the fraction of relevant documents that are retrieved

\[
Recall = \frac{\text{# relevant documents retrieved}}{\text{# of relevant documents in collection}}
\]
Search Evaluation – Cranfield Paradigm – Precision and Recall

• Precision and Recall tradeoff
  – How to get high precision?
  – How to get high recall?
  – More documents retrieved: high recall, low precision
Search Evaluation – Cranfield Paradigm – F-score

• F-score: balance between precision and recall

\[ F = \frac{1}{\alpha \frac{1}{P} + \frac{1-\alpha}{R}} \quad (\alpha \in [0,1]) \]

• Often use F1:

\[ F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (\alpha = 0.5) \]
Search Evaluation – Cranfield Paradigm – Precision, Recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>10</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>40</td>
<td>500</td>
<td></td>
</tr>
</tbody>
</table>

\[ P = \frac{10}{10+30} = \frac{1}{4} \]
\[ R = \frac{10}{10+40} = \frac{1}{5} \]
\[ F1 = \frac{2 \times \frac{1}{4} \times \frac{1}{5}}{\frac{1}{4} + \frac{1}{5}} = \frac{2}{9} \]
Search Evaluation – Cranfield Paradigm

• Measures
  – Unranked measures: for unranked documents
  – Ranked measures: for ranked documents
    • P@n, R@n, MAP(Mean Average Precision)
Search Evaluation – Cranfield Paradigm – P@n, R@n, MAP

• P@n, R@n
  – For each cutoff n, take top n docs as a set, and compute Precision/Recall for those n docs

• MAP (Mean average precision)
  – Average the precision at the positions of relevant docs

MAP = \frac{\sum_{i=1}^{n} P@i}{n}
Search Evaluation – Implicit Feedback

• User Behavior
  – Links a user clicks on the ranked results
  – Time a user spends reading a page
  – How a user reformulates a query
Outline

• Search
  – What’s search?
  – How to search?
  – Microblog Search
• Microblog Novelty Detection
• Search Evaluation
• Summary
Summary

• Search Engine
  – Inverted Index
• Ranking
  – Term-weighting
• Microblog Search Effectiveness
  – Document Expansion, Query Expansion
• Microblog Real-time Search
• Microblog Novelty Detection
  – Jaccard similarity
  – Clustering
• Search Evaluation
  – Cranfield Paradigm (test collection): Measures
Questions?