Lecture 19: Web Search (1)
What makes web search different?

• Users, information needs, intents
  • Query understanding
  • Search result diversification

• Document and corpus
  • HTML structure
  • Hyperlinks
  • Anchor texts
  • Spam

• Large-scale search logs
  • Inferring relevance labels: free training data
  • Personalization & contextualization
  • Diagnosing search quality
Web Search: Queries & Intents

Navigational Queries

- The immediate intent is to reach a particular site, for example:
  - Greyhound Bus: http://www.greyhound.com
  - Facebook: https://facebook.com
  - Don Knuth: https://www-cs-faculty.stanford.edu/~knuth/
- Simple intent
- Does not exist in classical IR (keyword & topical search)
- Users will be satisfied by one relevant result (and stop)
- Needs to rank the correct site as high as possible
- Ideal metrics for evaluation: reciprocal rank

Web Search: Queries & Intents

Informational Queries

• The intent is to acquire some information assumed to be present on one or more web pages.
• Most similar to classical IR (keyword & topical search)
• For example:
  • BM25
  • Inverted index
  • VA retirement plan
• Users will need to acquire a certain amount of information to satisfy
• But usually the purpose is not to find all relevant webpages!
• Ideal metrics for evaluation: nDCG@10, ERR@10, P@10

Web Search: Queries & Intents

Transactional Queries

• The intent is to perform a certain transaction.
• Usually requires a set of interactions with the relevant webpages (in contrast to getting information from static content).
• Examples:
  • flight ROA to CID
  • cheap dehumidifier
  • matlab 2018a trial download
• Users will be satisfied by getting things done
• Relevance judgments may be highly personal
• Sometimes a good result is only an entry point ...
• Does not existing in classical IR

How popular are these intents?

- Some results from AltaVista in 2001
- A pop-up survey of AltaVista users (3,190 responses)
- A random set of 1,000 queries from AltaVista search logs
- Over half of the queries are navigational or transactional
- Topical/informational search only makes a part of web search

<table>
<thead>
<tr>
<th>Type of query</th>
<th>User Survey</th>
<th>Query Log Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigational</td>
<td>24.5%</td>
<td>20%</td>
</tr>
<tr>
<td>Informational</td>
<td>?? (estimated 39%)</td>
<td>48%</td>
</tr>
<tr>
<td>Transactional</td>
<td>&gt; 22% (estimated 36%)</td>
<td>30%</td>
</tr>
</tbody>
</table>

### A finer-grained intent scheme by Rose & Levinson (2004)

<table>
<thead>
<tr>
<th>SEARCH GOAL</th>
<th>DESCRIPTION</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Navigational</td>
<td>My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.</td>
<td>aloha airlines duke university hospital kelly blue book</td>
</tr>
<tr>
<td>2. Informational</td>
<td>My goal is to learn something by reading or viewing web pages</td>
<td>what is a supercharger 2004 election dates baseball death and injury why are metals shiny</td>
</tr>
<tr>
<td>2.1 Directed</td>
<td>I want to learn something in particular about my topic</td>
<td>color blindness jfk jr</td>
</tr>
<tr>
<td>2.1.1 Closed</td>
<td>I want to get an answer to a question that has a single, unambiguous answer.</td>
<td>help quitting smoking walking with weights</td>
</tr>
<tr>
<td>2.1.2 Open</td>
<td>I want to get an answer to an open-ended question, or one with unconstrained depth.</td>
<td>pella windows phone card</td>
</tr>
<tr>
<td>2.2 Undirected</td>
<td>I want to learn anything/everything about my topic. A query for topic X might be interpreted as &quot;tell me about X.&quot;</td>
<td>travel amsterdam universities florida newspapers</td>
</tr>
<tr>
<td>2.3 Advice</td>
<td>I want to get advice, ideas, suggestions, or instructions.</td>
<td></td>
</tr>
<tr>
<td>2.4 Locate</td>
<td>My goal is to find out whether/where some real world service or product can be obtained</td>
<td>kazaaz lite mame roms</td>
</tr>
<tr>
<td>2.5 List</td>
<td>My goal is to get a list of plausible suggested web sites (i.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal</td>
<td>xxx porno movie free live camera in l.a. weather measure converter</td>
</tr>
<tr>
<td>3. Resource</td>
<td>My goal is to obtain a resource (not information) available on web pages</td>
<td>free jack o lantern patterns ellis island lesson plans house document no. 587</td>
</tr>
<tr>
<td>3.1 Download</td>
<td>My goal is to download a resource that must be on my computer or other device to be useful</td>
<td></td>
</tr>
<tr>
<td>3.2 Entertainment</td>
<td>My goal is to be entertained simply by viewing items available on the result page</td>
<td></td>
</tr>
<tr>
<td>3.3 Interact</td>
<td>My goal is to interact with a resource using another program/service available on the web site I find</td>
<td></td>
</tr>
</tbody>
</table>
How popular are these intents?

- Data source: 500 English queries from AltaVista search logs

<table>
<thead>
<tr>
<th>GOAL</th>
<th>SET 1</th>
<th>SET 2</th>
<th>SET 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>directed</td>
<td>2.70%</td>
<td>3.30%</td>
<td>7.30%</td>
</tr>
<tr>
<td>undirected</td>
<td>31.30%</td>
<td>26.50%</td>
<td>22.70%</td>
</tr>
<tr>
<td>advice</td>
<td>2.00%</td>
<td>2.70%</td>
<td>5.00%</td>
</tr>
<tr>
<td>locate</td>
<td>24.30%</td>
<td>25.90%</td>
<td>24.40%</td>
</tr>
<tr>
<td>list</td>
<td>2.70%</td>
<td>2.90%</td>
<td>2.10%</td>
</tr>
<tr>
<td>informational total</td>
<td>63.00%</td>
<td>61.30%</td>
<td>61.50%</td>
</tr>
<tr>
<td>download</td>
<td>4.30%</td>
<td>4.30%</td>
<td>5.60%</td>
</tr>
<tr>
<td>entertain</td>
<td>4.00%</td>
<td>8.20%</td>
<td>5.80%</td>
</tr>
<tr>
<td>interact</td>
<td>5.70%</td>
<td>4.30%</td>
<td>6.00%</td>
</tr>
<tr>
<td>obtain</td>
<td>7.70%</td>
<td>10.30%</td>
<td>7.70%</td>
</tr>
<tr>
<td>resource total</td>
<td>21.70%</td>
<td>27.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>navigational</td>
<td>15.30%</td>
<td>11.70%</td>
<td>13.50%</td>
</tr>
</tbody>
</table>

Yet another result

How popular are these intents?

- Data source: the Dogpile search engine transaction log
- The table shows results from automatic intent classification
- They manually checked 400 queries and the accuracy is 74%

<table>
<thead>
<tr>
<th>Classification</th>
<th>Occurrences</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informational</td>
<td>1,228,427</td>
<td>80.6%</td>
</tr>
<tr>
<td>Navigational</td>
<td>155,628</td>
<td>10.2%</td>
</tr>
<tr>
<td>Transactional</td>
<td>139,738</td>
<td>9.2%</td>
</tr>
<tr>
<td></td>
<td>1,523,793</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

How do search engines handle intent today?

- Define fine-grained intent types (no one knows the detailed types)
- Train fine-grained intent classifiers to predict the intent of queries
- Handling each intent in a particular way (a general trend today is to reduce the amount of interactions with the relevant webpages)
Search Result Diversification

Query ambiguity
- Some queries are ambiguous in nature
- Example: “apple” (company or fruit); “index” (database or IR)

Solutions
- Disambiguate queries? Possible but risky
  - When the prediction goes wrong, all results are wrong (and user may need to manually formulate a query to disambiguate)
- Search result diversification: if a query has several different aspects, we just show a few results for each possible aspect.
  - To reduce the risks of biasing towards a particular aspect.
  - To increase the novelty of search results (avoid showing too many results with redundant information)
  - Not only topic-wise, but usually also site-wise (do not show too many results from the same site)
What makes web search different?

• Users, information needs, intents
  • Query understanding
  • Search result diversification
• Document and corpus
  • HTML structure
  • Hyperlinks and anchor texts
  • Spam
• Large-scale search logs
  • Inferring relevance labels: free training data
  • Personalization & contextualization
  • Diagnosing search quality
HTML is semi-structured

• HTML usually has metadata field (but is optional)
  • <meta> tags

• Some HTML tags have semantics (but the webpage creators may not follow)
  • <title> and <body>
  • <p> for paragraph
  • <ul>, <ol>, and <li> for lists and list items
  • <h1> to <h7> headings
  • <em> for emphasize (usually looks as the same as <i> italic)

• A reasonable idea: weight texts in a web page differently by where they appear (in a particular HTML tag).
  • Texts in <title> and <h> tags are probably more important
  • Metadata texts are probably more important
BM25F: Field-weighted BM25

• Count a weighted combination of raw TF from different fields
  • Let $z$ be each field
  • $W_z$ and $tf_z$ are the weight and raw TF for each field

$$tf = \sum_z w_z tf_z$$

• $|d|$, avdl, $k_1$, and $b$ should be adjusted based on the weighting
• Apply the weighted TF and adjusted values of other factors to the original BM25 formula
• Practically needs to tune the weights for different fields based on relevance judgments in a test collection

---

BM25F: Does it work?

- Baseline (ScoreComb): compute BM25 scores for each field, and calculate a weighted combination of the scores.
Other Approach?

• Define matching features for different HTML tags manually
• Train a learning-to-rank model to rank search results
• Example: the first 10 features from the LETOR 4.0 dataset

<table>
<thead>
<tr>
<th>Column in Output</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TF(Term frequency) of body</td>
</tr>
<tr>
<td>2</td>
<td>TF of anchor</td>
</tr>
<tr>
<td>3</td>
<td>TF of title</td>
</tr>
<tr>
<td>4</td>
<td>TF of URL</td>
</tr>
<tr>
<td>5</td>
<td>TF of whole document</td>
</tr>
<tr>
<td>6</td>
<td>IDF(Inverse document frequency) of body</td>
</tr>
<tr>
<td>7</td>
<td>IDF of anchor</td>
</tr>
<tr>
<td>8</td>
<td>IDF of title</td>
</tr>
<tr>
<td>9</td>
<td>IDF of URL</td>
</tr>
<tr>
<td>10</td>
<td>IDF of whole document</td>
</tr>
</tbody>
</table>

The Latest Approach?

• Learn refined document representation by aggregating field-level representations in a deep neural network
• Outperform a hard LTR baseline a little bit
• Multi-field models outperform single-field ones by 5-10%

<table>
<thead>
<tr>
<th>Model</th>
<th>NDCG@1</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25-Field Concatenation</td>
<td>0.4281</td>
<td>0.5953</td>
</tr>
<tr>
<td>BM25F</td>
<td>0.4431</td>
<td>0.6020</td>
</tr>
<tr>
<td>LTR</td>
<td>0.4888</td>
<td>0.6341</td>
</tr>
<tr>
<td>NRM-Field Concatenation</td>
<td>0.4582</td>
<td>0.6110</td>
</tr>
<tr>
<td>NRM-Score Aggregation - Ind. Training</td>
<td>0.4729</td>
<td>0.6229</td>
</tr>
<tr>
<td>NRM-Score Aggregation - Co-training</td>
<td>0.4743</td>
<td>0.6279</td>
</tr>
<tr>
<td>NRM-F - Single Query Representation</td>
<td>0.4846</td>
<td>0.6345</td>
</tr>
<tr>
<td><strong>NRM-F</strong></td>
<td><strong>0.4906</strong></td>
<td><strong>0.6380</strong></td>
</tr>
</tbody>
</table>

What makes web search different?

• Users, information needs, intents
  • Query understanding
  • Search result diversification

• Document and corpus
  • HTML structure
  • Hyperlinks and anchor texts
  • Spam

• Large-scale search logs
  • Inferring relevance labels: free training data
  • Personalization & contextualization
  • Diagnosing search quality
Hyperlinks

• A reference to another webpage
• Hyperlinks in HTML: `<a href="URL">anchor</a>`
• Direction: inlink & outlink
• Anchor text

Computing

From Wikipedia, the free encyclopedia

*For the formal concept of computation, see the concept of a mathematical algorithm — e.g. through computers.*

In computing, a **hyperlink**, or simply a **link**, is a reference to data that the reader can directly follow either by clicking, tapping, or hovering. A hyperlink points to a whole document or to a specific element within a document. **Hypertext** is text
Link analysis

• The idea originates from citation analysis …
• A field widely examined in bibliometrics as early as 1970s …


The difference between the two approaches is not superficial, but is in fact quite important. Traditional probabilistic frameworks allow for arbitrarily rich and complex query models (e.g. [22]), whereas the recent language modeling approaches treat the query as a fixed sample of text with little room for modeling. Many popular techniques in Information Retrieval, such as relevance feedback and automatic query expansion have a very intuitive interpretation in the traditional probabilistic models, but are conceptually difficult to integrate into the language modeling framework of [16], because they involve augmenting the sample (as in [15]) rather than adjusting the model. Furthermore, explicit models of relevance are better suited to other information organization tasks, such as summarization [6], question answering [4], topic detection and tracking (TDT) [25, 12] and text segmentation [3].

A Language Modeling Approach to Information Retrieval

Jay M. Ponte and W. Bruce Croft
Computer Science Department
University of Massachusetts, Amherst
{jonte, croft}@cs.umass.edu
Citation Analysis: Two Purposes

Evaluative

• To assess the quality & impact of articles/journals/conferences or the productivity of individuals/groups/institutions/countries, etc.
• Citation count; journal impact factor; h-index

Relational

• To assess/infer the relevance of articles, journals, authors, etc.
• Direct citation
  • A->B: A and B are relevant
• Bibliographic coupling
  • A->B & A->C: B and C are relevant
• Co-citation
  • A->C & B->C: A and B are relevant
Link Analysis

Evaluative

• To assess the quality of webpages, websites, etc.
• PageRank, HITS, and many variants

Relational

• Anchor text helps refine document representations …
• Applying “direct citation”, “bibliographic coupling”, and “co-citation” to link analysis is less very popular in IR …
• An example: the anchor text helps match a query “ranking function”

Okapi BM25

From Wikipedia, the free encyclopedia

In information retrieval, Okapi BM25 a ranking function used by search engines.

Ranking (information retrieval)

From Wikipedia, the free encyclopedia

Ranking of query results is one of the fundamental problems scientific/engineering discipline behind search engines. Given match the query, the problem is to rank, that is, sort, the doc "best" results appear early in the result list displayed to the us of relevance of documents with respect to an information nee
PageRank scoring

• Imagine a user doing a random walk on web pages:
  • Start at a random page
  • At each step, go out of the current page along one of the links on that page, with equal probability
  • Iteratively doing the process
• “In the long run” each page has a long-term visit rate - use this as the page’s score.
  • Intuitively, a web page with many inlinks gets high visit rate
• So, in short, PageRank computes the probability of visiting a webpage following a random walk user model.
Not quite enough

• The web is full of dead-ends.
  • Random walk can get stuck in dead-ends (no outlinks)
  • Makes no sense to talk about long-term visit rates
“Teleporting”

• At a dead end, jump to a random web page.

• How would users “teleport” (in a real case)?
  • You can consider “teleport” as the case that users, instead of following outlinks of a webpage to visit the next page, directly enter the URL to visit a webpage.

• At any non-dead end, with probability $\alpha$, jump to a random web page.
  • With probability $(1 - \alpha)$, go out on a random link.
  • $\alpha$ is a parameter, e.g., 10%.
The full model

• The full model
  • Let: $L_i$ be the set of links on a page $i$; $N$ be the total # of pages
  • The chances of $i \rightarrow j$ is:
    \[
    P(i \rightarrow j) = (1 - \alpha) \cdot \frac{1}{|L_i|} + \alpha \cdot \frac{1}{N}
    \]

• Represent the transition probability as a $N \times N$ matrix
  • $P_{ij}$ stands for the transition probability of $i \rightarrow j$
  • The matrix only depends on the structure of web

• Represent the current probability of visiting a webpage as an N-dimensional row vector $A$
  • After transition, the new probability of visiting a webpage is $A' = AP$
The full model (cont.)

• Represent the current probability of visiting a webpage as an N-dimensional row vector $A$
  • After transition, the new probability of visiting a webpage is $A' = AP$
  • The probability vector will converge (details in CDM Ch.21)

• While it converges
  • We have $A = A'$
  • Transition would not change the probability of visiting a page
  • So we just solve the equation $AP = A$

• A simple iterative algorithm
  • Starts with uniform probability $A$; multiply $P$ until it converges
  • But requires more complex algorithm for large networks
So, who gets a high PageRank score?

- Following the user model, the probability of visiting a page depends on:
  - The number of inlinks
  - The probability (the PR score) of visiting the inlink pages
- So …
  - A page gets a high PR score if many pages link to the page.
  - A page gets a high PR score if other high PR-score pages link to the page.
- How to use PageRank in IR?
  - Filtering search results by PR scores (set up a threshold)
  - Include PR score as a ranking feature in learning-to-rank
  - Usually use PageRank percentile scores rather than raw scores, because the raw scores depend on N
Important Variants of PageRank

• Can we make PageRank …
  • topic-dependent & user-dependent?
  • Example 1: q = “quantum mechanics”, the “importance” of a webpage by links from sport and music websites are not relevant
  • Example 2: a software engineer types q = “apple” …

• The general idea
  • Use a topic/user-dependent web structure to compute scores, e.g., adjust the transition probability differently based on topics

• Topic-sensitive PageRank
  • For a query, predict its topic distribution, apply topic-sensitive PR

• Personalized PageRank
  • For a user, predict user’s interest topic distribution based on its profile and long-term search history, apply topic-sensitive PR
Anchor Text

- General idea
  - Use anchor text to refine document representation
- The simplest way of using anchor texts
  - Combine all inlinks’ anchor texts as a “pseudo” document
  - Build an index for the pseudo-documents
  - Ranking: combine the retrieval scores from original documents and pseudo-documents
Does it work?

- statMAP and MPC(30) are just some evaluation metrics
- Text, anchor, mix: text-only, anchor-only, combination runs
- In-degree & text + in-degree: using in degree as priors (the idea is kind of similar to PageRank)

<table>
<thead>
<tr>
<th>Run</th>
<th>Full collection</th>
<th>No Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>statMAP</td>
<td>MPC(30)</td>
</tr>
<tr>
<td>Text</td>
<td>0.1442</td>
<td>0.3079</td>
</tr>
<tr>
<td>Anchor</td>
<td>0.0567</td>
<td>0.5558</td>
</tr>
<tr>
<td>Mix</td>
<td>0.1643*</td>
<td>0.4812*</td>
</tr>
<tr>
<td>In-degree</td>
<td>0.0823</td>
<td>0.1876</td>
</tr>
<tr>
<td>Text * In-degree</td>
<td>0.1098</td>
<td>0.2694</td>
</tr>
<tr>
<td>UDWAxQEWeb</td>
<td>0.1999</td>
<td>0.5010</td>
</tr>
<tr>
<td>uogTrdphCEwP</td>
<td>0.2072</td>
<td>0.4966</td>
</tr>
<tr>
<td>ICTNETADRun4</td>
<td>0.1746</td>
<td>0.4368</td>
</tr>
</tbody>
</table>

Marijn Koolen and Jaap Kamps. The importance of anchor text for ad hoc search revisited. In SIGIR 2010: 122-129.
Does it work?

- $\alpha$-nDCG and IA-P@10: some variants of nDCG and P@10 for search result diversification
- Using anchor text alone have the best performance in both ad hoc search and search result diversification here

Table 5: Ad hoc and Diversity evaluation using the Diversity relevance judgements. Significance tests are with respect to the full text run, confidence levels are 0.95 (*), 0.99 (**) and 0.999 (***)

<table>
<thead>
<tr>
<th>Run</th>
<th>$\alpha$-nDCG@10</th>
<th>nDCG@10</th>
<th>IA-P@10</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>0.120</td>
<td>0.1564</td>
<td>0.054</td>
<td>0.1700</td>
</tr>
<tr>
<td>Anchor</td>
<td>0.257***</td>
<td>0.2780**</td>
<td>0.082**</td>
<td>0.2460***</td>
</tr>
<tr>
<td>Text + Anchor</td>
<td>0.223***</td>
<td>0.2459**</td>
<td>0.083**</td>
<td>0.2420***</td>
</tr>
<tr>
<td>uogTrDYCcsB</td>
<td>0.282</td>
<td>–</td>
<td>0.132</td>
<td>–</td>
</tr>
<tr>
<td>ICTNETDivR3</td>
<td>0.272</td>
<td>–</td>
<td>0.095</td>
<td>–</td>
</tr>
<tr>
<td>UamsDancTFb1</td>
<td>0.250</td>
<td>–</td>
<td>0.079</td>
<td>–</td>
</tr>
</tbody>
</table>

Marijn Koolen and Jaap Kamps. The importance of anchor text for ad hoc search revisited. In SIGIR 2010: 122-129.
So, Anchor Text

Important?

• Yes! Almost a must-have feature for web search!
• Sometimes (e.g., for popular webpages), using anchor text alone performs better than using the content alone
• But may be biased towards old web pages (because of more links); PageRank has the same issue

How to use?

• Design features & learning-to-rank
Thursday

- Users, information needs, intents
  - Query understanding
  - Search result diversification
- Document and corpus
  - HTML structure
  - Hyperlinks and anchor texts
  - Spam
- Large-scale search logs
  - Inferring relevance labels: free training data
  - Personalization & contextualization
  - Diagnosing search quality