Automatic Identification of User Goals in Web Search [WWW’05]

Uichin Lee @ UCLA
Zhenyu Liu @ UCLA
Junghoo Cho @ UCLA

Presenter: Emiran Curtmola @ UC San Diego — CSE 291
4/29/2008
Motivation

- Need to improve the quality of a SE’s results → improve the user’s browsing experience
  - Page ranking
  - Result clustering
  - Answer presentation
    - Snippet selection
  - Which hyperlinks to highlight and how much in order to indicate paths to search results (include the distance to relevant pages as the nr. clicks)
  - Query answers organized as to reflect the organization’s intranet structure (similar to a table of contents)
- …
Statement of Problem

- Understand user web search behavior
  - Understand how and for what people are searching
  - Understand why users are searching
    - What are the goals of users issuing queries on the Web?
Challenges

- Many users, many needs, many goals
  - Is there an agreed classification of query goals?
  - Do most queries have a detectable goal?
    - Is query goal classification feasible?

- Characterization of Web query goals

- Goal: Automatically identify query goals on the Web
  - A way for SE to associate user goals and queries
State-of-the-Art

- Taxonomy of query goals [Rose et al., WWW’04], [Broder, SIGIR’02]
  - Conceptual framework for user goals
    - Offline classification of web queries by manually investigating Web query logs
Contributions

- Show that the majority of Web queries have a detectable clear goal
- Propose a benchmark of queries and their goals
- Extract properties of Web queries to predict the query goals
  - Effectiveness measurements show 80%-90% query goal prediction accuracy
Roadmap

- Taxonomy of query goals
- Degree of identifiability of a query goal via a human subject study
- Automatic identification of query goals
- Experimental measurements
Taxonomy of Query Goals (Rose, Broder)

- Methodology
  - User surveys on AltaVista query-log dataset

- Taxonomy of Web query goals
  - **Navigational** = user has a particular Web page in mind
    E.g., “pubmed”, “citeseer”, “bestbuy”, “Stanford”

  - **Informational** = user intends to visit multiple pages to learn about a topic
    E.g., “hidden markov model”
### Table 1: The Search Goal Hierarchy. Queries are only assigned to leaf nodes.

All examples are taken from actual AltaVista queries.

<table>
<thead>
<tr>
<th>SEARCH GOAL</th>
<th>DESCRIPTION</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>17% 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>60% 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>3% 2.1 Directed</td>
<td>I want to learn something in particular about my topic</td>
<td></td>
</tr>
<tr>
<td>25% 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>color blindness jfk jr help quitting smoking walking with weights</td>
</tr>
<tr>
<td>3% 2.3 Advice</td>
<td>I want to get advice, ideas, suggestions, or instructions.</td>
<td>pella windows phone card travel</td>
</tr>
<tr>
<td>25% 2.4 Locate</td>
<td>My goal is to find out whether/where some real world service or product can be obtained</td>
<td>amsterdam universities florida newspapers</td>
</tr>
<tr>
<td>2% 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></td>
</tr>
<tr>
<td>23% 3. Resource</td>
<td>My goal is to obtain a resource (not information) available on web pages</td>
<td>kaza lite name roms xxx porno movie free live camera in l.a. weather measure converter</td>
</tr>
<tr>
<td>4.5% 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>6% 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>5% 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>free jack o lantern patterns ellis island lesson plans house document no. 587</td>
</tr>
<tr>
<td>8% 3.4 Obtain</td>
<td>My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.</td>
<td></td>
</tr>
</tbody>
</table>
Roadmap

- Taxonomy of query goals
- Degree of identifiability of a query goal via a human subject study
- Automatic identification of query goals
- Experimental measurements
Predictability of Query Goals

- User’s goal is subjective → can we associate a **query** with a particular **goal** without any user feedback?
  
  - Predictable queries: queries can be clearly classified as navigational or informational
    - How many are predictable?
  
  - Unpredictable queries: classification by human observers is bimodal
    - How many queries are unpredictable?
Human Subject Study Setup

- 50 most popular Google queries issued from UCLA CS department
- 28 grad students from CS dept. guess the most probable goal for each query

- Assumption: their consensus guess is correct
Questionnaire

- Use user intention descriptions to classify
  - **Choice 1:** I already have a particular Website in mind, and my major interest is just to reach that site through the search engine.
  - **Choice 2:** I know there’s a particular Website corresponding to this query. However, my interest is not only to reach that site, but to visit some other sites returned by the search engine.
  - **Choice 3:** I have no particular Website in mind. I am willing to click on multiple results returned by the search engine.
Manual Classification Results

- \( i(q) = \) percentage of users who indicate q’s goal is informational
Unpredictable Queries

- Breakdown: 23 queries
  - 17 queries belong to the following topics
    - Software names (e.g., “cygwin”, “spybot”, “ns2”)
    - Personal names (e.g., CS researchers)
  - 6 queries on diverse topics

- Distribution of software (12) and personal (8) names
Manual Classification Results (II)

- Clear separation of informational and navigational queries (after 20-query removal)
A large fraction of queries are predictable
- Can be associated with a particular goal
  \( \rightarrow \) Feasible for automatic goal classification:
  Next, use the 30 queries as a benchmark

Most of the unpredictable queries fall into certain topics
- Use topic detection methods to detect the topic
Roadmap

- Taxonomy of query goals
- Degree of identifiability of a query goal via a human subject study
- Automatic identification of query goals
- Experimental measurements
Methodology

- Detect the query goal based on the user-click behavior and the Web structure

- Propose metrics to measure
  - Click-behavior of other users who previously issued same query (leverage the SE query log)
    - Click distribution
    - Average number clicks per query
  - Distribution of HTML links (SE does not have enough stats about the query)

- Experimentally confirm the effectiveness of metrics
Click-Behavior on Previous Queries (I)

- Intuition: in the past, other users clicked most likely on a single (and same) answer for a navigational query

- Metric: **click distribution per query**
  - Measure number clicks per query answer
Click Distribution

- Sort answers for a query in descending order of the number of clicks they received from all users

- Navigational query: a skewed distribution
- Informational query: a flat distribution
Click-Behavior on Previous Queries (II)

- Intuition: number of results a user clicks per query
- Metric: average number of clicks per query

- Navigational query: user clicks once
- Informational queries: user clicks multiple times
Anchor-link Distribution

- Intuition: there is a direct correlation between
  - the query (e.g., query="pubmed")
  - the number of anchor destinations from Web links with the same anchor text as the query (e.g., <a href="www.ncbi.nih.gov">PubMed</a>)

- Metric: anchor-link distribution
  - Crawl the Web
  - Locate all links on the Web with anchor text as the query
  - Count the number of distinct anchor destinations
Anchor-link Distribution

- Sort the anchor destinations in descending order of their frequency

- Navigational query: a skewed distribution
  - A query is likely to be navigational if many links exist with the same destination and the same anchor text which is the query

- Informational queries: a flat distribution
Limitation of Anchor-Link Distribution

- Danger of Web Crawling: THE NOISE
  - Link spam
    - Artificial links not really related to the anchor text to increase the page rank
  - Mirror sites at multiple locations

- Solution: use spam and mirror detection techniques
Roadmap

- Taxonomy of query goals
- Degree of identifiability of a query goal via a human subject study
- Automatic identification of query goals
- Experimental measurements
Experimental Setup: UCLA Query Benchmark

- 50 most popular queries → 30 predictable queries

- Record all Google incoming/outgoing messages from the department
  - 6 months (2004);
  - 147,744 unique queries
  - 1.6 query frequency on average
  - pick only queries issued by at least 3 different IPs
    - on average, the benchmark queries were issued by 19.6 IPs each
Measurements

- User-click frequency for query q
  - Track “Referer” HTML field to collect click-data for q
  - On average, each of the 30 queries got 42 user clicks from ~20 different issuers

- Anchor data
  - crawling 60,824,009 Web pages starting from sites in the Open Directory Project
  - Found 3,169 matching anchor text on average per benchmark query
Results: Click Distribution

- User-click distribution: mean, median, Skewness, Kurtosis show similar predictive power

- Goal-prediction graph for the median

  - Threshold = 1.0 $\Rightarrow$ 80% accuracy:
    - 25 queries correctly classified
    - 5 incorrectly classified for informational queries

Figure 11: Median of click distribution
Results: Avg. Nr. of Clicks per Query

- Threshold = 1.5 $\Rightarrow$ 80% accuracy
  - Mid-point between one click per navigational queries and more than one for informational queries

![Figure 12: Avg # of clicks per query](image-url)
Results: Anchor Link Distribution

- Mean, median, Skewness, Kurtosis show similar predictive power

- Threshold $= 1.0 \Rightarrow 75\%$ accuracy:
  - incorrectly classified for navigational queries

![Graph showing median of anchor-link distribution](image)
Results: Combined Metrics

- Linear combination of metrics
  
  \[ m = w_1 \times m_1 + w_2 \times m_2 + \ldots + w_n \times m_n \]

- E.g., 90% accuracy for \( w_1 = w_2 = 1 \)
Comparison with Prior Work (I)

- [Kang et al., SIGIR’03]’s goal prediction is based on the content of the pages clicked on
  - Count the occurrence of query terms in the Web pages

- Anchor usage rate
  - Frequency of query terms in the anchor text vs.
  - Frequency of query terms in the overall document collection

- Navigational query: if the frequency in the anchor text is higher
Comparison with Prior Work (II)

- Query term distribution
  - Frequency of query terms in the home pages (top-level page of a website) vs. in the content-page collection
  - Navigational query: if the frequency in the home pages is higher

- Term dependence
  - Navigational query: if the co-occurrence of multiple query terms are more dependent in the homepage collection vs. the content-page collection
  - Measure for term dependence: mutual information
Comparison with Prior Work (III)

- Using prior metrics on UCLA benchmark

*Figure 15: Anchor usage rate*
*Figure 16: Query term distribution*
*Figure 17: Term dependence*

60% accuracy                    no clear separation
Summary

- 60% of queries have clearly defined goals
  - Amenable for automatic goal identification using
    - Past user-click behavior (leverage query log)
    - Anchor-link distribution (crawl the Web)
  - Up to 90% prediction accuracy

- 40% of queries are unpredictable
  - Try to detect topics instead of detecting goals
Future Work

- Automate identification of Web query goals in a more detailed taxonomy (e.g., hierarchical)

- More comprehensive query benchmark
  - Current benchmark: too narrow + too small
  - Maybe a narrow benchmark is appropriate in practice for geographically defined regions
References


4/29/2008
Questions?

Thanks!