

Characterizing the Influence of Domain Expertise on Web Search Behavior

Ryen White, Susan Dumais, Jaime Teevan

Microsoft Research

{ryenw, sdumais, teevan}@microsoft.com

Example to start

- A cardiologist and a newly-diagnosed patient get the same results for the query “heart disease”
- If we could estimate their level of expertise we could tailor the search experience to each of them
 - Cardiologist could get technical articles
 - Patient could get tutorial information
- **This paper is about characterizing and using such domain expertise to improve Web search**

Background

- Domain expertise = knowledge of subject area
- Domain expertise \neq search expertise
 - Search expertise is knowledge of search process
- Previous research has highlighted differences between domain experts and domain non-experts
 - Site selection and sequencing, task completion time, vocabulary and search expression, ...
 - Involve small numbers of subjects w/ controlled tasks
- We extend this work in breadth (\uparrow domains) and scale

Outline

- Studying Domain Expertise
 - Study overview
 - Log data
 - Automatically identifying domain experts
 - Differences between experts vs. non-experts
- Using Domain Expertise
 - Predicting domain expertise based on search interaction
 - Improving search experience via expertise information
- Conclusions



Studying Domain Expertise

Study

- Log-based study of Web search behavior
- Contrast strategies of experts and non-experts
- Large-scale analysis w/ greater diversity in vocabulary, web sites, and tasks than lab-based studies
- Four domains were studied
 - **Medical**, Legal, Financial, Computer Science
 - Large professional groups who use Web, of general interest
 - Just focus on Medical in this talk for time...

Data Sources

- Logs w/ querying *and* browsing behavior of many users
 - Three months from May 2007 through July 2007
 - > 10 billion URL visits from > 500K users
- Extracted browse trails and search sessions
 - *Browse trails* = sequence of URLs per tab/browser instance
 - *Search sessions* = sub-trails starting w/ search engine query and ending w/ 30 min. interaction timeout
- Search sessions let us compare domain experts and non-experts in and out of their domain of interest
- First need to differentiate experts from non-experts ...

Identifying Domain Experts

- Two steps in identifying domain experts from logs:
 - **Step 1:** Identify users with topical interest
 - Ensures that behavior relates to users interested in domain and helped control for topic differences
 - **Step 2:** Separate experts from non-experts
 - From user group in Step 1, separate experts based on whether they visit specialist Websites
- Simple, broadly-applicable method
- Lets us extend lab studies to real-world settings

Topical Interest

- Classified browse trails using Open Directory Project
 - Automatically assigned labels to URLs based on ODP with URL back-off as required
 - Filtered outliers and computed % pages in each domain
 - Medical = Health/Medicine
 - Financial = Business/Financial_Services
 - Legal = Society/Law/Legal_Information
 - Computer Science = Computers/Computer_Science


| Domain | # users | # sessions | # in-domain sessions |
|------------------|---------|------------|----------------------|
| Medical | 45,214 | 1,918,722 | 94,036 |
| Financial | 194,409 | 6,489,674 | 279,471 |
| Legal | 25,141 | 1,010,868 | 36,418 |
| Computer Science | 2,427 | 113,037 | 3,706 |

Dividing Experts & Non-Experts

- Surveys, interviews, etc. not viable at scale
- Divided experts/non-experts using observable behavior
- Filtered users by whether they visited specialist sites
 - Sites identified through discussion w/ domain experts

| Domain | Expert URL filters | Expert | Non-expert |
|-----------|--|---------------|------------------|
| Medical | ncbi.nlm.nih.gov/pubmed, pubmedcentral.nih.gov | 7,971 (17.6%) | 37,243 (82.4%) |
| Financial | bloomberg.com, edgar-online.com, hoovers.com, sec.gov | 8,850 (4.6%) | 185,559 (95.4 %) |
| Legal | lexis.com, westlaw.com | 2,501 (9.9%) | 22,640 (90.1 %) |
| CS | acm.org/dl, portal.acm.org | 949 (39.1%) | 1,478 (60.9%) |

- Most sites require subscription; assume visitors have *above average* domain knowledge



Differences between Domain Experts and Non-Experts

Domain Expertise Differences

- Behavior of experts/non-experts differs in many ways
- Some are obvious:
 - *Queries* (experts use more tech. vocab., longer queries)
 - *Source selection* (experts utilize more tech. sources)
 - URL-based analysis
 - Content-based analysis (judges rated page technicality)
 - *Search success* (experts more successful, based on CTR)
- Some are less obvious:
 - *Session features*, e.g.,
 - Branchiness of the sessions
 - Number of unique domains
 - Session length (queries, URLs, and time)

Branchiness & Unique Domains

- Session branchiness = $1 + (\# \text{ revisits to previous pages in the session followed by visit to new page})$

| Session Feature | Expert | | Non-expert | |
|------------------|----------|-----------|------------|-----------|
| | <u>M</u> | <u>SD</u> | <u>M</u> | <u>SD</u> |
| Branchiness | 9.91 | 12.11 | 8.54 | 11.07 |
| # unique domains | 8.98 | 8.13 | 7.57 | 6.78 |

- Expert sessions are more branchy and more diverse than non-experts
- Experts may have developed strategies to explore the space more broadly

Session Length

- Length measured in URLs, queries, time

| Session Length Feature | Expert | | Non-expert | |
|--------------------------------|----------|-----------|------------|-----------|
| | <u>M</u> | <u>SD</u> | <u>M</u> | <u>SD</u> |
| Page views (inc. result pages) | 39.70 | 47.30 | 27.68 | 45.68 |
| Query iterations | 13.93 | 19.14 | 9.90 | 15.14 |
| Time (seconds) | 1776.45 | 2129.32 | 1549.74 | 1914.86 |

- Greater investment in tasks by experts than non-experts
- Search targets may be more important to experts making them more likely to spend time and effort

Other Considerations

- Expert/non-expert diffs. hold across all four domains
- *Out of domain* search sessions are similar:

| Session Feature | Expert | | Non-expert | |
|--------------------------------|----------|-----------|------------|-----------|
| | <u>M</u> | <u>SD</u> | <u>M</u> | <u>SD</u> |
| Branchiness | 4.23 | 7.11 | 4.28 | 7.52 |
| Unique domains | 4.19 | 4.13 | 4.28 | 3.99 |
| Page views (inc. result pages) | 17.89 | 19.06 | 18.01 | 31.44 |
| Query iterations | 4.79 | 8.71 | 4.32 | 7.89 |
| Time (seconds) | 749.94 | 1227.51 | 753.96 | 1243.07 |

- Similarities in other features (e.g., queries)
- Observed differences attributable to domain



Using Domain Expertise

Predicting Domain Expertise

- Based on interaction behavior we can estimate a user's level of domain expertise
 - Rather than requiring offline tests
- Search experience can be tailored based on estimation
 - Just like we needed with the cardiologist and the patient
- Three prediction challenges:
 - **In-session**: After observing ≥ 1 action(s) in a session
 - Post-session: After observing a single session
 - User: After observing ≥ 1 sessions from same user

Within-Session Prediction

- Predicting domain expertise as the session proceeds
 - Used maximum margin averaged perceptron
 - Trained using features of queries, pages visited, both
 - Five-fold cross validation and ten experimental runs
- e.g., for CS, our best-performing predictor:

| Action type | Action number | | | | | Full session |
|-------------|---------------|--------|--------|--------|--------|--------------|
| | 1 | 2 | 3 | 4 | 5 | |
| All | .616* | .625* | .639** | .651** | .660** | .718** |
| Queries | .616* | .635** | .651** | .668** | .683** | .710** |
| Pages | .578 | .590* | .608* | .617* | .634** | .661** |

*,** = significant difference from maximal margin, always neg. (.566)

- Predict after just a few actions; Queries best – less noisy

Improving Search Experience

- Search engine or client-side application could bias results toward websites suitable for expertise level
 - Reinforces behavior rather than encouraging learning
- Help domain non-experts become experts over time
 - Provide non-expert definitions for related expert terms
 - e.g., search for [cancer] includes definition of [malignancy]
 - Help non-experts identify reliable expert sites or use the broader range of information that experts do

Conclusions

- Large-scale, log-based study of Web search behavior of domain experts and non-experts
- Showed that experts/non-experts search differently within their domain of expertise, and similarly otherwise
- Differences/similarities visible across four domains
 - Extending previous lab studies in breadth and scale
- Developed models to predict domain expertise
 - Can do this accurately for a user / post-session / in-session
- Domain expertise information can be used to tailor the search experience and help non-experts become experts