# Semantics of Query Rewriting Patterns in Search Logs

Sumio Fujita[1] Georges Dupret[2] Ricardo Baeza-Yates[3]

[1]Yahoo! JAPAN Research

[2]Yahoo! Labs

[3]Yahoo! Research

November 2, 2012 / ESAIR'12

# Semantics of query rewriting patterns

- Boldi *et al.* pointed out that the typically useful recommendations are either specializations or topic shifting, which they refer to as “parallel moves”.
- Specialization reformulations might be observed less frequently than topic shifting depending on the search contexts.
- Topic shifting occurs especially when users engage in complex tasks like researching for a new vehicle and comparing competing candidate models.
- For example, in our logs, the most frequent queries after “toyota” are “honda”, “nissan” and “lexus”.
- The most frequent query after “driver’s license renewal” is “slight violence of traffic laws”, which may jeopardize renewing their driver’s license.

# Assumptions

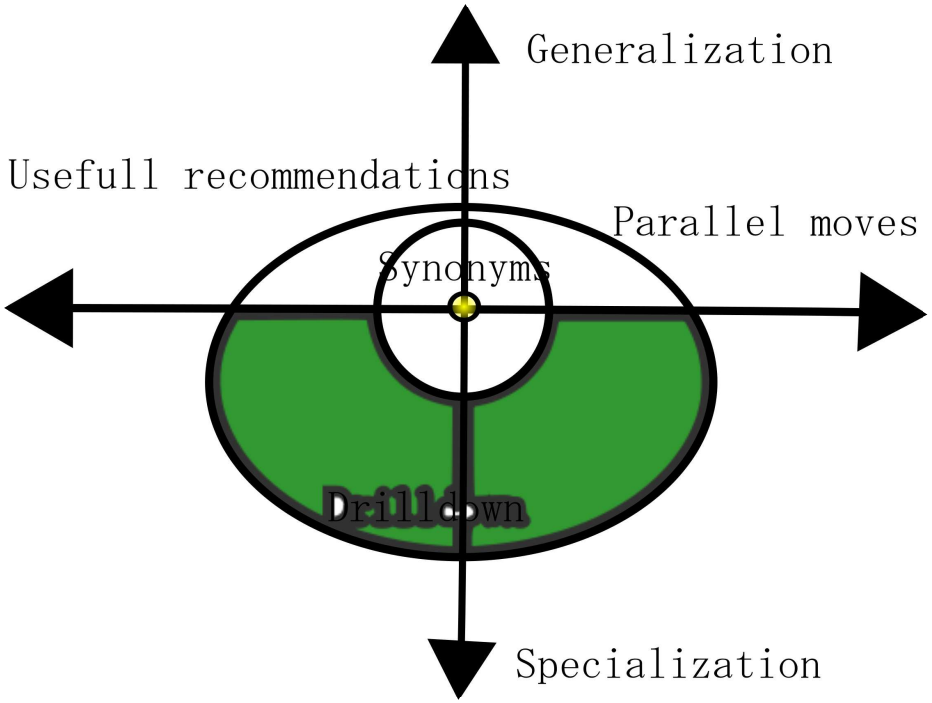


Figure: Schematic view of semantic relations of related queries.

# Assumptions

- Since most useful recommendations are either specializations or parallel moves, better to use distinct methods to cover both types.
- In the semantic hierarchy of information needs, locating the original user query at the center, specialization queries in the lower part are useful as specialization queries: co-click and co-topic based methods can identify such relations.
- The neighbouring queries in the semantic hierarchy are generally useful as the parallel moves: a co-session based method can identify such relations.
- Too close queries are not useful as recommendations.

# Queries in Best Rank Directed Co-Click Relations (CCQs)

$$CCQ_q \equiv \bigcup_{u \in UC_q} \arg\min_{q' \in QC_u} \mathsf{rank}_u(q') \ .$$

- where, the URL cover $UC_q$ of a query $q$ is the set of URLs clicked in response to query $q$,
- the query cover of URL $u$, $QC_u$ is the set of queries for which URL $u$ is clicked.
- CCQs are the queries that rank some clicked URLs at the best positions,
- mostly specialization queries but also some are web jargon like queries.
- They share some clicked URLs but may not share any keywords.

# Queries in Best Rank Directed Co-Click Relations (CCQs)

$$
\begin{aligned}
P_{CC}(q_2|q_1) &= \sum_{u \in UC_{q1}} P(q_2|u) \cdot P(u|q_1) \\
&= \sum_{u \in UC_{q1}} \frac{P(u|q_2) \cdot P(q_2)}{P(u)} \cdot P(u|q_1)
\end{aligned}
$$

with

$$
\begin{aligned}
P(u) &= \frac{cnt(u)}{\sum_{q \in Q} cnt(q)}, P(q) = \frac{cnt(q)}{\sum_{q' \in Q} cnt(q')}, \text{ and} \\
P(u|q) &= \frac{cnt(u,q)}{cnt(q)}.
\end{aligned}
$$

# Queries in Co-topic Relations (CTQs)

- Commercial search engines commonly use expansions of input queries in logs as recommendations, such as,
- “curry” vs “curry recipe”, “curry restaurant”,...
- CTQs are queries expanded by additional keywords,
- mainly representing specialization rewriting.
- They share some keywords but may not share any clicked URLs.

$$P_{CT}(q_2|q_1) = \frac{cnt(q_2)}{cnt(q_1) + \sum_{q_{2'} \in CTQ_{q_1}} cnt(q_{2'})} .$$

# Queries in Co-Session Relations (CSQs)

- CSQs are the queries submitted consecutively from the same user in a time interval no longer than 5 minutes,
- directly representing some users rewriting activities.
- CSQs include not only the reformulation or rewriting of queries, such as in the co-topic relation, but also queries that reflect a shift in information needs.
- They may not share neither keywords, nor clicked URLs.

$$P_{CS}(q_2|q_1) = \frac{cnt(q_2, q_1)}{cnt(q_1)} .$$

# Query Similarity Measures

- Baeza-Yates and Tiberi [2007] evaluated semantic relations between queries connected by an edge of their click cover graph:

$$Sim_{prefix}(D, D') = |P(D, D')|/max\{|D|, |D'|\} \ ,$$

- where $P(D, D')$ is the longest common prefix of the category paths $D$ and $D'$ where the queries $q$ and $q'$ belong to respectively, referring to the search directory. We slightly revised this as follows:

$$Sim_{substring}(D, D') = |C(D, D')|/max\{|D|, |D'|\},$$

- where $C(D, D')$ is the number of common sub-parts of two category paths, referring to the Yahoo! JAPAN Category hierarchy.

# Examples

Table: Examples of extracted queries by three methods. Queries are translated from Japanese.

| Original query: “ANA” (an airline company in Japan) | | | |
|---|---|---|---|
| Rank | Co-click | Co-topic | Co-session |
| 1 | ANA skyholiday | ANA time schedule | JAL |
| 2 | ANA mileage club | ANA steak-holder’s benefit coupon | Skymark |
| 3 | airplane | ANA wallpapers | JR |
| 4 | ANA domestic airline | ANA mileage | JTB |
| 5 | ANA timetables | ANA tour | HIS |

# Results

- Average similarities of original query - extracted query pairs by category matching.
- The CCQs showed slightly higher similarities than the CTQs because co-click relations assure semantically close relationship by the existence of co-clicked URLs.
- Clearly, the CSQ pairs are semantically less close to each other due to the topic shifting patterns.

| Sim. func. | CCQ | CTQ | CSQ |
|---|---|---|---|
| Prefix | **0.8217** | 0.8198 | 0.7785 |
| Substring | **0.8474** | 0.8355 | 0.7936 |
| Avg. | **0.8346** | 0.8277 | 0.7861 |

# Topological relations in Hierarchy

- Ratio of the topological relations in the semantic hierarchy of query pairs(%).
- Semantically close relationship is observed in the CCQ and CTQ pairs.

| Relation type | CCQ | CTQ | CSQ |
|---|---|---|---|
| $C_1$ and $C_2$ are Identical | 1.12 | 1.24 | 0.43 |
| $C_1$ and $C_2$ are Sibling | 1.32 | 1.41 | 1.00 |
| $C_2$ is descendant of $C_1$ | 1.02 | 1.76 | 0.45 |
| $C_1$ is descendant of $C_2$ | 0.58 | 0.58 | 0.36 |
| $C_1$ and $C_2$ share some hyper layers | 32.08 | 36.39 | 23.62 |

# Conclusions

- We examined three methods to explore query rewriting patterns from the search logs.
- The first method exploits the clicked document positions in the ranking and selects queries, which may return a higher rank for the clicked document.
- The second method is based on the observation that users often refine their queries by adding some terms.
- The third method uses the query sequences in search sessions and identifies some typical topic shifts from the query.
- We evaluated semantic relations of extracted query pairs by using a large scale semantic hierarchy of Japanese web page annotations.
- Our experiments showed that the semantic relations of the query pairs reflect the characteristics of each method that extracted rewriting patterns.