

Mining User-logs for Query and Term Clustering

- An exploration on Encarta encyclopedia

Jian-Yun Nie,
 RALI, DIRO, Univ. Montreal
 Ji-Rong Wen, Hong-Jiang Zhang,
 MSRCN

Context of this work

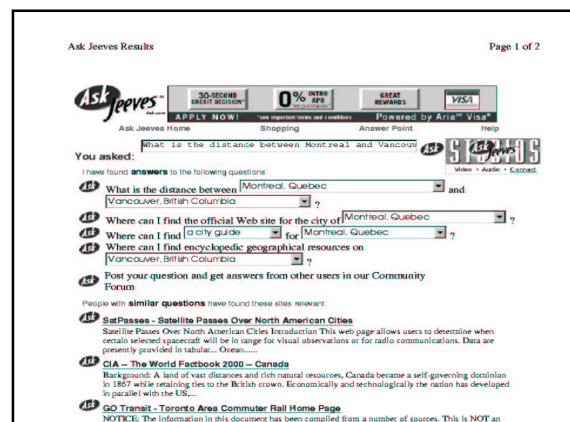
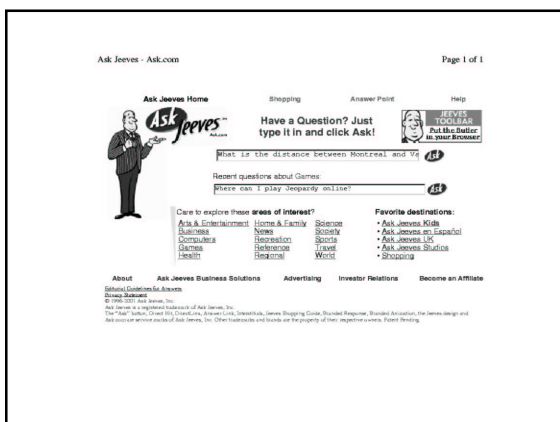
- Summer 2000 at MSR in China (Beijing)
- Needs of Encarta group at Microsoft
 - Human editors work on improvements by adding both new articles and new index
 - They wanted to identify the “hot topics”
 - Only manual analysis of user queries was being used
- Build a clustering tool that automatically groups similar queries.

Context

- Influence of AskJeeves
 - Provide more accurate answers (answer questions)
 - Hypothesis: many users are interested in a limited number of topics (hot topics).
 - If these queries (questions) can be answered correctly, most users can be satisfied.

FAQ-based Question-Answering

- A set of FAQs that have previously been answered, or whose answers have been manually checked.
- A user’s query is compared with the FAQs.
- Similar FAQs are proposed to the user.
- The user selects a FAQ
 - » Answers of higher quality



Distance result Page 1 of 1

Distance result

Distance between Montreal, Quebec, Canada and Vancouver, British Columbia, Canada, as the crow flies:

2295 miles (3694 km) (1995 nautical miles)

Initial heading from Montreal to Vancouver:
west-northwest (294.7 degrees)

Initial heading from Vancouver to Montreal:
east-northeast (77.2 degrees)

See driving distance and directions (courtesy MapBlag).

See these places on the map (courtesy Xerox PARC).

Montreal, Quebec, Canada

Location: 45:30:00N 73:34:48W

Vancouver, British Columbia, Canada

Location: 49:15:00N 123:04:48W

You may try a new search.

A service of Ball Online: The Ultimate Source of Ball find

How to cluster queries?


- Query similarity based on queries:
 - keywords
 - question form (templates, question words, etc.)
- Observations:
 - Queries are short
 - Forms and words vary
 - Difficulty to cluster truly similar queries

Using document links for ranking/similarity

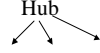
- Citation links have been explored in IR in 1970s, however, links were limited.
- PageRank in Google

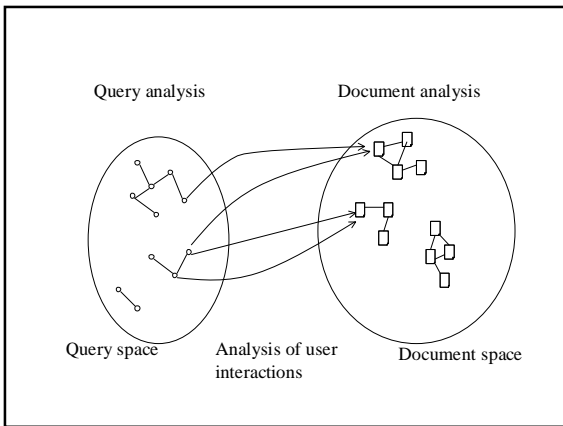
$$R(p) = \epsilon/n + (1 - \epsilon) \sum_{(q \rightarrow p)} R(q)/\text{outdegree}(q)$$
- HITS (Kleinberg)

Authority



Hub





Related work on cross-connections

- Relevance feedback in IR
 - Relevance judgement by user
 - Query reformulation
 - e.g. Rocchio: $Q' = \alpha * Q + \beta * R - \gamma * NR$
 - Document clustering
 - All the documents relevant to a query are put into the same cluster

Related work (cont'd)

- Citeseer, AskJeeves, Amazon, Lycos, etc.
 - Users who read this page (paper) also read other pages (papers).
- Su, Yang, et al. (2000)
 - URL co-occurrences within time windows in user-logs
- Beeferman & Berger (2000)
 - Exploring user-logs for document and query clustering

Our hypotheses

- Hypothesis 1:
Queries of similar form or words are similar.
 - Hypothesis 2:
Queries leading to common document clicks are similar.
- ⇒ Combining the two criteria

Expected advantages

- Benefit from user's own decision of click-throughs
 - implicit relevance judgements
- Bridge the gap between document space and query space
 - Different wording
 - Construct a "live thesaurus" linking user's words with document's words

Possible problems

- A clicked document is not necessarily relevant to the query.
 - Even in a relevant document, many words do not have semantic relations with the query.
- ⇒ May end up with noisy clusters.

Why may this idea work?

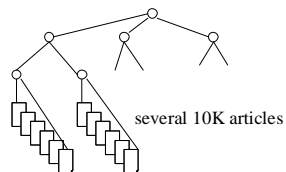
- Pseudo-relevance feedback in IR
 - IR system finds a set of documents
 - The top ranked documents are assumed to be "relevant" (many of them are not!)
 - Using these documents to reformulate the query
 - Improvements in effectiveness of 10-15%
- ⇒ Non strict feedback can help
- ⇒ User click-thoughts are better data:
top-documents + some user judgment
- ⇒ Large number of data

Encarta encyclopedia

- Document hierarchy

9 categories

93 sub-categories

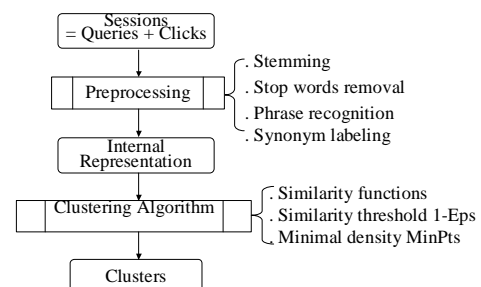


- Example:

Country (sub_cat)

- people (or population)
- culture
- geography
- etc.

Overview of query clustering



Query session

- Encarta interface:
 - The user submits a query (question)
 - Encarta search engine proposes a list of articles
 - The user selects (clicks on) some of the documents
- User-logs:
 - ... address, time, query, clicks, ...
- Extract Sessions:
 - query -> { ... document-click, ... }
- about 23 000 sessions/day from MSN

Choice of clustering algorithm

- A number of algorithms available
 - Hierarchy agglomerative clustering
 - k_means, etc.
- Our criteria
 - complexity (large amount of data)
 - incremental (daily change)
 - not too many manual settings (e.g. number)
 - No limit on max. cluster size

BDBScan

- complexity $O(n \cdot \log(n))$
- incremental (incremental BDBScan)
- clusters of different shapes and sizes
- requires
 - Eps = distance threshold
 - MinPts = minimal number in a cluster
- Principle:
 - scan elements only once
 - if an element is a core element ($\#Eps\text{-neighbor} \geq \text{MinPts}$) then expand the cluster by adding the neighbors.

Similarity based on queries

- Selection of keywords
 - stop-words
 - stemming
- $\text{Sim}(q1, q2) = \frac{\#common_kw(q1, q2)}{\text{Max}(\#kw(q1), \#kw(q2))}$
- $\text{Sim}(q1, q2) = \frac{\sum_{k_i \in q1, q2} (w_{q1}(k_i) + w_{q2}(k_i)) / 2}{\text{Max}(\sum_{k_i \in q1} (k_i), \sum_{k_i \in q2} (k_i))}$

Using query form

- Edit distance($q1, q2$)
- $\text{Sim}(q1, q2) = 1 - \text{edit_dist}(q1, q2)$

Examples:

- Where does silk come from?
- Where does lead come from?
- Where does dew come from?
- => Where does X come from?

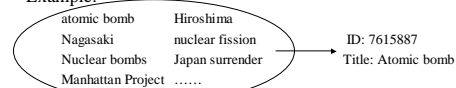
Using document clicks

$D_C(q_i) = \{ d_c_{i1}, d_c_{i2}, \dots, d_c_{ip} \} \subseteq D(q_i)$

$D_C(q_j) = \{ d_c_{j1}, d_c_{j2}, \dots, d_c_{jq} \} \subseteq D(q_j)$

- $\text{Sim}(q1, q2) = \frac{\#common_click(q1, q2)}{\text{Max}(\#click(q1), \#click(q2))}$

Example:



Using Encarta hierarchy

- Hypothesis: two documents within the same category (sub-category) are more similar.
- $S(d1, d2) = [\text{Level}(\text{common_parent}(d1, d2)) - 1] / 3$
- $\text{Sim}(q1, q2) = 1/2 * (\frac{\sum_i \text{Max}_j s(d_i, d_j)}{\#\text{doc_click}(q1)} + \frac{\sum_j \text{Max}_i s(d_i, d_j)}{\#\text{doc_click}(q2)})$

Example

Query 1: image processing

ID: 761558022 Title: Computer Graphics

Query 2: image rendering

ID: 761568805 Title: Computer Animation

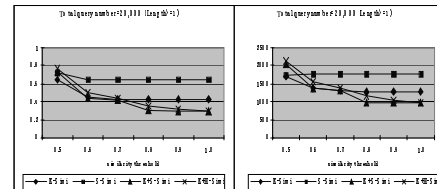
- Common parent: "computer science"
 - common keyword: "image"
- => similar

Combinations of criteria

- Linear combinations
 - $\text{Sim}(q1, q2) = \alpha * \text{Sim}_{\text{kw}}(q1, q2) + \beta * \text{Sim}_{\text{click}}(q1, q2)$
- How to set α and β ?
 - using training data (not available in our case)
 - set by the user (Encarta editors)

Preliminary experiments

- Impact of similarity (1-Eps) threshold (MinPts=3)



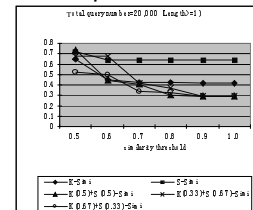
- Observations
 - at Sim=1.0, 42% (kw) and 64% (click) queries are included into 1291 and 1756 clusters
 - flat S-Simi (doc_click) curve (due to few clicks/query)
 - Seems to support FAQ and FAD

Top queries and document clicks

- Among 23 828 queries:
- $\text{Sim}_{\text{kw}} \geq 0.5$
 - world war (4 113), new X (249), X system (160),
 - G.Washington (156), tree X (104), nuclear, plant (73),
 - Egypt (66), Charles X (51), battle (50)
- $\text{Sim}_{\text{click}} \geq 0.5$
 - African American history (625) World war (424) Canada (321),
 - Agriculture (162), Birds (132), Civil war, American
 - (90) California (88) Anatomy (81), Mexico (80),
 - Arizona, pollution (73) American literature, Poetry (69)
 - England (57), Argentina (53), Atomic bomb (53),
 - Fish (51), Shakespeare (50)

Experiments (cont'd)

- Impact of α and β



- Different behaviors between 0.5 and 0.9

Ongoing experiments

- clustering quality
 - How good is a cluster? (precision)
 - What is the coverage of the clusters? (recall)
- requires desired results
- Rough evaluation:
 - combinations of keywords and clicks seem to be better than keywords only.

Example

- Keywords alone

Cluster 1:	thermodynamics	
Cluster 2:	law of thermodynamics	polygamy law
	family law	ohm's law watt's law
- Combining keywords and clicks

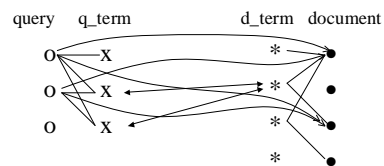
Cluster 1:	thermodynamics	law of thermodynamics
Cluster 2:	polygamy law	family law
Cluster 3:	ohm's law	watt's law

The status of the tool

- Encarta editors are using it to discover FAQs.
- 7 000 potential FAQs have been located among several millions
- eulogistic feedback from them

Extensions

- Construct a “live thesaurus”



Similar to translation problem (statistical model?)
 - authors and users speak different “languages”

Using “live thesaurus” for query expansion
 ≡ Finding answers according to previous user interactions

Extensions (cont'd)

- More generally, connections
 - among documents (inter-doc. connections)
 - among queries (inter-query connections)
 - between documents and queries (cross d&q)
- document (query) clustering using query (document) connections through d-q connections
- Derive new query-document connections (find new answers): usage-based retrieval

Conclusions

- User-logs (user interactions) provide useful information that are complementary to keyword and link-based similarity
- data available
- to be further exploited
- Ongoing projects
 - evaluation of clustering quality
 - live thesaurus from larger data set

Thanks