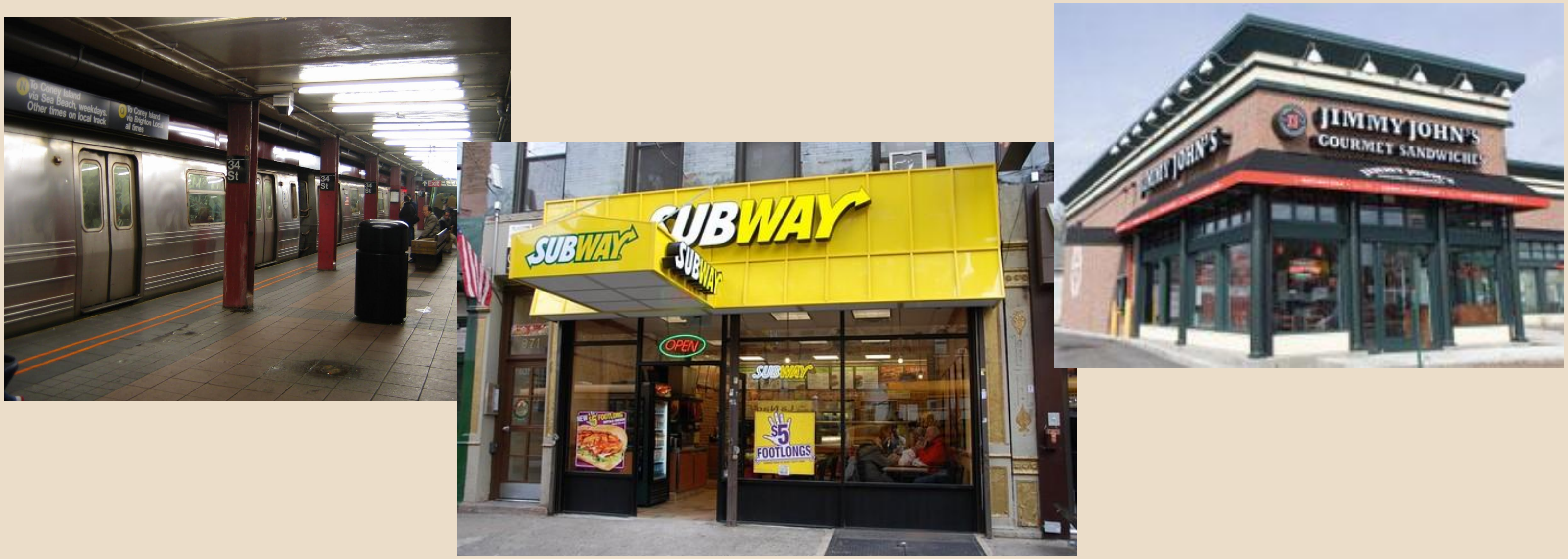
Collaborative Querying & Visualizing to improve on standard Information Retrieval

David Crowe

Improving Information Retrieval (IR)

- Problem:
  - missing/unhelpful results

- Suppose a user searches for "Subway"...


To Coney Island via Sea Beach weekdays. Other times on local track
To Coney Island via Brighton Local all times
34 St
SUBWAY
OPEN
$5 FOOTLONGS
JIMMY JOHN'S
GOURMET SANDWICHES

Improving Information Retrieval (IR)

- Problem:
  - missing/unhelpful results

- Causes:
  - users unfamiliar with Information Retrieval(IR) ops
  - vocab mismatches
  - context/semantics

- To Solve it:
  - Use past queries to identify strongly-related queries
  - Show these similar queries, let user explore/learn

Categorizing Similarity

- Term-Based: (query = "bag of terms")

  - query attributes ONLY *(Query Term Vectors)*

- Result-Based: (query = "result of executing")

  - result attributes ONLY *(Term Vectors, URLs)*

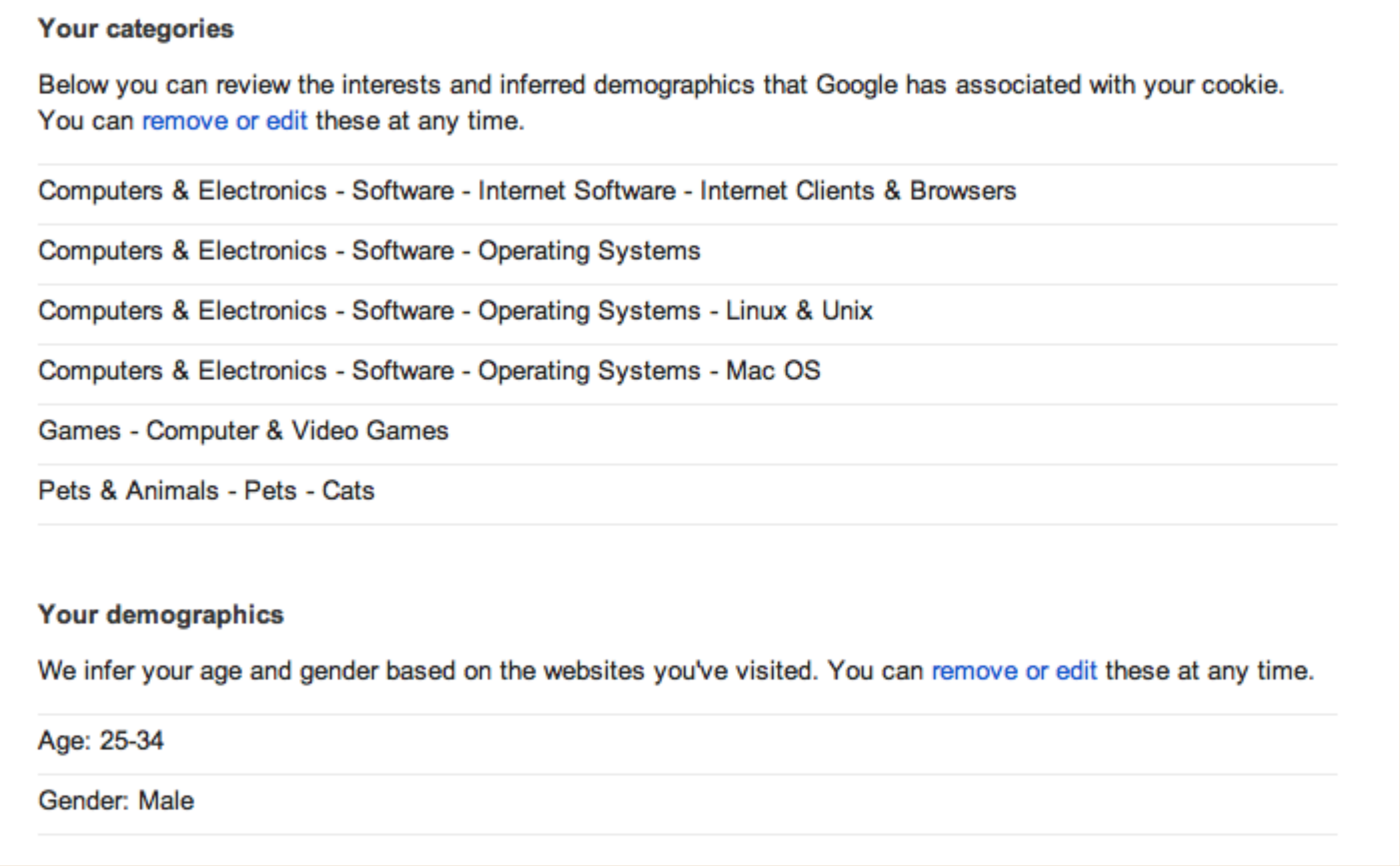
- Feedback-Based: (query = "*relevant* results")

  - result attributes AND clickthrough-data

- Community-Based: (consider interests of the user)

  - subsets of queries by user affiliation

Google's Search & Ad Data



Your categories

Below you can review the interests and inferred demographics that Google has associated with your cookie. You can remove or edit these at any time.

Computers & Electronics - Software - Internet Software - Internet Clients & Browsers

Computers & Electronics - Software - Operating Systems

Computers & Electronics - Software - Operating Systems - Linux & Unix

Computers & Electronics - Software - Operating Systems - Mac OS

Games - Computer & Video Games

Pets & Animals - Pets - Cats

Your demographics

We infer your age and gender based on the websites you've visited. You can remove or edit these at any time.

Age: 25-34

Gender: Male

Supposed to be at:
https://www.google.com/settings/ads/onweb/

How Query Graph Visualization does similarity + clustering

- To avoid term/result-based drawbacks, QGV defines 'hybrid_similarity' for queries Qi,Qj (ALPHA+BETA=1):
    - hybrid_similarity(Qi,Qj) =

      ALPHA*result_similarity(Qi,Qj) + BETA*term_similarity(Qi,Qj)
- A cluster on some node Qi is a list of nodes Qj that are similar by more than 'some number' THRESHOLD:
    - hybrid_similarity(Qi,Qj) ≥ THRESHOLD

- The team found the best result given when:
    - ALPHA = 0.75 | BETA = 0.25 | THRESHOLD = 0.9

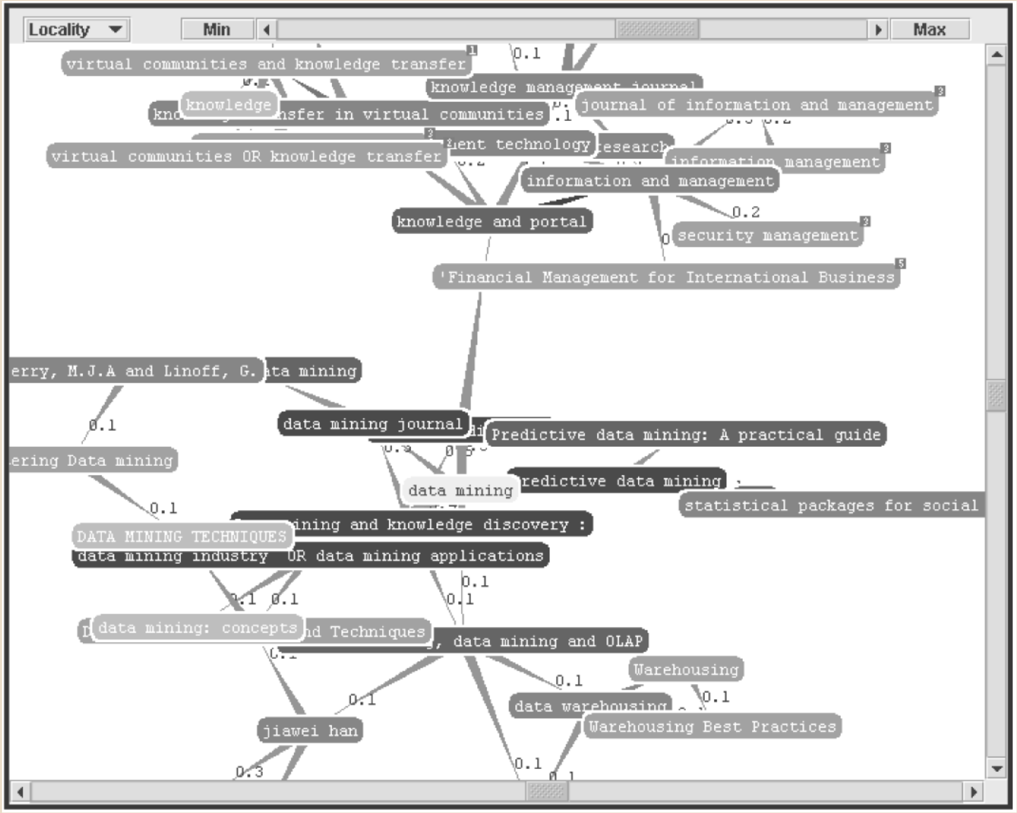
The Query Graph (QGV)

Functionality:

- ○ Generate clusters to form a Query Network graph
- ○ Allow users to explore the graph visually

Visualization:

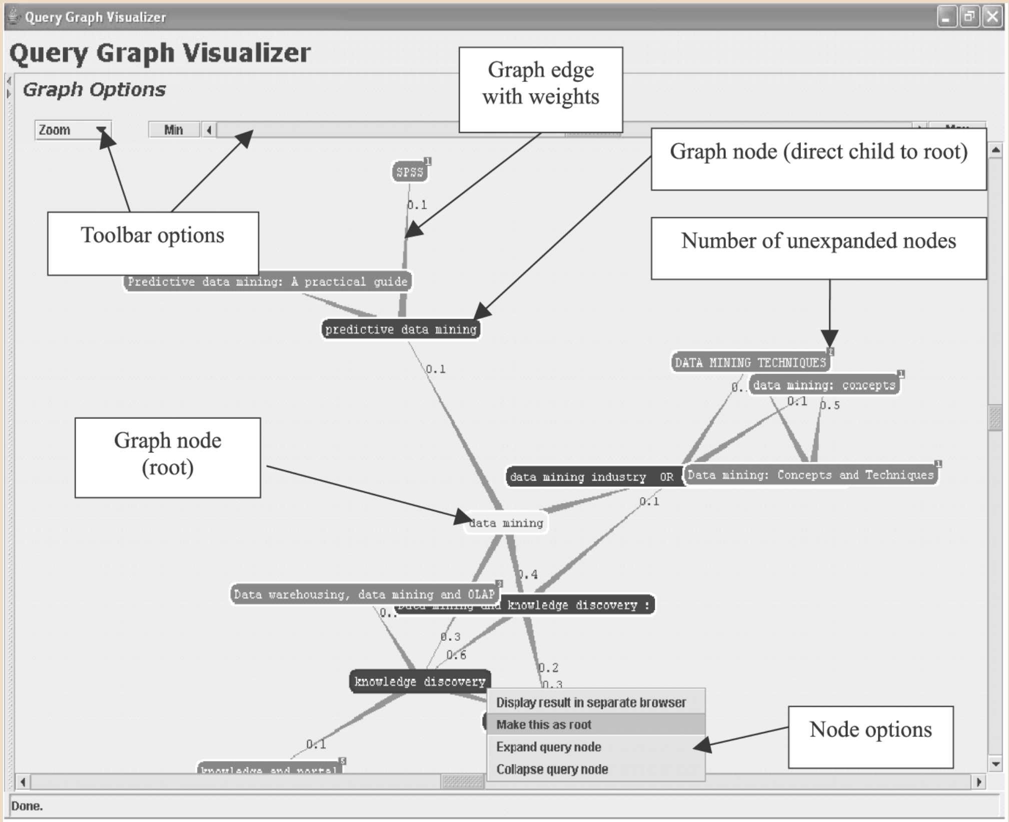
- ○ cluster = directly connected
- ○ root query = white
- ○ depth from root = lightness
- ○ similarity = edge coefficient


Locality
Min
Max
virtual communities and knowledge transfer
0.1
knowledge management journal
knowledge
kno sfer in virtual communities
journal of information and management
ent technology
esearch
virtual communities OR knowledge transfer
information management
information and management
knowledge and portal
0.2
security management
'Financial Management for International Business
erry, M.J.A and Linoff, G.
ata mining
0.1
data mining journal
Predictive data mining: A practical guide
ering Data mining
data mining
redictive data mining
statistical packages for social
0.1
ining and knowledge discovery :
DATA MINING TECHNIQUES
data mining industry OR data mining applications
0.1
0.1
0.1
0.1
data mining: concepts
nd Techniques
, data mining and OLAP
Warehousing
0.1
0.1
data warehousing
0.1
Warehousing Best Practices
jiawei han
0.3
0.1

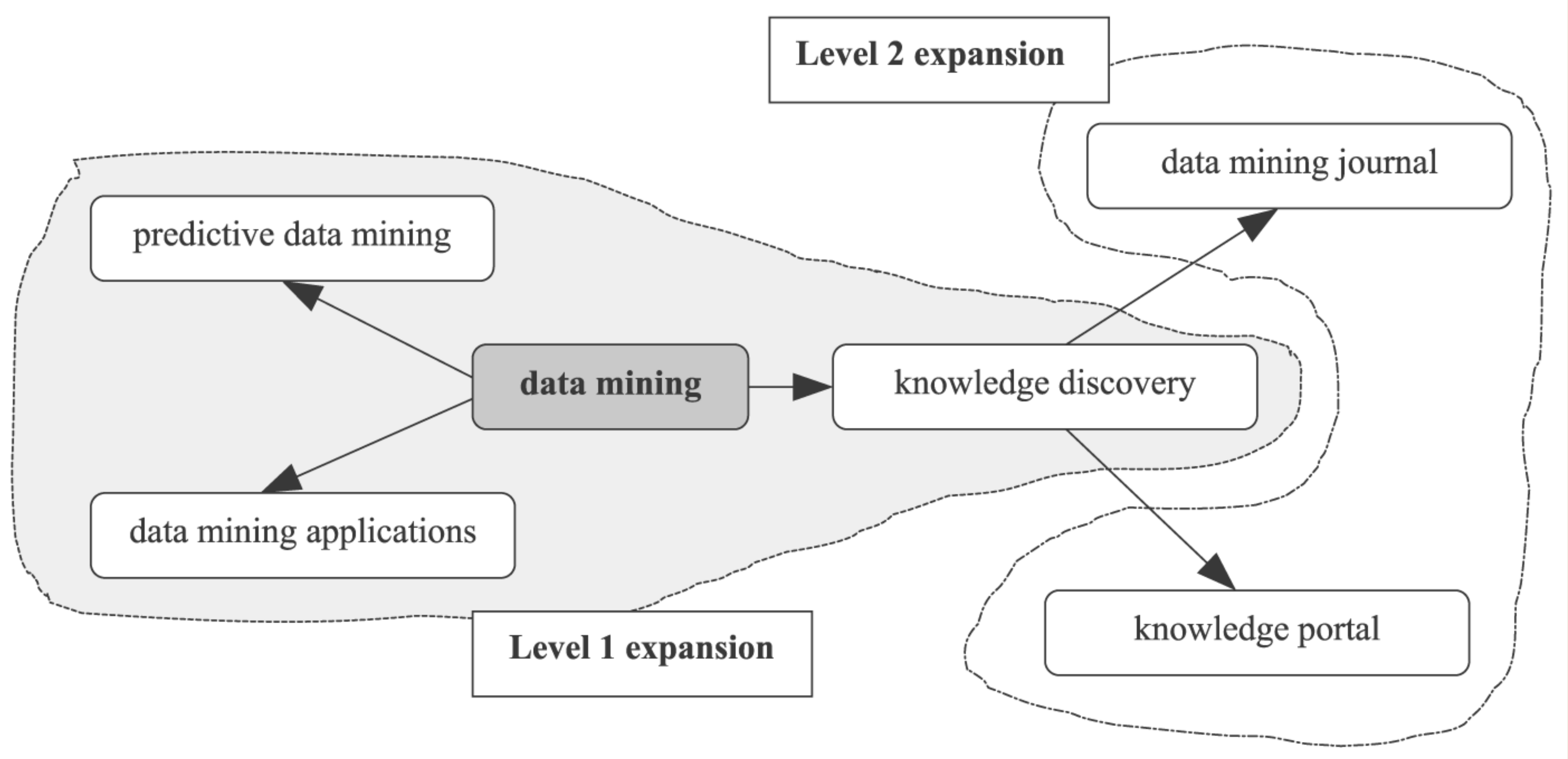
Navigating and Searching

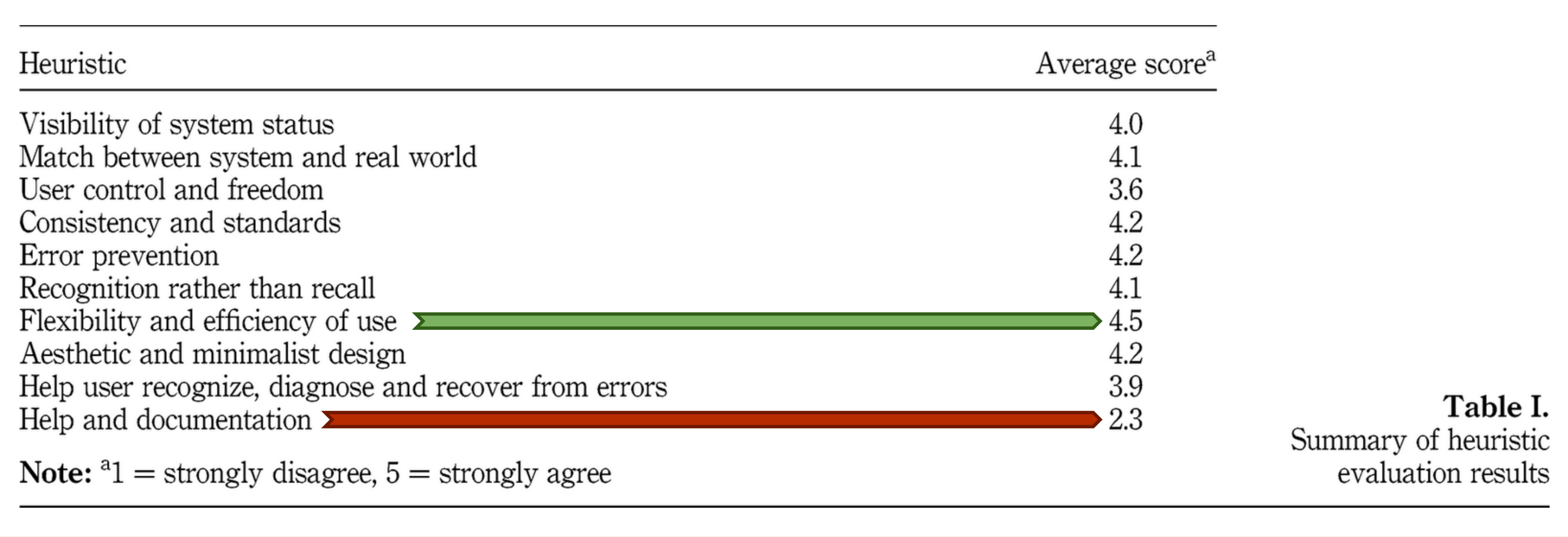
- Toolbar:
  - zoom: shrink/grow view
  - rotate: view from different directions
  - locality zooming: set network depth to draw

- Node Controls (popup):
  - Search via outside IR provider
  - Make Root Node
  - Expand/Collapse



QGV Displaying Clusters (Visual)





Heuristic	Average score[a]
Visibility of system status	4.0
Match between system and real world	4.1
User control and freedom	3.6
Consistency and standards	4.2
Error prevention	4.2
Recognition rather than recall	4.1
Flexibility and efficiency of use	4.5
Aesthetic and minimalist design	4.2
Help user recognize, diagnose and recover from errors	3.9
Help and documentation	2.3

Note: [a]1 = strongly disagree, 5 = strongly agree

Table I.
Summary of heuristic evaluation results

Evaluation

Do you think this could be extended to SQL?

Any application to your projects?

Can we take it further?

Did they achieve their goal?

Other thoughts

- Link to paper 'Nielsen' ratings (Evaluation):

  http://dl.acm.org/citation.cfm?doid=191666.191729

  I haven't read it, but it sounded interesting.

- Optimization:
  - Given a query Q in the DB that matches some result document, replace the document with Q [document surrogate] since it is a fair description. In tests this boosts performance by almost 30%. (Billerbeck et al. (2003))