Questions, Queries, and the Uncertainty Principle

Andy Berks
Merck & Co.
Background

• Inspired by Sandy Lawson talk at Chicago ACS meeting
• “Question, query, relevant response - pick any two”
• Attributed to Jim Seals, formerly of CAS
Basic idea

- Question
- Query
- Database Response
- Query bound to database

- Question
- Query
- Database Response
- Query bound to question
Value of data in databases

Important database elements

• Data source
  – capture
  – indexing
  – internal structure
    • Fields
    • indexes

• Search engine
  – Query language or mechanism
  – Precision of retrieval - are 100% of retrieved records relevant?
  – Relevance of retrieval - are 100% of relevant records retrieved?
Database structure

2 types of databases

• Structured
  - have a secondary index (SI) that orders the records and can be searched.
  - Additional objective of SI’s is added intelligence in database. Interpretation (machine or human) of concepts that can be classified.
  - Examples: CA. WPI, Beilstein

• Unstructured
  - no index.
  - Examples: Full text journals, patents, newspapers
Assumptions

- Boolean search engine can retrieve any records containing any combination of discrete data elements
  - (generally true)
- Data quality is perfect
  - (Data quality is not perfect, but in the limit of the ideal, assume that it is)
- Indexing is perfect
  - idealized case, but not really true
Technical disclosure databases

- Unambiguous databases: Customers and Orders, Roledex files
  - Information is present or absent.
- Technical databases are complex
  - Technical articles, meeting presentations, patents, etc.
  - The concepts are complex - main points, background material, experimental details, references.
  - Secondary indexes attempt to capture complex concepts uniformly
Problem

Higher order concepts are encompassed in the secondary indexing.

Effective users must know SI policies

- hierarchy
- philosophy
- search interface
- coding systems
  - MeSH, Derwent Manual Codes, CA Lexicon

Effective queries must be structured with these points in mind by the searcher
- Naïve users will not be effective searchers of SI based systems.
- SI based systems have added value not present in unstructured databases.
Human users have a question in mind, want a response of relevant records
Users may not know how to formulate a query for a database search system
Users may not have a question addressed by database indices
Query tightly bound to question
User asks a question that makes great sense to people, but no sense to the database. SI does not address the question. Response will be irrelevant to question (orthogonal).

Query tightly bound to database
User asks a question in the format and interface optimized to the database. The question is addressed by the SI. Response will be perfect - 100% relevance and 100% precision (ideal case). This constrains and taints the kinds of questions that can be asked.
In microscopic world described by quantum mechanics, human scale intuition fails. Classic Heisenberg uncertainty is that momentum and position cannot be simultaneously determined.

\[ \Delta p \geq \frac{h}{4\pi \Delta q} \quad \text{or} \quad \Delta p \cdot \Delta q \geq \frac{h}{4\pi} \]

A change in position \( p \) is inversely proportional to a change in momentum \( q \). The more precise the measurement of \( p \), the less precise the knowledge of \( q \).

Major implication: act of measuring one or the other variable changes the experiment

Also: human intuition fails
Uncertainty in the Macroscopic World

• In everyday life, many things have orthogonal relationships
  - Two or activities or concepts that occupy the same space and time but are different
  - Concentrating on one element distorts the space time field and adversely affects the second or $n$th element.

• Example: Driving a car and talking on a cell phone
  - Driving a car is generally safe
  - Talking on a cell phone is generally safe
  - Doing both at the same time is generally unsafe.
Uncertainty in databases

Change in scale affects human intuition. Consider a file (like a rolodex) with 100 cards. You can memorize every card there. Now consider the file getting larger. If the file grows to thousands and millions of records, you have no hope of personally understanding each record. The user is at the mercy of the search engine, indexes, quality of data. 100 cards is human scale. 1,000,000 cards is not human scale.

Human intuition is that any card can be retrieved at any time with known criteria (eg, alphabetical name order). But with a large database, the cards need to be converted to machine readable records. Users become dependent on a search interface to retrieve the cards.
Higher order concepts

Consider a database more complex than an address list, like records of journal articles or patents. The higher order concepts are the essential elements of the document, the main points. Database builders attempt to extract the main points into keywords, classification schemes, etc. (SI systems). When a user asks a question that is not addressed by SI, complexity in the query is required. Complexity is defined here as any question that is not directly addressed in SI’s.
Database complementarity

Proposal: in database queries, an uncertainty relationship exists. The complementary pair is question q and relevance r.

\[
\Delta r \geq \frac{V}{\Delta q}
\]

As the relevance of the response r increases, the relevance of the response to a question q decreases. The variable V changes with the relationship of the question to the indexes of the database. If the question perfectly matches an index, V will be large and \( \Delta r \) will be large. If the question does not match the indices, then V will be small, and \( \Delta r \) will be small.

Ideal search: large V (100%), large r (100%).
Relevance vs. Precision

Relevance and precision are inversely related:

\[ p = \frac{Z}{r} \]

where \( r \)=relevance
\( p \)=precision
\( Z \)=index quality (I.e., relevant to question). High \( Z \) means SI is more relevant; Low \( Z \) means SI is less relevant or no SI is present.

- SI system: precision \( \uparrow \), relevance \( \approx \downarrow \)
- Unindexed system: precision \( \downarrow \), relevance \( \approx \uparrow \)
Sandy Lawson’s solution to the Q-Q-RR problem was to develop a natural language interface.

- Natural language interface smoothes and balances the equation.
  - V declines relative to an ideal case.
  - Query formulation becomes trivial - users are insulated from the underlying database structure. Users depend on the search engine ability to parse a question and find relevant answers.
  - The cost is an inability to take full advantage of SI, so relevance and precision will decline. Irrelevant hits will increase, and retrieval precision will suffer relative to the ideal case.
  - Applying a natural language engine to a structured db will make it more unstructured like.
Unstructured databases

- Retrieval can’t distinguish between main points, tangential references, and background material.
- The uncertainty eqn is pushed to the right (smaller $r$) because there is no SI. $V$ becomes small.
- The precision eqn is pushed to the right. At relatively constant $r$, precision decreases (smaller $p$)
Conclusion

Effective use of structured db’s constrains the questions that can be asked.

Implications

Its impossible to build a database with indices to answer all possible questions (V approaches the ideal limit of 100%)

Its impossible to perform a search on a complex question with perfect retrieval and precision. The uncertainty principle gets in the way.

A perfect search is impossible
Shameless plug for mediated searches

• Full text, natural language, or other simplified user interfaces are useful, but for important decisions, decision makers cannot be expected to be database experts or achieve the full benefit of SI systems with value added indexing.

• Experts performing mediated searches are needed to provide confidence in retrieval and ensure that data and references behind decisions are the best possible.

• Experts interface between naïve but important questions and the best SI features of value added databases to maximize the overall information retrieval experience.
Author contact info

Andrew H. Berks, Ph.D.
Sr. Information Scientist
Merck & Co. Patent Dept. RY60-35
126 E. Lincoln Ave
Rahway NJ 07065 USA
732-594-1701 fax: 732-594-5832
andrew_berks@merck.com