Query Formulation Notes   Ryen W. White (ryen@umd.edu)   February 21, 2005

Overview
- Information Need
- Query Formulation
- Information Need Change

---

1. Information Need

A gap between what we know and what we want to know that motivates the search.

Taylor
- Discusses the "area of the question"
  - Generation, Information Retrieval (IR) system, user
- User's generate questions, convert them to queries and pose them to IR systems
  - "Indication of inadequacy"
  - Syntactic matching (too much emphasis)
- Question was ignored in development of early IR systems
- Not retrieving "information", we retrieve "stuff" (e.g., books, reports, papers)
  - Queries presenting to retrieve things not the information we want
  - What we want is related to what we expect to get
- Answers may not imply information need satisfaction
  - Perhaps only change in information need
  - IR systems do not handle continuity
    - Each submitted query independent from previous
    - User and system designers forced to think this way
- Four levels of question formation
  - Information need distinct/traceable
  - Ignores contextual factors
  - Progression towards concrete question
  - Visceral (what they want to know)
    - "ideal" question
  - Conscious (what they think they want to know)
    - Question as perceived by user
  - Formalized (e.g., TREC topics)
    - **We want IR system to answer this**
  - Compromised (e.g., TREC queries)
    - What the IR to be able to answer
    - What we understand the IR system is capable of
- Questions posed affected by many factors
  - User responsibility, distance to system, response time
  - Document type
  - Classification, indexing, depth, term usage
  - User role, definition of success (precision/recall)

- o Form of the output, not worth trouble
- Single correct answer/set of answers (acceptability)
  - o Multiple interpretations of question, indecision
  - o Different amounts of acceptability for different levels of question
  - o Role of question type
  - o Pain and trouble of getting relevant information
- Readiness
  - o "State of mind", allow selection of information
  - o Constantly changing, relevant/irrelevant information changes state

Belkin
- Information needs not precisely specifiable
- Anomalous State of Knowledge (ASK) underlies information need (visceral)
  - o ASK can be derived by eliciting problem statements
  - o Classes of ASK
  - o IR systems should be built based on these ASKs
- Problem drives IR situation
  - o Change with interaction with the system
  - o Situational requirements of response (e.g., info sources, info modes)
- Fundamental element in IR is development of information need from ASK
- THOMAS: IR without query formulation
- "best-match" principle not representative of reality
  - o Need IR techniques that take better account of the reality of IR
    - ▪ Represent user anomalies
    - ▪ Evaluation w.r.t. user problem
    - ▪ Iteration and Interaction in retrieval (e.g., Relevance Feedback)
- Represent state of knowledge as networks derived from association data
  - o No real/true representation of knowledge/representation
- Represent ASKs as narrative statements of problematic situation
- Use different retrieval strategy for different types of ASK

---

2. Query Formulation

- *Initial query (entered by user)*
  - o *Generally short (around 2 terms)*
    - ▪ *Ranked retrieval likes long queries, Boolean retrieval does not*
  - o *Straight from "Conscious need", bypass "Formalized need"*
  - o *Users perceive what system is capable of*
    - ▪ *What words should be used, query syntax*
  - o *Help them enter longer queries (larger query input boxes), describe what information is expected, provide examples, related terms*

- *Query reformulation (refine original query)*
  - o *Change query after initial request*
  - o *Use initial tentative query to better understand retrieval system*

- *Query-by-example ("More Like This")*
  - *Show systems examples of relevant information*
  - *Reinforcement learning problem*
    - *What features should the system learn? How should these be interpreted?*

- *Relevance Feedback (provide good/bad examples to improve query)*
  - *Users provide examples of good or bad documents*
  - *Initial query refined based on documents marked*

- *Users are reluctant to do anything except enter short queries and click Search!*
  - *How does this relate to your own experiences? What do you want done?*

Ruthven and Lalmas

- Users can have problems expressing needs (in queries)
  - Can still identify which documents are relevant (Relevance Feedback)
  - Process is physically and cognitively burdensome
- IR systems find information for which there is no definite answer
  - IR process is uncertain
    - Users express compromised information needs
    - Documents are represented as a "bag of words", no inter-word relationships
- Retrieval
  - Matching between query and document representation
  - Models
    - Boolean, Vector space, Probabilistic, Logical
- Evaluation
  - Measures (precision/recall)
  - Techniques (residual ranking, freezing, test and control groups)
- Query expansion versus Term reweighting
- RF extensions
  - Term dependence (relationship between terms e.g., in phrases)
  - Dynamism of information needs (adapt to changing needs)
  - Negative RF (items marked as not relevant and items that the user has not marked – RF systems assume unmarked items are non-relevant)
  - Combining evidence (multiple query representations, retrieval algorithms, RF techniques)
  - Implicit RF (RF captured implicitly from user interaction)
- Automatic/Interactive/Manual query modification
- Interfaces
  - Incremental RF, Ostensive browsing
- Users and their role in RF (characteristics, experience, provision of RF)

3. Information need change

- *Information needs are not fixed or static*
  - *Change in exposure to new information*
    - *User can reach new level of understanding (or confusion!)*
  - *Change can be slight or can be dramatic*
    - *Slight change e.g., seeing a document and becoming more sure of what your are looking for → keep following same path*
    - *Dramatic change e.g., seeing a document and realizing you've been mistaken → change of tack*

Baldonado and Winograd

- SenseMaker
- Information exploration tasks
- Heterogeneous sources
- Support evolution of user's information needs through contextual information
  - Approximate current context through selected references
  - Set of interactions within context, change in context
- New information, different types of information, change in characteristics of context
  - New query formulations/information seeking contexts
  - Information foraging/new information patches