

Text Technologies

Retrieval Models

Victor Lavrenko

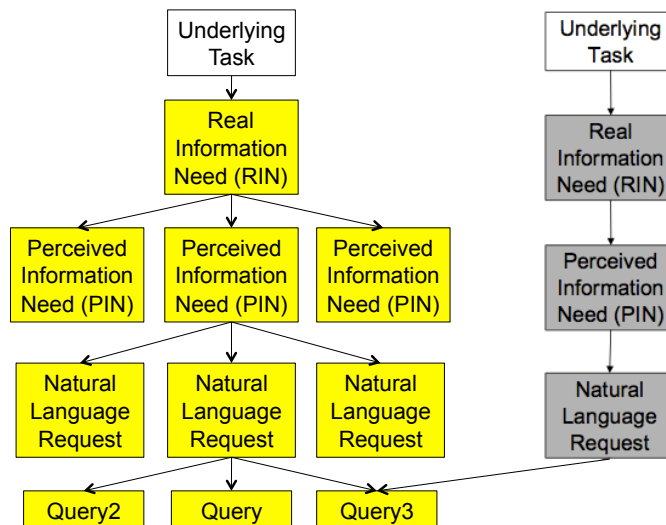
Copyright © 2011 Victor Lavrenko

Queries and Information Needs

- Information need
 - underlying cause for the search
 - related to a specific task the user needs to do
- Relevance
 - does the document satisfy the information need?
 - does it help the user solve their task effectively?
- Query:
 - one of the many expression of an information need
 - can be an expression of many different information needs
 - we never observe information needs directly
 - have to infer them from the observed query

Copyright © 2011 Victor Lavrenko

From information needs to queries



Copyright © 2011 Victor Lavrenko

Information needs: example

- Underlying task: crawling news
- Real information need:
 - a method for extracting content from web-pages
- Perceived information need:
 - a package that “knows” all the ways to mark up content and navigation sections in a webpage
- Natural language request:
 - “I need a good python library for parsing webpages”
- Query:
 - “python parser”

Copyright © 2011 Victor Lavrenko

Retrieval Models

- Mathematical formalism for the following processes
 - formulation: information need → query + refinement
 - indexing: documents → index terms
 - **retrieval: query + corpus → search results**
- Variables involved
 - documents (D), queries (Q), relevance $R(Q,D) \rightarrow \{0,1\}$
 - sometimes: user, task, context, display, search history, ...
- Usually involve an abstract analogy
 - document is an urn containing words
 - query is a logical formula that needs to be proved
 - user is a greedy memory-less stochastic process

Copyright © 2011 Victor Lavrenko

Types of Retrieval Models

- Why do we need models?
 - same reason there are theories in physics
 - explicit set of assumptions that can be compared, tested
 - a guide for developing new retrieval algorithms
- Exact-match models
 - query: precise set of match / non-match criteria
 - result = set of documents
 - dominant in the past, still used by professional users
- Best-match models
 - query: describes a good match
 - result = ranking of all documents in the corpus

Copyright © 2011 Victor Lavrenko

Prominent Retrieval Models

- Vector-space model
 - queries and documents are points in a vector-space
- Boolean retrieval (exact-match)
 - query is a logical formula, document can satisfy it
- Probabilistic model
 - IR as probabilistic classifier: $P(\text{relevance}|\text{document})$
- Inference-network model
 - document: source of evidence, query: belief network
- Language modeling
 - document is an urn; query is a sample drawn from an urn

Copyright © 2011 Victor Lavrenko

Boolean Retrieval Model

- Query: logical expression involving document features
 - example: **cryogenic AND (lab OR labs) AND-NOT cartoon**
 - systems also have proximity operators, reg-exp for spelling
- Exact-match: document either satisfies query or not
 - ranking by external criterion: date, PageRank

Pros

- very efficient (optimizations)
- predictable, easy to explain
- precisely pinpoint info. need
- best when you know domain

Cons

- users: hard to formulate queries
- no easy way to get R/P tradeoff
- consistently worse than ranking

Copyright © 2011 Victor Lavrenko

Boolean query development

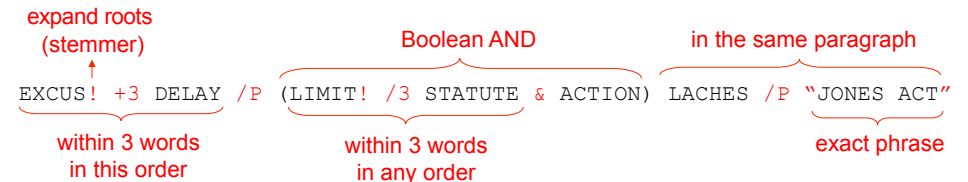
- Interactive process, driven by number of results
 - lincoln
 - "president lincoln"
 - president AND lincoln
 - president AND lincoln AND NOT transport
 - president AND lincoln AND NOT (car OR automobile)
 - president AND lincoln AND biography AND birthplace AND NOT (car OR automobile)
 - president AND lincoln AND (biography OR birthplace) AND NOT (cat OR automobile)
- successful if results are instant (cf. Google Instant)
 - users love control, feel empowered to interact

Copyright © 2011 Victor Lavrenko

Example: WESTLAW ® (legal)

- Part of Thompson-Reuters, operates since 1974
 - legal / financial market, 700k professional (*paying*) users
 - 47% of all legal searches, 8.4m transactions per day
 - 50TB in 17k databases, 10³ data centres in 51 countries
- Boolean engine, example query:

What cases have discussed the concept of excusable delay in the application of statutes of limitations or the doctrine of laches involving actions in admiralty or under the Jones Act?



Copyright © 2011 Victor Lavrenko

Example: Ovid ® (medical)

Database of medical studies

- 18+ million articles, public
- use of X to treat Y in cases Z

Uses Boolean Model

- restricted vocabulary (term/)
 - MeSH: Medical Subject Headings
 - automated expansion →
- Boolean query:
 - AND, OR, NOT
 - control where matches occur

MeSH Heading	Neck Pain
Tree Number	C10.597.617.576
Tree Number	C23.888.592.612.553
Tree Number	C23.888.646.501
Entry Term	Cervical Pain
Entry Term	Neckache
Entry Term	Anterior Cervical Pain
Entry Term	Anterior Neck Pain
Entry Term	Cervicalgia
Entry Term	Cervicodynia
Entry Term	Neck Ache
Entry Term	Posterior Cervical Pain
Entry Term	Posterior Neck Pain

Fish oils for asthma in children

- 1. fatty acids, omega-3/ or alpha-linolenic acid/ or docosahexaenoic acids/ or eicosapentaenoic acid/
- 2. fatty acids, omega-6/ or gamma-linolenic acid/ or linoleic acids/
- 3. Fish Oils/
- 4. exp Plant Oils/
- 5. omega oils.mp.
- 6. omega.mp.
- 7. 1 or 2 or 3 or 4 or 6
- 8. Hypersensitivity, Immediate/dh, pc, dt [Diet Therapy, Prevention & Control, Drug Therapy]
- 9. Asthma/pc, dh, dt [Prevention & Control, Diet Therapy, Drug Therapy]
- 10. Dermatitis, Atopic/pc, dh, dt [Prevention & Control, Diet Therapy, Drug Therapy]
- 11. Immunoglobulin E/
- 12. Rhinitis, Allergic, Perennial/dh, pc, dt [Diet Therapy, Prevention & Control, Drug Therapy]
- 13. Rhinitis, Allergic, Seasonal/dh, dt, pc [Diet Therapy, Drug Therapy, Prevention & Control]
- 14. 8 or 9 or 10 or 11 or 12 or 13
- 15. Pregnancy/
- 16. Infant, Newborn/
- 17. Primary Prevention/
- 18. Child, Preschool/
- 19. 15 or 16 or 17 or 18

Queries: collaborative, reusable

Copyright © 2011 Victor Lavrenko

Summary: Boolean model

- Boolean retrieval model: does D satisfy Q?
 - Q: logical formula based on occurrence of words, etc.
 - EXCUS! +3 DELAY /P (LIMIT! /3 STATUTE & ACTION) /P "JONES ACT"
 - D: set of predicates that either match Q, or not
 - efficient, explainable, based on interactive query formulation
- Deficiencies:
 - must know query language, database contents, terminology
 - no ranking → no recall / precision tradeoff
 - Boolean operators difficult to work with
 - AND constraint → no matches
 - OR constraint → millions of matches

Copyright © 2011 Victor Lavrenko