Chapter IR:VI

VI. Users and Queries

- ❑ Information Needs and Queries
- ❑ Query Transformation and Refinement
- ❑ Cross-Language Search

Query Transformation and Refinement

Query transformation

- In general, same operations on text as on query
- Some differences in stopping and stemming
- Some transformations not needed
  - Tokenization
  - Structure analysis (queries have no title, etc.)

Query Transformation and Refinement

Query-Based Stopping

- ❑ Stopword removal at query time
- ❑ Retain stopwords in index
  - – Flexibility to deal with queries that contain stopwords
- ❑ Stopwords in query can be
  - – Treated as normal words
  - – Removed
  - – Conditionally removed (not if prefixed by +)

Query Transformation and Refinement

Query-Based Stemming

- ❑ Make decision about stemming at query time rather than during indexing
  - – Improved flexibility and effectiveness
- ❑ Query is expanded using word variants
  - – Documents are not stemmed
  - – Thus, e.g., query `rock climbing` automatically expanded with `climb`
    - • But not stemmed to `climb`
- ❑ Alternative: Index word and its stem
  - – Increased efficiency
  - – But larger index

Query Transformation and Refinement

Stem Classes

- A stem class is the group of words that will be transformed into the same stem by the stemming algorithm
  - Generated by running stemmer on large corpus
  - E.g., Porter stemmer on TREC News collection might produce

    /bank banked banking bankings banks
    /ocean oceaneering oceanic oceanics oceanization oceans

    /polic polical polically police policeable policed policement policer policers polices policial policially policier policiers policies policing policization policize policly policy policying policys
  - Quite long classes – adds many words to query
  - Contain some errors (cf. the section on stemming texts)

Query Transformation and Refinement

Stem Classes

- Can be used for stemming or for expansion
  - Can drift to incorrect topics (banking → bank)
- Stem classes are often too big and inaccurate
- Modify using analysis of word co-occurrence
- Assumption:
  - Word variants that could substitute for each other should co-occur often in documents

Query Transformation and Refinement

Modifying Stem Classes

1. For all pairs of words in the stem classes, count how often they co-occur in text windows of k words, where k is typically in the range 50–100.
2. Compute a co-occurrence or association metric for each pair. This measures how strong the association is between the words.
3. Construct a graph where the vertices represent words and the edges are between words whose co-occurrence metric is above a threshold τ (set empirically).
4. Find the connected components of this graph. These are the new stem classes.

Query Transformation and Refinement

Modifying Stem Classes

- Dice's Coefficient is an example of a term association measure between terms a and b:
  - $2 \cdot n_{ab}/(n_a + n_b)$
  - where n_x is the number of windows containing the terms in x
- Proportion of term occurrences that are co-occurrences
- Remember: Two vertices are in the same connected component of a graph if there is a path between them
  - Forms word clusters
- Example output of modification

  /bank banking banks
  /policies policy

  /police policed policing

Query Transformation and Refinement

Spell Checking

- Important part of query processing
  - 10–15% of all web queries have spelling errors
  - Reliance on “Did you mean . . . ”
- Almost 600 ways to (mis-)spell Britney Spears were recognized at Google within 3 months in the early 2000’s (http://www.google.com/jobs/britney.html)

488941	britney spears	364	britey spears	109	brittant spears	54	britnye spears	29	britent spears	21	bratney spears
40134	brittany spears	364	brittiny spears	98	bittney spears	54	britt spears	29	brittnany spears	21	britani spears
36315	brittney spears	329	brtney spears	98	brithey spears	54	brttany spears	29	britttany spears	21	britanie spears
24342	britany spears	269	bretney spears	98	brittiany spears	48	bitany spears	29	btiney spears	21	briteany spears
7331	britny spears	269	britneys spears	98	btitney spears	48	briny spears	26	birttney spears	21	brittay spears
6633	briteny spears	244	britne spears	89	brietny spears	48	brirney spears	26	breitney spears	21	brittinay spears
2696	britteny spears	244	brytney spears	89	brinety spears	48	britant spears	26	brinity spears	21	brtany spears
1807	briney spears	220	breatney spears	89	brintny spears	48	britnety spears	26	britenay spears	21	brtiany spears
1635	brittny spears	220	britiany spears	89	britnie spears	48	brittanny spears	26	britneyt spears	19	birney spears
1479	brintey spears	199	britnney spears	89	brittey spears	48	brttney spears	26	brittan spears	19	brirtney spears
1479	britanny spears	163	britnry spears	89	brittnet spears	44	birttany spears	26	brittne spears	19	britnaey spears
1338	britiny spears	147	breatny spears	89	brity spears	44	brittani spears	26	btittany spears	19	britnee spears
1211	britnet spears	147	brittiney spears	89	ritney spears	44	brityney spears	24	beitney spears	19	britony spears
1096	britiney spears	147	britty spears	80	bretny spears	44	brtitney spears	24	birteny spears	19	brittanty spears
991	britaney spears	147	brotney spears	80	britnany spears	39	brienty spears	24	brightney spears	19	britttney spears
991	britnay spears	147	brutney spears	73	brinteny spears	39	brritney spears	24	brintiny spears	17	birtny spears
811	brithney spears	133	britteney spears	73	brittainy spears	36	bbritney spears	24	britanty spears	17	brieny spears
811	brtiney spears	133	briyney spears	73	pritney spears	36	briitany spears	24	britenny spears	17	brintty spears
664	birtney spears	121	bittany spears	66	brintany spears	36	britanney spears	24	britini spears	17	brithy spears
664	brintney spears	121	bridney spears	66	britnery spears	36	briterny spears	24	britnwy spears	17	brittanie spears
664	briteney spears	121	britainy spears	59	briitney spears	36	britneey spears	24	brittni spears	15	brinney spears
601	bitney spears	121	britmey spears	59	britinay spears	36	britnei spears	24	brittnie spears	15	briten spears
601	brinty spears	109	brietney spears	54	britneay spears	36	britniy spears	21	biritney spears	15	briterney spears
544	brittaney spears	109	brithny spears	54	britner spears	32	britbey spears	21	birtany spears	15	britheny spears
544	brittnay spears	109	britni spears	54	britney’s spears	32	britneu spears	21	biteny spears		. . .

Query Transformation and Refinement

Spell Checking

- Errors include typical word processing errors

  poiner sisters
  brimingham news
  catamarn sailing
  hair extenssions

  marshmellow world
  miniture golf courses
  psyhics
  home doceration

- But also many other types (for which terms and corrections cannot be found in common dictionaries)

  realstateisting.bc.com
  akia 1080i manunal
  ultimatwarcade

  mainscourcebank
  delottitouche

Query Transformation and Refinement

Spell Checking

- Basic approach: Suggest corrections for words not found in spelling dictionary
  - But `miniature golf curses` would not be corrected
- Suggestions found by comparing word to words in dictionary using similarity measure
- Most common similarity measure is edit distance
  - Minimum number of operations required to transform one word into the other

Query Transformation and Refinement

Edit Distance

- Damerau-Levenshtein distance
  - Counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
  - The often used Levenshtein-distance does not allow transpositions
  - Damerau-Levenshtein distance 1 (80% of the spelling errors)
    - extenssions → extensions (deletion)
    - poiner → pointer (insertion)
    - marshmellow → marshmallow (substitution)
    - brimingham → birmingham (transposition)
  - Damerau-Levenshtein distance 2
    - doceration → decoration (2 substitutions via deceration)

Query Transformation and Refinement

Edit Distance

- Techniques used to speed up calculation of edit distances
  - Restrict to words starting with same character
    - Spelling errors rarely occur in first letter
  - Restrict to words of same or similar length
    - Spelling errors rarely change length of words
    - Can be safe (if length exceeds threshold)
  - Restrict to words that sound the same
- Last option uses a phonetic code to group words
  - E.g. Soundex

Query Transformation and Refinement

Soundex Code

1. Keep the first letter (in upper case).
2. Replace these letters with hyphens: a, e, i, o, u, y, h, w.
3. Replace the other letters by numbers as follows:

1	b, f, p, v	4	l
2	c, g, j, k, q, s, x, z	5	m, n
3	d, t	6	r

4. Delete adjacent repeats of a number.
5. Delete the hyphens.
6. Keep the first three numbers or pad out with zeros.

Examples:

extenssions → E235	extensions → E235
marshmellow → M625	marshmallow → M625
brimingham → B655	birmingham → B655
poiner → P560	pointer → P536

Query Transformation and Refinement

Spelling Correction Issues

- In general, many corrections possible
  - lawers → lowers, lawyers, layers, lasers, lagers, . . .
- Ranking corrections
  - “Did you mean . . . ” feature requires accurate ranking of possible corrections
  - First idea: Rank by frequency
- Better idea: Use context
  - Choosing right suggestion depends on context (other words)
  - E.g., trial lawers → trial lawyers
- Run-on errors
  - E.g., mainscourcebank
  - Missing spaces can be considered another single character error in right framework

Query Transformation and Refinement

Noisy Channel Model

- Based on Shannon's theory of communication
- User chooses word w based on probability distribution $P(w)$
  - Called the language model
  - Can capture context information, e.g., $P(w_1|w_2)$
- User writes word, but noisy channel causes word e to be written instead with probability $P(e|w)$
  - Called error model
  - Represents information about the frequency of spelling errors
  - Probabilities for words within edit-distance will be high
  - Even $P(w|w) \leq 1$
    - Thus it is possible to correct
      miniature golf curses → miniature golf courses

Query Transformation and Refinement

Noisy Channel Model

- Need to estimate probability of correction
  - $P(w|e) = P(e|w)P(w)$
  - Works if one ignores context and run-on errors
- Estimate language model using context
  - $P(w) = \lambda P(w) + (1 - \lambda)P(w|w_p)$
  - w_p is previous word
  - λ specifies relative importance of probabilities
- Example
  - fish tink
  - “tank” and “think” both likely corrections (edit distance 1)
  - Both have high $P(w)$
  - But $P(\text{tank}|\text{fish}) > P(\text{think}|\text{fish})$ which makes “tank” the more likely correction

Query Transformation and Refinement

Noisy Channel Model

- Estimate $P(w)$: Language model probabilities estimated using corpus and query log
  - Query log useful, because it matches the task
    - And has fewer word pairs
  - Dictionary can help, too
- Estimate $P(e|w)$: Both simple and complex methods have been used for estimating the error model
  - Simple approach: Assume all words with same edit distance have same probability, only edit distance 1 and 2 considered
  - More complex approach: Incorporate estimates based on common typing errors
    - Keyboard layout

Query Transformation and Refinement

Noisy Channel Model

1. Tokenize the query
2. For each token, a set of alternative words and pairs of words is found using an edit distance modified by weighting certain types of errors as described above
   - The data structure that is searched for the alternatives contains words and pairs from both the query log and the trusted dictionary
3. Use noisy channel model to select the best corrections
4. Repeat from Step 2 until no better correction is found

Example:
miniture golfcurses
miniature golf curses
miniature golf courses

Query Transformation and Refinement

Query Expansion

- Thesaurus used in early search engines as a tool for indexing and query formulation
  - Manually specified preferred terms and relationships between them
  - Also called controlled vocabulary
- Add synonyms or more specific terms using query operators based on thesaurus
  - Improves search effectiveness
- Example: MeSH thesaurus (Medical Subject Headings)

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	Anteriro Neck Pain
Entry Term	Cervicalgia
Entry Term	Cervicodynia
Entry Term	Neck Ache
Entry Term	Posterior Cervical Pain
Entry Term	Posterior Neck Pain

Query Transformation and Refinement

Query Expansion

- ❑ Expansion based on explicit thesaurus (e.g., Wordnet or MeSH) rarely used since not very effective
  - Does not take context into account
- ❑ A variety of automatic or semi-automatic query expansion techniques have been developed
  - Goal: Improve effectiveness by matching related terms
  - Semi-automatic techniques require user interaction to select best expansion terms
- ❑ Query suggestion is a related technique
  - Alternative queries, not necessarily more terms
- ❑ Approaches usually based on an analysis of term co-occurrence
  - in the entire document collection,
  - in a large collection of queries,
  - or in the top-ranked documents in a result list
- ❑ Query-based stemming also an expansion technique

Query Transformation and Refinement

Term Association Measures

- Idea: Choose appropriate words from context
  - “tropical fish tanks” → expand “tank” with “aquarium”
  - vs. “armor” for “tanks”!
- Ideas for expansion
  - Consider all words holistically, rather than expanding individual words
  - Use relevance feedback
- Term association measures
  - Dice’s coefficient
  - Mutual information
  - Pearson’s Chi-squared (χ^2) measure

Query Transformation and Refinement

Term Association Measures

- Dice's Coefficient
  - Reminder: n_x is the number of windows containing terms in x

$$\frac{2n_{ab}}{n_a + n_b} \overset{\text{rank}}{=} \frac{n_{ab}}{n_a + n_b}$$

  - Rank equivalence: Produces same ranking / ordering
- Mutual Information

$$\log \frac{P(a,b)}{P(a)P(b)}$$

  - Measures extent to which words occur independently
  - Independent words: $P(a,b) = P(a)P(b) \Rightarrow$ mutual information = 0
  - Estimate $P(A) = n_A/N$

$$\log \frac{P(a,b)}{P(a)P(b)} = \log \left(N \frac{n_{ab}}{n_a n_b} \right) \overset{\text{rank}}{=} \frac{n_{ab}}{n_a n_b}$$

Query Transformation and Refinement

Term Association Measures

- Mutual Information Measure (MI) favors low frequency terms
  - Example: $n_a = n_b = 10$ and $n_{ab} = 5$ ⇒ 5/100
  - Example: $n_a = n_b = 1000$ and $n_{ab} = 500$ ⇒ 5/10000

- Expected Mutual Information Measure (EMI)
  - Weighting of MI with $P(a,b)$

$$P(a,b) \cdot \log \frac{P(a,b)}{P(a)P(b)} \quad = \quad \frac{n_{ab}}{N} \log \left(N \frac{n_{ab}}{n_a n_b} \right) \quad \overset{\text{rank}}{=} \quad n_{ab} \cdot \log \left(N \frac{n_{ab}}{n_a n_b} \right)$$

  - Previous example with N = 1 million: 23.5 vs. 1350
  - Problem: favors high-frequency terms

Query Transformation and Refinement

Term Association Measures

- Pearson's Chi-squared (χ^2) measure
  - Compares the number of co-occurrences of two words with the expected number of co-occurrences if the two words were independent

$$n_{ab} - N \cdot \frac{n_a}{N} \cdot \frac{n_b}{N}$$

  - Normalizes this comparison by the expected number

$$\frac{\left(n_{ab} - N \cdot \frac{n_a}{N} \cdot \frac{n_b}{N}\right)^2}{N \cdot \frac{n_a}{N} \cdot \frac{n_b}{N}} \overset{\text{rank}}{=} \frac{\left(n_{ab} - \frac{1}{N} n_a n_b\right)^2}{n_a n_b}$$

Query Transformation and Refinement

Term Association Measures

Measure	Acronym	Ranking
Dice's coefficient	Dice	$\frac{n_{ab}}{n_a+n_b}$
Mutual information	MI	$\frac{n_{ab}}{n_a n_b}$
Expected mutual information	EMI	$n_{ab} \cdot \log\left(N \frac{n_{ab}}{n_a n_b}\right)$
Chi-square	χ^2	$\frac{\left(n_{ab}-\frac{1}{N}n_a n_b\right)^2}{n_a n_b}$

Query Transformation and Refinement

Association Measure Example

Most strongly associated words for “tropical” in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

Dice	MI	EMI	χ^2
forest	trmm	forest	trmm
exotic	itto	tree	itto
timber	ortuno	rain	ortuno
rain	kuroshi	island	kuroshi
banana	ivirgarzama	like	ivirgarzama
deforestation	biofunction	fish	biofunction
plantation	kapiolani	most	kapiolani
coconut	bstilla	water	bstilla
jungle	almagreb	fruit	almagreb
tree	jackfruit	area	jackfruit
rainforest	adeo	world	adeo
palm	xishuangbanna	america	xishuangbanna
hardwood	frangipani	some	frangipani
greenhouse	yuca	live	yuca
logging	anthurium	plant	anthurium

- ❑ MI = χ^2
  - – favor low frequency terms
- ❑ Dice + EMI more general
  - – Sometimes too general (“most”)

Query Transformation and Refinement

Association Measure Example

Most strongly associated words for “fish” in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

Dice	MI	EMI	χ^2
species	zoologico	water	arlsq
wildlife	zapanta	species	happyman
fishery	wrint	wildlife	outerlimit
water	wpfmc	fishery	sportk
fisherman	weighout	sea	lingcod
boat	waterdog	fisherman	longfin
sea	longfin	boat	bontadelli
habitat	veracruzana	area	sportfisher
vessel	ungutt	habitat	billfish
marine	ulocentra	vessel	needlefish
endanger	needlefish	marine	damaliscu
conservation	tunaboat	land	bontebok
river	tsolwana	river	taucher
catch	olivacea	food	orangemouth
island	motoroller	endanger	sheepshead

- ❑ MI $\neq \chi^2$
  - – Because “fish” is high-frequency
  - – Still favor low frequency terms

Query Transformation and Refinement

Association Measure Example

Most strongly associated words for “fish” in a collection of TREC news stories. Co-occurrence counts are measured in windows of 5 words.

Dice	MI	EMI	χ^2
wildlife	zapanta	wildlife	gefilte
vessel	plar	vessel	mbmo
boat	mbmo	boat	zapanta
fishery	gefilte	fishery	plar
species	hapc	species	hapc
catch	odfw	tuna	odfw
water	southpoint	trout	southpoint
sea	anadromous	fisherman	anadromous
meat	taiffe	salmon	taiffe
interior	mollie	catch	mollie
fisherman	frampton	nmf	frampton
game	idfg	trawl	idfg
salmon	billingsgate	halibut	billingsgate
tuna	sealord	meat	sealory
caught	longline	shellfish	longline

- MI + χ^2
  - Still favor low frequency terms
- EMI
  - Somewhat improved
- Would you expand your query with any of these words?

Query Transformation and Refinement

Association Measures

- In general, associated words are of little use for expanding the query “tropical fish”
  - See previous tables
  - Terms associated with other contexts
    - Tropical forest, tropical fruit, fishing conservation
- Expansion based on whole query takes context into account
  - E.g., using Dice with term “tropical fish” gives the following highly associated words:
    goldfish, reptile, aquarium, coral, frog, exotic, stripe, regent, pet, wet
- Would have to find associations for every group of query terms
  - Impractical for all possible queries
  - Other approaches achieve this effect

Query Transformation and Refinement

Other Query Expansion Approaches

- ❑ Pseudo-relevance feedback
  - Expansion terms based on top retrieved documents (see next section)
- ❑ Context vectors
  - Represent each word by the words that co-occur with it
  - Create virtual document for that word
    - E.g., top 35 most strongly associated words for “aquarium” (Dice):
      zoology, cranmore, jouett, zoo, goldfish, fish, cannery, urchin, reptile, coral, animal, mollusk, marine, underwater, plankton, mussel, oceanography, mammal, species, exhibit, swim, biologist, cabrillo, saltwater, creature, reef, whale, oceanic, scuba, kelp, invertebrate, ark, crustacean, wild, tropical
  - Rank potential expansion terms by ranking their context vectors
  - If ranked high, it is a good candidate for expansion
    - Virtual document for aquarium contains tropical and fish (thus would be ranked high)
    - Virtual document for jungle contains only tropical (thus lower ranked)

Query Transformation and Refinement

Other Query Expansion Approaches

- Using document collection is expensive and depends on web page quality
- Query logs (query strings and click data)
  - Best source of information about queries and related terms
  - Example: Most frequent words in queries containing “tropical fish” from MSN log:

    stores, pictures, live, sale, types, clipart, blue, freshwater, aquarium, supplies
  - Query suggestion (not term expansion) based on finding similar queries
    - Suggest entire query: “tropical fish supplies,” not “supplies tropical fish”
  - Group queries based on click data (and not on query terms)
    - tropical fish ⇒ pet fish sales
    - Every query is represented by clicked-on pages
    - Similarity of pages is Dice’s coefficient based on clicked-on pages
      - n_{ab} is number of clicked-on pages for both queries
      - n_a and n_b is number of pages clicked on for individual queries

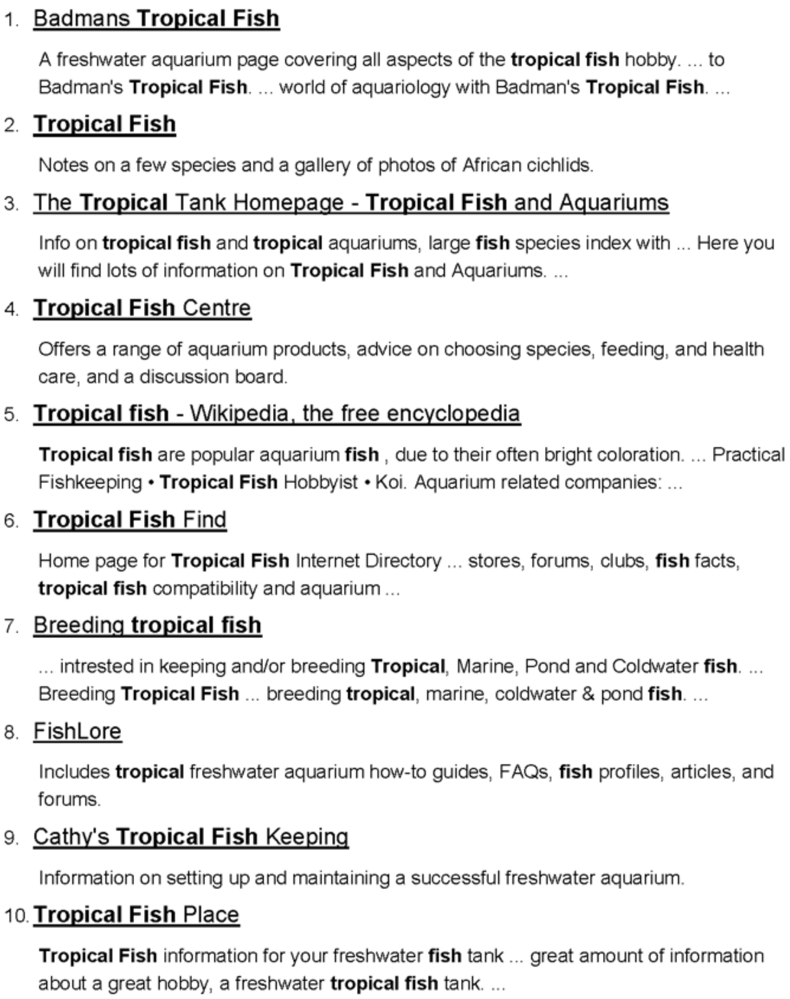
Query Transformation and Refinement

Relevance Feedback

- User marks relevant (+ maybe non-relevant) documents in the initial result
- System modifies query using terms from those documents and reranks
  - Example of simple machine learning algorithm using training data
    - Modifying the query = learning a classifier for relevant and non-relevant documents
    - But very little training data – just this query session
  - In general, queries are expanded with words that frequently occur in relevant documents
    - Or such words are weighted higher
  - Pseudo-relevance feedback just assumes top-ranked documents are relevant – no user input
    - Expansion terms depend on whole query (because it provided the initial ranking)
    - Quality of expansion depends on how many top 10 documents in initial ranking were indeed relevant

Query Transformation and Refinement

Relevance Feedback Example



1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

9. Cathy's **Tropical Fish** Keeping

   Information on setting up and maintaining a successful freshwater aquarium.

10. **Tropical Fish** Place

    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

Top 10 documents for “tropical fish”

Assume all are relevant

Query Transformation and Refinement

Relevance Feedback Example

- Assume top 10 are relevant
- Most frequent terms are (with frequency):
  - a (926), td (535), href (495), http (357), width (345), com (343), nbsp (316), www (260), tr (239), htm (233), class (225), jpg (221)
  - Too many stopwords and HTML expressions
- Use only snippets and remove stopwords
  - tropical (26), fish (28), aquarium (8), freshwater (5), breeding (4), information (3), species (3), tank (2), Badman’s (2), page (2), hobby (2), forums (2)
- Good expansion terms since they use context of multiple query terms

Query Transformation and Refinement

Relevance Feedback Example

- If Document 7 (“Breeding tropical fish”) was explicitly indicated to be relevant, the most frequent terms are:
  - breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2), keeping (1), interested (1)
  - Effect: Increases weight of expansion term breeding
- Specific weights and scoring methods used for relevance feedback depend on retrieval model

Query Transformation and Refinement

Relevance Feedback

- Both relevance feedback and pseudo-relevance feedback are effective, but not used in many applications
  - Pseudo-relevance feedback has reliability issues, especially with queries that do not retrieve many relevant documents
- Some applications use relevance feedback
  - “more like this” or “similar pages” functionalities
  - Relevance feedback to build profiles for filtering
- Query suggestion more popular
  - May be less accurate, but can work if initial query fails
  - Assumes user is looking for many relevant documents; otherwise initial result should be enough

Query Transformation and Refinement

Context and Personalization

- If a query has the same words as another query, should results be the same regardless of
  - who submitted the query,
  - why the query was submitted,
  - where the query was submitted,
  - what other queries were submitted in the same session?
- These other factors (the query context) could have a significant impact on relevance
  - But: Difficult to capture and successfully incorporate into ranking

Query Transformation and Refinement

User Models

- Generate user profiles based on documents that the person looks at
  - Web pages visited
  - Email messages
  - Word processing documents on the desktop
- Modify queries using words from profile
  - Software interest vs. buildings → query for “architect”
  - Users avoid providing explicit, specific profile (privacy)
    - Negative image for search engine using profiling
- Generally not effective
  - Imprecise, unspecific profiles (only snapshot)
  - Information needs can change significantly
    - What if I am generally interested in IT but now want to build a house?!

Query Transformation and Refinement

Query Logs

- Query logs provide important contextual information that can be used effectively
- Context in this case is
  - previous queries that are the same
  - previous queries that are similar
  - query sessions including the same query
- Based on entire user population
- Query history for individuals could be used for caching

Query Transformation and Refinement

Local Search

- Location is context
- Local search uses geographic information to modify the ranking of search results
  - Location derived from the query text
  - Location of the device where the query originated
- Examples
  - `fishing supplies cape cod`
  - `fishing supplies` from mobile device in Hyannis (city at Cape Cod)

Query Transformation and Refinement

Local Search

- ❑ Identify the geographic region associated with web pages
  - Use location metadata that has been manually added to the document
  - Identify locations such as place names, city names, or country names in the text
- ❑ Identify the geographic region associated with the query
  - at least 10–15% of queries contain some location reference
- ❑ Rank web pages using location information in addition to text and link-based features

Query Transformation and Refinement

Extracting Location Information

- Type of information extraction
  - Ambiguity and significance of locations are issues (toponyms)
  - E.g., cities of the same name etc.
- Location names are mapped to specific regions and coordinates

- Matching done by
  - Inclusion
  - Distance