



**Hasso
Plattner
Institut**

IT Systems Engineering | Universität Potsdam

Search Engines

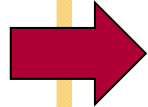
Chapter 6 – Queries and Interfaces

26.5.2011

Felix Naumann

Overview

2



- Information needs
- Query transformation & refinement
- Showing results
- Cross-language search



Information Needs

3

- An *information need* is the underlying cause of the query that a person submits to a search engine.
 - Sometimes called *information problem*: information need is generally related to a task
- Categorized using variety of dimensions
 - Number of relevant documents being sought
 - Type of information that is needed
 - Type of task that led to the requirement for information

Queries and Information Needs

4

- A query can represent very different information needs.
 - May require different search techniques and ranking algorithms to produce the best rankings (see Chapter 7)
- A query can be a poor representation of the information need.
 - User may find it difficult to express the information need
 - User is encouraged to enter short queries both by the search engine interface, and by the fact that long queries don't work

Interaction

5

- Interaction with the system occurs...
 - ... during query formulation and reformulation,...
 - ... and while browsing the result.
- Key aspect of effective retrieval
 - Users can't change ranking algorithm but can change results through interaction
 - Helps refine description of information need
 - ◇ e.g., same initial query, different information needs
 - ◇ how does user describe what they don't know?

ASK Hypothesis

6

- Belkin et al. (1982) proposed a model called Anomalous State of Knowledge
- ASK hypothesis:
 - Difficult for people to define exactly what their information need is, because that information is a gap in their knowledge
 - Search engine should look for information that fills those gaps
- Interesting ideas, little practical impact (yet)

Keyword Queries

7

- Query languages in the past were designed for professional searchers (*intermediaries*)

User query:

Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?

Intermediary query:

NEGLECT! FAIL! NEGLIG! /5 MAINT! REPAIR! /P NAVIGAT! /5 AID
EQUIP! LIGHT BUOY "CHANNEL MARKER"



Wildcard



OR semantics



Phrase



Same
paragraph

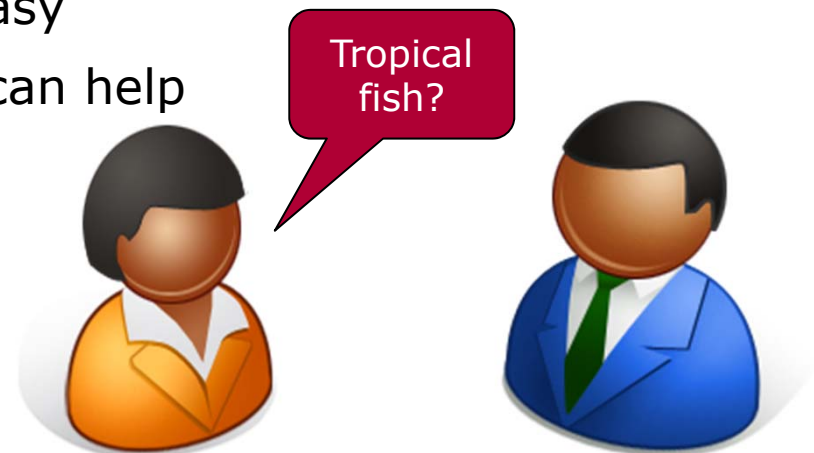


With 5
words

Keyword Queries

8

- Simple, *natural language* queries were designed to enable everyone to search.
- Current search engines do not perform well (in general) with natural language queries.
- People trained (in effect) to use keywords
 - Compare average of about 2.3 words/web query to average of 30 words/CQA query (*community-based question answering*)
- Keyword selection is not always easy
 - Query refinement techniques can help



Overview

9

- Information needs
- Query transformation & refinement
 - ➔ □ Stopping and stemming
 - Spell checking and suggestions
 - Query expansion
 - Relevance feedback
 - Context and personalization
- Showing results
- Cross-language search



Query transformation

10

- In general, same operations on text as on query
- Some differences in stopping and stemming
- Some transformations not needed
 - Tokenization
 - Structure analysis
- 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-based Stemming

11

- 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”,
 - ◇ not stemmed to “climb”
- Alternative: Index word and its stem
 - Increased efficiency
 - But larger index

Stem Classes

12

- 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
 - /bank banked banking bankings banks
 - /ocean oceaneering oceanic oceanics oceanization oceans
 - /polic polical polically police policeable policed
 - policement policer policers polices policial
 - policically policier policiers policies policing
 - policization policize policly policy policying policys
 - Quite long classes – adds many words to query
 - Contain some errors

Stem Classes

13

- 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



Modifying Stem Classes

14

1. For all pairs of words in the stem classes, count how often they co-occur in text windows of W words. W 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 T .
 - T is set empirically.
4. Find the *connected components* of this graph. These are the new stem classes.

Modifying Stem Classes

15

- For example: Dices' 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 x
- Proportion of term occurrences that are co-occurrences
- 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
 - /policies policy
 - /police policed policing
 - /bank banking banks

Overview

16

- Information needs
- Query transformation & refinement
 - Stopping and stemming
 - □ Spell checking and suggestions
 - Query expansion
 - Relevance feedback
 - Context and personalization
- Showing results
- Cross-language search



Spell Checking

17


- Important part of query processing
 - 10-15% of all web queries have spelling errors.
 - Reliance on “did you mean...”

Meinten Sie Databases Die ersten 2 angezeigten Ergebnisse

[Database - Wikipedia, the free encyclopedia](#) - [[Diese Seite übersetzen](#)]
 A **database** is a structured collection of records or data that is stored in a computer system. The structure is achieved by organizing the data according to ...
[en.wikipedia.org/wiki/Database](#) - 70k - [Im Cache](#) - [Ähnliche Seiten](#)

[About Databases: Microsoft Access, SQL Server, Oracle and More!](#) - [[Diese Seite übersetzen](#)]
 Your About.com Guide to **Databases** provides a comprehensive look at **database** systems with feature articles, the Net's best collection of links, ...
[Databases for Beginners - Learning SQL - Microsoft Access](#)
[databases.about.com/](#) - 28k - [Im Cache](#) - [Ähnliche Seiten](#)

Ergebnisse für: Datbases

 [Linksammlung zum Fachgebiet Rechtswissenschaft / Law Datenbanken ...](#)
 Dateiformat: PDF/Adobe Acrobat HTML Version

Error correction with Google

The data below shows some of the misspellings detected by our spelling correction system for the query [britney spears], and the count of how many different users spelled her name that way. Each of these variations was entered by at least two different unique users within a three month period, and was corrected to [britney spears] by our spelling correction system (data for the correctly spelled query is shown for comparison).

[Return to Google's jobs pages](#)

<http://www.google.com/jobs/britney.html>

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brizreny spears
40134 brittany spears	29 britttany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 brittany spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 britttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 britttney spears
7331 britny spears	26 birttney spears	9 britnw spears	5 btinyey spears	3 britnely spears	2 britain spears
6633 britney spears	26 brittney spears	9 britneym spears	5 btrittney spears	3 britnesy spears	2 britane spears
2696 britteny spears	26 brinity spears	9 britney spears	5 gritney spears	3 britnetty spears	2 britaney spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 britney spears	3 britnity spears	2 britanna spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britney spears	2 britannie spears
1338 britiny spears	26 btittany spears	9 rbitney spears	4 bxbritney spears	3 britneyey spears	2 britannt spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 brittezny spears	2 britannu spears
1096 britney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 britttney spears	2 britanyt spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brifitney spears	3 britttneyey spears	2 briteeny spears
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityey spears	2 britenany spears
811 britney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytny spears	2 britenet spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 britney spears	2 briteniy spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 britney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinary spears
544 brittany spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spears	2 britmy spears
544 brittnay spears	21 biteny spears	8 britney spears	4 britaby spears	3 britany spears	2 britnany spears
364 britey spears	21 bratney spears	8 britny spears	4 britay spears	3 briteny spears	2 britnat spears
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet spears	2 britnbey spears
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spears	2 britneh spears
269 britneys spears	21 brittay spears	7 birntey spears	4 britnewy spears	3 drittney spears	2 britneyey spears
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney spears	2 britney6 spears
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney spears	2 britneye spears
220 breatny spears	21 brtiany spears	7 breateny spears	4 britnewy spears	2 barittany spears	2 britneyh spears
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britneyey spears	19 brirtney spears	7 brintye spears	4 britnwey spears	2 bbitney spears	2 britneyyy spears
163 britny spears	19 britnaey spears	7 britianny spears	4 britnwey spears	2 bbritny spears	2 britney spears
147 breatny spears	19 britnee spears	7 britly spears	4 brittaby spears	2 bbrittany spears	2 britnjey spears
147 brittiney spears	19 britony spears	7 britnej spears	4 brittrey spears	2 beitany spears	2 britne spears
147 brinty spears	19 britttany spears	7 britneym spears	4 brittnay spears	2 beitny spears	2 britny spears

Spell Checking

19

- Errors include typical word processing errors
 - poiner sisters
 - brimingham news
 - catamarn sailing
 - hair extenssions
 - marshmellow world
 - miniture golf courses
 - psychics
 - home doceration
- but also many other types (terms and corrections not found in common dictionaries)
 - realstateisting.bc.com
 - akia 1080i manunal
 - ultimatwarcade
 - mainsourcebank
 - dellottitouche

Spell Checking

20

- 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

Edit Distance

21

- *Damerau-Levenshtein* distance
 - counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
 - Levenshtein-distance does not allow transpositions
 - e.g., Damerau-Levenshtein distance 1 (80% of spellign errors)
 - ◇ extenssions -> extensions (deletion)
 - ◇ poiner -> pointer (insertion)
 - ◇ marshmellow -> marshmallow (substitution)
 - ◇ brimingham -> birmingham (transposition)
 - Damerau-Levenshtein distance 2
 - ◇ doceration -> decoration (2 substitutions)



Edit Distance

22

- 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 word
 - ◇ Can be safe (if length exceed threshold)
 - Restrict to words that sound the same
- Last option uses a *phonetic code* to group words
 - e.g. Soundex

Soundex Code

23

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
 2. c,g,j,k,q,s,x,z
 3. d,t
 4. l
 5. m,n
 6. r

extensions → E235; extensions → E235
marshmallow → M625; marshmallow → M625
brimingham → B655; birmingham → B655
poiner → P560; pointer → P536
4. Delete adjacent repeats of a number.
5. Delete the hyphens.
6. Keep the first three numbers or pad out with zeros.
- In Germany: Kölner Phonetik
(http://de.wikipedia.org/wiki/Kölner_Phonetik)

Example: „STADT“

Spelling Correction Issues

24

- 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., *lawers* → *lowers, lawyers, layers, lasers, lagers*
but *trial lawers* → *trial lawyers*
- Run-on errors
 - e.g., “mainsourcebank”
 - missing spaces can be considered another single character error in right framework

Noisy Channel Model

25

- 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

Noisy Channel Model

26

- 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 $P(\text{tank}|\text{fish}) > P(\text{think}|\text{fish}) \Rightarrow$ tank more likely correction

Noisy Channel Model

27

- 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 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

Example Spellcheck Process

28

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 correction.
 4. Repeat from Step 2 until no better correction is found.
- Example
1. *"miniture golfcurses"*
 2. *miniature golfcourses*
 3. *miniature golf courses*

Overview

29

- Information needs
- Query transformation & refinement
 - Stopping and stemming
 - Spell checking and suggestions
 - □ Query expansion
 - Relevance feedback
 - Context and personalization
- Showing results
- Cross-language search



- 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*
- Particularly useful for *query expansion*
 - Add synonyms or more specific terms using query operators based on thesaurus
 - Improves search effectiveness
- 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	Anterior Neck Pain
Entry Term	Cervicalgia
Entry Term	Cervicodynia
Entry Term	Neck Ache
Entry Term	Posterior Cervical Pain
Entry Term	Posterior Neck Pain

Query Expansion

31

- Expansion based on explicit thesaurus (e.g., Wordnet or MeSH) seldom used – 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

Term Association Measures

32

- 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

Term Association Measures

33

- *Dice's Coefficient*

- Reminder: n_x is the number of windows containing x

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

- 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: $P(a,b) = P(a)P(b)$
=> mutual information = 0
- Estimate $P(A) = n_a/N$:

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

Term Association Measures

34

- Mutual Information Measure (MIM) favors low frequency terms

- Example: $n_a = n_b = 10$ and $n_{ab} = 5 \Rightarrow 5/100$

- Example: $n_a = n_b = 1000$ and $n_{ab} = 500 \Rightarrow 5/10000$

$$\frac{n_{ab}}{n_a n_b}$$

- *Expected Mutual Information Measure* (EMIM)

- Weighting of MIM with $P(a,b)$:

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

- Previous example with $N = 1$ million: 23,5 vs. 1350

- Problem: favors high-frequency terms

Term Association Measures

35

- *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.

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

Association Measure Summary




36

<i>Measure</i>	<i>Formula</i>
Mutual information (<i>MIM</i>)	$\frac{n_{ab}}{n_a \cdot n_b}$
Expected Mutual Information (<i>EMIM</i>)	$n_{ab} \cdot \log\left(N \cdot \frac{n_{ab}}{n_a \cdot n_b}\right)$
Chi-square (χ^2)	$\frac{(n_{ab} - \frac{1}{N} \cdot n_a \cdot n_b)^2}{n_a \cdot n_b}$
Dice's coefficient (<i>Dice</i>)	$\frac{n_{ab}}{n_a + n_b}$

Association Measure Example

37

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

[Xishuangbanna Tropical Botanical Garden](#)  
 Garden in Yunnan province administered by the Chinese Academy of Sciences, featuring thousands of tropical and subtropical plants from China and abroad, ...
[en.xtbg.ac.cn/](#) - [Cached](#) - [Similar pages](#) - 

[The International Tropical Timber Organization\(ITTO\)](#)  
 The International Tropical Timber Organization (ITTO) Official Web Site.
[www.itto.int/](#) - [Cached](#) - [Similar pages](#) - 

- $\chi^2 = \text{MIM}$
- Both favor low frequency terms

- EMIM and Dice more general
- Sometimes too general („most“)

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

Association Measure Example

38

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

• $\chi^2 \neq MIM$ because „fish“ is high-frequency
 • Both still favor low frequency terms

- Most strongly associated words for “fish” in a collection of TREC news stories.

Association Measure Example

39

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

• EMIM and Dice more specific

Would you expand your query with any of these words?

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

Association Measures

40

- 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.

Other Approaches

41

- Pseudo-relevance feedback
 - Expansion terms based on top retrieved documents for initial query (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*” (using Dice’s coefficient):
 - ◇ *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 for a query by ranking their context vectors
 - If ranked high, it is a good candidate for expansion
 - ◇ Document for *aquarium* contains high ranking for other query terms (*tropical* and *fish*)
 - ◇ Document for *jungle* contains high ranking for *tropical*, but not for *fish*

Other Approaches

42

- Using document collection is expensive and depends on varying quality of web pages.
- Query logs
 - Best source of information about queries and related terms
 - ◇ Short pieces of text and click data
 - 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 logs

43

Related searches for **Hasso Plattner**:

hasso plattner institute	hasso plattner stanford	sap	design thinking
hasso plattner net worth	hasso plattner cordevalle	dietmar hopp	klaus tschira
hasso plattner ventures	hasso plattner biography	marc benioff	leo apotheker
hasso plattner sail	hasso plattner 505	greg reyes	hasso plattner fortune
hasso plattner bib	hasso plattner familie	louise deputron	hasso plattner golf

Related searches for **Databases**:

education databases	examples of databases	access databases	microsoft access
database software	article databases	free databases	encyclopedia
online databases	relational databases	research databases	oracle
sql databases	types of databases	list of databases	ebSCO
medical databases	history databases	library databases	mysql

Overview

44

- Information needs
- Query transformation & refinement
 - Stopping and stemming
 - Spell checking and suggestions
 - Query expansion
 - □ Relevance feedback
 - Context and personalization
- Showing results
- Cross-language search



Relevance Feedback

45

- User identifies relevant (and maybe non-relevant) documents in the initial result list.
- System modifies query using terms from those documents and re-ranks documents.
 - 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.

Relevance Feedback Example

46

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.

Relevance Feedback Example

47

- 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, use context of multiple query terms

Relevance Feedback Example

48




- If Document 7 ("*Breeding tropical fish*") is *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.

Relevance Feedback

49

- 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”, “similar pages”
 - Relevance feedback to build profiles for filtering



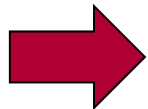
[Hasso Plattner - Wikipedia, the free encyclopedia](#)  
 Hasso Plattner (born 21 January 1944 in Berlin, Germany) is a German entrepreneur. He is married, has two children and lives in Schriesheim-Altenbach near ...
[en.wikipedia.org/wiki/Hasso_Plattner](#) - [Cached](#) - [Similar pages](#) - 

- 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.

Overview

50

- Information needs
- Query transformation & refinement
 - Stopping and stemming
 - Spell checking and suggestions
 - Query expansion
 - Relevance feedback
 - Context and personalization
- Showing results
- Cross-language search



Context and Personalization

51

- If a query has the same words as another query, results will 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

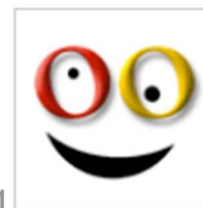
User Models

52

- 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
 - Sports interest -> query for "vikings"
 - 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
 - ◇ Sports and history ("vikings")



Profile



You don't yet have a public profile. [Learn more](#)

[Create a profile](#) or [edit your personal info](#) without creating a public profile.

Query Logs

53

- 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

- 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
- e.g.,
 - *"Fishing supplies cape cod"*
 - *"Fishing supplies"* from mobile device in Hyannis



Local Search

55

- 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 text.
- Identify the geographic region associated with the query.
 - 10-15% of queries contain some location reference.
- Rank web pages using location information in addition to text and link-based features

Extracting Location Information

56

- Type of information extraction
 - Ambiguity and significance of locations are issues (toponyms)
- Location names are mapped to specific regions and coordinates



- Matching done by
 - Inclusion
 - Distance

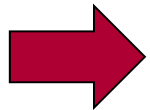
United States

- Springfield, California
- Springfield, Colorado
- Springfield, Florida
- Springfield, Jacksonville, Florida, a neighborhood of Jacksonville
- Springfield, Georgia
- Springfield, Illinois, the state capital of Illinois
 - Springfield, Illinois metropolitan area
- Springfield, Indiana, an unincorporated place
- Springfield, Kentucky
- Springfield, Louisiana
- Springfield, Maine
- Springfield, Massachusetts, the first Springfield in the United States, established 1636
 - Springfield, Massachusetts metropolitan area
- Springfield, Michigan, a city in Calhoun County
- Springfield, Minnesota, in Brown County
- Springfield, Missouri, as of 2007, the largest city in the United States named Springfield
 - Springfield, Missouri Metropolitan Area
- Springfield, Nebraska
- Springfield, New Hampshire
- Springfield Township, Union County, New Jersey, site of the Battle of Springfield (1780)
- Springfield Township, Burlington County, New Jersey
- Springfield/Belmont neighborhood of Newark, New Jersey
- Springfield, New York
- Springfield, Ohio
 - Springfield, Ohio metropolitan area (see Clark County, Ohio)
- Springfield, Oregon
- Springfield, Pennsylvania
- Springfield, South Carolina
- Springfield, South Dakota
- Springfield, Tennessee
- Springfield, Texas
- Springfield, Vermont
- Springfield, Virginia
- Springfield, West Virginia
- Springfield, Dane County, Wisconsin
- Springfield, Jackson County, Wisconsin
- Springfield, Marquette County, Wisconsin
- Springfield, St. Croix County, Wisconsin
- Springfield, Walworth County, Wisconsin

Overview

57

- Information needs
- Query transformation & refinement
- Showing results
 - Snippets
 - Advertising
 - Result clustering
- Cross-language search



Tropical Fish

One of the U.K.s Leading suppliers of **Tropical**, Coldwater, Marine **Fish** and Invertebrates plus.. . next day **fish** delivery service ...

www.tropicalfish.org.uk/tropical_fish.htm [Cached page](#)

- Query-dependent document summary
 - Link to web page and cached version
 - Title and URL
 - Short text summary (snippet)
 - ◇ Sometimes full sentences, sometimes not
 - Some query-independent features may be used.
- Simple summarization approach
 - first proposed by Luhn in 50's
 1. Rank each sentence in a document using a *significance factor*.
 2. Select the top sentences for the summary.

Sentence Selection

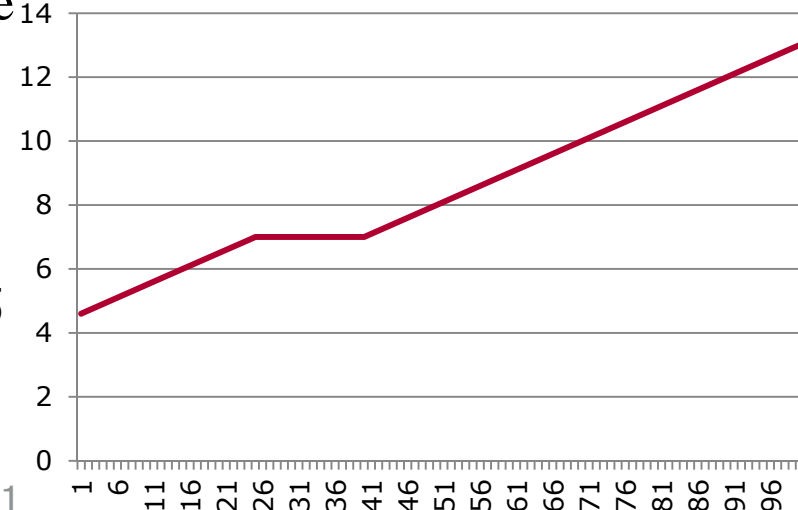
59

- Significance factor for a sentence is calculated based on the occurrence of *significant words*.
 - Significant words are of medium frequency
 - If $f_{d,w}$ is the frequency of word w in document d , then w is a significant word if it is not a stopword and

$$f_{d,w} \geq \begin{cases} 7 - 0.1(25 - s_d) & \text{if } s_d < 25 \\ 7 & \text{if } 25 \leq s_d < 40 \\ 7 + 0.1(s_d - 40) & \text{otherwise} \end{cases}$$

where s_d is the number of sentences in document d .

- Example: $s_d = 20$
 $\Rightarrow f_{d,w} \geq 7 - 0.1(25 - 20) = 6.5$



Sentence Selection

60

- Text is *bracketed* by significant words
 - Limit on number of non-significant words between two significant ones.
 - ◇ Usually 4 non-significant words
- Significance factor for bracketed text spans is computed by dividing the square of the number of significant words in the span by the total number of words
 - Initial sentence: w w w w w w w w w w w .
 - Significant words: w w s w s s w w s w w .
 - Bracketed: w w [s w s s w w s] w w .
 - Significance factor = $4^2/7 = 2.3$
- Significance factor for entire text is maximum significance factor for any bracket.

Snippet Generation

61

- Improvements based on better selection of significant words and sentence fragments.
 - In particular: Query dependent.
- Involves more features than just significance factor
- e.g. for a news story, could use
 - whether the sentence is a *heading*
 - whether it is the *first or second line* of the document
 - the total *number of query terms* occurring in the sentence
 - the number of *unique query terms* in the sentence
 - the longest *contiguous run of query words* in the sentence
 - a *density measure* of query words (significance factor)
- Weighted *combination* of features used to rank sentences

Snippet Generation

62

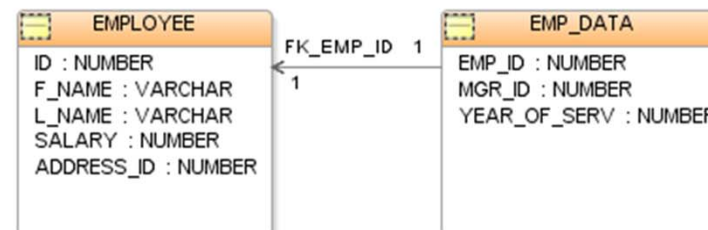
- Web pages are less structured than news stories.
 - Can be difficult to find good summary sentences
- Snippet sentences are often selected from other sources
 - Metadata associated with the web page
 - ◇ e.g., <meta name="description" content= ...>
 - External sources such as web directories
 - ◇ e.g., Open Directory Project, http://www.dmoz.org
- Certain pages, such as Wikipedia have better structure
 - Snippet generation easier

Database

From Wikipedia, the free encyclopedia
 (Redirected from [Databases](#))

This article is mainly about managing and structuring the collections of data held on computers. For a fuller discussion of DBMS software, see [Database management system](#). For databased content libraries, see [Online database](#)

A **database** is a structured collection of records or **data** that is stored in a **computer system**. The structure is achieved by organizing the data according to a **database model**. The model in most common use today is the **relational model**. Other models such as the **hierarchical model** and the **network model** use a more explicit representation of relationships.



Snippet Guidelines

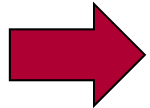
63

- Derived from analysis of clickthrough data
 - All query terms should appear in the summary, showing their relationship to the retrieved page.
 - When query terms are present in the title, they need not be repeated.
 - ◇ Allows snippets that do not contain query terms
 - Highlight query terms in URLs.
 - Snippets should be readable text, not lists of keywords.
 - ◇ Sentences or contiguous sentence fragments
- Snippet generation should be fast
 - Local document store

Overview

64

- Information needs
- Query transformation & refinement
- Showing results
 - Snippets
 - Advertising
 - Result clustering
- Cross-language search



Advertising

65

- Two kinds of advertising
 - *Sponsored search* – advertising presented with search results
 - *Contextual advertising* – advertising presented when browsing web pages
- Both involve finding the most relevant advertisements in a database
 - An advertisement usually consists of a short text description and a link to a web page describing the product or service in more detail.
 - Special case of text search

Sponsored Links

[Database](#)
Now with partitioning
Get new MySQL Enterprise 5.1 now
www.mysql.com

[Free Database Download](#)
Get the **database** that powers the world's most complex applications.
www.Objectivity.com

[Easy Database Design Tool](#)
Make sure your **database** design software has these 10 features.
www.modelright.com

[One Tool - All databases](#)
Manage 12 different relational DBs with one tool - AlligatorSQL
www.alligatorsql.com

Searching Advertisements

66

- Factors involved in ranking advertisements
 - Similarity of text content to query
 - Bids for keywords in query
 - Popularity of advertisement
- Who defines factors and weighting?
 - Payment model
 - Economics and game theory
- Small amount of text in advertisement
 - Dealing with vocabulary mismatch is important
 - Expansion techniques are effective
 - ◇ Both for query and for document (=advertisement)

Searching Advertisements

67

- Query reformulation based on search sessions
 - 50% of queries are reformulations
 - ◇ I.e., rich repository of associations
 - Learn associations between words and phrases based on co-occurrence in search sessions
 - ◇ “*Aquarium*” followed by “*fish tank*” in same session

- Pseudo-relevance feedback
 - Expand query and/or document using the Web
 - Use ad text or query for pseudo-relevance feedback
 - Effective ranking order
 1. Exact matches
 2. Stem matches
 3. Expansion matches

Example Advertisements

68

fish tanks at Target

Find **fish tanks** Online. Shop & Save at Target.com Today.
www.target.com

Aquariums

540+ Aquariums at Great Prices.
fishbowls.pronto.com

Not obvious,
but relevant

Freshwater **Fish Species**

Everything you need to know to keep your setup clean and beautiful
www.FishChannel.com

Pet Supplies at Shop.com

Shop millions of products and buy from our trusted merchants.
shop.com

Probably based
on keyword bid

Custom **Fish Tanks**

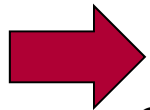
Choose From 6,500+ Pet Supplies. Save On Custom **Fish Tanks**!
shopzilla.com

- Advertisements retrieved for query "fish tank"

Overview

69

- Information needs
- Query transformation & refinement
- Showing results
 - Snippets
 - Advertising
 - Result clustering
- Cross-language search



Clustering Results

70

- Result lists often contain documents related to different *aspects* of the query topic.
 - “jaguar”
- *Clustering* is used to group related documents to simplify browsing.
 - See Chapter 9

Example clusters for query “tropical fish”

Pictures (38)

Aquarium Fish (28)

Tropical Fish Aquarium (26)

Exporter (31)

Supplies (32)

Plants, Aquatic (18)

Fish Tank (15)

Breeding (16)

Marine Fish (16)

Aquaria (9)

Clustering Results

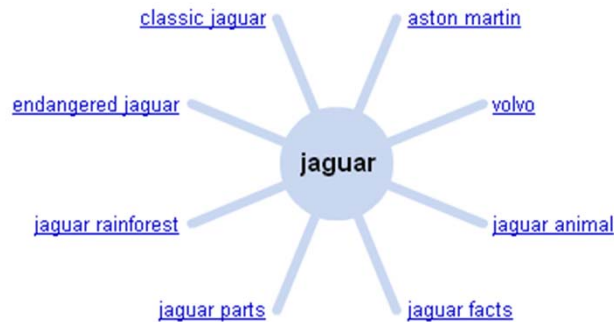
71



jaguar [Advanced Search](#) [Preferences](#)

Web > Wonder wheel [Hide options](#)

- › All results
 - [Videos](#)
 - [Forums](#)
 - [Reviews](#)
- › Any time
 - [Recent results](#)
 - [Past 24 hours](#)
 - [Past week](#)
 - [Past year](#)
- › Standard results
 - [Images from the page](#)
 - [More text](#)
- › Wonder wheel
 - [Timeline](#)



Refine Search

by keyword exclude

by date

by source

by domain

Selected Sources [2 of 3] [Add/Remove](#)

Yahoo! Wikipedia Amazon Books

Databases

250 total results

[Expand Outline](#) | [Collapse Outline](#)

Databases (250 results)

- Library (51)
- Model (24)
- Transaction (25)
 - Delete (14)
- Center (10)
 - Software (10)
- Collections (9)
- Books (9)
- Key Foreign (8)
- Reference Databases (8)
- Join (3)
- Key (5)
 - Application Programming (5)
 - Aka Triple-stores (4)
 - Ordered Flat Files (4)
 - Office (4)
 - EBSCO (4)
 - News (4)
 - Maintenance (3)
 - Statistics (3)
 - Forms of Data-structure Mentioned (3)
 - Open Directory (3)
 - Research (3)
 - Searchable Databases (3)
 - National (3)
 - v d e (3)
 - Central (3)
 - Retrieve the Information Stored in a Database (2)

Clustering Results – Requirements

73

- Efficiency
 - Must be specific to each query and are based on the top-ranked documents for that query
 - Typically based on snippets, not full text
 - ◇ Snippets focus on query-relevant part of text, not on entire text
- Easy to understand
 - Can be difficult to assign good labels to groups
 - Monothetic vs. polythetic classification

Types of Classification

74

- Monothetic
 - Every member of a class has the property that defines the class
 - Typical assumption made by users
 - Easy to understand, because easy to explain
- Polythetic
 - Members of classes share many properties but there is no single defining property
 - Most clustering algorithms (e.g. K-means) produce this type of output

Classification Example

75

- $D_1 = \{a, b, c\}$
- $D_2 = \{a, d, e\}$
- $D_3 = \{d, e, f, g\}$
- $D_4 = \{f, g\}$
- Possible monothetic classification
 - Not necessarily disjoint
 - $\{D_1, D_2\}$ (labeled using a) and $\{D_2, D_3\}$ (labeled e)
- Possible polythetic classification
 - Based on term overlap
 - $\{D_2, D_3, D_4\}, D_1$
 - No single term in common.
 - Labels?

Result Clusters

76

- Simple algorithm
 - Group based on words in snippets
 - Use all non-stop-terms that appear in at least two snippets
 - ◇ *aquarium* (5) (Documents 1, 3, 4, 5, 8)
 - ◇ *freshwater* (4) (1, 8, 9, 10)
 - ◇ *species* (3) (2, 3, 4)
 - ◇ *hobby* (3) (1, 5, 10)
 - ◇ *forums* (2) (6, 8)
- Refinements
 - Use phrases
 - Use more features
 - ◇ whether phrases occurred in titles or snippets
 - ◇ length of the phrase
 - ◇ collection frequency of the phrase
 - ◇ overlap of the resulting clusters

Faceted Classification

77

- A set of categories, usually organized into a hierarchy, together with a set of *facets* that describe the important properties associated with the category
 - Document can have value in every facet
- Manually defined
 - Potentially less adaptable than dynamic classification
 - Tedious
- Easy to understand
 - Commonly used in e-commerce

Example Faceted Classification

78

- Books (7,845)
- Home & Garden (2,477)
- Apparel (236)
- Home Improvement (169)
- Jewelry & Watches (76)
- Sports & Outdoors (71)
- Office Products (68)
- Toys & Games (62)
- Everything Else (44)
- Electronics (26)
- Baby (25)

- DVD (12)
- Music (11)
- Software (10)
- Gourmet Food (6)
- Beauty (4)
- Automotive (4)
- Magazine Subscription (4)
- Health & Personal Care (4)
- Wireless Accessories (1)
- Video Games (1)

Home & Garden

- Kitchen & Dining (149)
- Furniture & Décor (1,776)
- Pet Supplies (368)
- Bedding & Bath (51)
- Patio & Garden (22)
- Art & Craft Supplies (12)
- Home Appliances (2)
- Vacuums, Cleaning & Storage (107)

Brand

<brand names>

Seller

<vendor names>

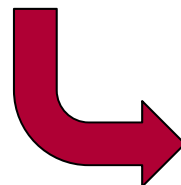
Discount

- Up to 25% off (563)
- 25% - 50% off (472)
- 50% - 70% off (46)
- 70% off or more (46)

Price

- \$0-\$24 (1,032)
- \$25-\$49 (394)
- \$50-\$99 (797)
- \$100-\$199 (206)
- \$200-\$499 (39)
- \$500-\$999 (9)
- \$1000-\$1999 (5)
- \$5000-\$9999 (7)

Categories for *"tropical fish"*

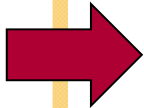


Subcategories and facets
for "Home & Garden"

Overview

79

- Information needs
- Query transformation & refinement
- Showing results
- Cross-language search



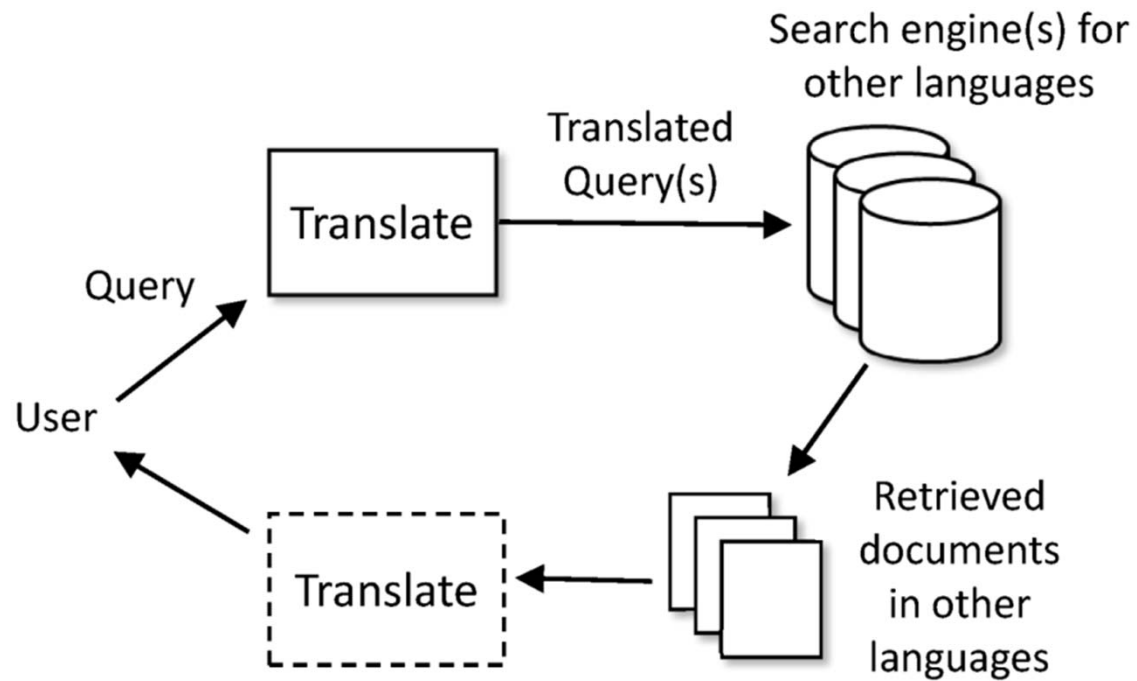
Cross-Language Search

80

- Query in one language, retrieve documents in multiple other languages.
- Involves query translation, and probably document translation.
- Query translation can be done using bilingual dictionaries.
- Document translation requires more sophisticated *statistical translation* models.
 - Similar to some retrieval models
 - To ensure grammatical correctness

Cross-Language Search

81



Statistical Translation Models

82

- Models require *parallel corpora* for training
 - Probability estimates based on *aligned* sentences
- Translation of unusual words and phrases is a problem
 - Also use *transliteration* techniques
 - ◇ e.g., Qathafi, Kaddafi, Qadafi, Gadafi, Gaddafi, Kathafi, Kadhafi, Qadhafi, Qazzafi, Kazafi, Qaddafi, Qadafy, Quadhaffi, Gadhdhafi, al-Qaddafi, Al-Qaddafi
 - ◇ Similar variations of Bill Clinton on arabic pages

Translation

83

- Web search engines also use translation

- e.g. for query "*pecheur france*"

Le pêcheur de France archives @ peche poissons - [[Translate this page](#)]

Le **pêcheur** de **France** Les média Revues de pêche Revue de presse Archives de la revue
Le **pêcheur** de **France** janvier 2003 n°234 Le **pêcheur** de **France** mars 2003 ...

- Translation link translates web page
- Uses statistical machine translation models
 - ◇ Choose most likely translation