Queries and Interfaces Part I

406.424 Internet Applications

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

Seoul National University

9/1/2010

information needs

- an information need is the **underlying cause of the query** that a person submits to a search engine
  - sometimes called information problem to emphasize that information need is generally related to a task
- categorized using variety of dimensions
  - e.g., 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

- a query can represent very different information needs
  - may require different search techniques and ranking algorithms to produce the best rankings
- 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

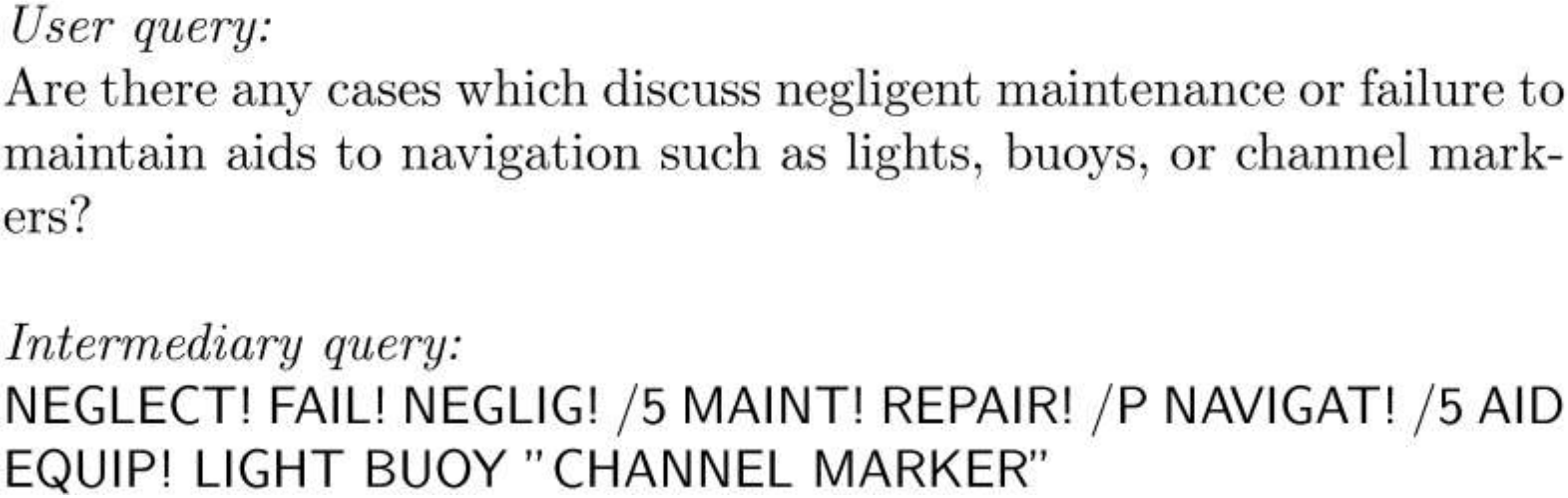
- interaction with the system occurs
  - during query formulation and reformulation
  - while browsing the result
- key aspect of effective retrieval
  - users can’t change ranking algorithm but can change results through interaction
  - helps refine description of information need
    - e.g., same initial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

ASK hypothesis

- 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

- 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"

keyword queries

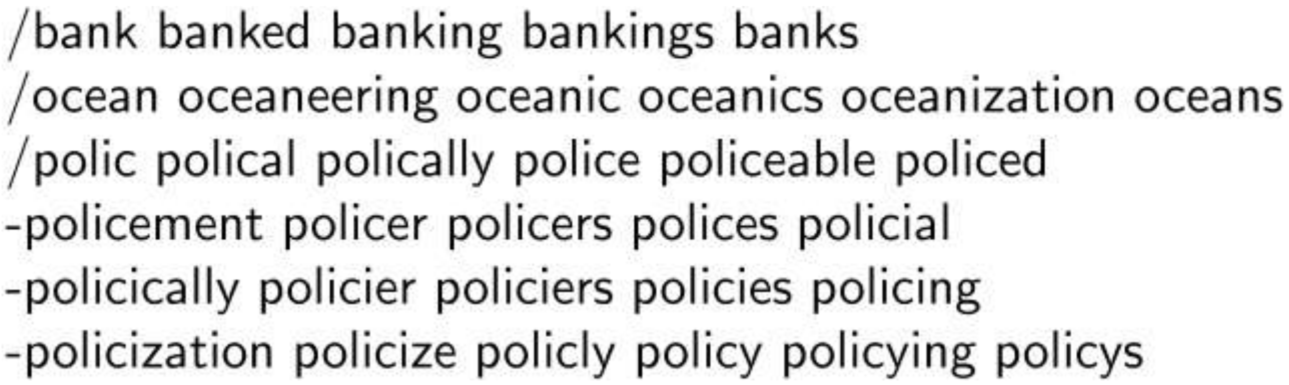
- simple, natural language queries were designed to enable everyone to search
- current search engines do not perform well (in general) with natural language queries
- people trained (in effect) to use keywords
  - compare average of about 2.3 words/web query to average of 30 words/CQA (community-based Q&A) query
- keyword selection is not always easy
  - query refinement techniques can help

query-based stemming

- make decision about stemming at query time rather than during indexing
  - improved flexibility, effectiveness
- query is expanded using word variants
  - documents are not stemmed
  - e.g., “rock climbing” expanded with “climb”, not stemmed to “climb”

stem classes

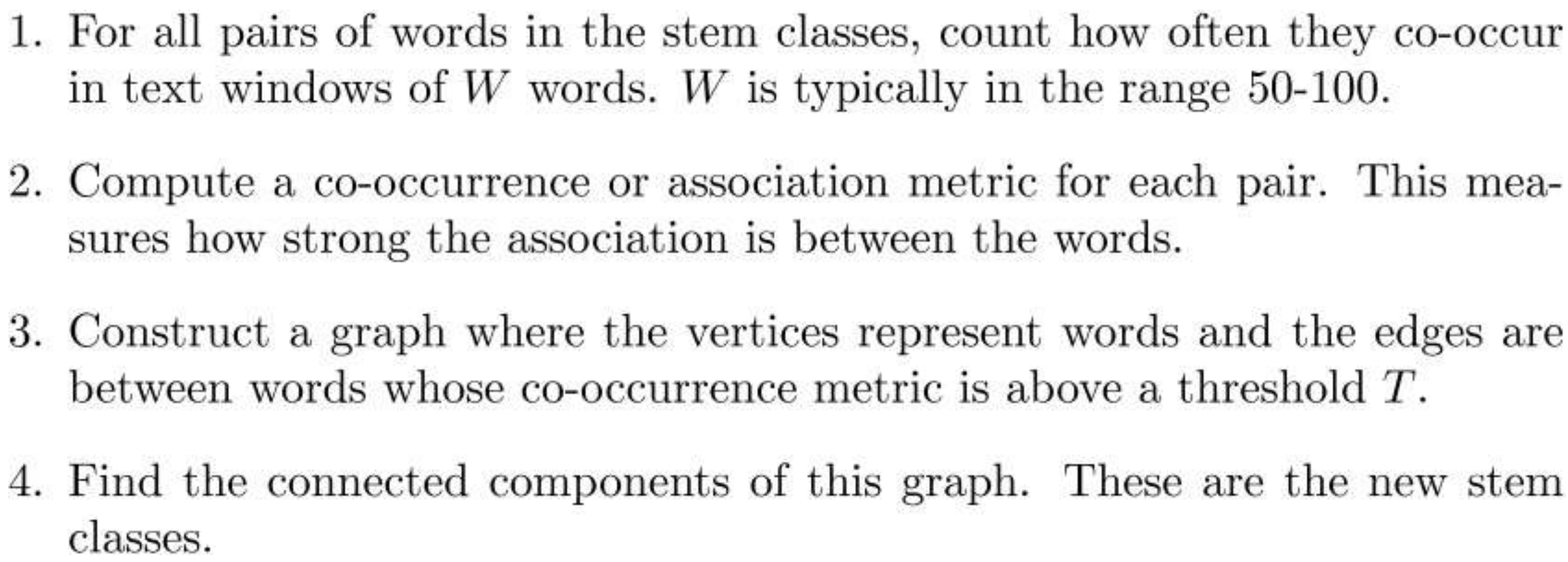
- a stem class is the group of words that will be transformed into the same stem by the stemming algorithm
  - generated by running stemmer on large corpus
  - e.g., Porter stemmer on TREC News



/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

stem classes

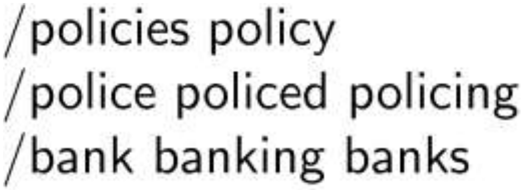
- stem classes are often **too big** and **inaccurate**
- modify using analysis of word co-occurrence
- assumption:
  - word variants that could substitute for each other should **co-occur often** in documents



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.
4. Find the connected components of this graph. These are the new stem classes.

modifying stem classes

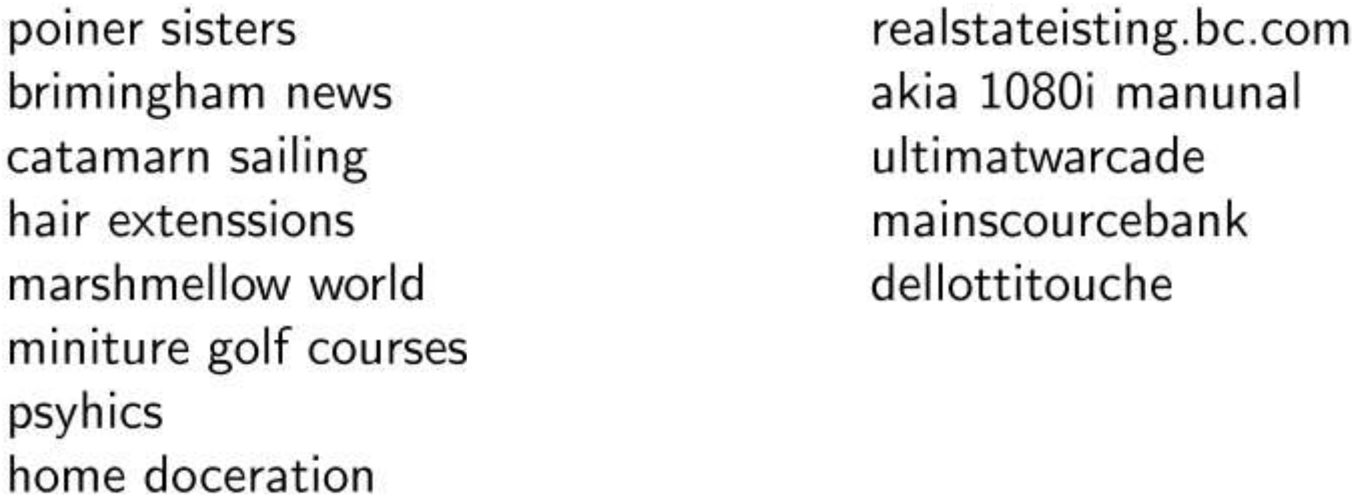
- Dices’ coefficient is an example of a term association measure
    - $2.n_{ab}/(n_a + n_b)$
    - where n_x is the number of windows containing *x*
- 2 vertices are in the same connected component of a graph if there is a path between them
    - forms word *clusters*
- example output of modification



```
/policies policy
/police policed policing
/bank banking banks
```

spell checking

- important part of query processing
  - 10-15% of all web queries have spelling errors
- “did you mean …” feature
- errors include typical word processing errors but also many other types, e.g.



poiner sisters
brimingham news
catamarn sailing
hair extenssions
marshmellow world
miniture golf courses
psyhics
home doceration

realstateisting.bc.com
akia 1080i manunal
ultimatwarcade
mainscourcebank
dellottitouche

spell checking

- basic approach: suggest corrections for words not found in **spelling dictionary**
- suggestions found by comparing word to words in dictionary using **similarity measure**
- most common similarity measure is **edit distance**
  - number of operations required to transform one word into the other

edit distance

- **Damerau-Levenshtein** distance
  - counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required
  - e.g., Damerau-Levenshtein distance 1

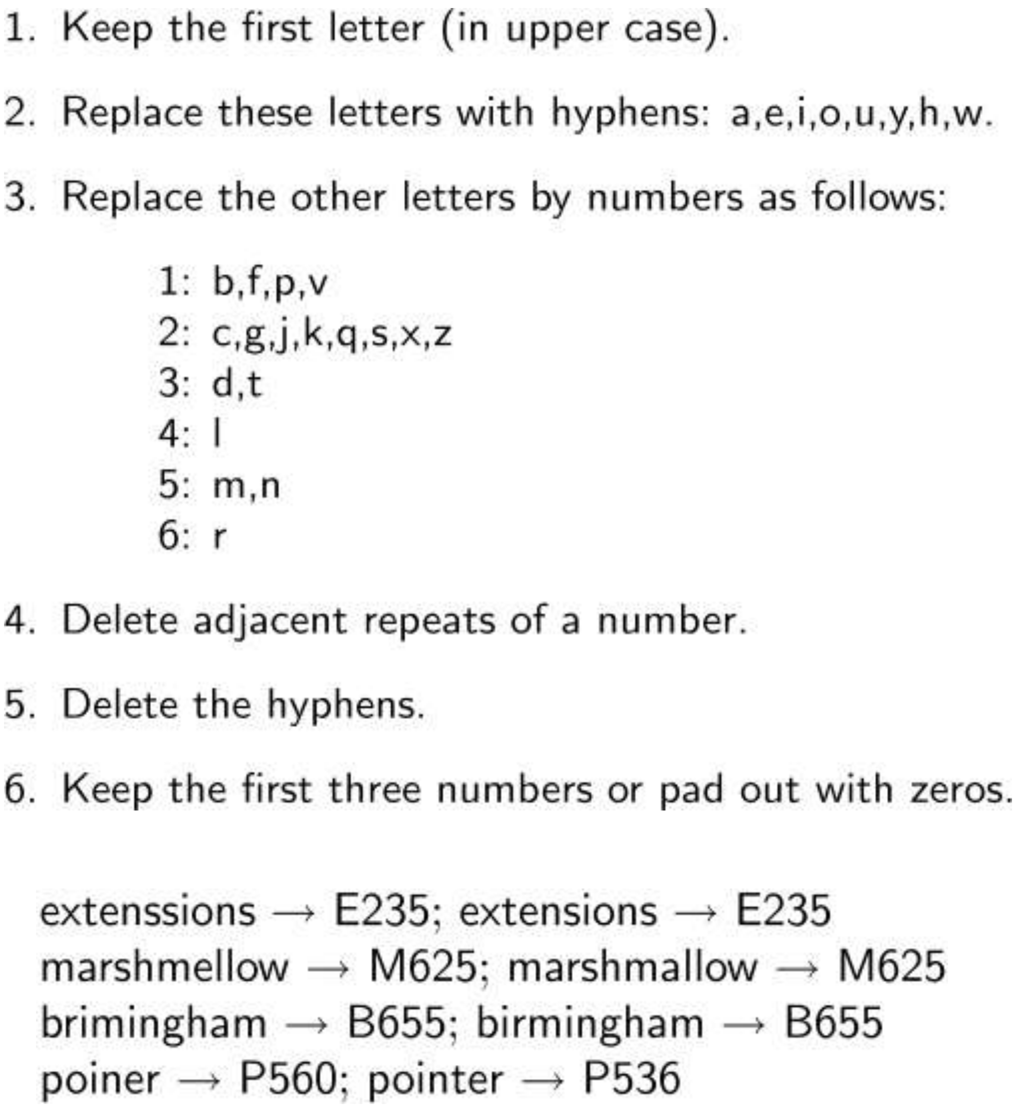
    extenssions → extensions(insertion error)
    poiner → pointer (deletion error)
    marshmellow → marshmallow (substitution error)
    brimingham → birmingham (transposition error)
  - distance 2

    doceration → deceration
    deceration → decoration

edit distance

- number of techniques used to **speed up calculation** of edit distances
  - restrict to words starting with same character
  - restrict to words of same or similar length
  - restrict to words that sound the same
- last option uses a **phonetic code** to group words
  - e.g. Soundex code, GNU Aspell checker

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

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

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

spelling correction issues

- ranking corrections
  - “did you mean...” feature requires accurate **ranking of possible corrections**
- 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., “mainscourcebank”
  - missing spaces can be considered another **single character error** in right framework

noisy channel model

- 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

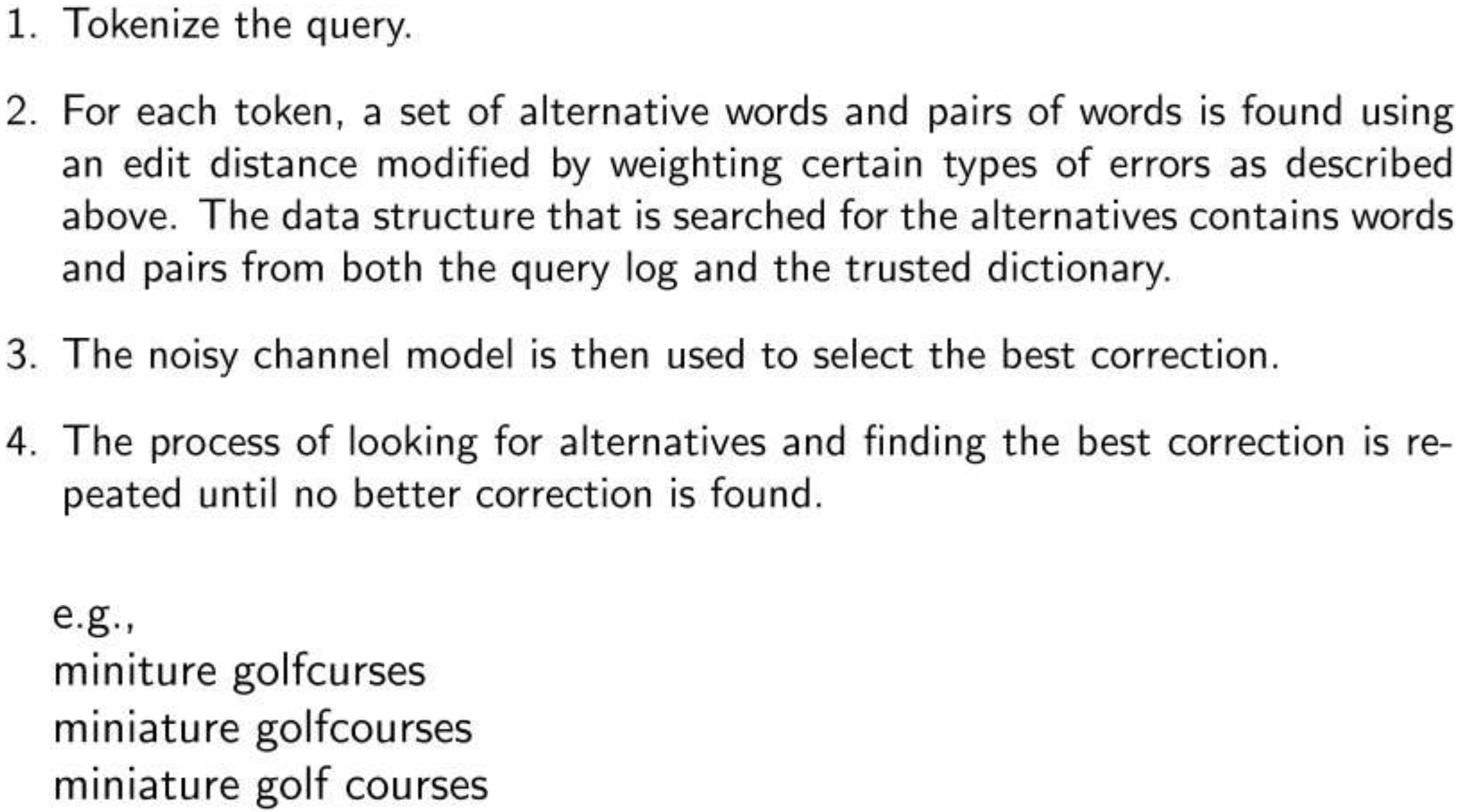
noisy channel model

- need to estimate probability of correction
  - $P(w|e)$: probability that the correct word is w given that a person wrote e
  - $P(w|e) \sim P(e|w)P(w)$ (Bayes' rule)
- estimate language model using context
  - e.g., $P(w) = \lambda P(w) + (1 - \lambda)P(w|w_p)$
  - w_p is previous word
- e.g.,
  - "fish tink"
  - "tank" and "think" both likely corrections, but $P(\text{tank}|\text{fish}) > P(\text{think}|\text{fish})$

noisy channel model

- language model probabilities estimated using corpus and query log
- 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

example spellcheck process



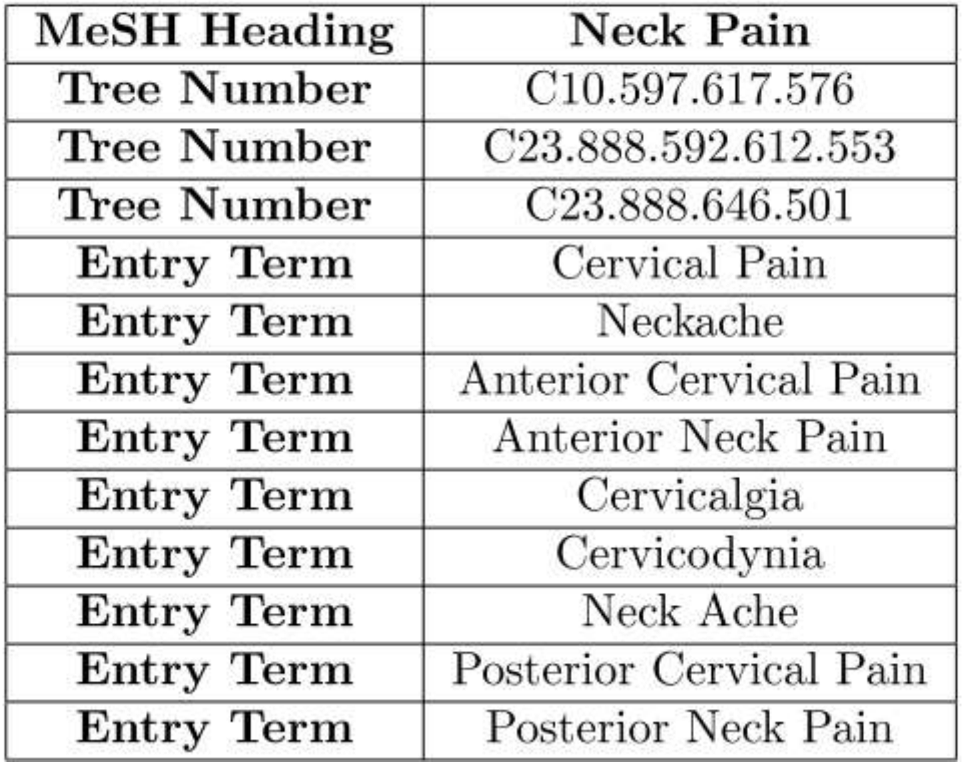
1. Tokenize the query.
2. For each token, a set of alternative words and pairs of words is found using an edit distance modified by weighting certain types of errors as described above. The data structure that is searched for the alternatives contains words and pairs from both the query log and the trusted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

e.g.,
miniture golfcurses
miniature golfcourses
miniature golf courses

thesaurus

- used in early search engines as a tool for **indexing** and **query formulation**
  - specified preferred terms and relationships between them
  - also called **controlled vocabulary**
- e.g., Wordnet (wordnet.princeton.edu)
- particularly useful for **query expansion**
  - adding synonyms or more specific terms using query operators based on thesaurus
  - improves search effectiveness

MeSH thesaurus



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

- a variety of automatic or semi-automatic query expansion techniques have been developed
  - goal is to 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

query expansion

- approaches usually based on an analysis of **term co-occurrence**
  - either in the entire document collection, a large collection of queries, or the top-ranked documents in a result list
  - query-based stemming also an expansion technique
- automatic expansion based on general thesaurus not effective
  - does not take context into account

term association measures

- Dice's coefficient

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

- MI (mutual information)
  - *P*(*a*): the probability that word *a* occurs in a text window of a given size
  - *P*(*a*,*b*): the probability that *a* and *b* occur in the same text window

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

term association measures

- MI measure favors low frequency terms
  - $n_a = n_b = 10$, $n_{ab} = 5$ => MI = 5×10^{-2}
  - $n_a = n_b = 1000$, $n_{ab} = 500$ => MI = 5×10^{-4}
- Expected Mutual Information Measure (EMIM)
  - addresses the problem of MI by weighting the MI value using the probability $P(a,b)$

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

term association measures

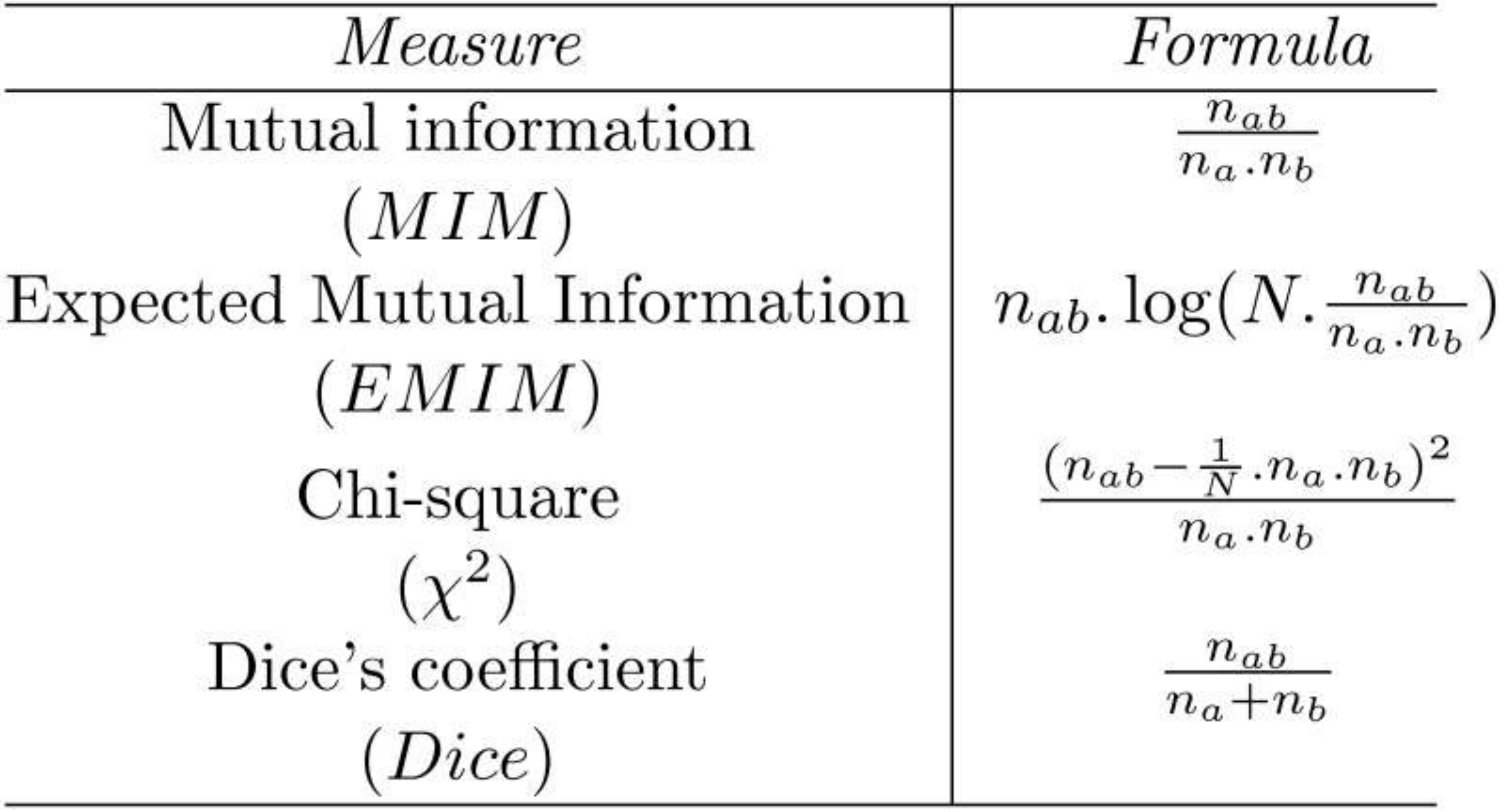
- Pearson’s Chi-squared (χ^2) measure
  - compares the **number of co-occurrences of two words** with the **expected number of co-occurrences** if the two words were independent
  - expected number of co-occurrences if two terms occur independently:

$$NP(a)P(b) = N\frac{n_a}{N}\frac{n_b}{N}$$

  - **normalizes** this comparison by the expected number

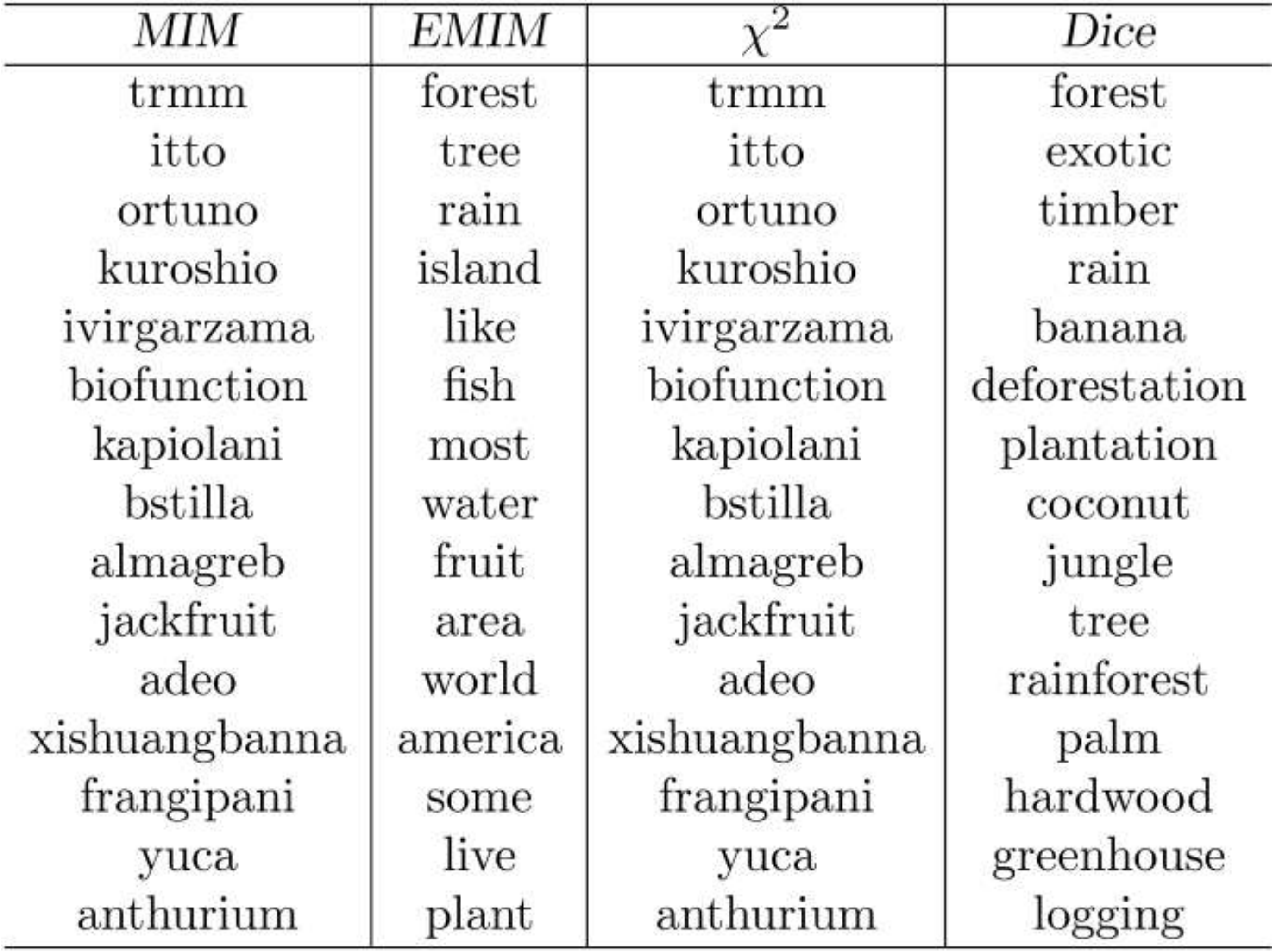
$$\frac{(n_{ab} - N.\frac{n_a}{N}.\frac{n_b}{N})^2}{N.\frac{n_a}{N}.\frac{n_b}{N}} \overset{rank}{=} \frac{(n_{ab} - \frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$$

association measure summary



Measure	*Formula*
Mutual information (MIM)	$\frac{n_{ab}}{n_a.n_b}$
Expected Mutual Information ($EMIM$)	$n_{ab}.\log(N.\frac{n_{ab}}{n_a.n_b})$
Chi-square (χ^2)	$\frac{(n_{ab}-\frac{1}{N}.n_a.n_b)^2}{n_a.n_b}$
Dice's coefficient ($Dice$)	$\frac{n_{ab}}{n_a+n_b}$

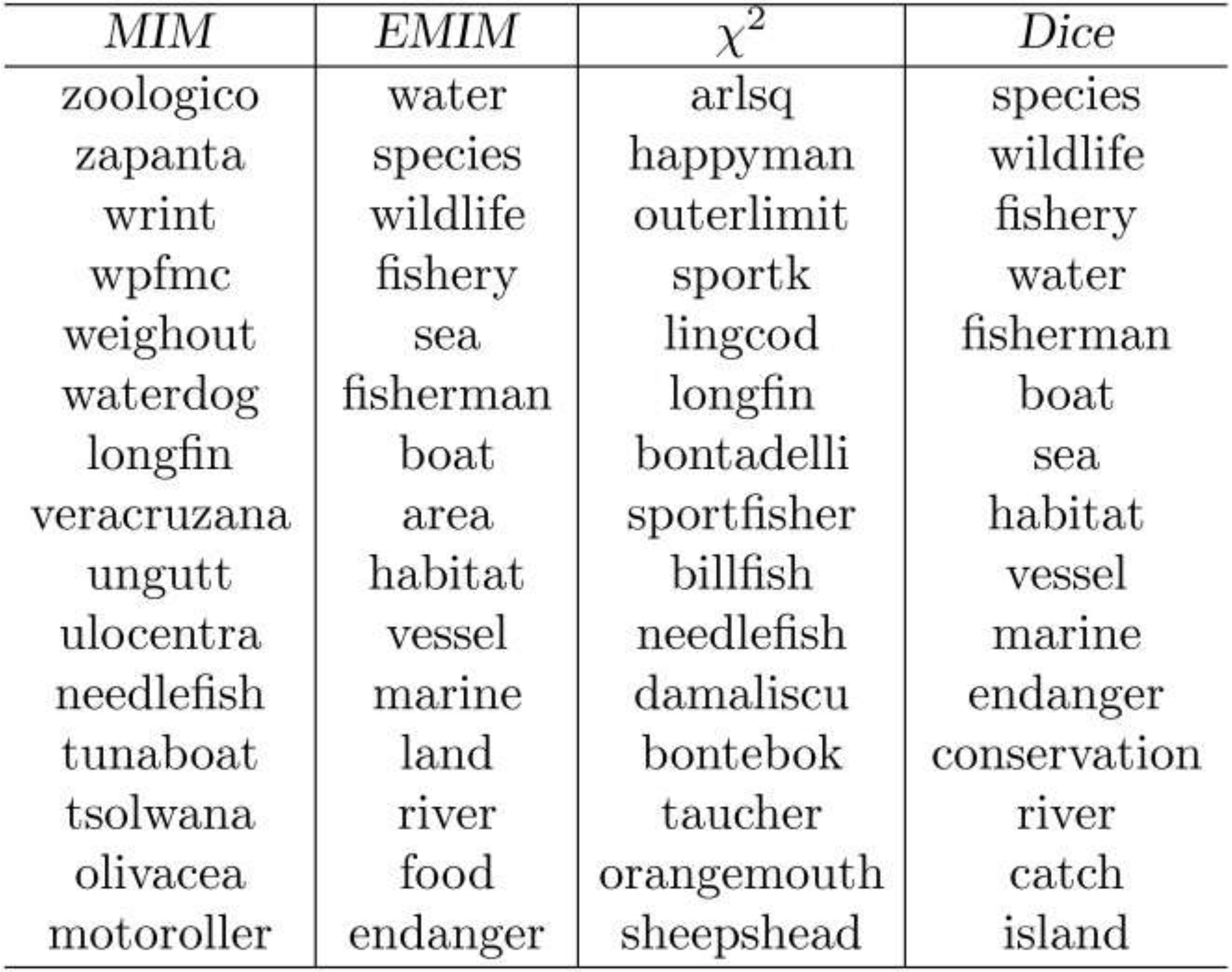
association measure example



MIM	*EMIM*	χ^2	*Dice*
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

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

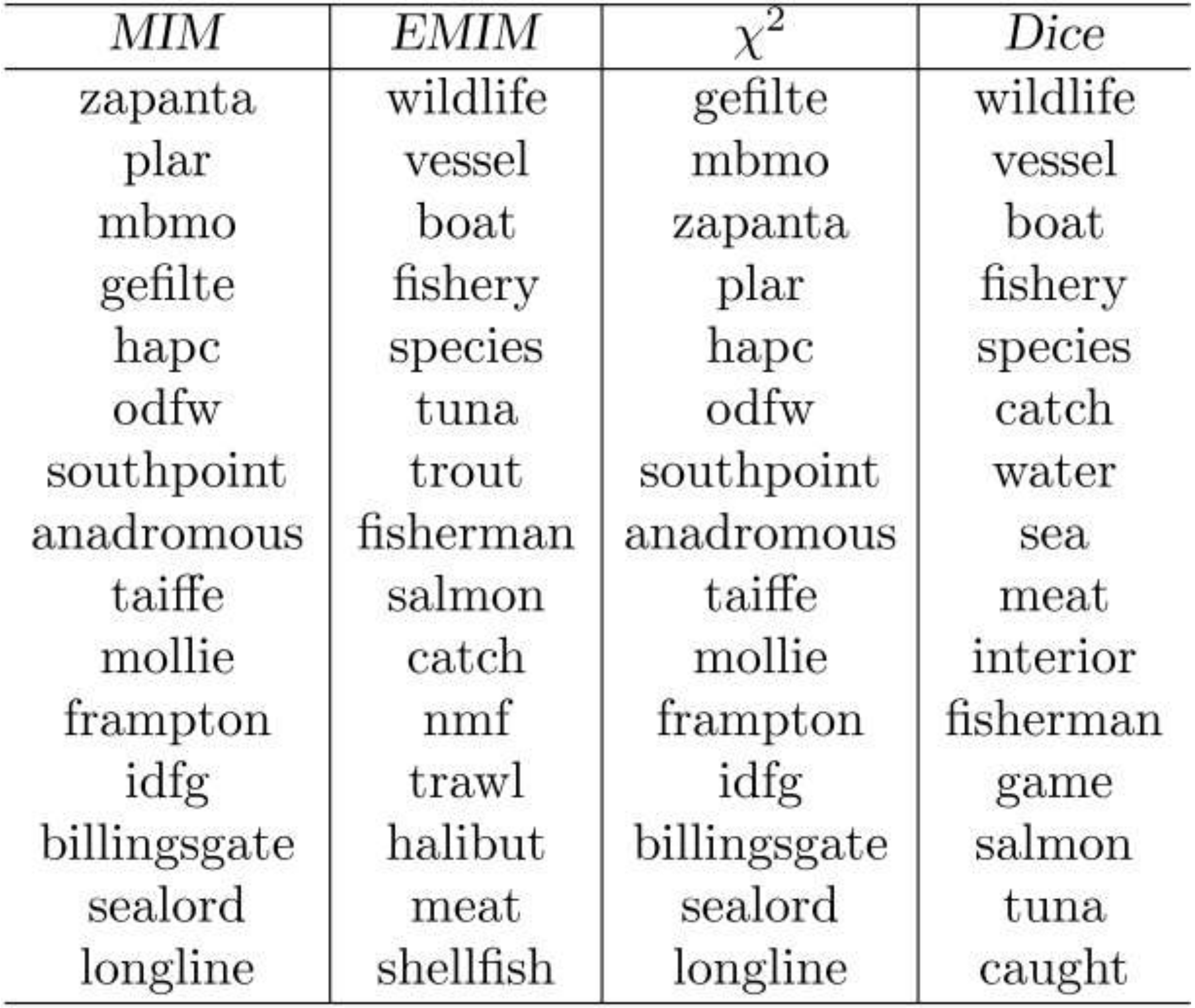
association measure example



MIM	*EMIM*	χ^2	*Dice*
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

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

Association Measure Example



MIM	*EMIM*	χ^2	*Dice*
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

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

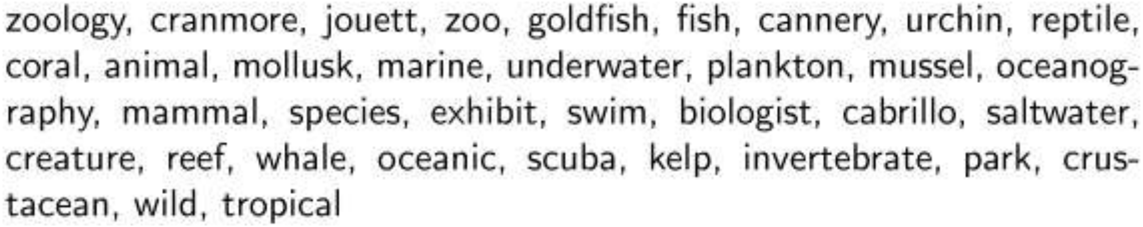
association measures

- individual associated words are of **little use for expanding the query** “tropical fish”
- 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
- **impractical for every group of words that could be used in a query**, other approaches used to achieve this effect

other approaches

- pseudo-relevance feedback
  - expansion terms based on terms from top retrieved documents for initial query
- context vectors
  - represent words by the words that co-occur with them
  - 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, park, crustacean, wild, tropical

  - “aquarium” would be a highly ranked expansion term for the query “tropical fish” since the context vector of “aquarium” contains “tropical” and “fish”

other approaches

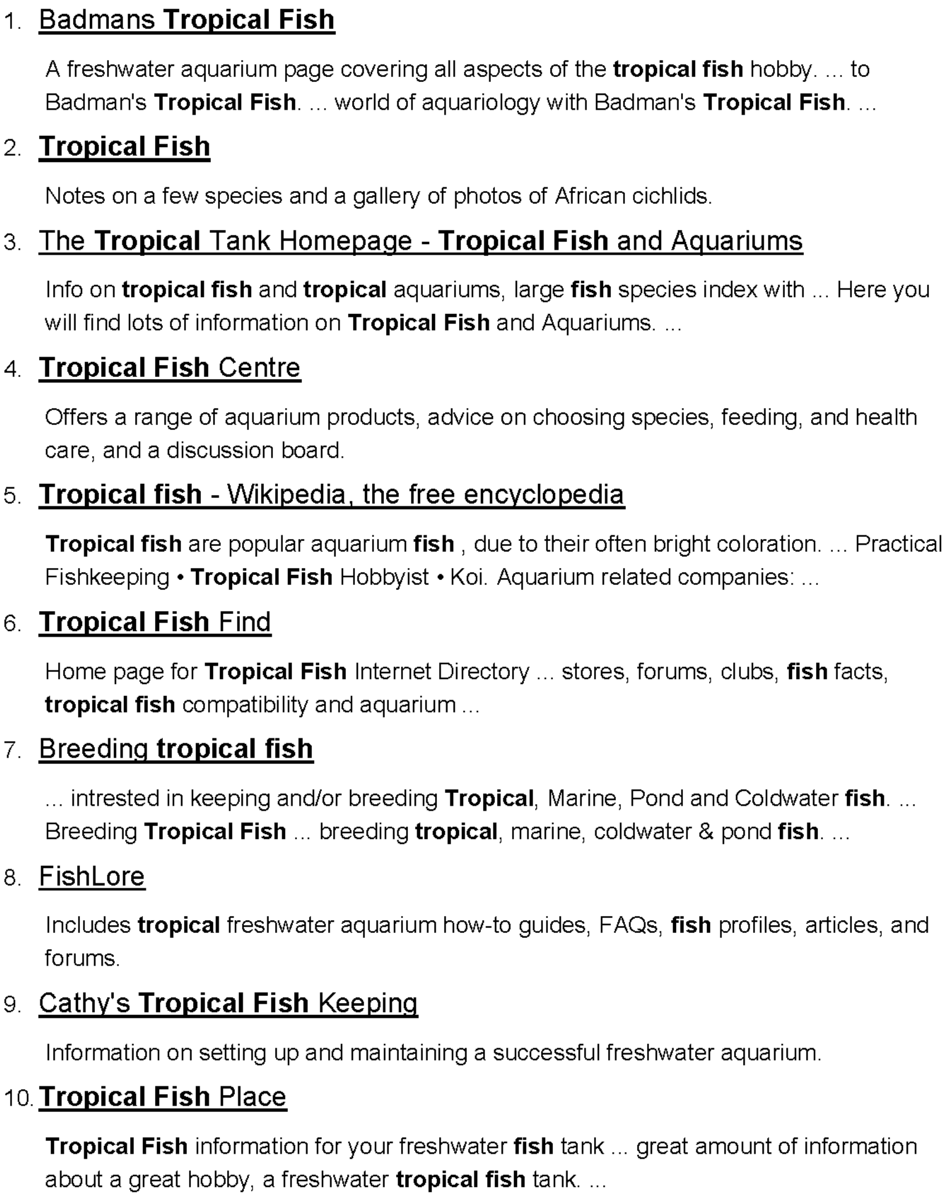
- query logs
  - best source of information about queries and related terms
    - short pieces of query text and click data
  - e.g., most frequent words in the queries containing “tropical fish” from MSN log:
    - stores, pictures, live, sale, types, clipart, blue, freshwater, aquarium, supplies
  - query suggestion based on finding similar queries rather than expansion terms

relevance feedback

- **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
  - but, very little training data
- pseudo-relevance feedback just assumes **top-ranked documents are relevant**
  - no user input
  - words that occur frequently in these documents may then be used to expand the initial query

relevance feedback example

top 10 documents for “tropical fish”



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. ...

relevance feedback example

- if we 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)

relevance feedback example

- 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)
- specific weights and scoring methods used for relevance feedback depend on retrieval model

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 don’t retrieve many relevant documents**
- some applications use relevance feedback
  - filtering, “more like this”
- query suggestion is more popular
  - may be less accurate, but can work if initial query fails

context and personalization

- 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 context) could have a significant **impact on relevance**
  - difficult to incorporate into ranking

user models

- generate **user profiles** based on documents that the person looks at
  - such as web pages visited, email messages, or word processing documents on the desktop
- modify queries using words from profile
- generally not effective
  - imprecise profiles, information needs can change significantly

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
- query history for individuals could be used for caching

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
- e.g.,
  - “흑룡강”

local search

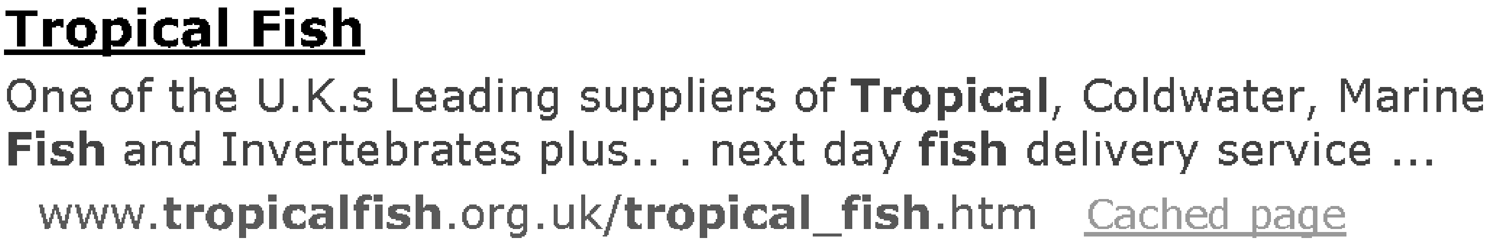
- identify the **geographic region associated with web pages**
  - use location metadata that has been manually added to the document
  - or identify locations such as place names, city names, or country names in text
- identify the **geographic region associated with the query**
  - 10-15% of queries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

extracting location information

- type of information extraction
  - ambiguity and significance of locations are issues
- location names are mapped to specific regions and coordinates

- matching done by inclusion, distance

snippet generation



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
- simple summarization approach
  - rank each sentence in a document using a significance factor
  - select the **top sentences for the summary**
  - first proposed by Luhn in 50's

sentence selection

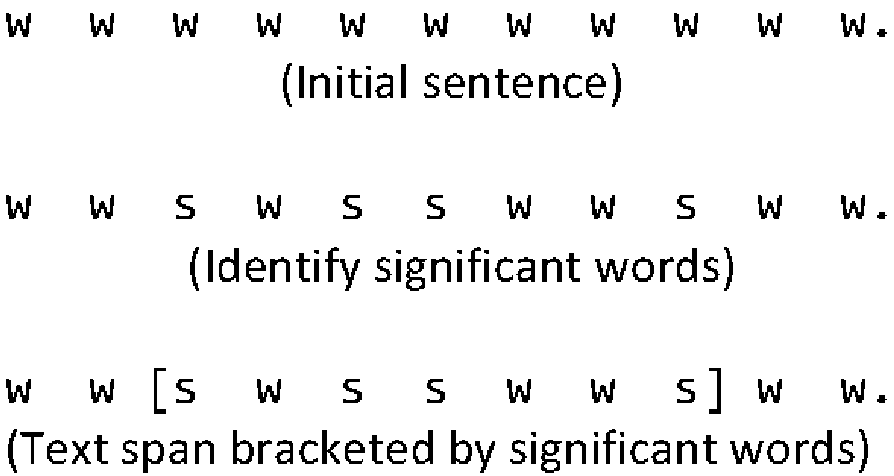
- significance factor for a sentence is calculated based on the **occurrence of significant words**
  - If $f_{d,w}$ is the frequency of word *w* in document *d*, then *w* is a significant word if it is **not a stopword** and

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

  - where s_d is the number of sentences in document *d*
  - text is bracketed by significant words (limit on number of non-significant words in bracket)

sentence selection

- 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
- e.g.,



```
w   w   w   w   w   w   w   w   w   w   w.
```

(Initial sentence)

```
w   w   s   w   s   s   w   w   s   w   w.
```

(Identify significant words)

```
w   w  [s   w   s   s   w   w   s]  w   w.
```

(Text span bracketed by significant words)

- significance factor = $4^2/7$ = 2.3

snippet generation

- 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

- 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
- snippets can be generated from text of pages like Wikipedia

snippet guidelines

- **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

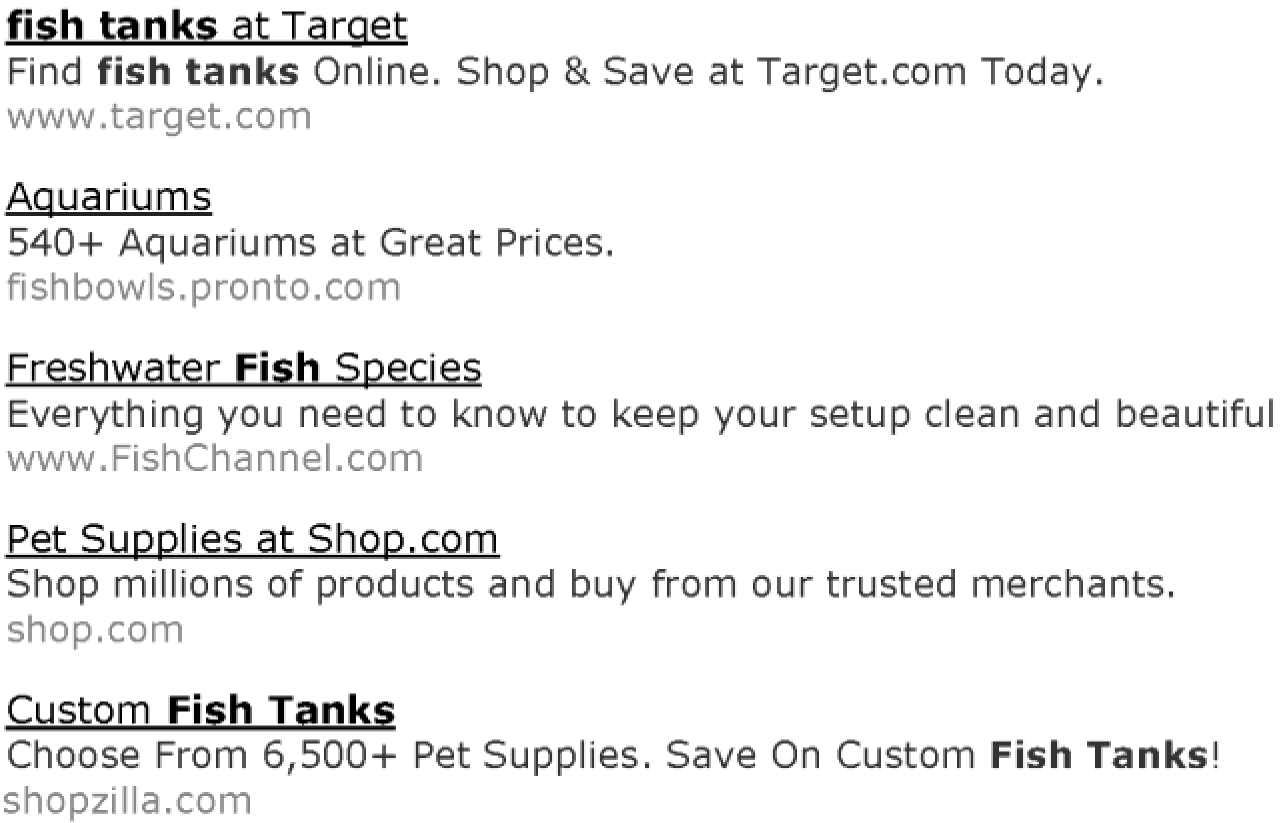
advertising

- 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

searching advertisements

- factors involved in ranking advertisements
  - similarity of text content to query
  - bids for keywords in query
  - popularity of advertisement
- small amount of text in advertisement
  - **dealing with vocabulary mismatch** is important
    - semantic similarity
  - expansion techniques are effective

example advertisements



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

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

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

advertisements retrieved for query “fish tank”

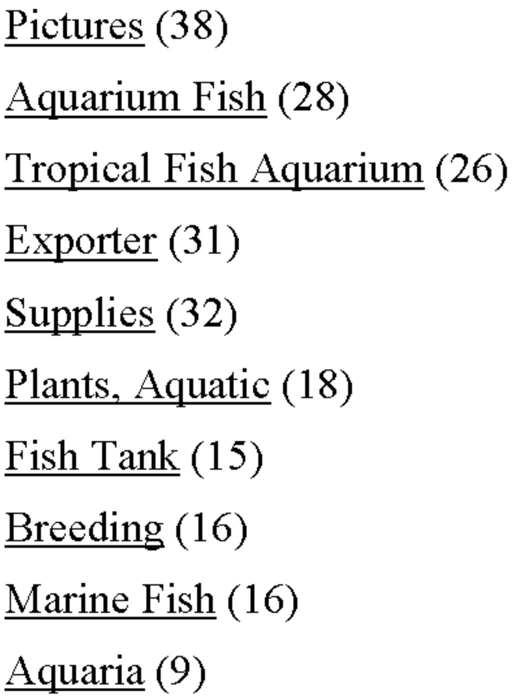
searching advertisements

- pseudo-relevance feedback
  - use ad text for pseudo-relevance feedback
- ranking of advertisements
  - rank **exact matches** first, followed by **stem matches**, followed by **expansion matches**
- query reformulation based on search sessions
  - learn associations between words and phrases based on co-occurrence in search sessions

clustering results

- result lists often contain documents related to different **aspects** of the query topic
- **clustering** is used to group related documents to simplify browsing

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)

requirements for clustering results

- efficiency
  - must be specific to each query and are based on the top-ranked documents for that query
  - typically based on snippets
- easy to understand
  - can be difficult to assign good labels to groups
  - monothetic vs. polythetic classification

types of classification

- monothetic
  - every member of a class has the property that defines the class
  - typical assumption made by users
  - easy to understand
- 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

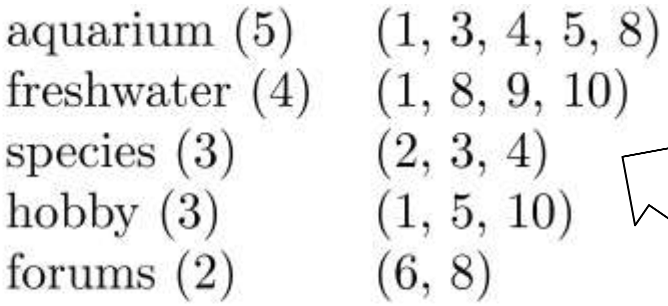
classification example

$$
\begin{aligned}
D_1 &= \{a, b, c\} \\
D_2 &= \{a, d, e\} \\
D_3 &= \{d, e, f, g\} \\
D_4 &= \{f, g\}
\end{aligned}
$$

- possible monothetic classification
  - – $\{D_1,D_2\}$ (labeled using *a*) and $\{D_2,D_3\}$ (labeled *e*)
- possible polythetic classification
  - – $\{D_2,D_3,D_4\}$, D_1
  - – labels?

result clusters

- simple algorithm
  - cluster documents based on the non-stopwords that occur in more than one snippets



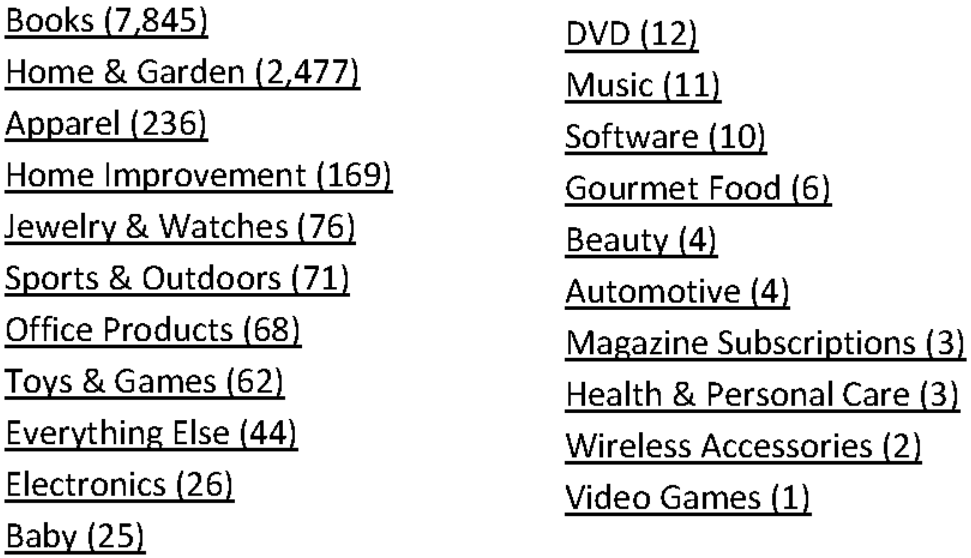
aquarium (5)	(1, 3, 4, 5, 8)
freshwater (4)	(1, 8, 9, 10)
species (3)	(2, 3, 4)
hobby (3)	(1, 5, 10)
forums (2)	(6, 8)

document numbers

- refinements
  - use phrases rather than words for clustering
  - 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

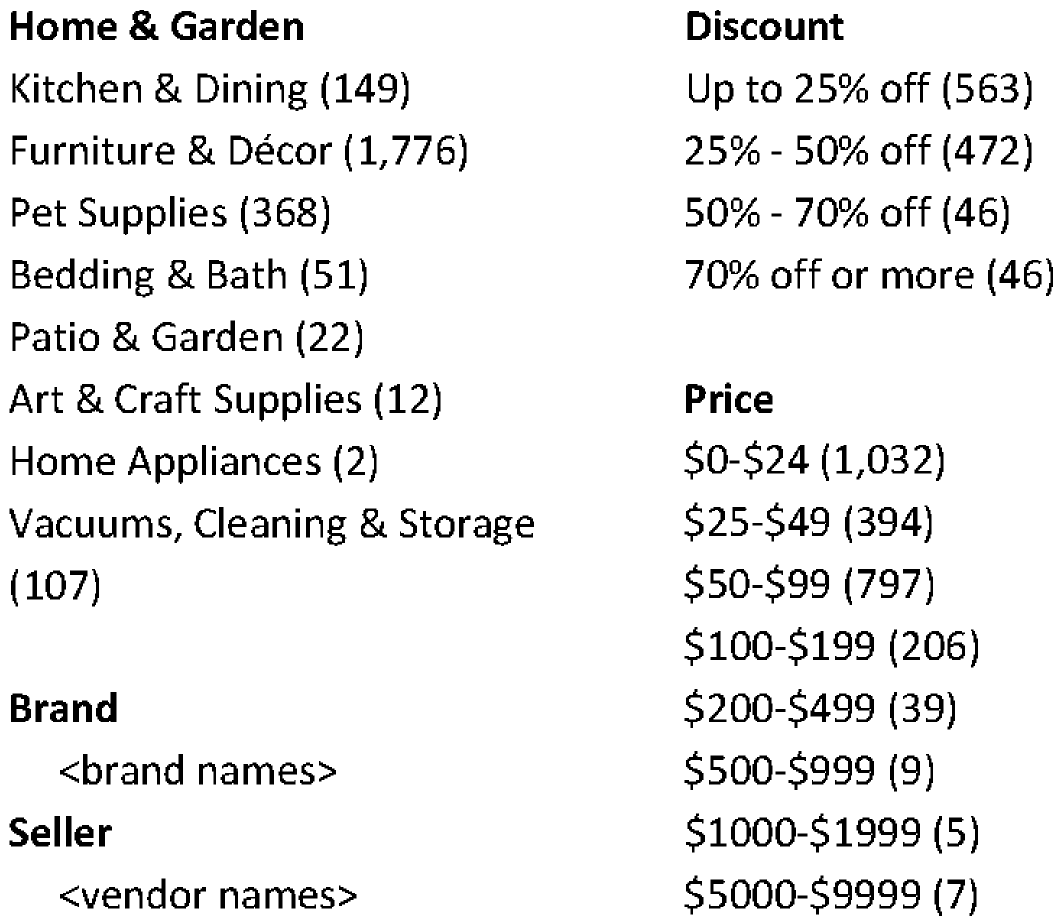
- consists of a **set of categories**, usually organized into a hierarchy, together with a set of facets that describe the important properties associated with the category
- manually defined
  - potentially less adaptable than dynamic classification
- easy to understand
  - commonly used in e-commerce



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 Subscriptions (3)
Health & Personal Care (3)
Wireless Accessories (2)
Video Games (1)

categories for "tropical fish"

example faceted classification



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)

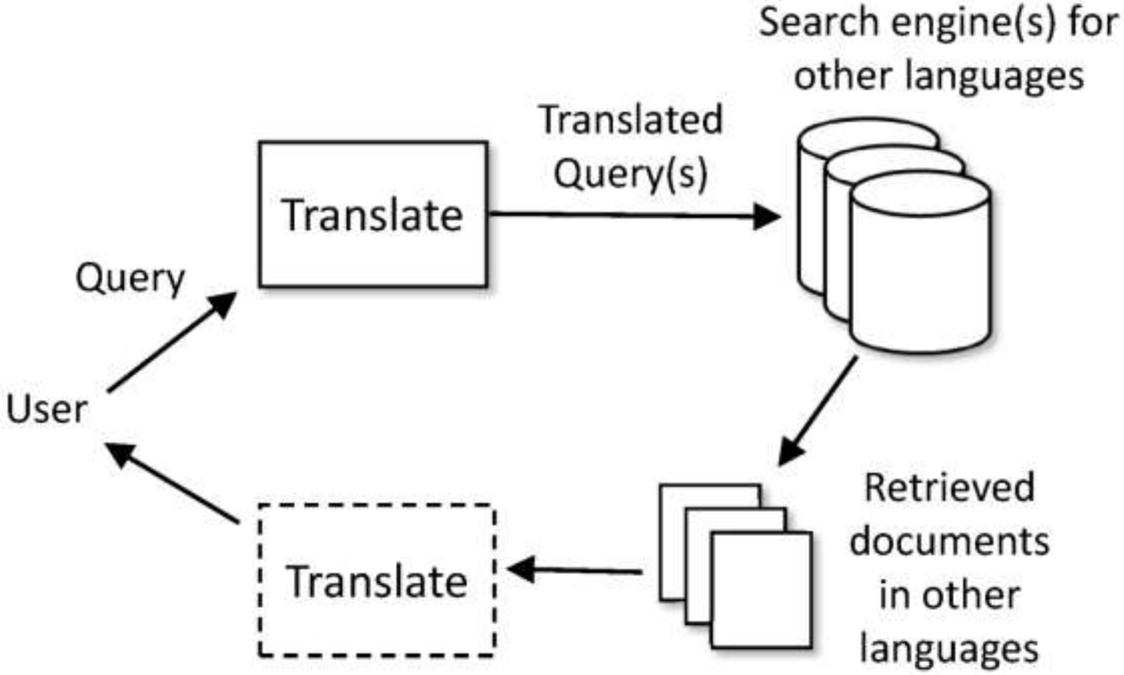
$5000-$9999 (7)

subcategories and facets for “Home & Garden”

cross-language search

- 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

cross-language search


Search engine(s) for other languages
Translated Query(s)
Translate
Query
User
Retrieved documents in other languages
Translate

statistical translation models

- models require **parallel corpora** for training
  - parallel corpora: collections of documents in one language together with the translations into one or more other languages
  - sentences in the parallel corpora are aligned, which means that sentences are **paired** with their translations
- translation of unusual words (e.g., proper names) and phrases is a problem
  - also use **transliteration** techniques: instead of translating, word is written in the characters of another language according to certain rules or based on similar sounds
  - e.g., Qathafi, Kaddafi, Qadafi, Gadafi, Gaddafi, Kathafi, Kadhafi, Qadhafi, Qazzafi, Kazafi, Qaddafy, Qadafy, Quadhaffi, Gadhdhafi, al-Qaddafi, Al-Qaddafi