Information Retrieval
INFO/CS 4300

- Instructor: Chris Buckley
  - Office hours Wednesdays 11am Gates 231
- Piazza will be the main communication tool
  - https://piazza.com/cornell/fall2014/info4300/home
  - Lecture notes will appear there.
  - TA office hours and locations appear there.

Course Admin

- Critique 1, Homework 1, Homework 2 – Graded, grades available on-line through CMS, hard copy can be picked up in the homework return room in Gates 216, open Mon-Fri noon-4pm. Hard copy for Homework 2 should be available Friday afternoon.
- Reminder of the Academic Code of Conduct - No copying! A few of you worked too closely with others; make sure what you turn in reflects **your** understanding of the problem. Just working too closely got just warnings this time.
- We also have one HW2 without a name; see Lu to claim (be prepared to show your Microsoft Word file or CMS submission).
- Project 1 – Due October 30.
- Critique 2 – Due today.

Previous Lectures

- Overview
- Evaluation 1
- Indexing
- Retrieval
  - Models
  - Weighting
  - Implementations
- TREC 1 - background, goals, and impact of TREC
- Evaluation 2
- TREC 2 – current research topics in TREC and other conferences

Today's Lecture

- Query Expansion
  - Thesaurus
  - Related Terms (Automatic)
  - Relevance Feedback
  - Pseudo-Relevance Feedback

Query Expansion

- Users typically supply very little of their information need, for good reasons
  - Not wanting to waste time
  - Not knowing their true information need
  - Not knowing what the system can make use of
  - Not knowing the vocabulary the collection documents use
- Can the system help, either with interaction or automatically?

The Thesaurus (either manual or automatic)

- Used in early search engines as a tool for indexing and query formulation
  - specified preferred terms and relationships between them
  - also called ***controlled vocabulary***
- 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 normally (domain thesaurus may be useful)
  - does not take context into account

Term Association Measures

- *Dice's Coefficient*

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

- *Mutual Information*

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

  - *N* number of text windows in the collection (documents, paragraphs)
  - *P(a)* probability that word *a* occurs in a given window of text
  - *P(a,b)* probability that *a* and *b* occur in the same window of text
  - Measures the extent to which 2 words occur independently

Term Association Measures

- Mutual Information measure favors low frequency terms
- ***Expected Mutual Information Measure*** (EMIM)

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

  - actually only 1 part of full EMIM, focused on word occurrence

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
  - normalizes this comparison by the expected number
  - also limited form focused on word co-occurrence

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

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.

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
mollic	catch	mollic	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

- 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 all possible queries, other approaches used to achieve this effect

Other Approaches

- Pseudo-relevance feedback
  - expansion terms based on top retrieved documents for initial query
  - Discussed shortly
- 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
  - Rank words for a query by ranking context vectors

Other Approaches

- Query logs
  - Best source of information about queries and related terms
    - short pieces of text and click data
  - e.g., most frequent words in queries containing "tropical fish" from MSN log:

    stores, pictures, live, sale, types, clipart, blue, freshwater, aquarium, supplies
  - query suggestion based on finding similar queries
    - group based on click data

Relevance Feedback

- User identifies relevant (and maybe non-relevant) documents in the initial result list
- System modifies query using terms from those documents and reranks documents
  - example of simple machine learning algorithm using training data
  - but, very little training data

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"

Relevance Feedback Example

- If document 7 ("Breeding tropical fish") is *explicitly* indicated to be relevant, the most frequent terms are:

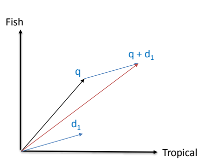
  breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2), keeping (1), interested (1)
- Specific weights and scoring methods used for relevance feedback depend on retrieval model

Rocchio Feedback

- Originally defined (1960's) in the vector space model.

Given query q and relevant document d_1 move the new query in the direction of d_1 by (weighted) vector addition. Similarly, given a non-relevant document d_2, move the query away from d_2.

$$Qnew = A * Qold + B * Drel/|Drel| - C * Dnonrel/|Dnonrel|$$



Rocchio Feedback

- Expressed in term weights:

$$q_j{}' = A * q_j + B * \frac{1}{|Rel|} * \sum_{D_i \text{ in } Rel} d_{ij} - C * \frac{1}{|Nonrel|} * \sum_{D_i \text{ in } Nonrel} d_{ij}$$

  - where typically
    - *A=8, B=16, C=4*
    - The set of Non-relevant documents is the entire collection
    - Only the 20-50 terms in the relevant documents are added
- Assumes query weights and document weights are commensurate

Rocchio Feedback – Which terms to add?

- Some of the same techniques for query expansion can be used
  - *EMIM*, χ^2
  - In practice, want many of the same features as are wanted for term weighting
    - Very common terms are not helpful
    - Terms occurring once in document often not helpful
- Use the 20-30 terms with highest average weights in the relevant documents works well

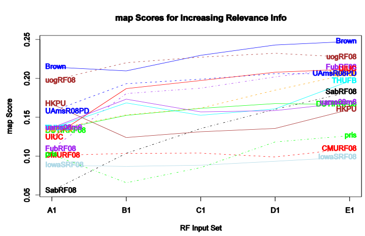
Overall Relevance Feedback

- Long history of success
  - Rocchio algorithm was first proposed in 1965
  - Still works on large collections
- Automatic RF has concentrated on statistical approaches
  - Vector Space
  - Probabilistic
  - Probabilistic dependency
  - Machine learning / classification
  - Language modeling

TREC 2008 Results: MAP on 31 topics

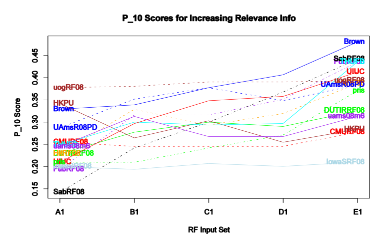
A1: baseline
B1: 1 rel doc (same all systems)
C1: 3 rel docs, 3 nonrel docs
D1: 10 docs (>= 3 rel docs)
E1: all previously judged docs

Terabyte collection (25 million documents)



TREC 2008: P(10) on 31 Topics

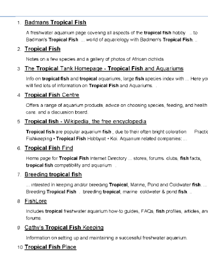
Most groups found between 1 and 1 ½ more relevant docs in the top 10 with RF. Sabir found 3 more, but with a low baseline.



Relevance Feedback Evaluation

- Uses standard measures, but…
  - What do you do with the documents already seen (the RF input)?
    - Unfair to have their improvement influence the measure
  - Several solutions (and many debates) in the past
- Residual collection evaluation is now used
  - Remove all seen docs from the collection (relevance judgments and runs) for both the original pre-feedback run and the feedback run.
  - Previous graphs had the Set E documents removed from all runs.

Relevance Feedback – can we do more?



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"

Pseudo-Relevance Feedback Example

- Pseudo-relevance feedback just assumes top-ranked documents are relevant – no user input
- 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)

LMs for Retrieval

- 3 possibilities:
  - probability of generating the query text from a document language model Query-Likelihood Model
  - probability of generating the document text from a query language model Difficult to use in practice
  - **comparing the language models representing the query and document topics** Let's explore this

Pseudo-Relevance Feedback – Language Model

- Estimate relevance model from query and top-ranked documents
- Rank documents by similarity of document model to relevance model
- ***Kullback-Leibler divergence*** (KL-divergence) is a well-known measure of the difference between two probability distributions

KL-Divergence

- Given the ***true*** probability distribution *P* and another distribution *Q* that is an ***approximation*** to *P*,

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

  - Use negative KL-divergence for ranking, and assume relevance model *R* is the true distribution (not symmetric),

$$\sum_{w \in V} P(w|R) \log P(w|D) - \sum_{w \in V} P(w|R) \log P(w|R)$$

KL-Divergence

- Given a simple maximum likelihood estimate for *P(w|R),* based on the frequency in the query text, ranking score is

$$\sum_{w \in V} \frac{f_{w,Q}}{|Q|} \log P(w|D)$$

  - rank-equivalent to query likelihood score
- Query likelihood model is a special case of retrieval based on relevance model

Estimating the Relevance Model

- Probability of pulling a word *w* out of the "bucket" representing the relevance model depends on the *n* query words we have just pulled out

- By definition

$$P(w|R) \approx P(w|q_1 \ldots q_n)$$

$$P(w|R) \approx \frac{P(w, q_1 \ldots q_n)}{P(q_1 \ldots q_n)}$$

Estimating the Relevance Model

- Joint probability is

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} p(D) P(w, q_1 \dots q_n | D)$$

- Assume

$$P(w, q_1 \dots q_n | D) = P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

- Gives

$$P(w, q_1 \dots q_n) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \prod_{i=1}^{n} P(q_i|D)$$

Estimating the Relevance Model

- *P(D)* usually assumed to be uniform
- *P(w, q_1 . . . q_n)* is simply a weighted average of the language model probabilities for *w* in a set of documents, where the weights are the query likelihood scores for those documents
- Formal model for pseudo-relevance feedback
  - query expansion technique

Ranking based on the Relevance Model

1. Rank documents using the query likelihood score for query Q.
2. Select some number of the top-ranked documents to be the set $\mathcal{C}$.
3. Calculate the relevance model probabilities $P(w|R)$.
4. Rank documents again using the KL-divergence score

$$\sum_w P(w|R) \log P(w|D)$$

Example from Top 10 Docs

president lincoln	*abraham lincoln*	*fishing*	*tropical fish*
lincoln	lincoln	fish	fish
president	america	farm	tropic
room	president	salmon	japan
bedroom	faith	new	aquarium
house	guest	wild	water
white	abraham	water	species
america	new	caught	aquatic
guest	room	catch	fair
serve	christian	tag	china
bed	history	time	coral
washington	public	eat	source
old	bedroom	raise	tank
office	war	city	reef
war	politics	people	animal
long	old	fishermen	tarpon
abraham	national	boat	fishery

Example from Top 50 Docs

president lincoln	*abraham lincoln*	*fishing*	*tropical fish*
lincoln	lincoln	fish	fish
president	president	water	tropic
america	america	catch	water
new	abraham	reef	storm
national	war	fishermen	species
great	man	river	boat
white	civil	new	sea
war	new	year	river
washington	history	time	country
clinton	two	bass	tuna
house	room	boat	world
history	booth	world	million
time	time	farm	state
center	politics	angle	time
kennedy	public	fly	japan
room	guest	trout	mile

Relevance Feedback Impact

- 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
  - Pure relevance feedback is reliable, but not user-interface friendly
    - Perhaps voice interfaces might help in the future?
- Some applications use a form of relevance feedback
  - filtering, "more like this"
- Query suggestion more popular
  - may be less accurate, but can work if initial query fails