
Natural Language Processing and Information Retrieval

Performance Evaluation
Query Expansion

Alessandro Moschitti

Department of Computer Science and Information
Engineering
University of Trento
Email: moschitti@disi.unitn.it



Measures for a search engine

- How fast does it index
 - Number of documents/hour
 - (Average document size)
- How fast does it search
 - Latency as a function of index size
- Expressiveness of query language
 - Ability to express complex information needs
 - Speed on complex queries
- Uncluttered UI
- Is it free?



Measures for a search engine

- All of the preceding criteria are *measurable*: we can quantify speed/size
 - we can make expressiveness precise
- The key measure: user happiness
 - What is this?
 - Speed of response/size of index are factors
 - But blindingly fast, useless answers won't make a user happy
- Need a way of quantifying user happiness



Measuring user happiness

- Issue: who is the user we are trying to make happy?
 - Depends on the setting
- Web engine:
 - User finds what s/he wants and returns to the engine
 - Can measure rate of return users
 - User completes task – search as a means, not end
 - See Russell <http://dmrussell.googlepages.com/JCDL-talk-June-2007-short.pdf>
- eCommerce site: user finds what s/he wants and buys
 - Is it the end-user, or the eCommerce site, whose happiness we measure?
 - Measure time to purchase, or fraction of searchers who become buyers?



Measuring user happiness

- Enterprise (company/govt/academic): Care about “user productivity”
 - How much time do my users save when looking for information?
 - Many other criteria having to do with breadth of access, secure access, etc.



Happiness: elusive to measure

- Most common proxy: *relevance* of search results
- But how do you measure relevance?
- We will detail a methodology here, then examine its issues
- Relevance measurement requires 3 elements:
 1. A benchmark document collection
 2. A benchmark suite of queries
 3. A usually binary assessment of either Relevant or Nonrelevant for each query and each document
 - Some work on more-than-binary, but not the standard



Evaluating an IR system

- Note: the **information need** is translated into a **query**
- Relevance is assessed relative to the **information need** *not* the **query**
- E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- Query: **wine red white heart attack effective**
- Evaluate whether the doc addresses the information need, not whether it has these words



Standard relevance benchmarks

- TREC - National Institute of Standards and Technology (NIST) has run a large IR test bed for many years
- Reuters and other benchmark doc collections used
- “Retrieval tasks” specified
 - sometimes as queries
- Human experts mark, for each query and for each doc, Relevant or Nonrelevant
 - or at least for subset of docs that some system returned for that query



Unranked retrieval evaluation:

Precision and Recall

- **Precision:** fraction of retrieved docs that are relevant
= $P(\text{relevant} | \text{retrieved})$
- **Recall:** fraction of relevant docs that are retrieved
= $P(\text{retrieved} | \text{relevant})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision $P = tp / (tp + fp)$
- Recall $R = tp / (tp + fn)$



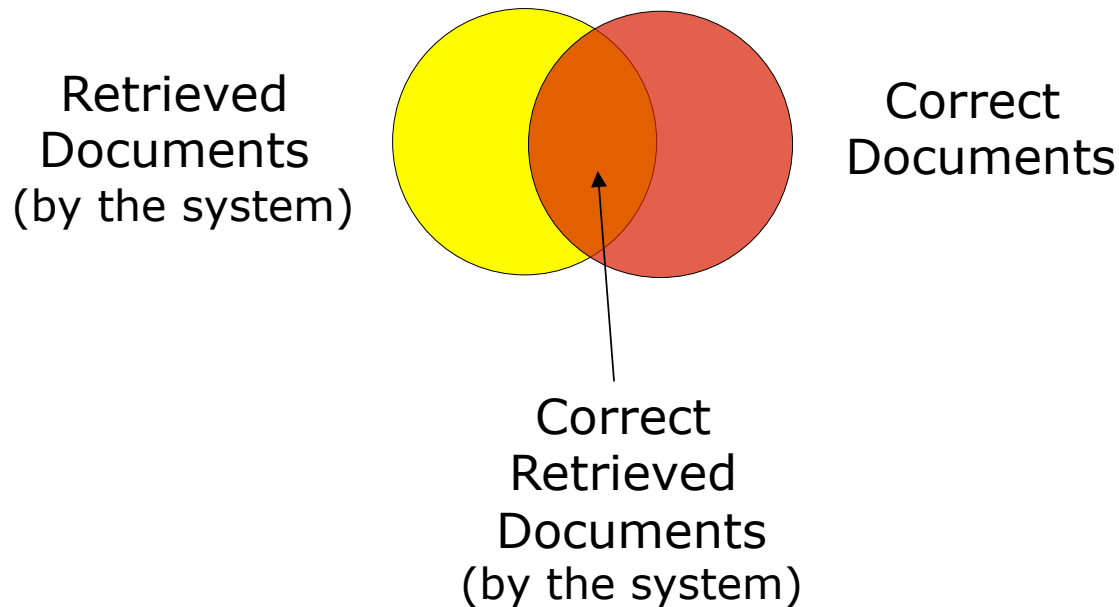
Should we instead use the accuracy measure for evaluation?

- Given a query, an engine classifies each doc as “Relevant” or “Nonrelevant”
- The **accuracy** of an engine: the fraction of these classifications that are correct
 - $(tp + tn) / (tp + fp + fn + tn)$
- **Accuracy** is a evaluation measure in often used in machine learning classification work
- Why is this not a very useful evaluation measure in IR?



Performance Measurements

- Given a set of document T
- Precision = # Correct Retrieved Document / # Retrieved Documents
- Recall = # Correct Retrieved Document / # Correct Documents



Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget....



snoogle.com

Search for:

0 matching results found.

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

Precision/Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved
- In a good system, precision decreases as either the number of docs retrieved or recall increases
 - This is not a theorem, but a result with strong empirical confirmation



Difficulties in using precision/recall

- Should average over large document collection/query ensembles
- Need human relevance assessments
 - People aren't reliable assessors
 - **Complete Oracle (CO)**
- Assessments have to be binary
 - Nuanced assessments?
- Heavily skewed by collection/authorship
 - Results may not translate from one domain to another



A combined measure: F

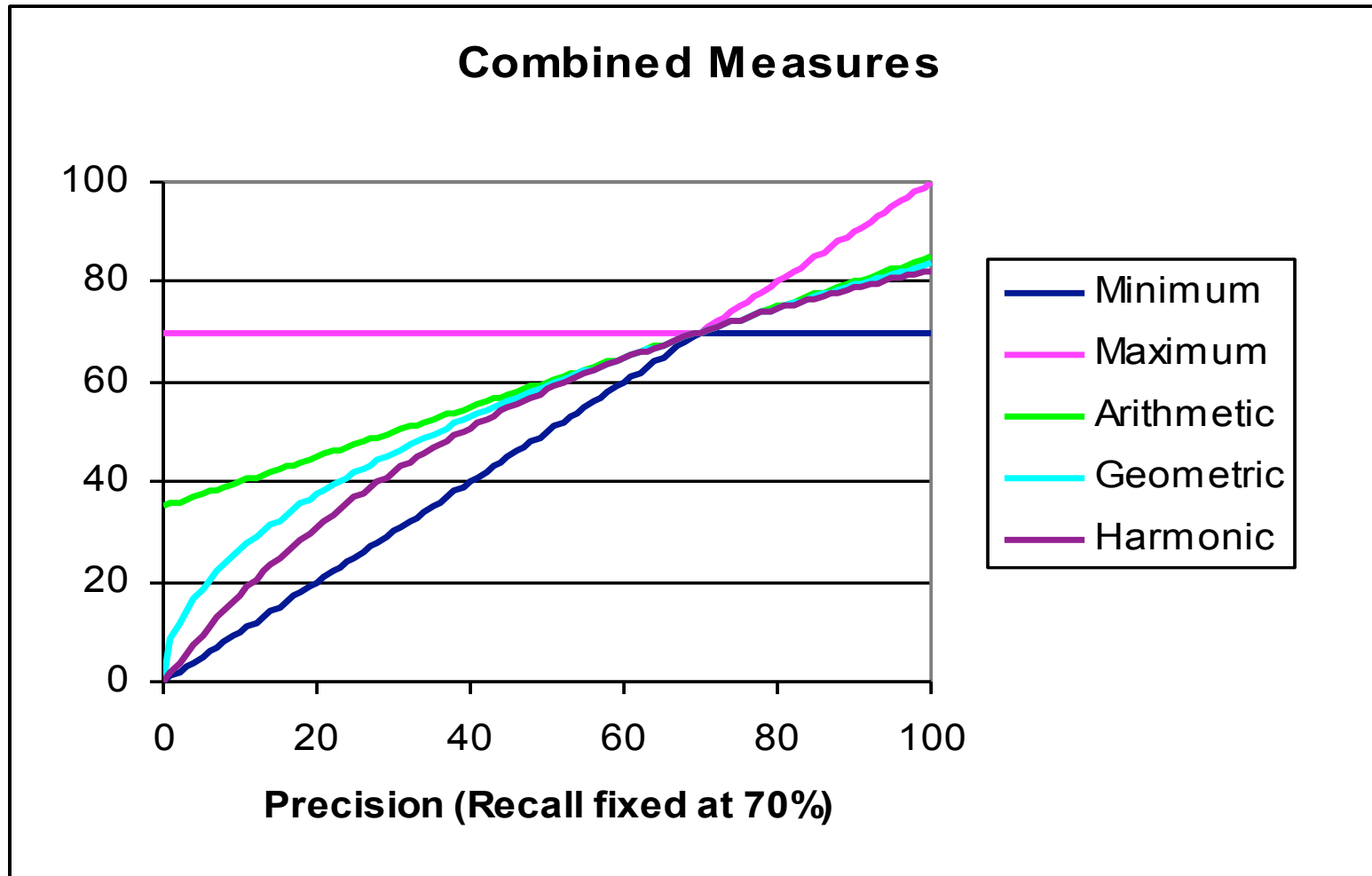
- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
 - i.e., with $\beta = 1$ or $\alpha = \frac{1}{2}$
- Harmonic mean is a conservative average
 - See CJ van Rijsbergen, *Information Retrieval*



F_1 and other averages

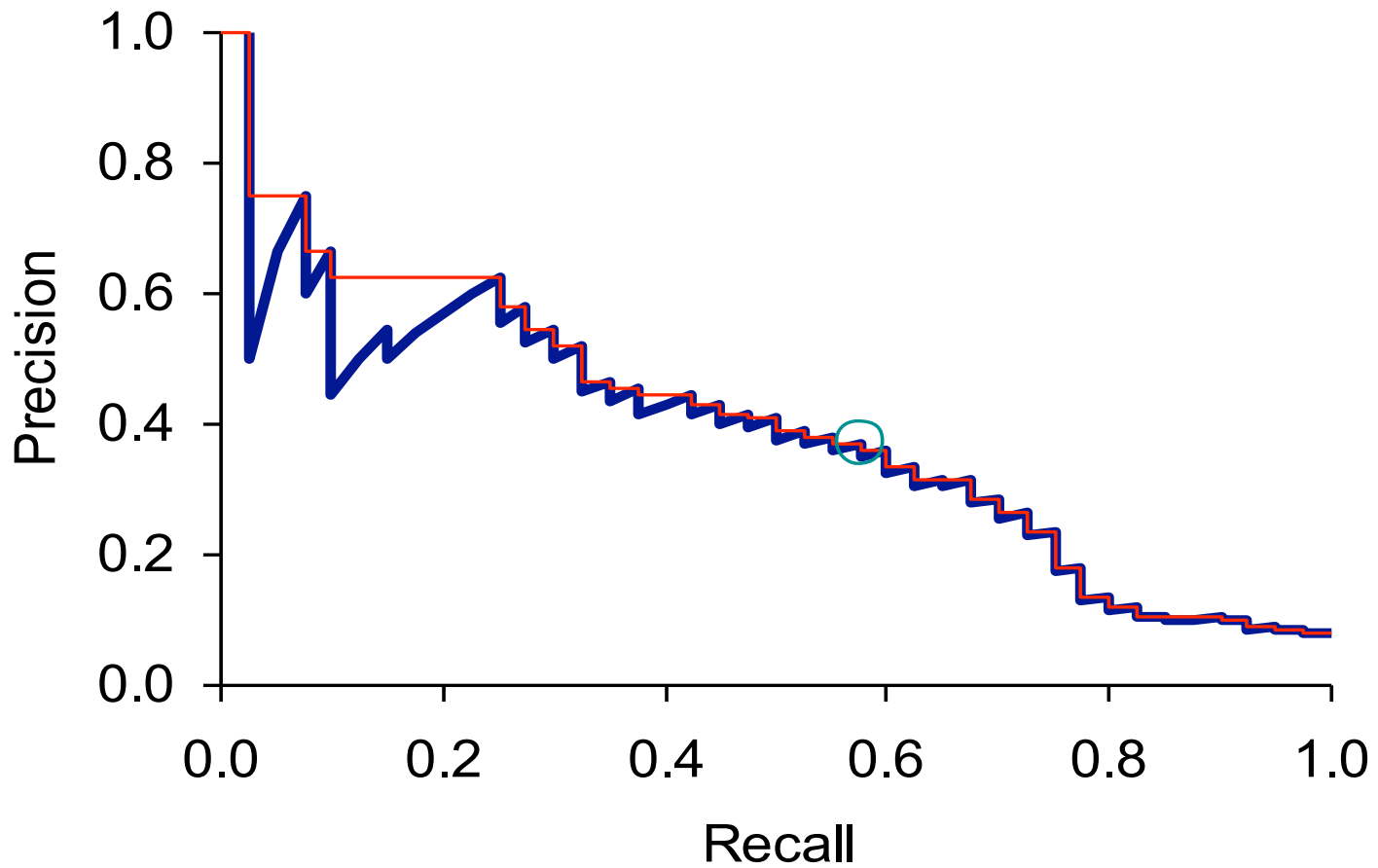


Evaluating ranked results

- Evaluation of ranked results:
 - The system can return any number of results
 - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*



A precision-recall curve



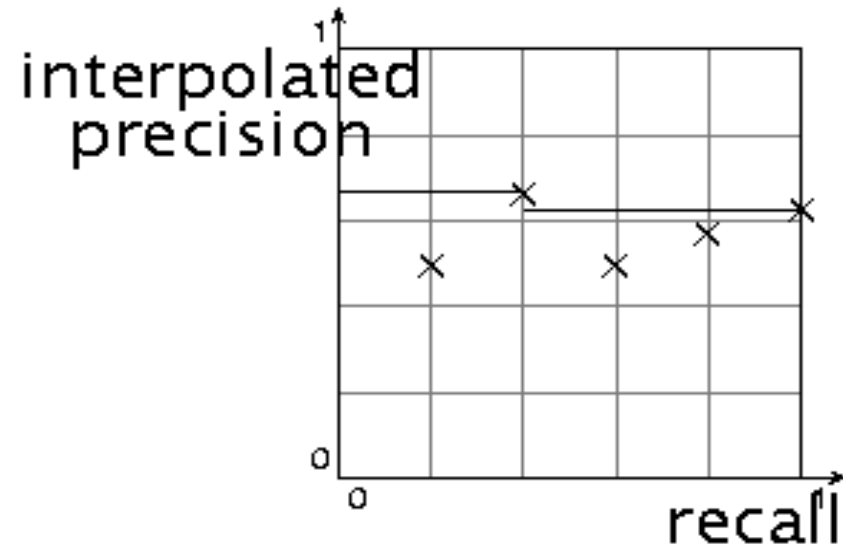
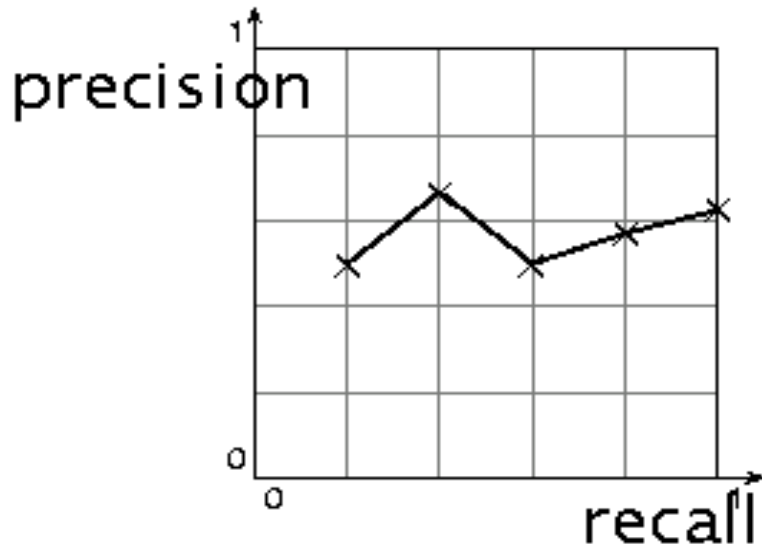
Averaging over queries

- A precision-recall graph for one query isn't a very sensible thing to look at
- You need to average performance over a whole bunch of queries.
- But there's a technical issue:
 - Precision-recall calculations place some points on the graph
 - How do you determine a value (interpolate) between the points?



Interpolated precision

- Idea: If locally precision increases with increasing recall, then you should get to count that...
- So you take the max of precisions to right of value



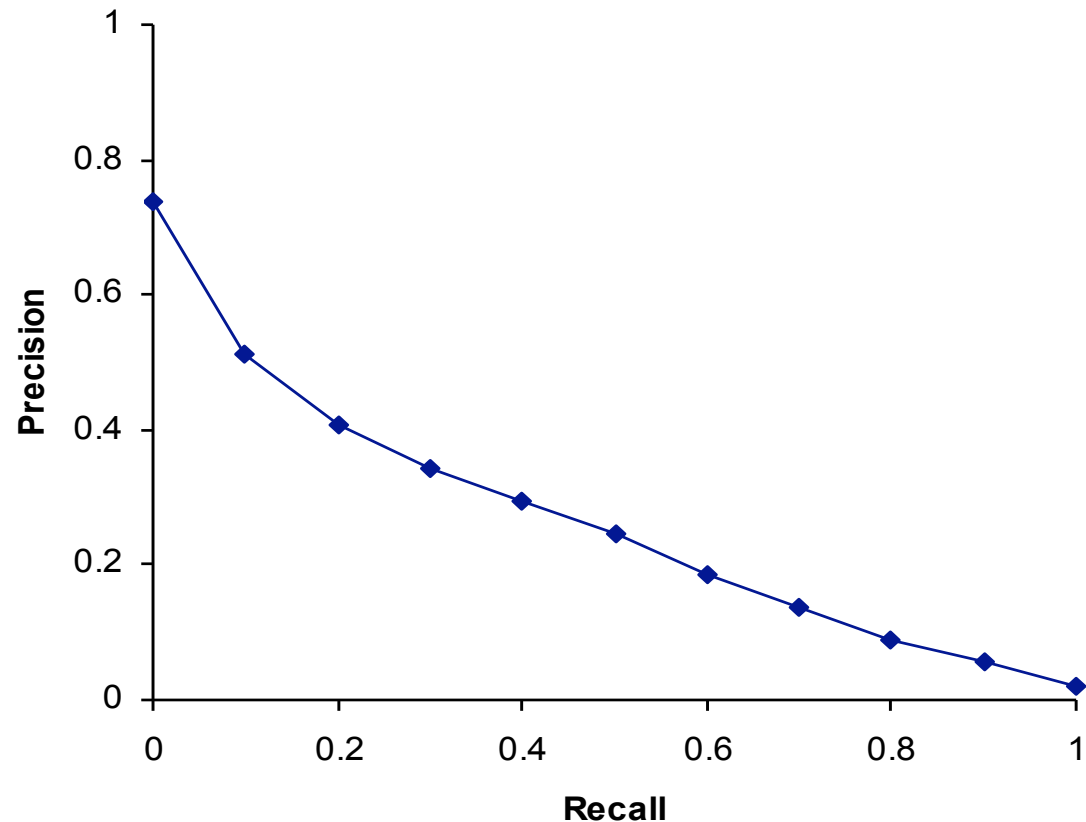
Evaluation

- Graphs are good, but people want summary measures!
 - Precision at fixed retrieval level (**no CO**)
 - ◆ Precision-at- k : Precision of top k results
 - ◆ Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages
 - ◆ But: averages badly and has an arbitrary parameter of k
 - 11-point interpolated average precision (**CO**)
 - ◆ The standard measure in the early TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (**the value for 0 is always interpolated!**), and average them
 - ◆ Evaluates performance at all recall levels



Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



Yet more evaluation measures...

- Mean average precision (MAP) (**no CO**)
 - Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved
 - Avoids interpolation, use of fixed recall levels
 - MAP for query collection is arithmetic ave.
 - ◆ Macro-averaging: each query counts equally
- R-precision (**no CO – just R relevant documents**)
 - If we have a known (though perhaps incomplete) set of relevant documents of size Rel , then calculate precision of the top Rel docs returned
 - Perfect system could score 1.0.



Variance

- For a test collection, it is usual that a system does crummily on some information needs (e.g., MAP = 0.1) and excellently on others (e.g., MAP = 0.7)
- Indeed, it is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query.
- That is, there are easy information needs and hard ones!



CREATING TEST COLLECTIONS FOR IR EVALUATION



Test Collections

TABLE 4.3 Common Test Corpora

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000



From document collections to test collections

- Still need
 - Test queries
 - Relevance assessments
- Test queries
 - Must be germane to docs available
 - Best designed by domain experts
 - Random query terms generally not a good idea
- Relevance assessments
 - Human judges, time-consuming
 - Are human panels perfect?



Kappa measure for inter-judge (dis)agreement

- Kappa measure
 - Agreement measure among judges
 - Designed for categorical judgments
 - Corrects for chance agreement
- $\text{Kappa} = [P(A) - P(E)] / [1 - P(E)]$
- $P(A)$ – proportion of time judges agree
- $P(E)$ – what agreement would be by chance
- Kappa = 0 for chance agreement, 1 for total agreement.



• $P(A)$? $P(E)$?

Kappa Measure: Example

Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant



Kappa Example

- $P(A) = 370/400 = 0.925$
- $P(\text{nonrelevant}) = (10+20+70+70)/800 = 0.2125$
- $P(\text{relevant}) = (10+20+300+300)/800 = 0.7878$
- $P(E) = 0.2125^2 + 0.7878^2 = 0.665$
- $\text{Kappa} = (0.925 - 0.665)/(1-0.665) = 0.776$

- $\text{Kappa} > 0.8$ = good agreement
- $0.67 < \text{Kappa} < 0.8$ -> “tentative conclusions” (Carletta '96)
- Depends on purpose of study
- For >2 judges: average pairwise kappas



TREC

- TREC Ad Hoc task from first 8 TRECs is standard IR task
 - 50 detailed information needs a year
 - Human evaluation of pooled results returned
 - More recently other related things: Web track, HARD

- A TREC query (TREC 5)

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>



Standard relevance benchmarks: Others

- GOV2
 - Another TREC/NIST collection
 - 25 million web pages
 - Largest collection that is easily available
 - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
 - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
 - This evaluation series has concentrated on European languages and cross-language information retrieval.
- Many others



Impact of Inter-judge Agreement

- Impact on **absolute** performance measure can be significant (0.32 vs 0.39)
- Little impact on ranking of different systems or **relative** performance
- Suppose we want to know if algorithm A is better than algorithm B
- A standard information retrieval experiment will give us a reliable answer to this question.



Critique of pure relevance

- Relevance vs Marginal Relevance
 - A document can be redundant even if it is highly relevant
 - Duplicates
 - The same information from different sources
 - Marginal relevance is a better measure of utility for the user.
- Using facts/entities as evaluation units more directly measures true relevance.
- But harder to create evaluation set
- See Carbonell reference



Can we avoid human judgment?

- No
- Makes experimental work hard
 - Especially on a large scale
- In some very specific settings, can use proxies
 - E.g.: for approximate vector space retrieval, we can compare the cosine distance closeness of the closest docs to those found by an approximate retrieval algorithm
- But once we have test collections, we can reuse them (so long as we don't overtrain too badly)



Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k , e.g., $k = 10$
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
 - NDCG (Normalized Cumulative Discounted Gain)
- Search engines also use non-relevance-based measures.
 - Clickthrough on first result
 - ◆ Not very reliable if you look at a single clickthrough ... but pretty reliable in the aggregate.
 - Studies of user behavior in the lab
 - A/B testing



A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most
- In principle less powerful than doing a multivariate regression analysis, but easier to understand



RESULTS PRESENTATION



Result Summaries

- Having ranked the documents matching a query, we wish to present a results list
- Most commonly, a list of the **document titles** plus a **short summary**, aka “10 blue links”

[John McCain](#)

John McCain 2008 - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.johnmccain.com · [Cached page](#)

[JohnMcCain.com - McCain-Palin 2008](#)

John McCain 2008 - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.johnmccain.com/Informing/Issues · [Cached page](#)

[John McCain News- msnbc.com](#)

Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...
www.msnbc.msn.com/id/16438320 · [Cached page](#)

[John McCain | Facebook](#)

Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...
www.facebook.com/johnmccain · [Cached page](#)



Summaries

- The title is often automatically extracted from document metadata. What about the summaries?
 - This description is crucial.
 - User can identify good/relevant hits based on description.
- Two basic kinds:
 - Static
 - Dynamic
- A **static summary** of a document is always the same, regardless of the query that hit the doc
- A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand




Static summaries


- In typical systems, the static summary is a subset of the document
- Simplest heuristic: the first 50 (or so – this can be varied) words of the document
 - Summary cached at indexing time
- More sophisticated: extract from each document a set of “key” sentences
 - Simple NLP heuristics to score each sentence
 - Summary is made up of top-scoring sentences.
- Most sophisticated: NLP used to synthesize a summary
 - Seldom used in IR; cf. text summarization work




Dynamic summaries

- Present one or more “windows” within the document that contain several of the query terms
 - “KWIC” snippets: Keyword in Context presentation

 [Christopher Manning, Stanford NLP](http://nlp.stanford.edu/~manning/)
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University.
nlp.stanford.edu/~manning/ - 12k - [Cached](#) - [Similar pages](#)

 [Christopher Manning, Stanford NLP](http://nlp.stanford.edu/~manning/)
Christopher Manning, Associate Professor of Computer Science and Linguistics, ... computational semantics, **machine translation**, grammar induction, ...
nlp.stanford.edu/~manning/ - 12k - [Cached](#) - [Similar pages](#)

 [Christopher Manning, Stanford NLP](http://nlp.stanford.edu/~manning/)
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University ... **Chris Manning** works on systems and formalisms that can ...
nlp.stanford.edu/~manning - [Cached](#)



Techniques for dynamic summaries

- Find small windows in doc that contain query terms
 - Requires fast window lookup in a document cache
- Score each window wrt query
 - Use various features such as window width, position in document, etc.
 - Combine features through a scoring function
- Challenges in evaluation: judging summaries
 - Easier to do pairwise comparisons rather than binary relevance assessments



Quicklinks

- For a *navigational query* such as **united airlines** user's need likely satisfied on www.united.com
- Quicklinks provide navigational cues on that home page



Web [+ Show options...](#)

[United Airlines Flights](#)
www.OneTravel.com/United-Airlines Save \$10 Instantly on **United Airlines** Airfares.

[United Airlines - Airline Tickets, Airline Reservations, Flight ...](#)
Airline tickets, **airline** reservations, flight airfare from **United Airlines**. Online reservation **airline** ticket purchase, electronic tickets, flight search, ... [+ Show stock quote for UUA](#)
www.united.com/ - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)

Search options	Baggage
EasyCheck-in Online	Services & information
Mileage Plus	Itineraries & check-in
My itineraries	Planning & booking

[More results from united.com »](#)



united airlines

Search Pad

SearchScan - On

102,000,000 results for united airlines:

Show All

United Air Lines

Wikipedia

Also try: [united airlines reservations](#), [united airlines flight](#), [More...](#)

[United Airlines - Airline Tickets, Airline Reservations ...](#) (Nasdaq: [UAUA](#))

Official site for **United Airlines**, commercial air carrier transporting people, property, and mail across the U.S. and worldwide.

[www.united.com](#) - 65k - [Cached](#)

[Planning & Booking](#)

[Shop for Flights](#)

[Itineraries & Check-in](#)

[Special Deals](#)

[Mileage Plus](#)

[Flight Status](#)

[Services & Information](#)

[Customer Service](#)

[more results from united.com »](#)

united airlines

UNITED AIRLINES

[United Airline Fleet](#)

[United Airline Schedule](#)

[United Airlines Reservations](#)

[United Airline Jobs](#)

[Reference](#)

ALL RESULTS

[Cheap Flight Tickets](#) · [www.CheapOair.com](#)

CheapOair - The Only Way to Go!! Find Over 18 Million Exclusive Fares.

[Fly United Airlines](#) · [www.OneTravel.com/United-Airline](#)

Save \$10 Instantly on **United Airlines** Flights. Book Now, Hurry!

Best match

[United Airlines - Airline Tickets, Airline Reservations, Flight ...](#)

[www.united.com](#) · Official site

Airline tickets, **airline** reservations, flight airfare from **United Airlines**. Online reservations, **airline** ticket purchase, electronic tickets, flight search, fares and availability ...

[Flights](#)

[Redeem miles](#)

[Check In Online](#)

[Children, pets, & assistance](#)

[My itineraries](#)

[Change your travel plans](#)

[Baggage](#)

[Special deals](#)

Customer service 800-864-8331

RELATED SEARCHES

[United Airlines Flight Status](#)

[US Airways](#)

[Continental Airlines](#)




Alternative results presentations?

YAHOO!

Web Images Video Local Shopping News more ▾

uni Search

- united airlines** ▶
- univision
- university of phoenix
- asian unicorn
- universal studios
- united states postal service
- united healthcare

UNITED AIRLINES - AIRLINE TICKETS,...
Airline tickets, airline reservations, flight airfare from United Airlines.
Online reservations, ...
 www.united.com

MORE INFO

Flights	Check In Online
Mileage Plus	My Itineraries
Baggage	Redeem Miles



Resources for this lecture

- IIR 8
- Carbonell and Goldstein 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR 21.



Relevance feedback

- We will use *ad hoc retrieval* to refer to regular retrieval without relevance feedback.
- We now look at four examples of relevance feedback that highlight different aspects.



Similar pages

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

[Web](#) [Video](#) [Music](#)

[Sarah Brightman Official Website - Home Page](#)

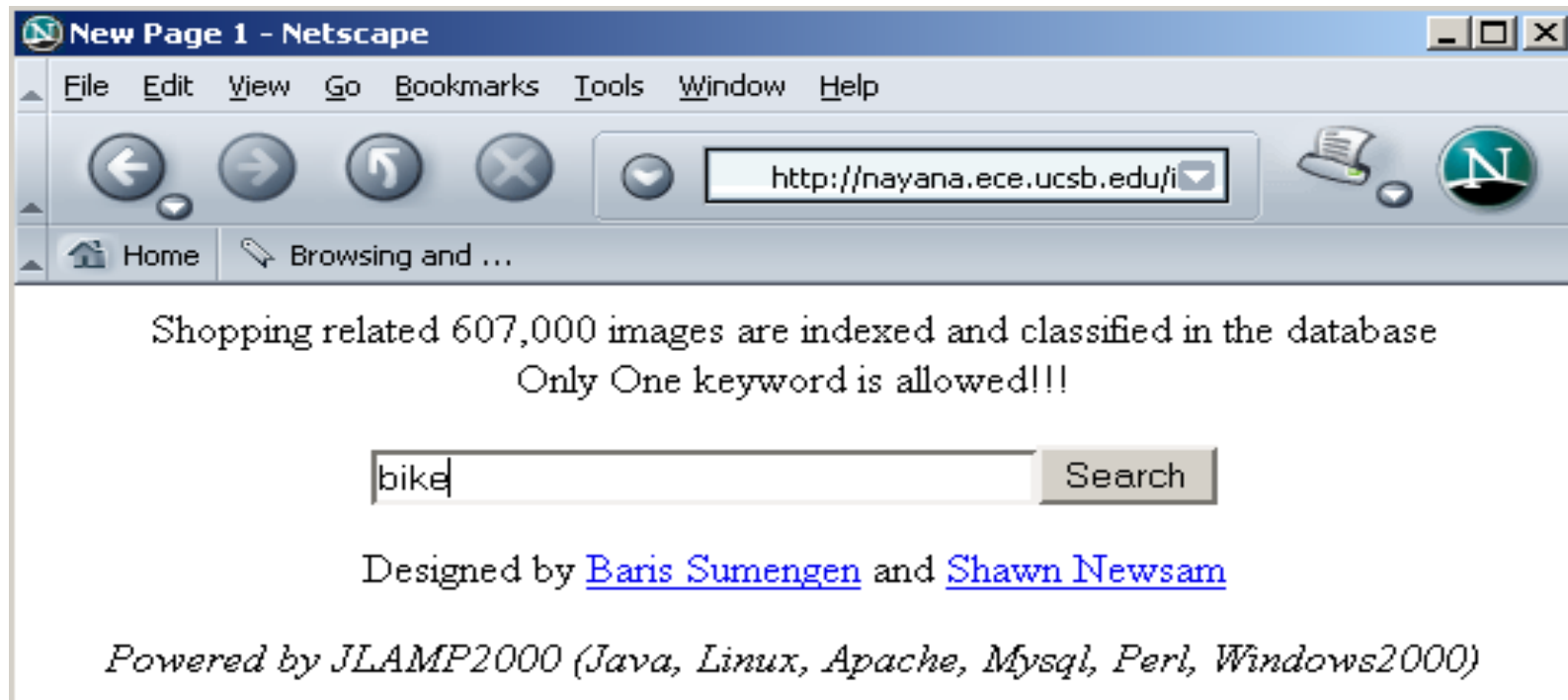
Official site of world's best-selling soprano. Join FAN AREA free to access exclusive perks, photo diaries, a global forum community and more...

www.sarah-brightman.com/ - 4k - [Cached](#) [Similar pages](#)















Relevance Feedback: Example

- Image search engine <http://nayana.ece.ucsb.edu/imsearch/imsearch.html>



Results for Initial Query






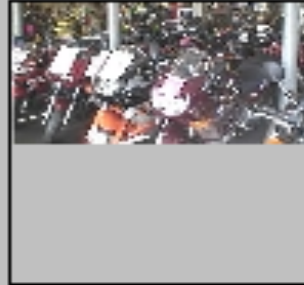






Browse Search Prev Next Random

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0



Relevance Feedback

Navigation buttons: [Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0



Results after Relevance Feedback

[Browse](#)[Search](#)[Prev](#)[Next](#)[Random](#)

(144538, 523493)
0.54182
0.231944
0.309876



(144538, 523835)
0.56319296
0.267304
0.295889



(144538, 523529)
0.584279
0.280881
0.303398



(144456, 253569)
0.64501
0.351395
0.293615



(144456, 253568)
0.650275
0.411745
0.23853



(144538, 523799)
0.66709197
0.358033
0.309059



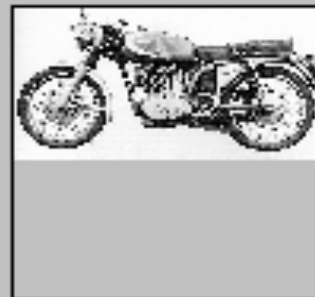
(144473, 16249)
0.6721
0.393922
0.278178



(144456, 249634)
0.675018
0.4639
0.211118



(144456, 253693)
0.676901
0.47645
0.200451



(144473, 16328)
0.700339
0.309002
0.391337



(144483, 265264)
0.70170796
0.36176
0.339948



(144478, 512410)
0.70297
0.469111
0.233859

Initial query/results

■ Initial query: *New space satellite applications*

- + 1. 0.539, 08/13/91, [NASA Hasn't Scrapped Imaging Spectrometer](#)
- + 2. 0.533, 07/09/91, [NASA Scratches Environment Gear From Satellite Plan](#)
- 3. 0.528, 04/04/90, [Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes](#)
- 4. 0.526, 09/09/91, [A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget](#)
- 5. 0.525, 07/24/90, [Scientist Who Exposed Global Warming Proposes Satellites for Climate Research](#)
- 6. 0.524, 08/22/90, [Report Provides Support for the Critics Of Using Big Satellites to Study Climate](#)
- 7. 0.516, 04/13/87, [Arianespace Receives Satellite Launch Pact From Telesat Canada](#)
- + 8. 0.509, 12/02/87, [Telecommunications Tale of Two Companies](#)

■ User then marks relevant documents with “+”.



Expanded query after relevance feedback

- 2.074 new
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil
- 15.106 space
- 5.660 application
- 5.196 eos
- 3.972 aster
- 3.446 arianespace
- 2.806 ss
- 2.053 scientist
- 1.172 earth
- 0.646 measure



Results for expanded query

- 2 1. 0.513, 07/09/91, [NASA Scratches Environment Gear From Satellite Plan](#)
- 1 2. 0.500, 08/13/91, [NASA Hasn't Scrapped Imaging Spectrometer](#)
3. 0.493, 08/07/89, [When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own](#)
4. 0.493, 07/31/89, [NASA Uses 'Warm' Superconductors For Fast Circuit](#)
- 8 5. 0.492, 12/02/87, [Telecommunications Tale of Two Companies](#)
6. 0.491, 07/09/91, [Soviets May Adapt Parts of SS-20 Missile For Commercial Use](#)
7. 0.490, 07/12/88, [Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers](#)
8. 0.490, 06/14/90, [Rescue of Satellite By Space Agency To Cost \\$90 Million](#)



Key concept: Centroid

- The centroid is the center of mass of a set of points
- Recall that we represent documents as points in a high-dimensional space
- Definition: Centroid

$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

where C is a set of documents.



Rocchio Algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance feedback query
- Rocchio seeks the query \vec{q}_{opt} that maximizes

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

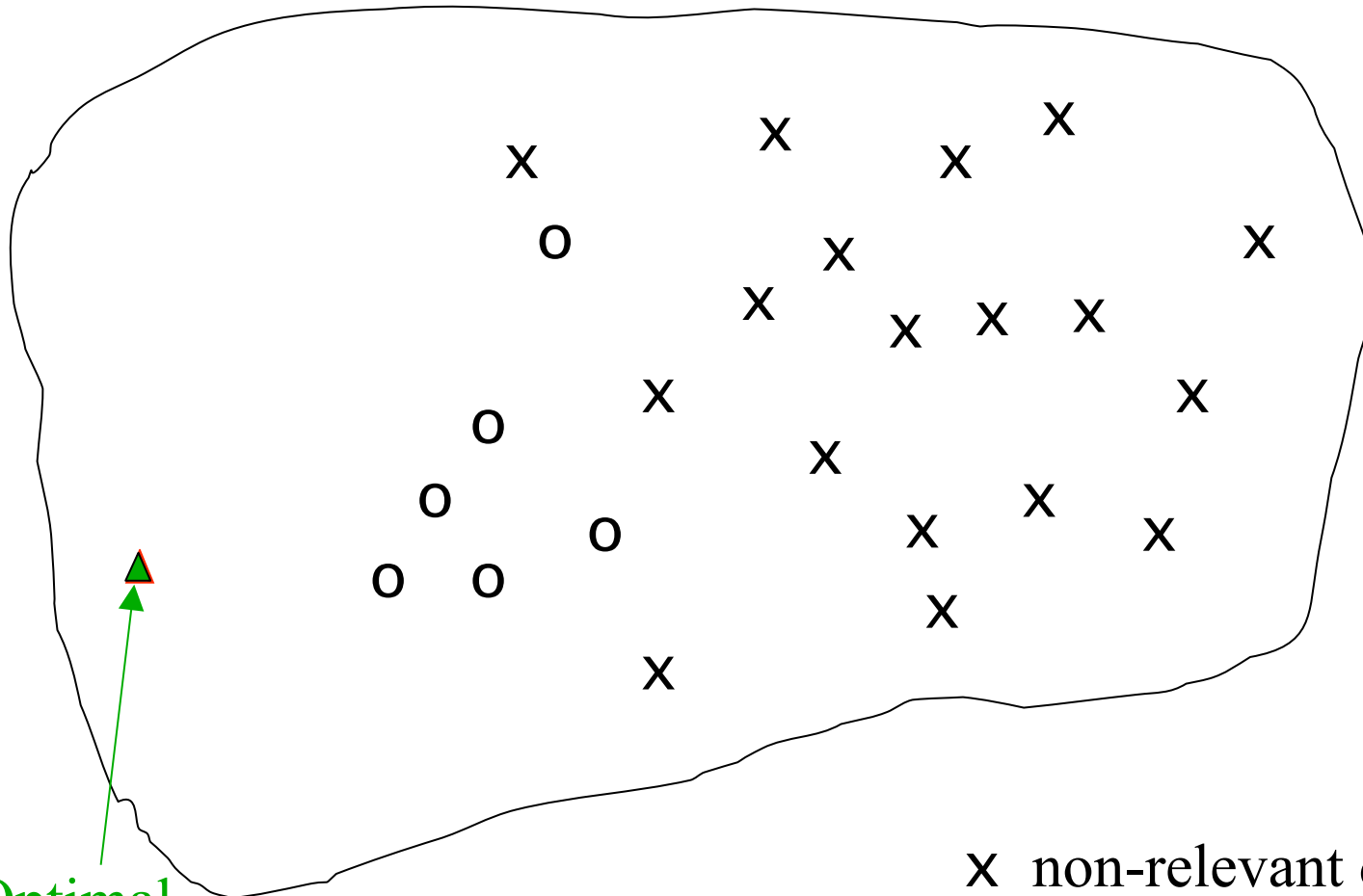
- Tries to separate docs marked relevant and non-relevant

$$\vec{q}_{opt} = \alpha \vec{q}_0 + \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

- Problem: we don't know the truly relevant docs



The Theoretically Best Query



Optimal
query

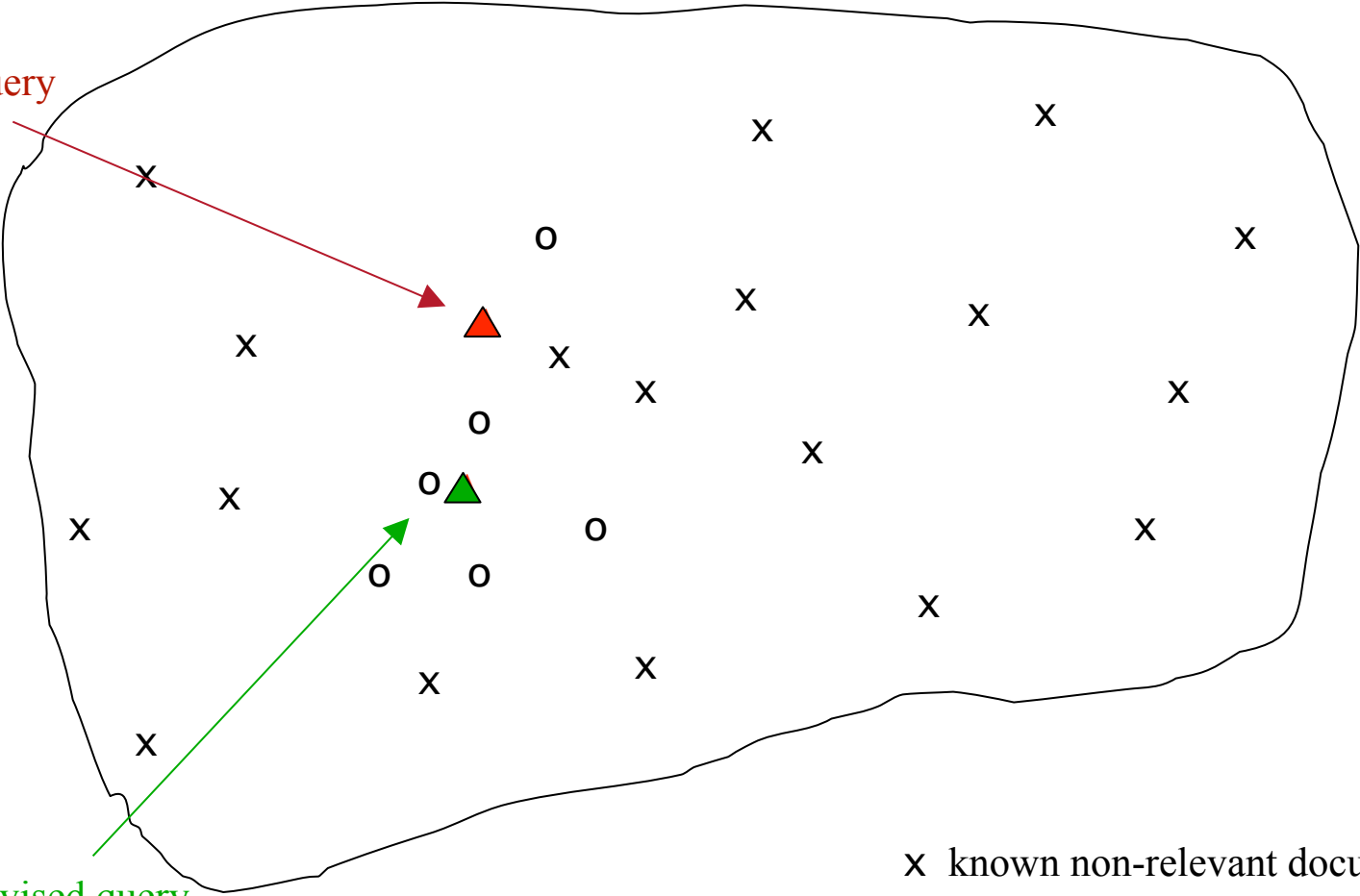
X non-relevant documents

O relevant documents



Relevance feedback on initial query

Initial query



Revised query

x known non-relevant documents
o known relevant documents



Weighting schemes for documents

- N , the overall number of documents,
- N_f , the number of documents that contain the feature f
- O_f^d the occurrences of the features f in the document d
- The weight f in a document is:

$$\omega_f^d = \left(\log \frac{N}{N_f} \right) \times o_f^d = IDF(f) \times o_f^d$$

- The weight can be normalized:

$$\omega'_f{}^d = \frac{\omega_f^d}{\sqrt{\sum_{t \in d} (\omega_t^d)^2}}$$



Relevance Feedback and query expansion: the Rocchio's formula

- ω_f^d , the weight of f in d
 - Several weighting schemes (e.g. TF * IDF, Salton 91')
- \vec{q}_f , the profile weights of f in C_i :

$$\vec{q}_f = \alpha \vec{q}_{0f} + \max \left\{ 0, \frac{\beta}{|T|} \sum_{d \in T} \omega_f^d - \frac{\gamma}{|\bar{T}|} \sum_{d \in \bar{T}} \omega_f^d \right\}$$

- T_i , the training documents in q



Similarity estimation between query and documents

- Given the document and the category representation

$$\vec{d} = \langle \omega_{f_1}^d, \dots, \omega_{f_n}^d \rangle, \quad \vec{q} = \langle \Omega_{f_1}, \dots, \Omega_{f_n} \rangle$$

- It can be defined the following similarity function (cosine measure)

$$s_{d,i} = \cos(\vec{d}, \vec{q}) = \frac{\vec{d} \cdot \vec{q}}{\|\vec{d}\| \times \|\vec{q}\|} = \frac{\sum_f \omega_f^d \times \Omega_f^i}{\|\vec{d}\| \times \|\vec{q}\|}$$

- d is retrieved for \vec{q} if $\vec{d} \cdot \vec{q} > \sigma$

