



SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**COURSE NAME : 19CS732 INFORMATION RETRIEVAL
TECHNIQUES**

IVYEAR / VII SEMESTER

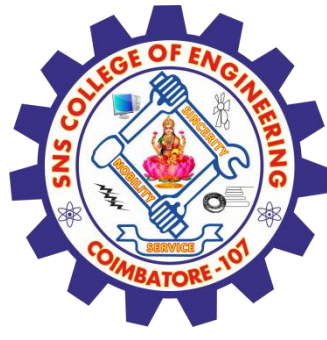
Unit 2- MODELING AND RETRIEVAL EVALUATION

Topic 9 : Relevance Feedback and Query Expansion



Problem

- Main topic today: two ways of improving recall: relevance feedback and query expansion
- As an example consider query q : [aircraft] ...
- ... and document d containing “plane”, but not containing “aircraft”
- A simple IR system will not return d for q .
- Even if d is the most relevant document for q !
- We want to change this:
- Return relevant documents even if there is no term match with the (original) query



Importance of Recall



Academic importance § Not only of academic importance

- Uncertainty about availability of information: are the returned documents relevant at all?
- Query words may return small number of documents, none so relevant
- Relevance is not graded, but documents missed out could be more useful to the user in practice § What could have gone wrong? – Many things, for instance ...
 - Some other choice of query words would have worked better
 - Searched for aircraft, results containing only plane were not returned



The gap between the user and the system



- The gap The retrieval system can only rely on the query words (in the simple setting) Wish: if the system could get another chance
- If the system gets another chance Modify the query to fill the gap better
Usually more query terms are added à query expansion The whole framework is called relevance feedback.



Relevance Feedback



- User issues a query – Usually short and simple query
- The system returns some results
- The user marks some results as relevant or nonrelevant
- The system computes a better representation of the information need based on feedback
- Relevance feedback can go through one or more iterations.
- It may be difficult to formulate a good query when you don't know the collection well, so iterate















Relevance Feedback -Cont..




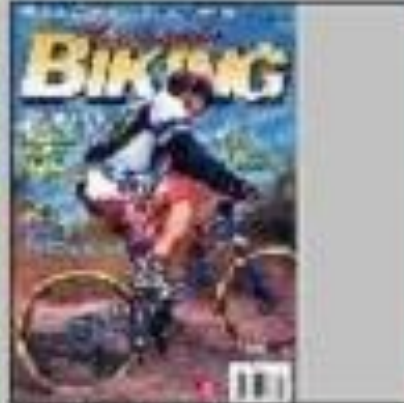










Relevance Feedback -Cont..

Browse Search Prev Next Random

					
(144473, 16459)	(144457, 252140)	(144456, 262031)	(144456, 262063)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264544)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0







Relevance Feedback -Cont..

Browse Search Prev Next Random

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 240611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	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 0.267364 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 0.309039
					
(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



Key concept for relevance feedback: Centroid

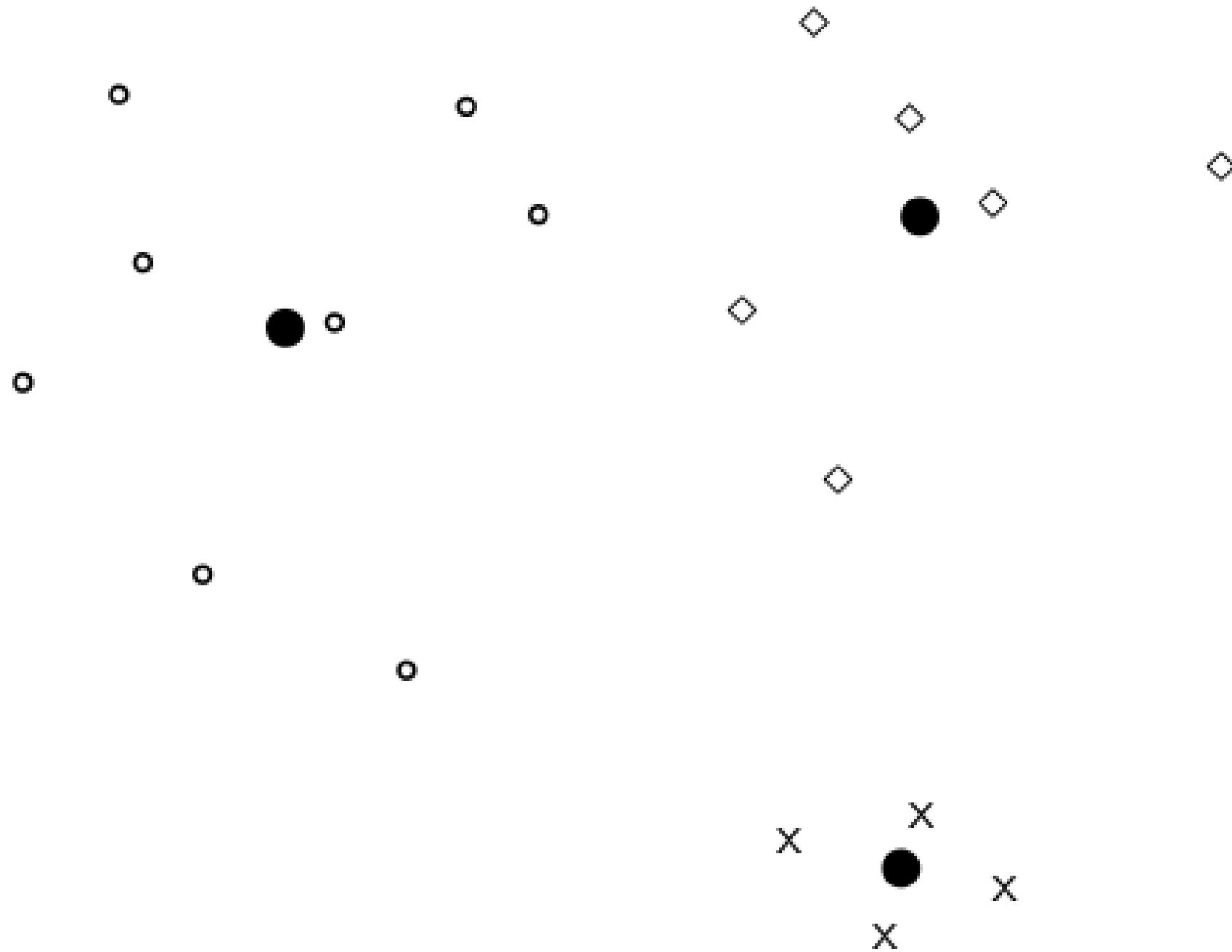
- The centroid is the center of mass of a set of points.
- Recall that we represent documents as points in a high-dimensional space.
- Thus: we can compute centroids of documents.
- Definition:

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

where D is a set of documents and $\vec{v}(d) = \vec{d}$ is the vector we use to represent document d .



Centroid: Example





Query expansion

- Query expansion is another method for **increasing recall**.
- We use “global query expansion” to refer to “global methods for query reformulation”.
- In global query expansion, the query is modified based on some global resource, i.e. a resource that is not query-dependent.
- Main information we use: (near-)synonymy
- A publication or database that collects (near-)synonyms is called a **thesaurus**.
- We will look at two types of thesauri: manually created and automatically created.



Types of query expansion



- Manual thesaurus (maintained by editors, e.g., PubMed)
- Automatically derived thesaurus (e.g., based on co-occurrence statistics)
- Query-equivalence based on query log mining (common on the web as in the “palm” example)



Types of query expansion

- For each term t in the query, expand the query with words the thesaurus lists as semantically related with t .
- Example from earlier: HOSPITAL → MEDICAL
- Generally increases recall
- May significantly decrease precision, particularly with ambiguous terms
 - INTEREST RATE → INTEREST RATE FASCINATE
- Widely used in specialized search engines for science and engineering
- It's very expensive to create a manual thesaurus and to maintain it over time.
- A manual thesaurus has an effect roughly equivalent to annotation with a **controlled vocabulary**.



Activity



Disadvantages



- 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



Advantages



- User gives feedback on **documents**.
 - More common in relevance feedback
- User gives feedback on **words** or **phrases**.
 - More common in query expansion



Assessment 1



1. List out the Advantages of Relevance Feedback and Query Expansion

- a) _____
- b) _____
- c) _____
- d) _____



2. Identify the disadvantages of Relevance Feedback and Query Expansion Collection

- a) _____
- b) _____
- c) _____
- d) _____



TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, –Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.
2. Ricci, F, Rokach, L. Shapira, B.Kantor, –Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, –Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, –Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU