# Sistemas de Recuperação de Informação

`https://github.com/fccoelho/curso-IRI`

## IRI 11: Recuperação de Informação Probabilística

Flávio Codeço Coelho

Escola de Matemática Aplicada, Fundação Getúlio Vargas

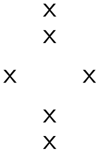
# Sumário da Aula

1. Recap

# Relevance feedback: Basic idea

- The user issues a (short, simple) query.
- The search engine returns a set of documents.
- User marks some docs as relevant, some as nonrelevant.
- Search engine computes a new representation of the information need – should be better than the initial query.
- Search engine runs new query and returns new results.
- New results have (hopefully) better recall.

# Rocchio illustrated



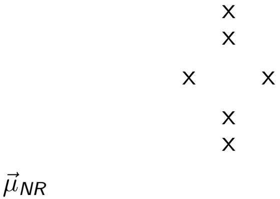
# Rocchio illustrated

x
x
x
x
x
x
$\vec{\mu}_R$

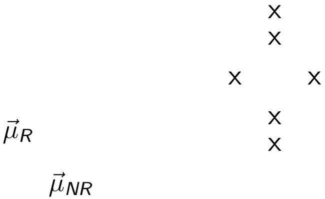
# Rocchio illustrated

x
x
x x
x
x
$\vec{\mu}_R$

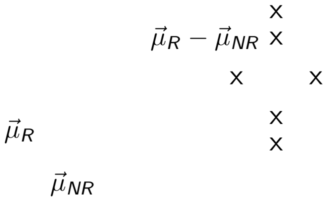
# Rocchio illustrated



# Rocchio illustrated



# Rocchio illustrated



X

$\vec{\mu}_R - \vec{\mu}_{NR}$ X

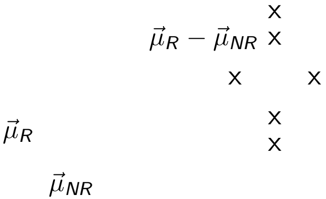
X X

X

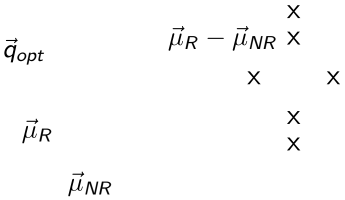
$\vec{\mu}_R$ X

$\vec{\mu}_{NR}$

# Rocchio illustrated



# Rocchio illustrated



X

$\vec{\mu}_R - \vec{\mu}_{NR}$ X

$\vec{q}_{opt}$

X X

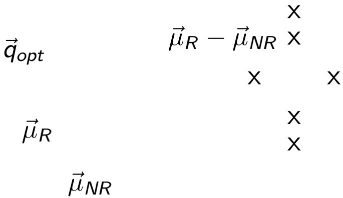
X

$\vec{\mu}_R$

X

$\vec{\mu}_{NR}$

# Rocchio illustrated



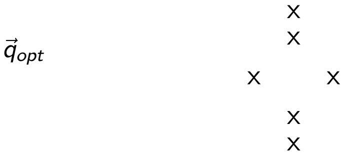
$\vec{q}_{opt}$

$\vec{\mu}_R - \vec{\mu}_{NR}$

$\vec{\mu}_R$

$\vec{\mu}_{NR}$

# Rocchio illustrated



# 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)

# Query expansion at search engines

- Main source of query expansion at search engines: query logs
- Example 1: After issuing the query [herbs], users frequently search for [herbal remedies].
  - → "herbal remedies" is potential expansion of "herb".
- Example 2: Users searching for [flower pix] frequently click on the URL photobucket.com/flower. Users searching for [flower clipart] frequently click on the same URL.
  - → "flower clipart" and "flower pix" are potential expansions of each other.

# Take-away today

- Probabilistically grounded approach to IR
- Probability Ranking Principle
- Models: BIM, BM25
- Assumptions these models make