Sistemas de Recuperação de Informação

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

IRI 11: Recuperação de Informação Probabilística

Flávio Codeço Coelho

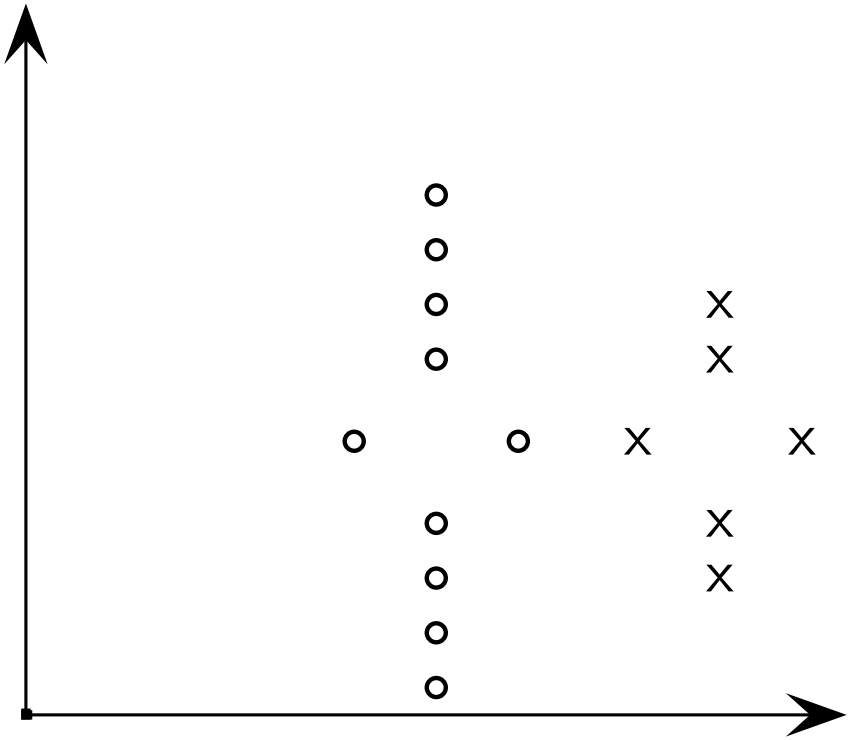
Escola de Matemática Aplicada, Fundação Getúlio Vargas

Sumário da Aula

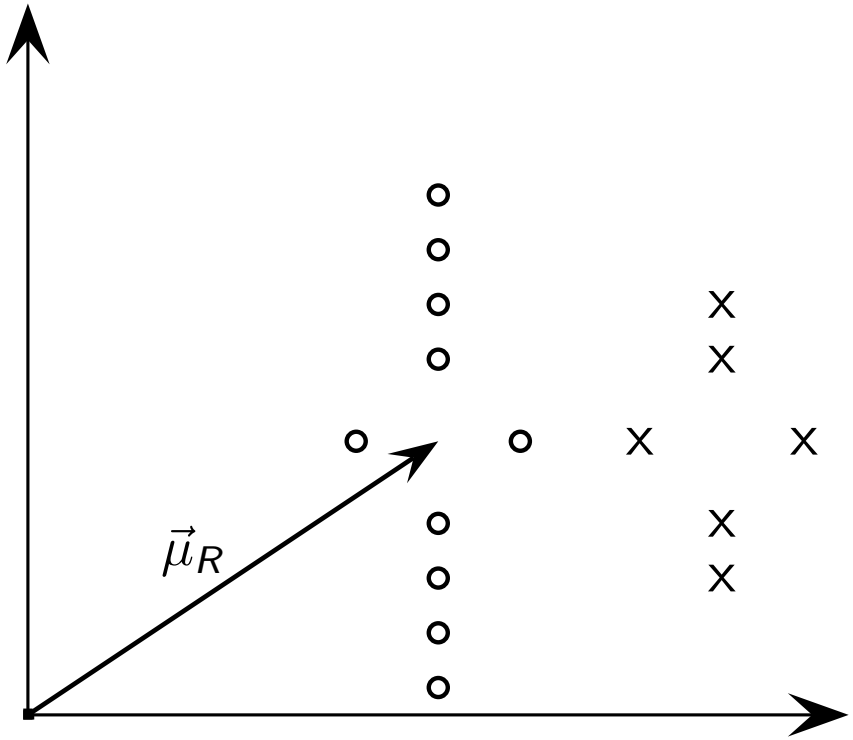
Revisão de relevância: Ideia básica

- O usuário faz uma consulta simples, curta.
- O buscador retorna um conjunto de documentos.
- O usuário marca alguns documentos como relevantes outros não.
- Buscador computa nova representação da informação requerida – deve ser melhor que a consulta inicial.
- Buscador executa nova consulta e retorna resultados.
- Novos resultados apresentação melhor revocação (espera-se).

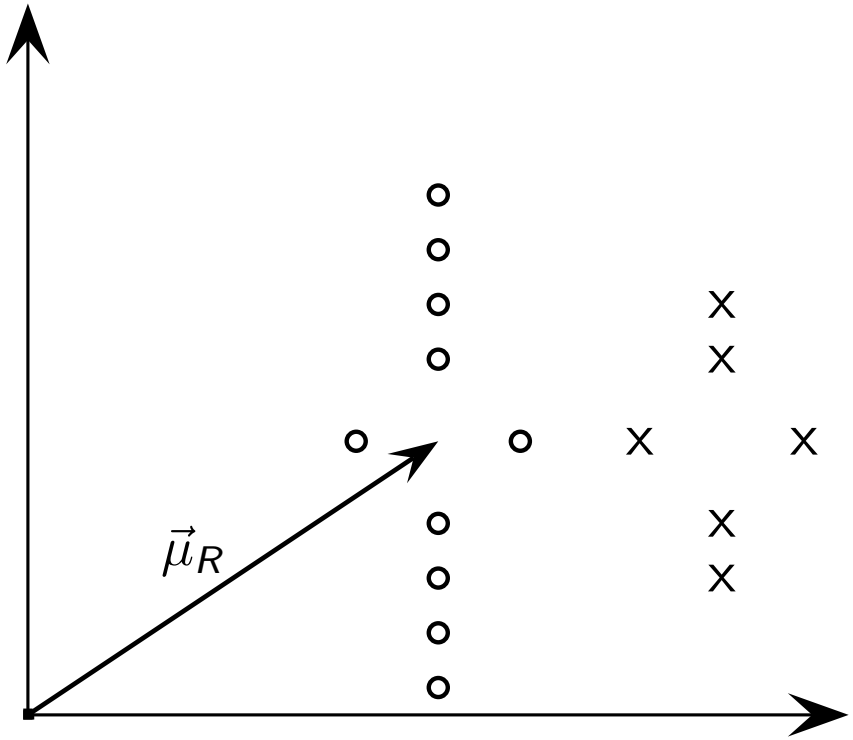
Rocchio



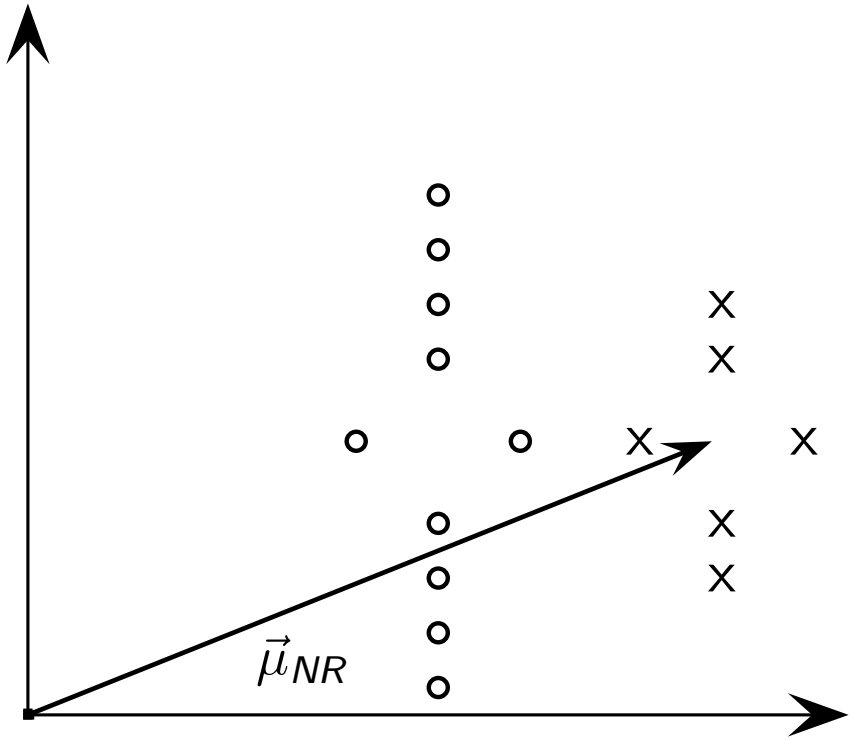
Rocchio



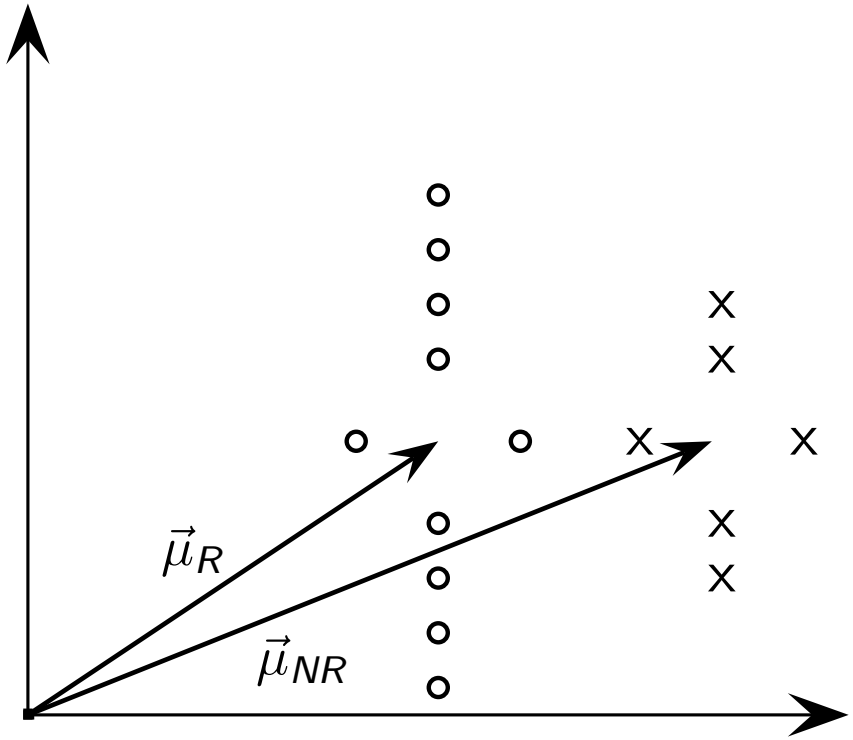
Rocchio



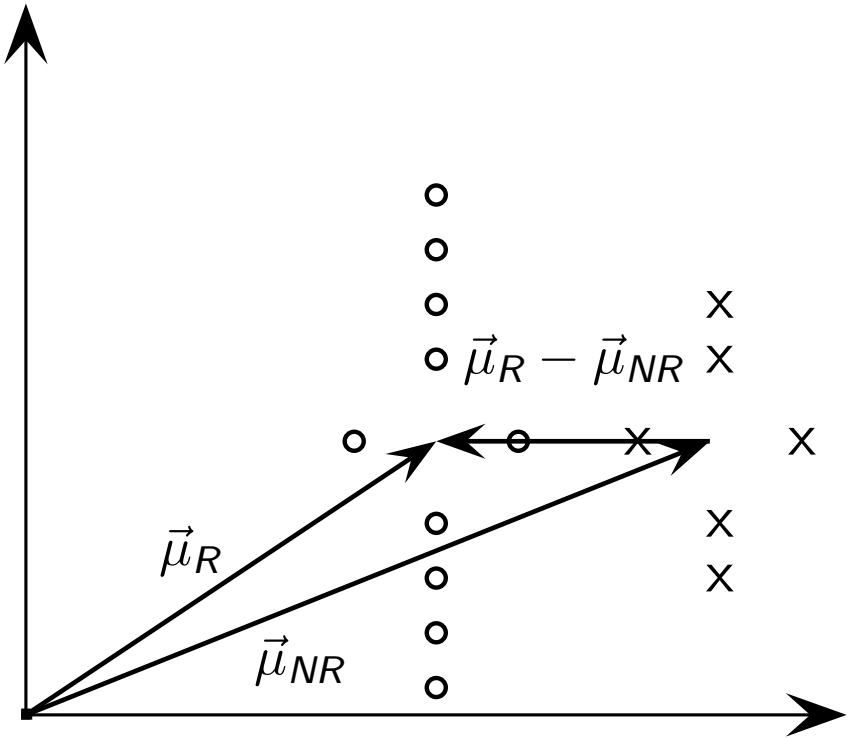
Rocchio



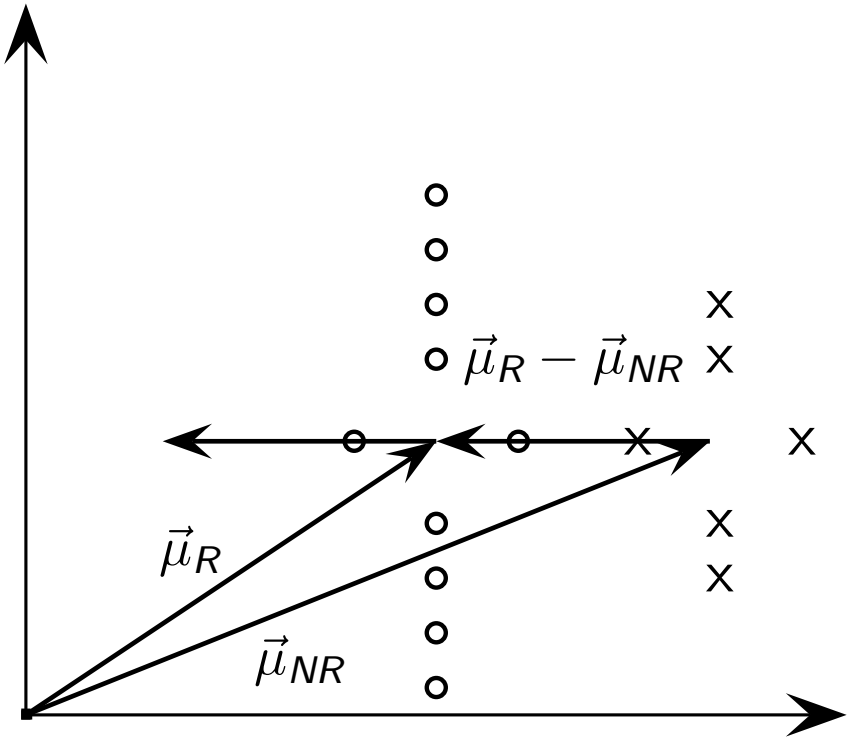
Rocchio



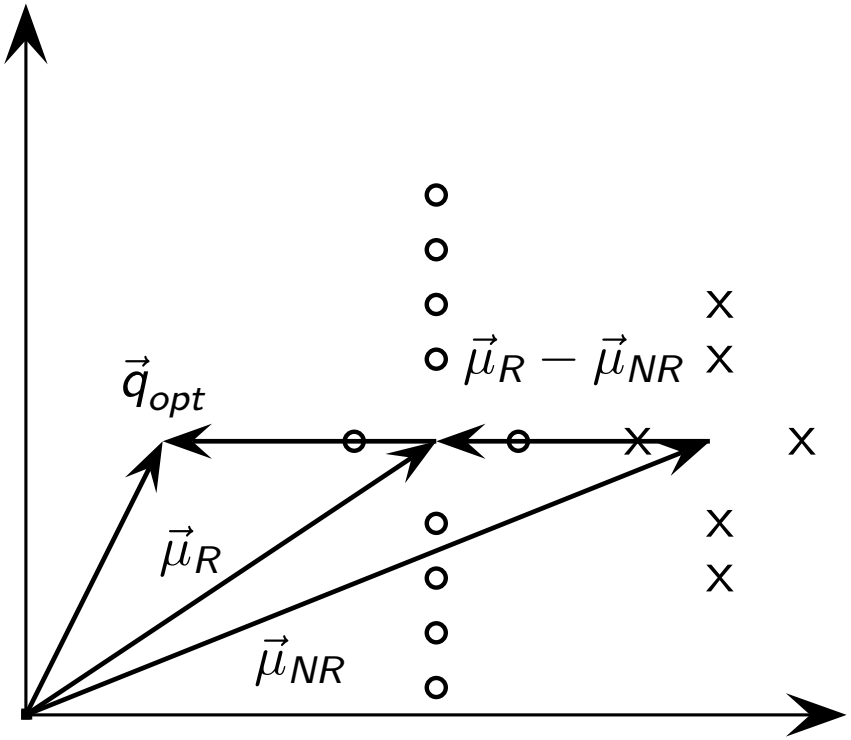
Rocchio



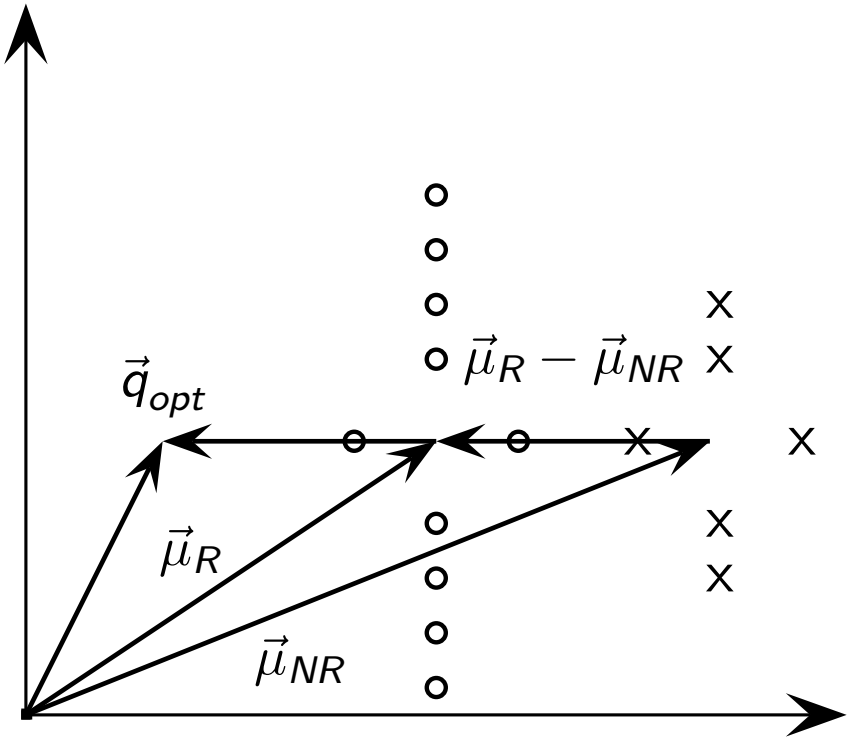
Rocchio



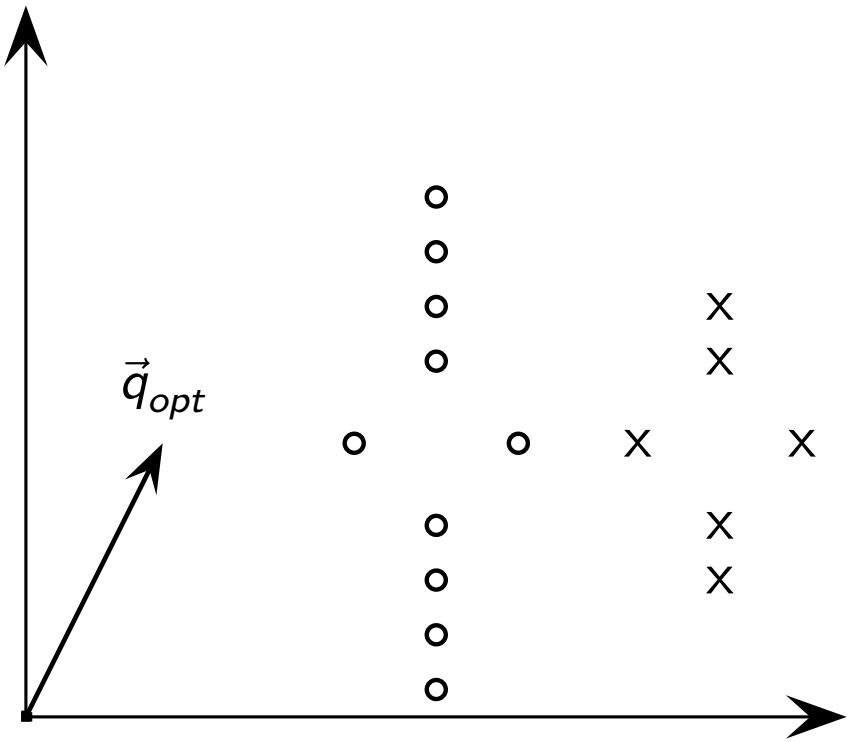
Rocchio



Rocchio



Rocchio



Tipos de expansão de consulta

- Tesauro manual (mantido por editores, p.ex., PubMed)
- Tesauro derivado automaticamente (p.ex., baseado em estatísticas de co-ocurrence statistics)
- Query-equivalence based on query log mining (common on the web as in the “palm” example)

Expansão de Consulta em Buscadores

- Fonte principal de expansões de consulta em buscadores: logs de consulta
- Exemplo 1: Depois de consultar por [herbal], usuários frequentemente buscam por [remédio herbal].
  - $\rightarrow$ "remédio herbal" é uma expansão em potencial para "herbal" ou "erva".
- Exemplo 2: Usuários buscando por [fotos de flores] frequentemente clicam na URL photobucket.com/flor. Usuários buscando por [desenhos de flor] frequentemente clicam na mesma URL.
  - $\rightarrow$ "desenhos de flor" e "fotos de flor" São potencialmente extensões uma da outra.

Conclusão de Hoje

- Abordagem probabilistica a RI
- Principio de Rankeamento de probabilidade
- Modelos: BIM, BM25
- Pressupostos destes modelos