Organização e Recuperação de Informação: Avaliação de sistemas de busca e resumos

Marcelo K. A.

Faculdade de Computação - UFU

Avaliação de um sistema de ORI

Medidas objetivas

- Rapidez de indexação
  - exemplo: bytes por hora
- Rapidez de busca
  - tempo médio de resposta ao usuário em milisegundos
  - número de consultas respondidas por hora
- Custo operacional por consulta em R$

Medidas subjetivas ou indiretas

- Satisfação do usuário
- O que é importante para a satisfação?
  - Tempo de resposta
  - Tamanho do índice disponível
  - Interface fácil de usar
  - Relevância
- Como medir a satisfação do usuário?

Quem são os usuários?

Quem são os usuários que buscamos prover satisfação?

- Usuário de buscador web.
  - Sucesso: encontrar o que estava procurando.
  - Medida: taxa de retorno de uso
- Anunciante de buscador web
  - Sucesso: usuário clica no anúncio
  - Medida: taxa de cliques
- E-comércio: comprador
  - Sucesso: comprador adquire produto
  - Medida: tempo para aquisição, fração de usuários compradores
- E-comércio: vendedor
  - Sucesso: venda de produto
  - Medida: lucro por produto
- Empresa: executivo
  - Sucesso: empregados são mais produtivos
  - Medida: aumento de lucro da companhia

Definição de satisfação

- Satisfação é frequentemente considerada resultado da relevância dos resultados ao interesse do usuário
- Metodologia:
  - Coleção de documentos preparada previamente para avaliação
  - Coleção de consultas dos usuários
  - Medições de relevância de cada par consulta-documento

Relevância: como medir?

- Coleção de documentos e consultas de referência *benchmark*
- Uma avaliação da relevância do par consulta-documento
  - "a resposta certa" - binário ou contínua (raro)

Exemplo: TREC Ad hoc (*Text Retrieval Conference*)

`http://trec.nist.gov/data/qrels_noneng/index.html`

```
TOPIC        ITERATION      DOCUMENT#          RELEVANCY
1            0              AP880212-0161          0
1            0              AP880216-0139          1
```

- TOPIC: assunto "alvo"
- ITERATION: informação experimental (não usado)
- DOCUMENT#: docId
- REVEVANCY: o documento fornece informação sobre o tópico considerado

Precisão e recuperação

- A taxa de precisão (P) é a fração de documentos recuperados que são relevantes

$$\text{Precisão} = \frac{|\text{itens relevantes recuperados}|}{|\text{itens recuperados}|}$$

- A taxa de recuperação (R) é a fração de documentos relevantes que foram recuperados

$$\text{Recuperação} = \frac{|\text{itens relevantes recuperados}|}{|\text{itens relevantes}|}$$

Precisão e recuperação

	Relevante	Não relevante
Recuperado	verdadeiros-positivos (VP)	falsos-positivos (FP)
Não recuperado	falsos-negativos(FN)	verdadeiros-negativos (FN)

$$\text{Precisão: } P = \frac{VP}{(VP + FP)}$$

$$\text{Recuperação: } R = \frac{VP}{(VP + FN)}$$

Equilíbrio entre precisão e recuperação

- É possível aumentar recuperação ao listar mais documentos
  - Para atingir recuperação máxima é só retornar todos os documentos
  - Mas prejudica precisão
- É possível aumentar precisão ao reduzir a recuperação
  - Retornar apenas um documento e se esse for relevante, obtém-se precisão máxima
  - Mas prejudica recuperação

Equilíbrio entre precisão e recuperação

- É possível aumentar recuperação ao listar mais documentos
  - Para atingir recuperação máxima é só retornar todos os documentos
  - Mas prejudica precisão
- É possível aumentar precisão ao reduzir a recuperação
  - Retornar apenas um documento e se esse for relevante, obtém-se precisão máxima
  - Mas prejudica recuperação
- É necessário equilibrar precisão e recuperação

Equilíbrio entre precisão e recuperação

$$F_\alpha = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

tal que $\alpha \in (0, 1)$
Com $\alpha = 1/2$, a medida F é conhecida como média harmônica e pode ser usada como um resumo de P e R:

$$F_{\alpha=1/2} = 2\frac{PR}{P+R}$$

Exemplo

	relevantes	não relevantes	
recuperados	20	40	60
não recuperados	60	1000000	1000060
	80	1000040	1000120

$$P = 20/(20+40) = 1/3 \approx 0.33$$

$$R = 20/(20+60) = 1/4 \approx 0.25$$

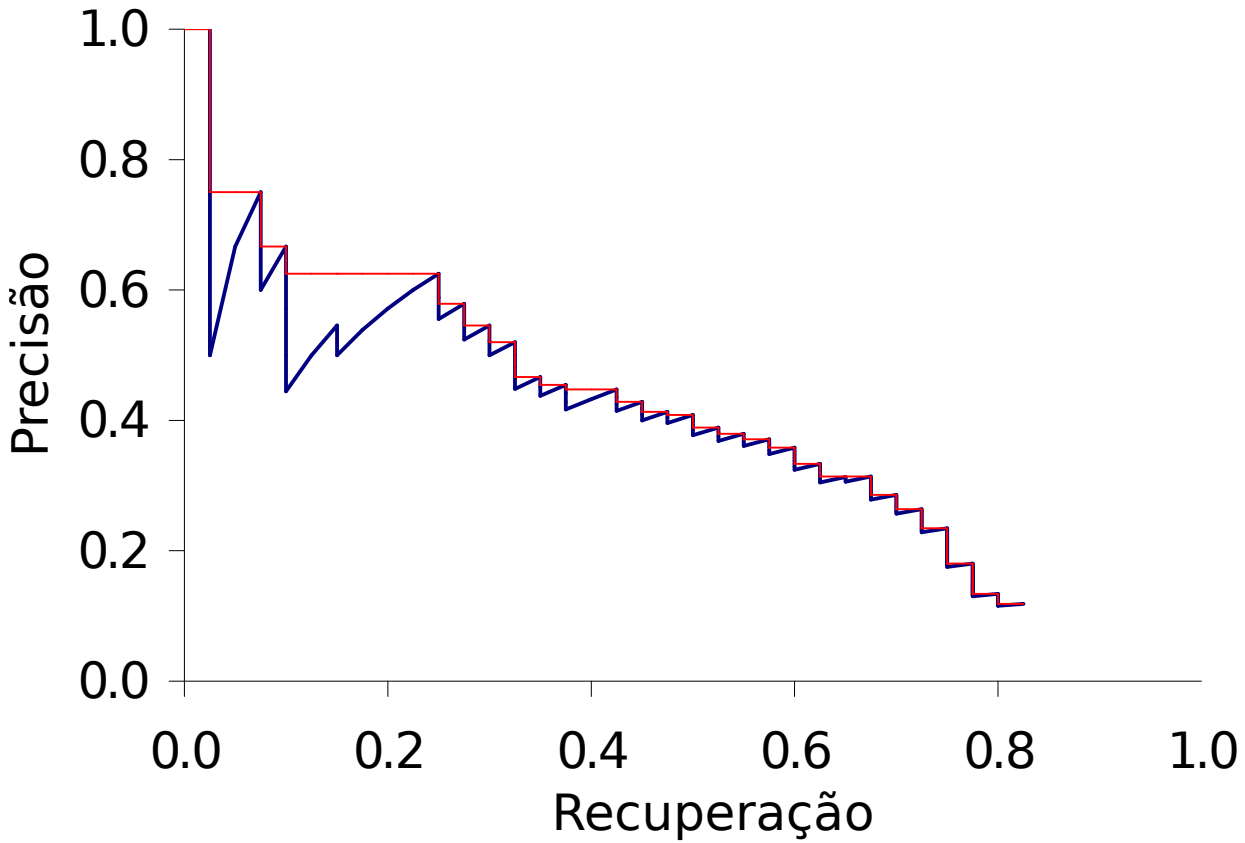
$$F_{\alpha=0.5} = 2\frac{1}{\frac{1}{1/3} + \frac{1}{1/4}} = \frac{2}{7} \approx 0.28$$

Porquê não usar a acurácia: $A = \frac{VP+VN}{VP+FP+FN+VN}$?

A curva de precisão-recuperação

- Precisão e recuperação são medidas para conjuntos não rankeados.
- Transformar medidas de conjunto em medidas de listas ordenadas:
- Calcular essas medidas para resultados top 1, top 2, top 3 ...
- Obtém-se a curva de precisão-recuperação.

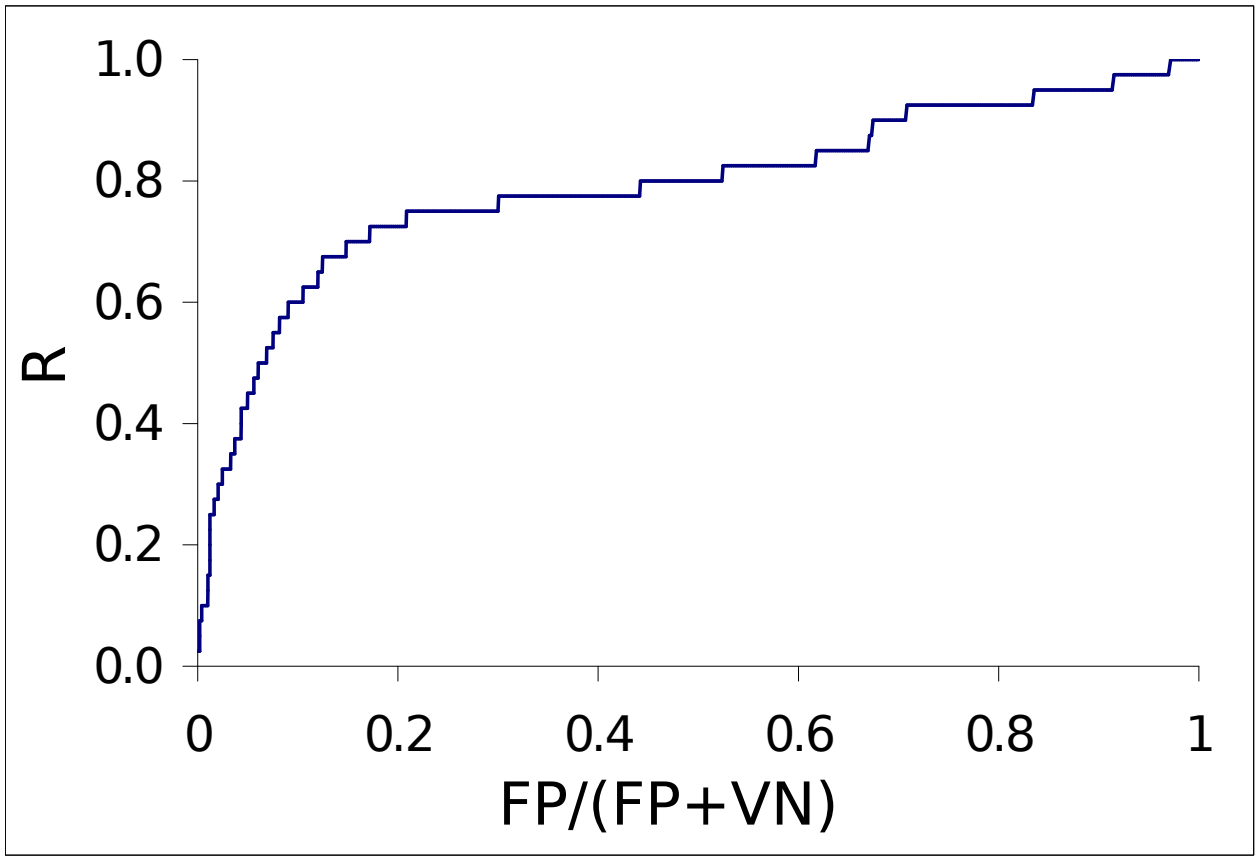
Uma curva de precisão-recuperação



- Cada ponto corresponde a um resultado top-k
- Interpolação (em vermelho): máximo dos pontos futuros

A curva ROC

Curva *Receiver Operating Characteristics*: avalia relação entre taxa de recuperação e taxa de falsos positivos.



Impacto da discordância entre juízes

- Juízes discordam frequentemente. Quer dizer que o resultado dos experimentos são inválidos?

Impacto da discordância entre juízes

- Juízes discordam frequentemente. Quer dizer que o resultado dos experimentos são inválidos?
- Não

Impacto da discordância entre juízes

- Juízes discordam frequentemente. Quer dizer que o resultado dos experimentos são inválidos?
- Não
- Suponha que queremos saber se algoritmo A é melhor que o B

Impacto da discordância entre juízes

- Juízes discordam frequentemente. Quer dizer que o resultado dos experimentos são inválidos?
- Não
- Suponha que queremos saber se algoritmo A é melhor que o B
- Um experimento pode oferecer uma resposta mesmo se juízes discordam muito

Avaliação de discordância

- Avaliação da concordância na avaliação humana de relevância
  - Juízes J1 e J2 responderam Sim/Não para "é relevante?"

	J2 Sim	J2 Não	J2 Total
J1 Sim	300	20	320
J1 Não	10	70	80
J1 Total	310	90	400

- Usar a medida Kappa $\kappa = \frac{P(A)-P(E)}{1-P(E)} \in (0,1)$
  - Quanto maior Kappa, maior concordância
- $P(A)$ = proporção de concordância entre "juízes"
  $P(A) = \frac{300+70}{400}$
- $P(E)$ = proporção de concordância aleatória
- $P(E) = P(\text{relevante})^2 + P(\text{não relevante})^2$
  - $P(\text{relevante}) = \frac{\text{número de respostas como relevante}}{\text{total de respostas}}$
  - $P(\text{relevante}) = \frac{320+310}{400+400}$ e $P(\text{não relevante}) = \frac{80+90}{400+400}$

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados
- ou usam medidas que previlegiam conseguir ranking 1

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados
- ou usam medidas que previlegiam conseguir ranking 1
- Buscadores também usam medidas não baseadas em relevância

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados
- ou usam medidas que previlegiam conseguir ranking 1
- Buscadores também usam medidas não baseadas em relevância
  - Exemplo 1: taxa de cliques no primeiro resultado em relação a outros resultados

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados
- ou usam medidas que previlegiam conseguir ranking 1
- Buscadores também usam medidas não baseadas em relevância
  - Exemplo 1: taxa de cliques no primeiro resultado em relação a outros resultados
  - Exemplo 2: estudos empíricos com usuários (aula passada)

Avaliação em grandes sistemas

- Taxa de recuperação é difícil medir na web
- Buscadores frequentemente usam precisão nos top-10 resultados
- ou usam medidas que previlegiam conseguir ranking 1
- Buscadores também usam medidas não baseadas em relevância
  - Exemplo 1: taxa de cliques no primeiro resultado em relação a outros resultados
  - Exemplo 2: estudos empíricos com usuários (aula passada)
  - Exemplo 3: testes A/B

Testes A/B

- Objetivo: testar uma única inovação

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando
- Maior parte dos usuários usam o sistema antigo

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando
- Maior parte dos usuários usam o sistema antigo
- Redirecionar uma parte dos usuários (por exemplo 1%) para o sistema com a inovação

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando
- Maior parte dos usuários usam o sistema antigo
- Redirecionar uma parte dos usuários (por exemplo 1%) para o sistema com a inovação
- Comparar uma medida de qualidade o sistema com e sem a inovação

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando
- Maior parte dos usuários usam o sistema antigo
- Redirecionar uma parte dos usuários (por exemplo 1%) para o sistema com a inovação
- Comparar uma medida de qualidade o sistema com e sem a inovação
- Provavelmente o método mais usado em buscadores

Testes A/B

- Objetivo: testar uma única inovação
- Pré-requisito: ter um sistema funcionando
- Maior parte dos usuários usam o sistema antigo
- Redirecionar uma parte dos usuários (por exemplo 1%) para o sistema com a inovação
- Comparar uma medida de qualidade o sistema com e sem a inovação
- Provavelmente o método mais usado em buscadores
- Variante: oferecer o novo sistema ao usuário

Usar mais que somente relevância

- Relevância foi definida para um par consulta-documento

Usar mais que somente relevância

- Relevância foi definida para um par consulta-documento
- Definição alternativa: relevância marginal

Usar mais que somente relevância

- Relevância foi definida para um par consulta-documento
- Definição alternativa: relevância marginal
- A relevância marginal de um documento na posição k na lista de resultados é a informação adicional que ele provê em relação ao documentos anteriores $d_1 \ldots d_{k-1}$.

Como apresentamos os resultados ao usuário?

- Mais comum: como uma lista de links
- Como cada documento deve ser descrito e apresentado?
- Importante para usuário identificar rapidamente os docs relevantes

Descrição de documentos na lista de resultados

- Mais comum: título, url, meta dados e um resumo
- Como fazer o resumo?
  - Tipo 1: estático
  - Tipo 2: dinâmico

Resumos

- Resumo estático: sempre o mesmo, independe da consulta
- Resumo dinâmico: depende da consulta. Tentar explicar porque o documento foi recuperado

Resumos estáticos

- Em sistemas típicos, o resumo estático é um subconjunto do documento
- Heurística simples: as 50 primeiras palavras do documento
- Mais sofisticado. Extrair as sentenças-chaves do documento:
  - Heurísticas de Processamento de Linguagem natural para pontuar sentenças
  - Sumário é composto das sentenças mais pontuadas
  - Aprendizado de Máquina
- Mais sofisticado: utilizar interpretador e sintetizador de resumo

Resumos dinâmicos

- Apresentar uma ou mais trechos (em inglês `snippets`) do documento
- Preferir trechos nos quais os termos da busca apareceram, se possível mais de um
- O resumo que é construído desta maneira apresenta o conteúdo completo do trecho

Resumo dinâmico

Consulta: “los roques economia”

Archipiélago Los Roques www.venezueladigital.net/roques/
`El Archipiélago Los Roques se encuentra ubicado en las coordenadas geográficas: ... La economía de Los Roques esta fundamentada en dos actividades: el ...`

Geração de resumos dinâmicos

- Como obter os outros termos do trecho do documento?
- Não podemos eficientemente construir um resumo dinâmico a partir do índice posicional invertido
- Necessário armazenar cópias de documentos: cache
- O índice posicional tem a posição de termos no documento
- Não armazenar documentos longos

Resumos dinâmicos

- Espaço na página de resultados é limitado $\rightarrow$ trechos devem ser curtos . . .
- . . . mas trechos devem ser longos o suficiente para fazerem sentido
- Ideal:

Resumos dinâmicos

- Espaço na página de resultados é limitado $\rightarrow$ trechos devem ser curtos . . .
- . . . mas trechos devem ser longos o suficiente para fazerem sentido
- Ideal: trechos linguísticamente bem construídos
- Ideal: trecho deve responder à consulta, assim não é necessário abrir o documento
- Resumos dinâmicos são importantes para os usuários . . .
  - . . . para escolher rapidamente o documento relevante
  - . . . economizar tempo em abrir documentos