Avaliação da Recuperação

Wendel Melo
Faculdade de Computação
Universidade Federal de Uberlândia

Recuperação da Informação
Adaptado do Material da Profª Vanessa Braganholo - IC/UFF

Avaliação de sistemas de busca

- Podemos avaliar um sistema de busca quanto a uma série de fatores:

Avaliação de sistemas de busca

- Podemos avaliar um sistema de busca quanto a uma série de fatores:
  - Desempenho;
  - Escalabilidade;
  - Utilização de recursos;
  - Interfaces de entrada e saída;
  - ...

Avaliação de sistemas de busca

- Podemos avaliar um sistema de busca quanto a uma série de fatores:
  - Desempenho;
  - Escalabilidade;
  - Utilização de recursos;
  - Interfaces de entrada e saída;
  - ...
- No entanto, os termos *avaliação de sistemas de busca* e/ou *avaliação de recuperação de informação* são utilizados com o significado de aferir a qualidade da resposta de sistemas de busca.

Avaliação da Recuperação da Informação

- É um processo sistemático no qual se associa uma métrica quantitativa aos resultados produzidos por um sistema de RI (busca) em resposta a um conjunto de consultas de usuário;
- A métrica associada deve estar diretamente associada à relevância dos resultados para os usuários;
- Assim, embora a métrica seja quantitativa, o objetivo final é avaliar a qualidade do sistema com respeito às respostas fornecidas.

Para que avaliar um sistema de RI?

Para que avaliar um sistema de RI?

- Para saber se o mesmo está cumprindo seu papel e como está desempenhando;

Para que avaliar um sistema de RI?

- Para saber se o mesmo está cumprindo seu papel e como está desempenhando;
- Para comparar com outros sistemas de RI;

Para que avaliar um sistema de RI?

- Para saber se o mesmo está cumprindo seu papel e como está desempenhando;
- Para comparar com outros sistemas de RI;
- Para avaliar se modificações no modelo de RI e/ou no ranqueamento trazem melhorias ao sistema;

Para que avaliar um sistema de RI?

- Para saber se o mesmo está cumprindo seu papel e como está desempenhando;
- Para comparar com outros sistemas de RI;
- Para avaliar se modificações no modelo de RI e/ou no ranqueamento trazem melhorias ao sistema;
- Para saber com quais tipos de base de dados/consultas o sistema/modelo pode funcionar melhor.

Avaliação da RI

- Na prática, pode ser difícil avaliar um sistema de RI, pois a relevância pode depender de muitos fatores subjetivos para cada usuário;
- Ainda assim, costuma-se associar métricas aos resultados da consulta por:
  - Simplicidade;
  - Poder repetir experimentos diversas vezes;
  - Custo relativamente baixo.

Avaliação da RI

Qual seria a metodologia mais intuitiva para a avaliação de sistemas de RI?

Avaliação da RI

Qual seria a metodologia mais intuitiva para a avaliação de sistemas de RI?

- Convocar especialistas para analisar a resposta produzida pelo sistema de RI?

Avaliação da RI

Qual seria a metodologia mais intuitiva para a avaliação de sistemas de RI?

- Convocar especialistas para analisar a resposta produzida pelo sistema de RI?
  - Seria custoso, demorado e não prático para avaliar diversas configurações distintas.

Avaliação da RI

Qual seria a metodologia mais intuitiva para a avaliação de sistemas de RI?

- ~~Convocar especialistas para analisar a resposta produzida pelo sistema de RI?~~
  - ~~Seria custoso, demorado e não prático para avaliar diversas configurações distintas.~~
- Comparar o resultado produzido por um sistema de RI com o produzido por especialistas humanos;

Avaliação da RI

Qual seria a metodologia mais intuitiva para a avaliação de sistemas de RI?

- Comparar o resultado produzido por um sistema de RI com o produzido por especialistas humanos;
- É comum então a adoção das chamadas **coleções de referência**.
- Coleções de referência: bases de documentos de referência onde especialistas já apontaram quais seriam os documentos relevantes para determinadas consultas específicas.

Exemplos de coleções de referência

- TREC (Text Retrieval Conference)
- CF (Cystic Fibrosis Database) na MEDLINE
- http://www.search-engines-book.com/collections/
  - CACM (Communications of the ACM)
  - Wikipedia

Métricas de Recuperação

- Dada uma coleção de referência, com uma conhecida requisição de informação (consulta) I, sejam:
  - *R*: conjunto de documentos relevantes (apontados por especialistas);
  - *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- A partir desses dois conjuntos, definimos as métricas de precisão e revocação (cobertura).

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.

- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

- A precisão se remete ao quanto os resultados da busca são uteis.

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

- A precisão se remete ao quanto os resultados da busca são uteis.
- A precisão por si só é uma medida completa para avaliação?

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

- A precisão se remete ao quanto os resultados da busca são uteis.
- A precisão por si só é uma medida completa para avaliação?
- E se houverem 100 docs relevantes e o algoritmo recuperar 10 docs, sendo 9 relevantes?

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

- A precisão se remete ao quanto os resultados da busca são uteis.
- A precisão por si só é uma medida completa para avaliação?
- E se houverem 100 docs relevantes e o algoritmo recuperar 10 docs, sendo 9 relevantes?
- A precisão será de 90%! Mas só se recuperou 9 de 100 relevantes...

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.

- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

Métricas de Recuperação

- *R*: conjunto de documentos relevantes (apontado por especialistas);
- *A*: conjunto resposta do algoritmo de RI sendo avaliado.
- **Precisão:** fração dos documentos recuperados que é relevante:

$$precisão = p = \frac{|R \cap A|}{|A|}$$

- **Revocação (cobertura):** fração dos documentos relevantes que é recuperada

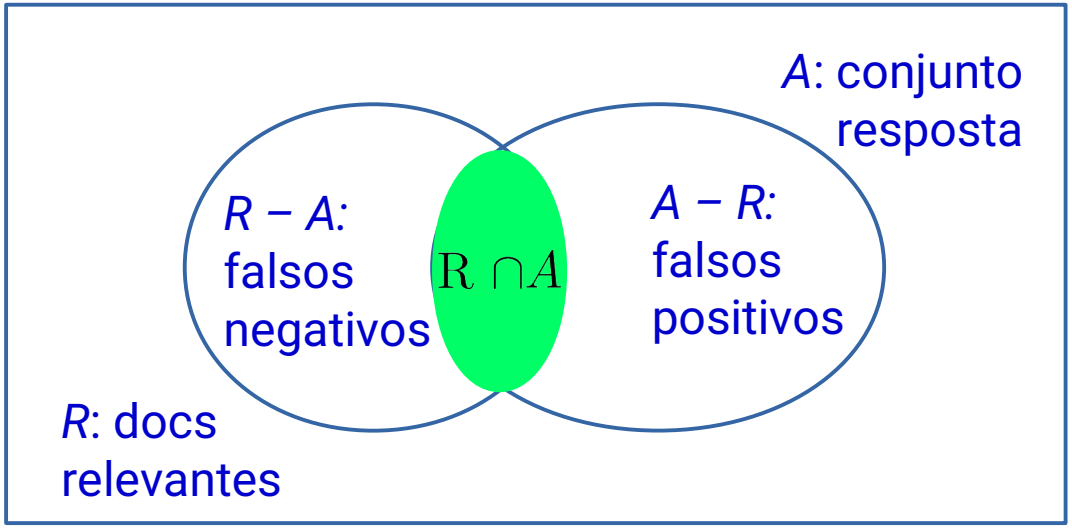
$$revocação = r = \frac{|R \cap A|}{|R|}$$

- A revocação se refere a quão completos os resultados estão

Precisão e Revocação

- Pode-se obter 100% de revocação se todos os documentos forem retornados em todas as consultas!
  - Todavia, a precisão será baixa

Precisão e Revocação



- A precisão é máxima quando o conjunto *A-R* (falsos positivos) é vazio;
- A revocação é máxima quando o conjunto *R-A* (falsos negativos) é vazio

Precisão e Revocação

- Observe que as métricas de precisão e revocação supõem que todos os documentos do conjunto resposta foram avaliados;
- Na prática, muitas vezes o usuário examina uma parte da resposta de acordo com a ordem de ranqueamento;
- Assim, as medidas de precisão e revocação variam conforme o usuário vai avaliando as respostas, e costuma-se plotar uma curva de precisão versus revocação.

Precisão e Revocação

- Suponha que, para uma consulta q de uma coleção de referência, o conjunto *R* tenha 10 docs relevantes e que os docs recuperados pelo algoritmo, já com ranqueamento, foram:

1 - d_{12}*	5 - d_{8}	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_{9}*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_{3}*
4 - d_{6}	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

- Quando o usuário olhar apenas o 1º doc, pelo fato do mesmo ser relevante, teremos 10% de revocação (1 doc relevante observado de um total 10) e 100% de precisão (todos os docs vistos até então são relevantes).

Precisão e Revocação

1 - d_{12}*	5 - d_{8}	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_{9}*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_{3}*
4 - d_{6}	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

- O próximo documento relevante é o terceiro. Nesse ponto, teremos 20% de revocação (dois docs relevantes observados de um total de 10) e 67% de precisão (dos três docs vistos ate então, dois são relevantes).

Precisão e Revocação

1 - d_{12}*	5 - d_8	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_9*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_3*
4 - d_6	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
10%	100%
20%	67%
30%	50%
40%	40%
50%	33%

Só é necessário tabelar nas revocações relativas a documentos relevantes!

Precisão e Revocação

1 - d_{12}*	5 - d_8	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_9*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_3*
4 - d_6	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
10%	100%
20%	67%
30%	50%
40%	40%
50%	33%

Montamos então um gráfico com a precisão em cada nível de revocação.

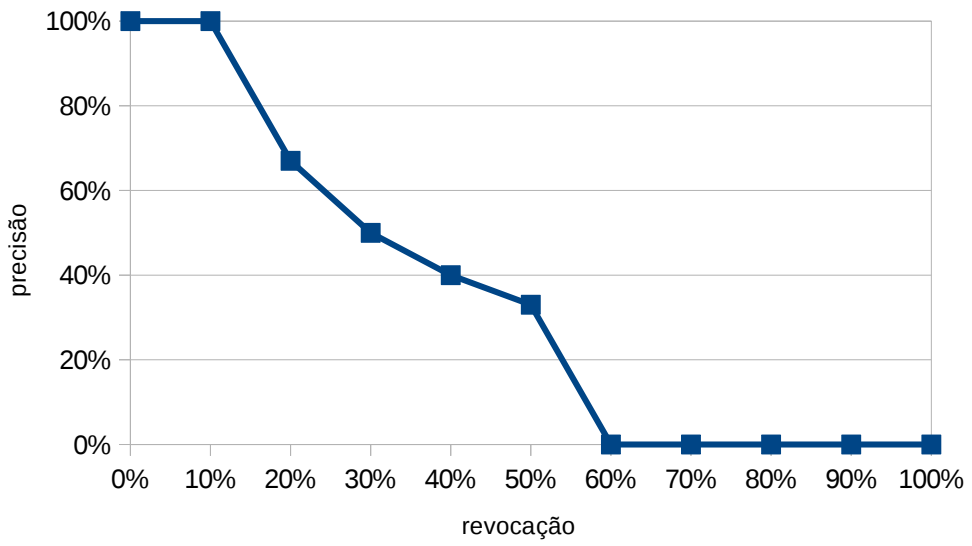
Precisão e Revocação

1 - d_{12}*	5 - d_{8}	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_{9}*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_{3}*
4 - d_{6}	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
10%	100%
20%	67%
30%	50%
40%	40%
50%	33%



Precisão e Revocação

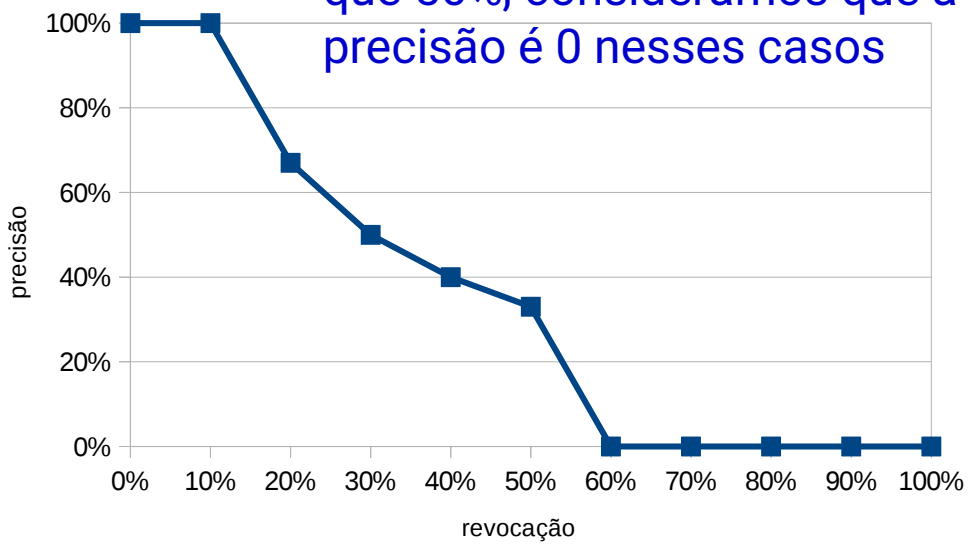
1 - d_{12}*	5 - d_8	9 - d_{18}	13 - d_{27}
2 - d_{84}	6 - d_9*	10 - d_{25}*	14 - d_{11}
3 - d_{56}*	7 - d_{51}	11 - d_{38}	15 - d_3*
4 - d_6	8 - d_{19}	12 - d_{48}	

Documentos com * também pertencem a *R*, isto é, são relevantes

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
10%	100%
20%	67%
30%	50%
40%	40%
50%	33%

Como não há revocação maior que 50%, consideramos que a precisão é 0 nesses casos



Precisão e Revocação

- Em geral, consideramos 11 níveis padrão de revocação:

$$r_0 = 0\%, \quad r_1 = 10\%, \quad r_2 = 20\%, \quad \ldots \quad r_{10} = 100\%$$

- Para evitar picos nos gráficos, consideramos que a precisão $p(r_j)$ no nível r_j é dada por:

$$p(r_j) = \max_{k \geq j} p(r_k)$$

Regra de interpolação

- Assim, teremos uma curva não crescente, pois, em cada ponto, considera-se a precisão como o maior valor tabelado daquele ponto em diante.

Precisão e Revocação

- **Exemplo**: Suponha que, para uma consulta q_2 conhecida, hajam 3 documentos relevantes e que a resposta retornada pelo algoritmo avaliado foi:

1 - d_7	5 - d_{79}
2 - d_6 *	6 - d_{30}
3 - d_2	7 - d_1 *
4 - d_{13}	8 - d_{15} *

Documentos com * também pertencem a *R*, isto é, são relevantes

- O segundo resultado é o primeiro relevante. Assim, para a revocação de 33%, temos 1/2 = 50% de precisão.

- O sétimo resultado é o segundo relevante. Assim, para revocação de 66%, temos 2/7 = 29% de precisão.

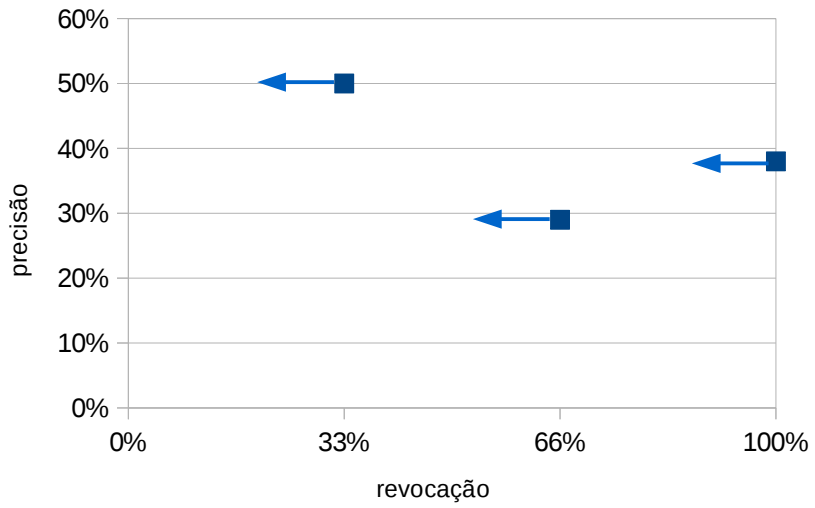
Precisão e Revocação

1 - d_7	5 - d_{79}
2 - d_6 *	6 - d_{30}
3 - d_2	7 - d_1 *
4 - d_{13}	8 - d_{15} *

Documentos com * também pertencem a *R*, isto é, são relevantes

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
33%	50%
66%	29%
100%	38%



Precisão e Revocação

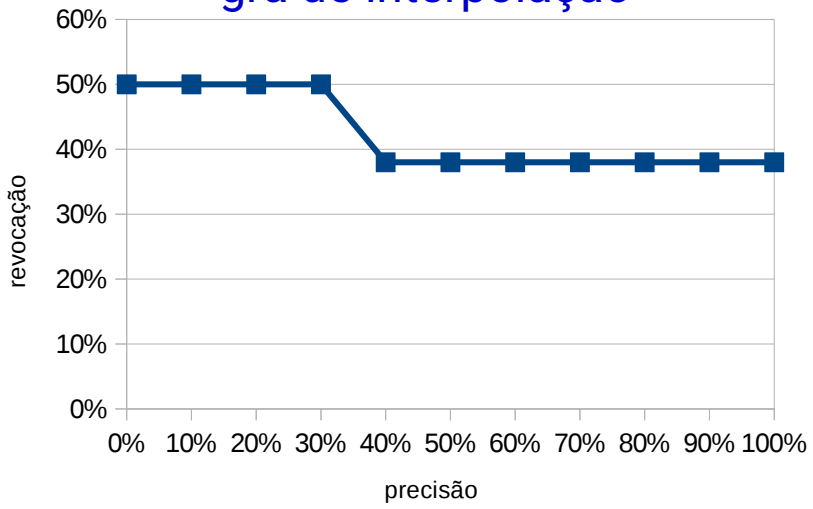
1 - d_7	5 - d_{79}
2 - d_6 *	6 - d_{30}
3 - d_2	7 - d_1 *
4 - d_{13}	8 - d_{15} *

Documentos com * também pertencem a *R*, isto é, são relevantes

Considerando os níveis padrão de revocação e aplicando a regra de interpolação

Estendendo o raciocínio, temos a tabela:

Revocação	Precisão
33%	50%
66%	29%
100%	38%



Precisão e Revocação

- Como uma só consulta não é suficiente para avaliar o desempenho de um sistema, na prática, toma-se um conjunto de N_q consultas de teste e calcula-se a média das precisões $\bar{p}(r_j)$ para cada nível de revocação r_j:

$$\bar{p}(r_j) = \sum_{i=1}^{N_q} \frac{p_i(r_j)}{N_q}$$

Precisão na consulta *i* para nível de revocação r_j

Nº total de consultas de teste

Precisão e Revocação

- É comum comparar diferentes sistemas plotando suas curvas de revocação-precisão média no mesmo gráfico;
- Em alguns casos, costuma-se adotar a área abaixo da curva (AVC) como medida para se comparar os sistemas. Valores maiores para a área indicam maior qualidade;
- Empiricamente, tem-se observado que, ao se aumentar o nível de revocação, o nível de precisão diminui;
- Algoritmos com maiores níveis de precisão costumam ser preferíveis para a WEB. Em contextos mais específicos, como área médica ou jurídica, níveis maiores de revocação podem ser preferíveis.