Evaluación de algoritmos de RI

Los algoritmos de RI utilizan diferentes técnicas:

Slide 1

- Tesauros
- Ponderación de términos
- Medidas de similaridad
- Pattern matching
- Relevance feedback

¿Cómo se compara el comportamiento de los diferentes algoritmos?

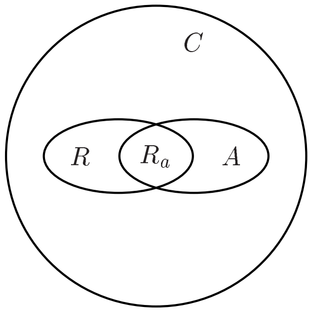
Hoy por hoy no hay base teórica. Evaluación experimental: Medidas y bancos de pruebas.

Slide 2

Horizonte: Tenemos un conjunto R de documentos relevantes respecto de una interrogación q a una coleción C.

Usando un algoritmo de RI obtenemos un conjunto A en respuesta a q.

El conjunto de documentos relevantes recuperados es $R_a = A \cap R$.



Medidas

- Recall es la fraccion de documentos relevantes recuperados.

$$Recall = \frac{|R_a|}{|R|}$$

- Precisión es la fracción de documentos recuperados que son relevantes.

Slide 3

$$Precision = \frac{|R_a|}{|A|}$$

En la definición se asume que el usuario clasifica todos los documento en A como relevantes o no.

En la práctica el usuario examina una lista ordenada, por similaridad, de los documentos recuperados.

En vez de utilizar solo los dos valores, se dibuja una gráfica precisión-recall.

Un ejemplo

$R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

Documentos recuperados en orden:

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56}	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3

Slide 4

Si cortamos la lista en cada posición i en la que aparece algún documento relevante podemos calcular los niveles de precisión y recall.

i	1	3	6	10	15
recall	10%	20%	30%	40%	50%
precision	100%	66%	50%	40%	33%

i	1	3	6	10	15
recall	10%	20%	30%	40%	50%
precision	100%	66%	50%	40%	33%

Estos valores son los que representan en la gráfica.

Slide 5

- Se pueden fijar los niveles de recall en los que evaluar la precisión típicamente 11: 0% 10% ... 100%
- Se pueden fijar intervalos en la lista (nivel de precisión) en los que se evalúa recall.
- Por interpolación se calcula precisión/recall en cualquier nivel deseado.

Slide 6

Evaluación

Los algoritmos de recuperación se evalúan frente a un conjunto de interrogaciones.

- Se combinan las gráfica precisión-recall en una única curva con los valores medios.
- Para ello se fijan unos niveles de recall (típicamente 11: 0% 10% … 100%) y se calcula

Slide 7

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q},$$

  donde N_q es el número de interrogaciones y $P_i(r)$ es la precisión de la interrogación i-ésima en el nivel de recall r.

- Así se obtiene una curva de precision versus recall que representa el comportamiento medio del algoritmo de RI.

Problemas:

- Hay que unificar las curvas, por ejemplo en los niveles de recall standard.
- Hay que comparar curvas.
- Puede que el cálculo de valores medios enmascare algún comportamiento malo.

Slide 8

Medidas resúmen: valor único

- Sintetizar la curva precisión-recall correspondiente a una interrogación en un único valor.
- Permite una comparación directa. Un algoritmo A es mejor que otro B si el valor asociado es mayor.
- Permite evaluación bajo diferentes hipótesis reflejadas en el conjunto de interrogaciones.

Slide 9

Precisión media en los documentos relevantes visualizados

Se obtine los valores de precisión cada vez que se observa un documento relevante y se calcula la media de estos valores.

R-Precision

es la precisión cuando se han recuperado R documentos, donde $R = |R_q|$.

Medidas resúmen: estadísticas

Histogramas de precisión

Sean $RP_A(i)$ y $RP_B(i)$ los valores de R-precision de dos algoritmos A y B.

Definimos $RP_{A/B}(i) = RP_A(i) - RP_B(i)$.

Se representa mediante un histograma, un rectángulo por cada pregunta.

Tablas resúmen

Slide 10

Tradicional en estadística:

- número de preguntas
- número de documentos recuperados
- número de documentos relevantes
- ...

Críticas

- Se necesita conocer toda la colección.
- Son medidas que están correlacionadas.
- No permiten evaluación de un sistema interactivo.
- ¿Qué ocurre en modelos en los que no hay puntuación?

Slide 11

Medidas alternativas

Combinación de precisión y recall.

- Media armónica

$$F(j) = \frac{2}{\frac{1}{R(j)} + \frac{1}{p(j)}}$$

donde j es el documento en la posición j del ranking.

F alcanza un valor mayor cuando las dos medidas son altas.

Slide 12

- Medida E

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{R(j)} + \frac{1}{p(j)}}$$

b es un parámetro especificado por el usuario que mide el interés en valorar más o menos precisión o recall

para $b = 1$ es el complemento de $F(j)$.

para $b > 1$ el usuario está más interesado en precisión que en recall

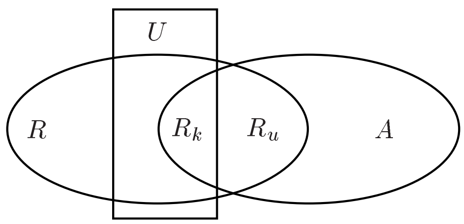
Slide 13

para $b < 1$ el usuario está más interesado en recall que en precisión

Medidas orientadas al usuario

R conjunto documentos relevantes y U conjunto de documentos conocidos por el usuario $U \subseteq C$.

Slide 14



Donde R_k es el conjunto de documentos recuperados relevantes que ya eran conocidos por el usuario y R_u es el conjunto de documentos relevantes nuevos.

Medidas

- Tasa de cobertura $coverage = \frac{|R_k|}{|U|}$
- Tasa de novedad $novelty = \frac{|R_u|}{|R_u| + |R_k|}$
- Recall relativo se usa un parámetro adicional e el número de documentos relevantes que espera el usuario

Slide 15

$$relative - recall = \frac{|R|}{e}$$

  se puede utilizar como medida de satisfación, si no es 1, el usuario continuará interrogando al sistema.
- Recall effort cociente entre e y el número de documentos que ha examinado.

Colecciones de referencia

No hay una base teórica en la evaluación. Por ello se necesitan disponer de colecciones en las que eveluar los algoritmos.

Como todo trabajo experimental debe diseñarse cuidadosamente el conjunto de pruebas y la coleccion que mejor se adapte al tipo de colección (o usuario) en la que se utilizara el algoritmo.

Slide 16

- TREC Text REtrieval Conference
- CACM Communications of the ACM
- ISI Institute of Scientific Information