Fundamentos de la Minería de Datos Web

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Métricas en IR y Búsqueda Web

Evaluación de los sistemas de recuperación de la información

UCO ONLINE

Evaluación desde el punto de vista del diseñador

- Eficacia en la ejecución
  - Tiempo que tarda el sistema en llevar a cabo una operación
  - Medida importante cuando los sistemas son interactivos
- Eficiencia en el almacenamiento
  - Cantidad de memoria necesaria para almacenar los datos
- Efectividad en la recuperación
  - Relevancia de los documentos recuperados respecto a la necesidad de información del usuario

Evaluación desde el punto de vista del usuario

- Exhaustividad
  - Habilidad para presentar todos los documentos relevantes
- Precisión
  - Habilidad para presentar solo documentos relevantes
- Esfuerzo
  - Para formulación de consultas, examinar los resultados, añadir feedback…
- Intervalo de tiempo entre petición y respuestas
- Presentación de los resultados de búsqueda
- Alcance o cobertura de la colección documental
  - Proporción en la que se incluyen en la recuperación todos los documentos relevantes ya conocidos por el usuario

Evaluación: corriente cognitiva

- La corriente cognitiva es un modelo de evaluación que considera el rol del usuario en el proceso de IR
  - Proporción de cobertura o alcance (*coverage ratio*)
  - Proporción de novedad (*novelty ratio*)
  - Satisfacción del usuario
  - Beneficio y frustraciones
  - Utilidad

Evaluación: corriente algorítmica

- La corriente algorítmica es el modelo tradicional de evaluación
- Las medidas más empleadas son el recall y la precisión
- Otras medidas son: precisión media, cobertura, F-measure, Fall-out, etc.
- Otras medidas dependerán de la ordenación de los documentos recuperados

Métricas en IR y Búsqueda Web

Medidas de evaluación

UCO ONLINE

Formalización

- **D** – Conjunto de documentos
- **N** – Número total de documentos en **D**
- **R** – Conjunto de documentos relevantes (también D_q – conjunto de documentos relevantes de la consulta **q**)
- $\overline{R} = D - R$ – Conjunto de documentos no relevantes
- **A** – Conjunto de documentos recuperados
- $A \cap R$ – Conjunto de documentos relevantes recuperados

El algoritmo de IR, primero computa la puntuación de relevancia para todos los documentos y luego produce el ranking R_q de los documentos en base a esta de relevancia, siendo d_1^q el más relevante para **q**

$$R_q : <d_1^q, d_2^q, \ldots, d_N^q>$$

Recall y precisión

- **Recall** en una posición de ranking **i** o documento d_i^q, **r(i)**, es la fracción de los documentos relevantes desde d_1^q a d_i^q en R_q

Porción de documentos relevantes recuperados

$$r(i) = \frac{s_i}{|D_q|}$$

Número de documentos relevantes en el intervalo $s_i < |D_q|$

- **Precisión** en una posición de ranking **i** o documento d_i^q, **p(i)**, es la fracción de documentos desde d_1^q a d_i^q en R_q que son relevantes

Porción de documentos recuperados que son relevantes

$$p(i) = \frac{s_i}{i}$$

Recall y precisión

Rank i	+/–	$p(i)$	$r(i)$
1	+	1/1 = 100%	1/8 = 13%
2	+	2/2 = 100%	2/8 = 25%
3	+	3/3 = 100%	3/8 = 38%
4	–	3/4 = 75%	3/8 = 38%
5	+	4/5 = 80%	4/8 = 50%
6	–	4/6 = 67%	4/8 = 50%
7	+	5/7 = 71%	5/8 = 63%
8	–	5/8 = 63%	5/8 = 63%
9	+	6/9 = 67%	6/8 = 75%
10	+	7/10 = 70%	7/8 = 88%
11	–	7/11 = 63%	7/8 = 88%
12	–	7/12 = 58%	7/8 = 88%
13	+	8/13 = 62%	8/8 = 100%
14	–	8/14 = 57%	8/8 = 100%
15	–	8/15 = 53%	8/8 = 100%
16	–	8/16 = 50%	8/8 = 100%
17	–	8/17 = 53%	8/8 = 100%
18	–	8/18 = 44%	8/8 = 100%
19	–	8/19 = 42%	8/8 = 100%
20	–	8/20 = 40%	8/8 = 100%

Mayor rango

Doc. relevante

Doc. irrelevante

Menor rango

Documentos 1, 2, 3, 5, 7

Posición 8

p(8) = 5 / 8 = 0,63

r(9) = 6 / 8 = 0,75

Documentos 1, 2, 3, 5, 7, 9

Todos los documentos relevantes:
1, 2, 3, 5, 7, 9, 10, 13

F-measure y Fall-out

- **F-measure** combina la precisión y la exhaustividad (*recall*) según parámetro *alpha* (α)

$$F-measure = \frac{(1+\alpha)*recall*precision}{recall+\alpha*precision}$$

Habitualmente, $\alpha = 1$, lo que se conoce como ***F_1*** o ***F-score balanceado***

- **Fall-out** es la proporción de documentos no relevantes que son recuperados de entre todos los documentos no relevantes

$$Fallout = \frac{|A \cap \overline{R}|}{|\overline{R}|}$$

Precisión media, R-precisión y precisión interpolada

- **Precisión media** es un valor único de precisión útil para la comparación de algoritmos de RI para una consulta **q**, y se calcula en base a la precisión de cada documento relevante del ranking

$$p_{avg} = \frac{\sum_{d_i^q \in D_q} p(i)}{|D_q|}$$

- **R-precision** mide la precisión obtenida cuando se ha recuperado un número de documentos igual al número de documentos relevantes
- **Precisión interpolada en 11 puntos** calcula la media de las precisiones en los puntos en que se alcanza el 0%, 10%, 20%, ..., 100% de los documentos relevantes

Precisión media, R-precisión y precisión interpolada

	Ranking 1	Ranking 2	Ranking 3
	d1	d10	d6
	d2	d9	d1
	d3	d8	d2
	d4	d7	d7
	d5	d6	d8
	d6	d5	d3
	d7	d4	d4
	d8	d3	d5
	d9	d2	d9
	d10	d1	d10
Precisión	0.5	0.5	0.5
Precisión-R	1	0	0.4
Precisión no interpolada	1	0.3544	0.5726
Precisión interpolada 11 pt	1	0.5	0.6440

Curva Precisión-*Recall*

- Basado en la precisión y exhaustividad (*recall*) de cada posición de ranking
- Dibujamos una curva en la que el eje ***x*** es el *recall* y el eje ***y*** es la precisión
- La curva se suele trazar utilizando 11 niveles de *recall* estándar: 0%, 10%, ..., 100%

  Si los niveles exactos de *recall* no se ajustan al ranking, es necesario interpolar el valor de la precisión a esos niveles concretos

  Sea r_i un nivel de *recall*, i = {0,1,2..,10} y $p(r_i)$ la precisión en ese nivel que se calcula como:

$$p(r_i) = \max_{r_i \leq r \leq r_{10}} p(r)$$

  Para interpolar la precisión a un nivel particular r_i, tomamos el valor máximo de precisión entre r_i y r_{10}

- af

Curva Precisión-*Recall*

i	$p(r_i)$	r_i
0	100%	0%
1	100%	10%
2	100%	20%
3	100%	30%
4	80%	40%
5	80%	50%
6	71%	60%
7	70%	70%
8	70%	80%
9	62%	90%
10	62%	100%

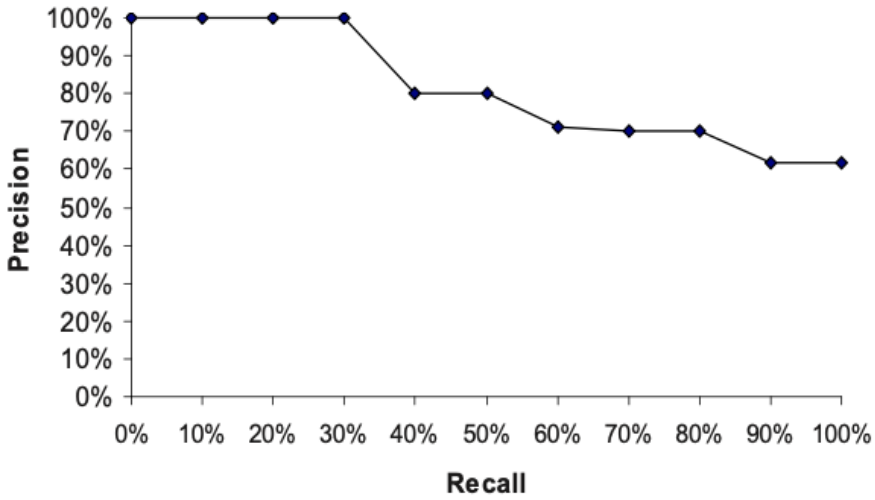
Interpolación de $p(r_i)$

Rank i	+/–	$p(i)$	$r(i)$
1	+	1/1 = 100%	1/8 = 13%
2	+	2/2 = 100%	2/8 = 25%
3	+	3/3 = 100%	3/8 = 38%
4	–	3/4 = 75%	3/8 = 38%
5	+	4/5 = 80%	4/8 = 50%
6	–	4/6 = 67%	4/8 = 50%
7	+	5/7 = 71%	5/8 = 63%
8	–	5/8 = 63%	5/8 = 63%
9	+	6/9 = 67%	6/8 = 75%
10	+	7/10 = 70%	7/8 = 88%
11	–	7/11 = 63%	7/8 = 88%
12	–	7/12 = 58%	7/8 = 88%
13	+	8/13 = 62%	8/8 = 100%
14	–	8/14 = 57%	8/8 = 100%
15	–	8/15 = 53%	8/8 = 100%
16	–	8/16 = 50%	8/8 = 100%
17	–	8/17 = 53%	8/8 = 100%
18	–	8/18 = 44%	8/8 = 100%
19	–	8/19 = 42%	8/8 = 100%
20	–	8/20 = 40%	8/8 = 100%

Curva Precisión-*Recall*

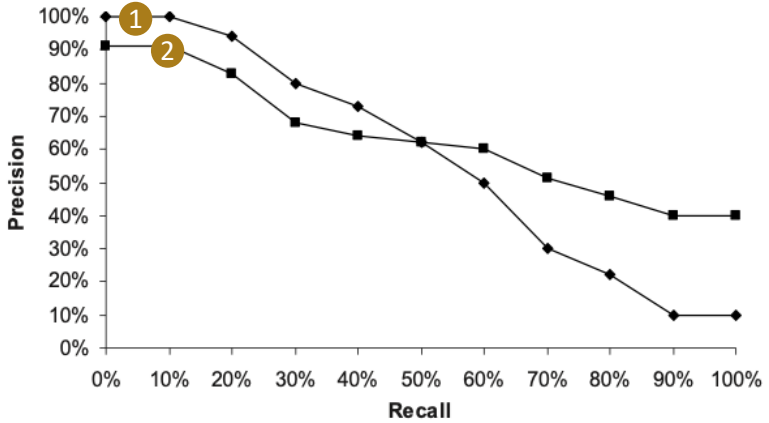
i	$p(r_i)$	r_i
0	100%	0%
1	100%	10%
2	100%	20%
3	100%	30%
4	80%	40%
5	80%	50%
6	71%	60%
7	70%	70%
8	70%	80%
9	62%	90%
10	62%	100%

Dibujo de curva



Curva Precisión-*Recall*

- Las curvas de *precisión-recall* son un mecanismo para comparar dos algoritmos sobre la misma consulta y el mismo conjunto de documentos



- En el ejemplo, vemos que la **curva 1** tiene valores de precisión más altos para niveles bajos de *recall*, pero luego son peores que la **curva 2** con los niveles de *recall* elevados

Evaluación de múltiples consultas

- Lo habitual es que el rendimiento de un algoritmo se quiera evaluar sobre un elevado número de consultas diferentes
- La **precisión global**, $\overline{p}(r_i)$, en cada nivel de *recall* r_i, se calcula como la media de las precisiones de los distintos individuos para ese nivel:

$$\overline{p}(r_i) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} p_j(r_i)$$

Q es el conjunto de todas las consultas

$p_j r_i$ es la precisión de la consulta **j** a nivel de recall r_i

- Utilizando la precisión media de cada nivel, se puede dibujar la curva P-R del algoritmo

Consideraciones sobre precisión y *recall*

- Aunque en teoría son independientes, en la práctica un *recall* alto implicará casi siempre decrementar la precisión, y viceversa → hay que encontrar el compromiso (dependerá de la aplicación en concreto)
- Un **problema** de precisión y *recall* es que resulta difícil determinar $\mathbf{D_q}$ para cada q
  - Este problema es evidente en la Web, en la que es imposible determinar $\mathbf{D_q}$ debido a que existen demasiadas páginas como para inspeccionarlas manualmente (y comprobar si son relevantes)
  - Sin $\mathbf{D_q}$, el *recall* no puede calcularse → *recall* no tiene sentido en búsquedas Web
  - En búsquedas Web, la precisión es crítica → puede ser estimada para los documentos "top ranked" (inspeccionar los *n* mejores documentos sí es razonable)
- Los motores de búsqueda suelen calcular la precisión top-5, 10, 15, 20, 25 y 30

¡Gracias!

UCO ONLINE