Análisis y Recuperación de Información

1er Cuatrimestre 2017

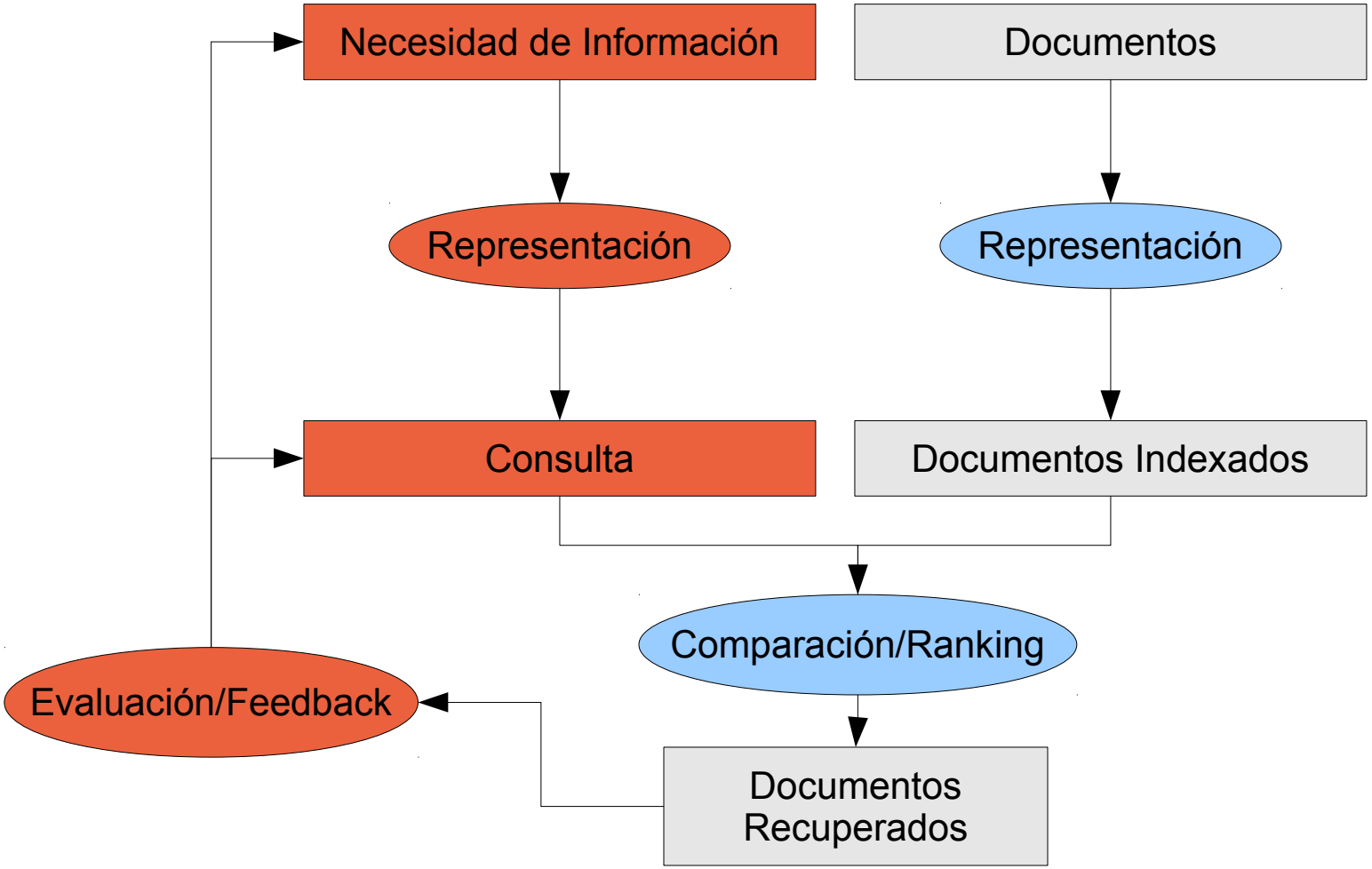
Página Web

http://www.exa.unicen.edu.ar/catedras/ayrdatos/

Prof. Dra. Daniela Godoy

ISISTAN Research Institute
UNICEN University
Tandil, Bs. As., Argentina
http://www.exa.unicen.edu.ar/~dgodoy
dgodoy@exa.unicen.edu.ar

Recuperación de Información



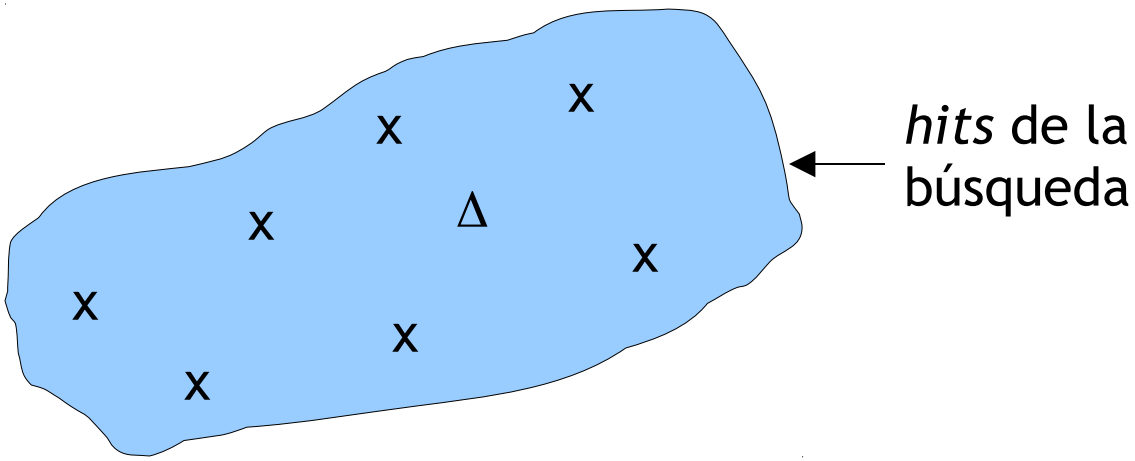
Reformulación de Consultas

- Feedback de relevancia
  - directo
  - pseudo feedback
- Expansión de consultas
  - con diccionarios reales
  - con diccionarios artificiales

Feedback de Relevancia

- Feedback de relevancia:
  - el usuario plantea una consulta simple
  - el usuario marca los resultados resultantes como relevantes o no relevantes
  - el sistema calcula la mejor representación para una necesidad de información en base al feedback
  - el feedback de relevancia se toma en una o más interacciones
- Idea fundamental: puede ser dificultoso para un usuario formular una buena consulta cuando no se conocen bien los documentos

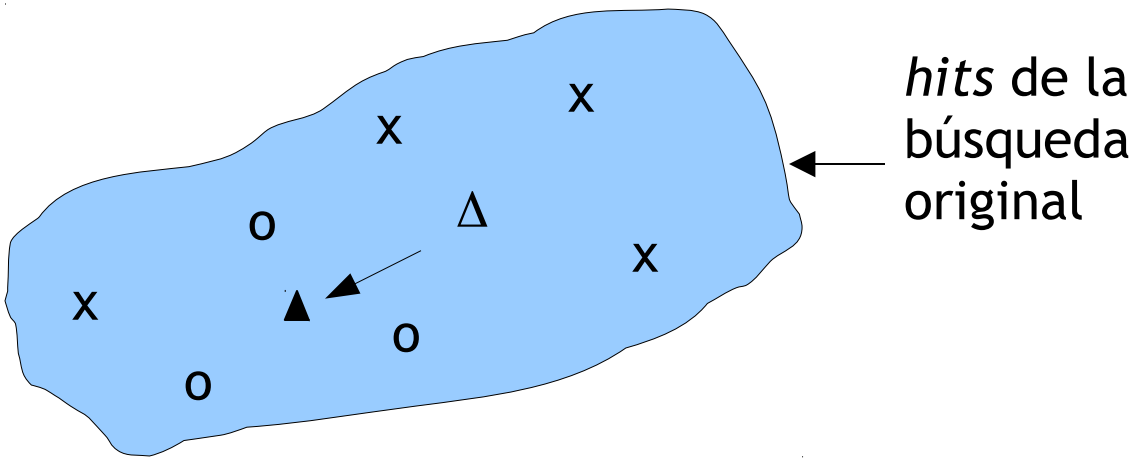
Feedback de Relevancia



x documentos devueltos por la búsqueda

Δ consulta

Feedback de Relevancia



x documentos identificados por el usuario como no relevantes

o documentos identificados por el usuario como relevantes

Δ consulta original

▲ consulta reformulada

Feedback de Relevancia

- Para una consulta q:
  - D_R es el conjunto de todos los documentos relevantes
  - D_{NR} es el conjunto de todos los documentos no relevantes
- $sim(q, D_R)$ es la similitud media entre la consulta q y los documentos en D_R
- $sim(q, D_{NR})$ es la similitud media entre la consulta q y los documentos en D_{NR}
- La mejor consulta sería la que maximize:

  $sim(q, D_R) - sim(q, D_{NR})$

Feedback de Relevancia

- En la práctica D_R y D_{NR} no se conocen (el objetivo es encontrarlos)
- Los resultados de la consulta inicial puede usarse para estimar $sim(q, D_R) - sim(q, D_{NR})$

Feedback de Relevancia

- Modificar una consulta existente en base a juicios de relevancia
  - extracción de términos de los documentos relevantes para incluir en la consulta
  - eliminación de términos de la consulta que aparecen en los documentos no relevantes
- En el modelo de espacio de vectores esto se puede hacer usando álgebra de vectores:
  - agregar los vectores de los documentos relevantes al vector de la consulta
  - substraer los vectores de los documentos irrelevantes del vector de la consulta
  - además de agregar o substraer palabras cambia los pesos de las palabras iniciales

Feedback de Relevancia

- El algoritmo de Rocchio incorpora los juicios de relevancia en el modelo de espacio de vectores
- Intenta maximizar $sim(q, D_R) - sim(q, D_{NR})$
- Feedback positivo y negativo:
  - el feedback positivo mueve la consulta hacia los documentos relevantes
  - el feedback negativo mueve la consulta lejos de los documentos irrelevantes (no necesariamente cercano a los relevantes)
    - el feedback negativo no siempre mejora la efectividad
    - algunos sistemas sólo se basan en feedback positivo

Feedback de Relevancia

$$Q_1 = \alpha Q_0 + \frac{\beta}{D_r} \sum_{d_j \in D_r} d_j - \frac{\gamma}{D_{nr}} \sum_{d_j \in D_{nr}} d_j$$

- Fórmula de Rocchio:
  - Q_0 es la consulta inicial y Q_1 la consulta luego de una iteración
  - D_r es el conjunto de documentos relevantes
  - D_n es el conjunto de documentos no relevantes

Feedback de Relevancia

$$Q_1 = \alpha Q_0 + \frac{\beta}{D_r} \sum_{d_j \in D_r} d_j - \frac{\gamma}{D_{nr}} \sum_{d_j \in D_{nr}} d_j$$

- Valores para α, β y γ?
  - α recibe usualmente el valor 0.75. Después de un número de iteraciones, los pesos originales de los términos pueden reducirse mucho
  - si β y γ tienen igual peso, los documentos relevantes y no relevantes contribuyen de la misma manera a la nueva consulta
  - si α=1 y γ=0, solo los documentos relevantes son usados en la nueva consulta
  - usualmente se usan β=0.75 y γ=0.25

Feedback de Relevancia

	T1	T2	T3	T4	T5
Q0	3	0	0	2	0
D1 (re)	2	4	0	0	2
D2 (re)	1	3	0	0	0
D3 (nr)	0	0	4	3	3

Peso de términos y juicios de relevancia para 3 documentos devueltos como resultado de la consulta Q_0

Asume que β y γ son 0.25

$$\boldsymbol{Q_1} = (3, 0, 0, 2, 0) + 0.25*(2+1, 4+3, 0, 0, 2) - 0.25*(0, 0, 4, 3, 2)$$
$$= \mathbf{(3.75, 1.75, 0, 1.25, 0)}$$

(Las entradas negativas se pasan a 0)

Feedback de Relevancia

$$Q_0 = (3, 0, 0, 2, 0) \longrightarrow Q_1 = (3.75, 1.75, 0, 1.25, 0)$$

Usando la nueva consulta y calculando similitudes, da los siguientes resultados →

	D1	D2	D3
Q0	6	3	6
Q1	11.5	7.5	3.25

- La consulta inicial dio un alto resultado para D3, aunque era irrelevante para el usuario (dado al peso del término 4)
  - en general, cuanto menos términos en la consulta es más probable que un término particular pueda resultar en resultados no relevantes
  - la nueva consulta decrementa el resultado para D3 e incrementa los de D1 y D2
- La nueva consulta agrega un peso para el término 2
  - inicialmente puede no haber estado en el vocabulario del usuario
  - fue agregado porque apareció como significante en un número suficiente de documentos relevantes

Feedback de Relevancia

- Desventajas:
  - los usuarios muchas veces se niegan a proveer feedback explícito
  - es más difícil de ver porqué un documento en particular fue recuperado
  - los usuarios no siempre tienen suficiente conocimiento para hacer la consulta inicial
  - la distribución de los términos en los documentos relevantes e irrelevantes es similar
  - las consultas largas son ineficientes en la mayoría de los buscadores

Feedback de Relevancia

- Observaciones:
  - por su construcción, la consulta reformulada va a rankear los documentos explícitamente indicados como relevantes más alto y los marcados como irrelevantes más abajo
  - el método no debe evaluarse por las mejoras en los documentos que recibieron feedback, sería equivalente al testing sobre los datos de entrenamiento en aprendizaje de máquina
  - la evaluación debe enfocarse en la generalización hacia otros documentos de relevancia desconocida

Feedback de Relevancia

- Pseudo-feedback:
  - usar los métodos de feedback de relevancia sin entradas explícitas del usuario
  - se asume que los top *m* documentos recuperados son relevantes y se usan para reformular la consulta
  - permite la expansión de consultas que incluyen términos que están correlacionados con los términos de la consulta

Feedback de Relevancia

- Feedback de relevancia en la Web:
  - algunos buscadores ofrecen páginas relacionadas o similares, la forma más simple de feedback de relevancia
  - otros consideran esto difícil de explicar para los usuarios
  - Excite tenía inicialmente feedback de relevancia real, pero los abandonó por falta de uso
    - sólo en el 4% de las sesiones se usaba el feedback de relevancia (“More like this”)
    - 70% de los usuarios miraba la primera página y no continuaba la búsqueda
    - el feedback de relevancia mejoraba los resultados en 2/3 veces que se lo usaba

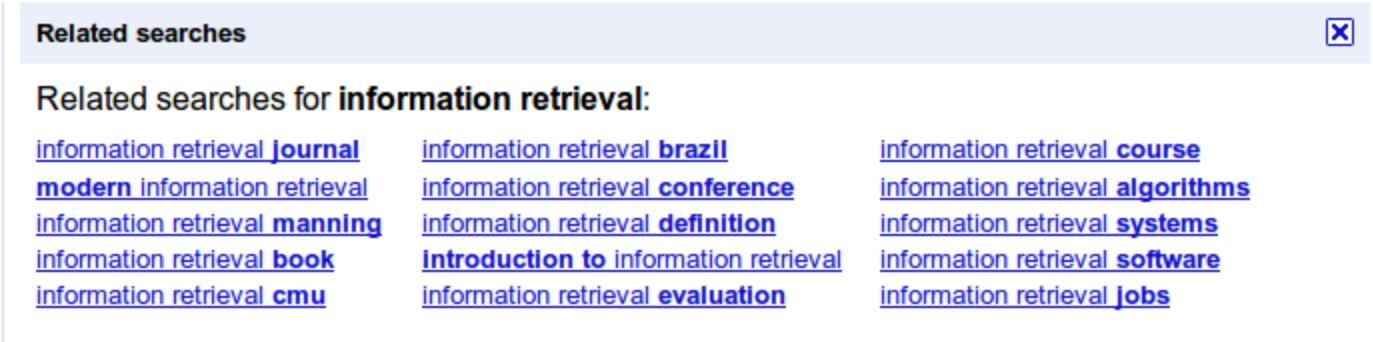
Expansión de Consultas

- En el feedback de relevancia los usuarios dan información adicional sobre la relevancia o no de los documentos

- En la expansión de consultas los usuarios dan información adicional (utilidad) sobre las palabras o frases de la consulta

Expansión de Consultas

- Expansión de consultas:
  - Feedback
    - información acerca de stop-words, stemming, etc.
    - número de *hits* de cada término o frase
  - Sugerencias
    - diccionario
    - diccionario derivado automáticamente (co-ocurrencia)
    - basado en análisis de los *logs* de las sesiones



Expansión de Consultas

- Expansión basada en diccionarios:
  - no requiere input del usuario
  - para cada término *t* en una consulta, se expande la consulta con sinónimos o palabras relacionadas del diccionario
    - feline → feline cat
  - es posible dar menor peso en la consulta a los términos adicionados que a los originales
  - mejora el recall
  - puede disminuir en mucho la precisión, particularmente con términos ambiguos
    - interest rate → interest rate fascinate evaluate
  - costo de acceder al diccionario

Expansión de Consultas

- Un diccionario provee información sobre sinónimos y palabras y frases semánticamente relacionadas
- Ejemplos con WordNet:
  - agregar sinónimos
    - physician → doc, doctor, MD, medical, mediciner, medico, sawbones
  - agregar hipónimos para especializar
    - fruit → apple (especialización)
  - agregar hipérnimos para generalizar
    - tree → plant (generalización)
  - agregar términos relacionados
    - physician → medic, general practitioner, surgeon

Expansión de Consultas

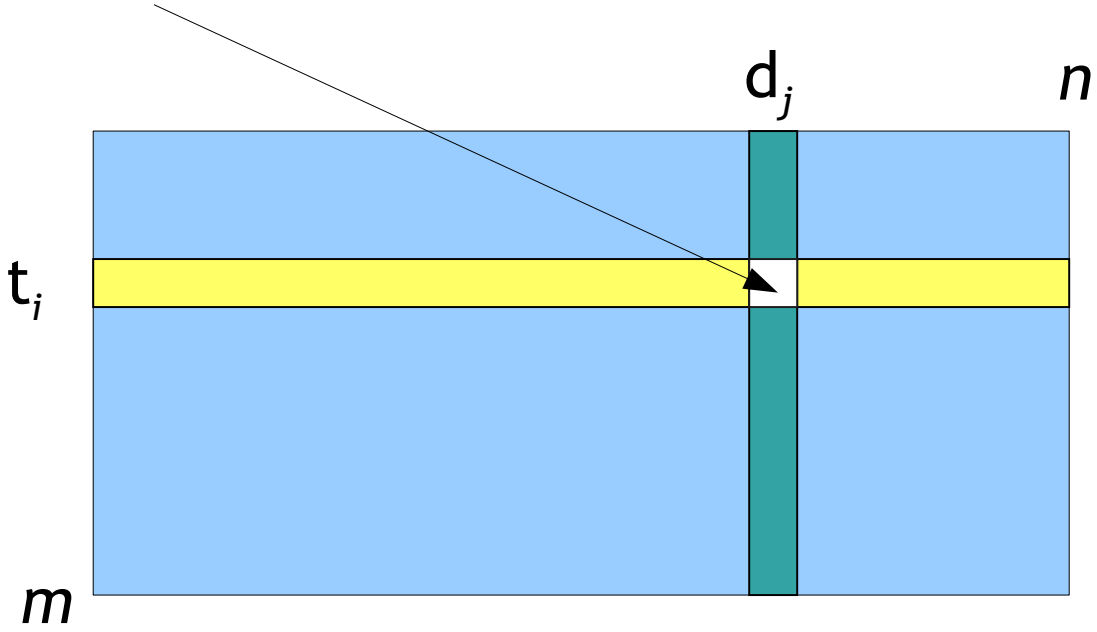
- La generación automática de diccionarios:
  - intenta generar un diccionario automáticamente analizando la colección de documentos
  - dos enfoques:
    - basado en co-ocurrencia (palabras que co-ocurren son probablemente similares entre sí)
    - análisis de relaciones gramaticales (por ejemplo las entidades que se cocinan y se comen son probablemente comestibles)
  - la co-ocurrencia es más robusta mientras que las relaciones gramaticales tienen mayor exactitud

Expansión de Consultas

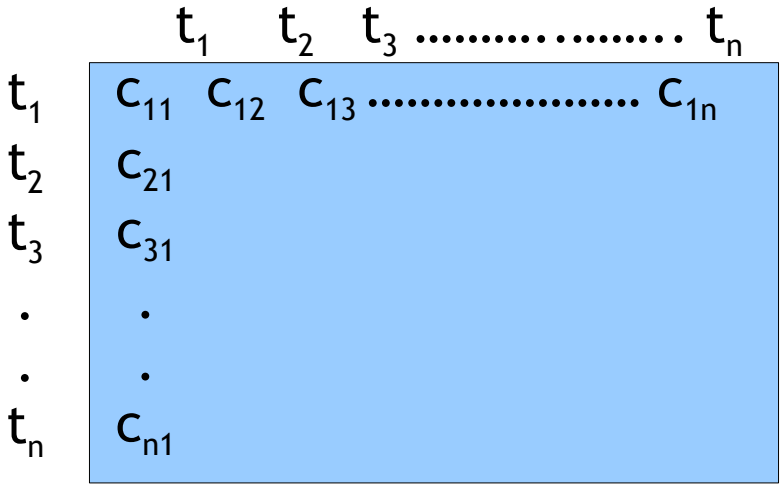
word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazec
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

Expansión de Consultas

- Co-ocurrencia: puede calcularse en base a las similitudes término-término en $C = AA^T$ donde A es una matriz término-documento
  - f_{ij} *es* el peso de t_i *en* d_j (normalizado)



Expansión de Consultas



$$c_{ij} = \sum_{d_k \in D} f_{ik} \times f_{jk}$$

$$s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}$$

- Matriz de asociaciones:
  - c_{ij} factor de correlación entre el término *i* y el término *j*
  - la correlación basada en frecuencia favorece los términos más frecuentes
  - la normalización le da un score de 1 si los dos términos tienen la misma frecuencia en todos los documentos

Expansión de Consultas

- La correlación basada en asociación no tiene en cuenta la proximidad de los términos en los documentos, sino co-ocurrencias de frecuencias
- Correlación métrica:

$$c_{ij} = \sum_{k_u \in V_i} \sum_{k_v \in V_j} \frac{1}{r(k_u, k_v)} \qquad s_{ij} = \frac{c_{ij}}{|V_i| \times |V_j|}$$

  - V_i conjunto de todas las ocurrencias del término *i* en cualquier documento
  - r(k_u,k_v) distancia en palabras entre la ocurrencia de k_u y la de k_v (∞ si k_u y k_v ocurren en diferentes documentos)

Expansión de Consultas

- Para cada término *i*, expander la consulta con los n términos con el valor más alto de c_{ij} (s_{ij})
- Esto agrega términos semánticamente relacionados que están en el "vecindario" de los términos de la consulta

Expansión de Consultas

- Si los términos están muy altamente correlacionados la expansión no va a recuperar mucha información adicional
- La calidad de las asociaciones es cuestionable, posibles errores:
  - falsos positivos: palabras que aparentan ser similares y no lo son
  - falsos negativos: palabras que aparentan ser distintas y son similares

Expansión de Consultas

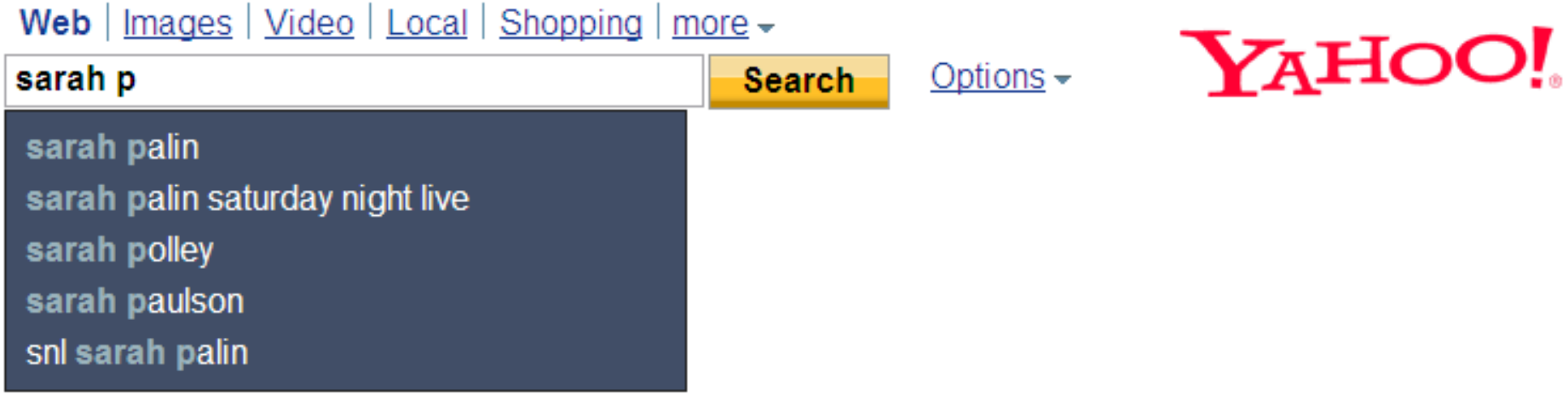
- La ambigüedad de los términos puede introducir términos correlacionados que son estadísticamente irrelevantes
  - Apple computer → Apple red fruit computer
- Refinamiento: solo agregar términos que son similares a todos los términos de la consulta

$$sim(k_i, Q) = \sum_{k_j \in Q} c_{ij}$$

  - *fruit* no se agregar a *Apple computer* porque es distante de *computer*
  - *fruit* se agrega a *apple pie* porque *fruit* es similar a ambas palabras *apple* y *pie*

Expansión de Consultas

- Minería de los logs de query
- Recomienda consultas recientes hechas con frecuencia que contienen el string tipeado por el usuario de manera parcial
- Se analiza cada consulta como un documento y se rankea de acuerdo al matching parcial



Expansión de Consultas

- La expansión de consultas es muchas veces efectiva en aumentar el recall
- En muchos casos la precisión disminuye drásticamente
- En general, la correlación no es tan buena como el feedback de relevancia, pero si puede llegar a ser igual de bueno que el pseudo-feedback de relevancia