webir

Clase 8

Temas

- Estructura de Internet
- Relevance Feedback
- Expansión de consultas
- Intención del usuario
  - Filtrado colaborativo

Estructura de Internet

- Medidas de relevancia independientes del contenido
  - HITS
    "Authoritative Sources in a Hyperlinked Environment" de Jon M. Kleinberg, 1997
  - PageRank
    "The Anatomy of a Large-Scale Hypertextual Web Search Engine" de Sergey Brin, Larry Page, 1998
- Los enlaces de una página a otra representan un especie de aval

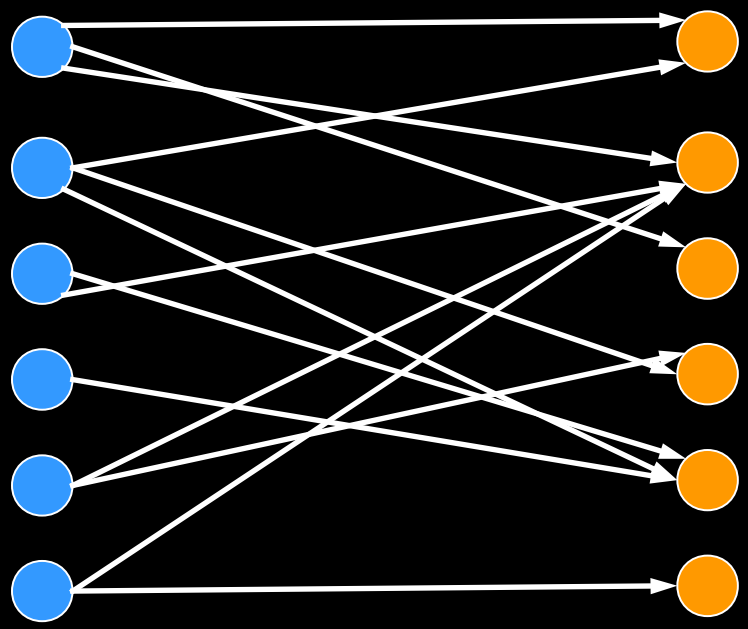
Estructura de Internet

- Los enlaces de una página a otra representan un especie de aval

- No siempre es cierto
  - Navegacionales
  - Organizacionales
  - Spam

HITS – Kleinberg "Hyperlink-Induced Topic Search"

- Dos tipos de nodos



Hubs

Autoridades

HITS - Kleinberg

- Funciones de actualización de los coeficiente de autoridad $\mathbf{a}^p$ y coeficiente de hub $\mathbf{h}^p$ asociados a cada página:

$$a^p \leftarrow \sum_{(q,p)\in E} (h^q)$$

$$h^p \leftarrow \sum_{(p,q)\in E} (a^q)$$

- "Buena" autoridad
  - Si p es apuntado por muchas páginas con alto valor $\mathbf{h}^p$ debería recibir un alto coeficiente $\mathbf{a}^p$
- "Buen" hub
  - Si p apunta a muchas páginas con alto valor $\mathbf{a}^p$ debería recibir un alto coeficiente $\mathbf{h}^p$
- Aplicar de forma iterativa, alternada las funciones de actualización y normalizar
- Inicializar todos los coeficientes a y h con el mismo valor

HITS - Kleinberg

- Devolver las k mejores autoridades y hubs
- ¿Cuantas iteraciones?
  - Se demuestra que mediante las iteraciones se converge a un punto fijo a^* e h^*
  - Convergencia rápida
- No es necesario inicializar todo los coeficiente al mismo valor
- Páginas similares
  - Construir el conjunto base a partir de la página de interés

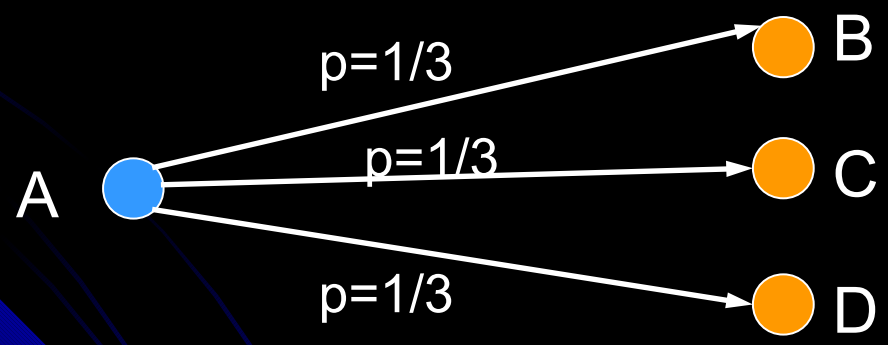
HITS - Kleinberg

- Conjunto base
  - Páginas que contienen el término o concepto – conjunto raíz
  - Agregar aquellas páginas que son apuntadas por el conjunto raíz o que apuntan al conjunto raíz
- Páginas similares
  - Construir el conjunto base a partir de la página de interés

PageRank – Brin y Page

- Coeficiente en el rango [0,1] calculado para cada página

- Un visitante que recorre el grafo al azar



PageRank – Brin y Page

- El visitante que recorre el grafo de acuerdo a probabilidades a priori no conocidas de transición
  - Suponemos equiprobabilidad
- Visitará un mayor número de veces aquellas páginas que
  - tengan más enlaces entrantes
  - son más importantes

PageRank – Brin y Page

- Teleportación
  - Se elige “saltar” con equiprobabilidad a alguna de todas las páginas, $p=1/N$
  - Si no hay enlaces de salida de una página el visitante no podría avanzar
  - Desde cualquier página con enlaces de salida, con probabilidad $p=\alpha$, $0<\alpha<1$
    - Comunmente $\alpha=0.1$

- PageRank es $\pi(v)$, es el tiempo total que el visitante está en el nodo v, depende de N y α

PageRank – Brin y Page

- Modelamos mediante un proceso estocástico de tiempo discreto
  - Períodos de tiempo fijos
  - Decisión en cada uno de esos momentos
- Cadenas de Markov
  - N estados, uno por cada página o nodo
  - Matriz de transición *P,* entre estados, de dimensión $N \times N$
    - $P_{i,j}$ es la probabilidad de transición de un estado *i* a otro *j*
    - $P_{i,j} \in [0,1]$
    - *Propiedad de Markov*, $P_{i,j}$ sólo depende de *i*

PageRank – Brin y Page

- La distribución de probabilidad de una cadena de Markov
  - Vector con *N* componentes, en el rango [0,1], que suman 1
  - Probabilidad de estar en cada uno de los *N* estados

PageRank – Brin y Page

- Para construir la matriz de transición *P*
- Tomar la matriz de adjacencia *A* del grafo de páginas
  - Los valores de las columnas que no tienen 1s se reemplazan por *1/N* (teletransportación pura)
  - Los valores de las columnas que tienen algún 1 se reemplazan por *1/(suma de 1s de la columna)*
  - Multiplicar la matríz por *(1- α)*
  - Sumar *α/N* a cada coeficiente

PageRank – Brin y Page

- La posición de visitante se describe en cada momento con el vector $x = (x_1, x_2, \ldots x_N)$
- La posición del visitante en $t=0$ es el vector x con todas las coordenadas en 0, salvo la que corresponde a la página/nodo donde comienza $x = (0, 1, 0, \ldots 0)$
- La posición de visitante en $t=1$ es xP
- La posición de visitante en $t=2$ es $xPP=xP^2$

PageRank – Brin y Page

- Una cadena de Markov es ergódica si existe un valor entero positivo T_0, tal que para todas las parejas de estados *i, j* de la cadena, si se empieza en tiempo 0 en el estado *i*, para todo $t>T_0$ la probabilidad de llegar al estado *j* es mayor que 0

- Condiciones necesarias y suficientes
  - Irreducible: secuencia de transiciones de probabilidad mayor que 0 para toda pareja de estados *i, j*
  - Aperiódica: no hay subconjuntos de estados que sólo tienen transiciones con probablidad positiva entre ellos

PageRank – Brin y Page

- **Teorema:** una cadena de Markov ergódica tiene un único vector de probabilidades de estados para el estado estacionario π
- Si $\eta(i,t)$ es el número de visitas al estado i

$$\lim_{t\to\infty} \frac{\eta(i,t)}{t} = \pi(i)$$

- $\pi(i)$ es la probabilidad en el estado estacionario para el estado i
- Coincide con el vector propio principal de P
- Se define PageRank de i como el valor $\pi(i)$

PageRank – Brin y Page

- ¿La cadenas definidas anteriormente con teletransportación son ergódicas?

- Condiciones necesarias y suficientes
  - Irreductible, secuencia de transiciones de probabilidad mayor que 0 para toda pareja de estados i, j
  - Aperiodica, no hay subconjuntos de estados que sólo tienen transiciones con probablidad positiva entre ellos

- $\pi(i)$ representará el PageRank de cada *i*

PageRank – Brin y Page

- Cálculo de PageRank
  - Calcular el valor propio izquierdo de P que es el vector π, tal que $\pi P = \lambda \pi$
  - Simular el camino visitado mediante sucesivas iteraciones
    - Hasta que hayan pocos/imperceptibles cambios
    - Menos de un umbral μ

PageRank – Brin y Page

- Independiente de la consulta
  - **Medida global y estática**
- Se usa junto con otras medidas
- Aprendizaje automático para ver la importancia de cada una de las medidas

- PageRank por temas
- ¿Como se podría incluir información de la consulta?

PageRank – Brin y Page

- Independiente de la consulta
  - **Medida global y estática**
- Se usa junto con otras medidas
- Aprendizaje automático para ver la importancia de cada una de las medidas

- PageRank por temas
- ¿Como se podría incluir información de la consulta?
  - En el conjunto base
  - En las probabilidades de transición – por ej. saltos (solo deportes, noticias, etc.)

Problema de Sinónimos

- aircraft = {plane, airplane}
- **Métodos locales**
  - Dependiente de los documentos devueltos
  - Se usan los resultados de la consulta
  - **Relevance Feedback (RF), Pseudo RF, Global RF**
- **Métodos globales**
  - Independientes de la consulta y resultados
  - Expansión de consultas mediante sinónimos
  - Corrección ortográfica

Relevance Feedback

(1) Consulta
(2) Conjunto de documentos como resultado
(3) Evaluación del usuario
(4) Relevante/no relevante
(5) Revaluar la necesidad de información del usuario
(6) Nuevo conjunto de documentos como resultado

Eventualmente se repite

Relevance Feedback – Rocchio

- Algoritmo de los 70’
- Agregar RF al modelo de espacio vectorial de los documentos y las consultas
- Se quiere encontrar una nueva consulta $q_{opt} = \max_q [\text{sim}(q, C_r) - \text{sim}(q, C_{nr})]$
  - sim puede ser similitud en coseno
  - Maximizar similitud con docs relevantes
  - Minimizar similitud con docs no relevantes
  - C_r y C_{nr} en realidad no son conocidos

Relevance Feedback – Rocchio

- Similitud en coseno

$$\overrightarrow{q_{opt}} = \frac{1}{|C_r|}\sum_{\vec{d}_j \in C_r}\vec{d}_j - \frac{1}{|C_{nr}|}\sum_{\vec{d}_j \in C_{nr}}\vec{d}_j$$

- Diferencia entre los centroides de C_r y C_{nr}
- C_r y C_{nr} en realidad no son conocidos

$$\overrightarrow{q_m} = \alpha\overrightarrow{q_0} + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

- Se usa el conjunto de documentos evaluados (D_r y D_{nr}) y la consulta original para producir variaciones de la consulta

Relevance Feedback – Bayes

- Clasificador mediante regla de Bayes – peso de t
- $P(x_t=1 \mid R)$ probabilidad de que el término *t* aparezca en el documento
  - R es la variable que indica la probabilidad de que un documento sea relevante

$$P(x_t = 1|R = 1) = |VR_t|/|VR|$$

$$P(x_t = 1|R = 0) = (df_t - |VR_t|)/(N - |VR|)$$

  - *N* número total de documentos
  - df_t − número de docs que contienen a *t*
  - *VR* es el conjunto de docs relevantes <u>conocidos</u>
  - VR_t es el subconjunto de *VR* que contienen a *t*
- Indica como modificar los pesos de los términos de la consulta

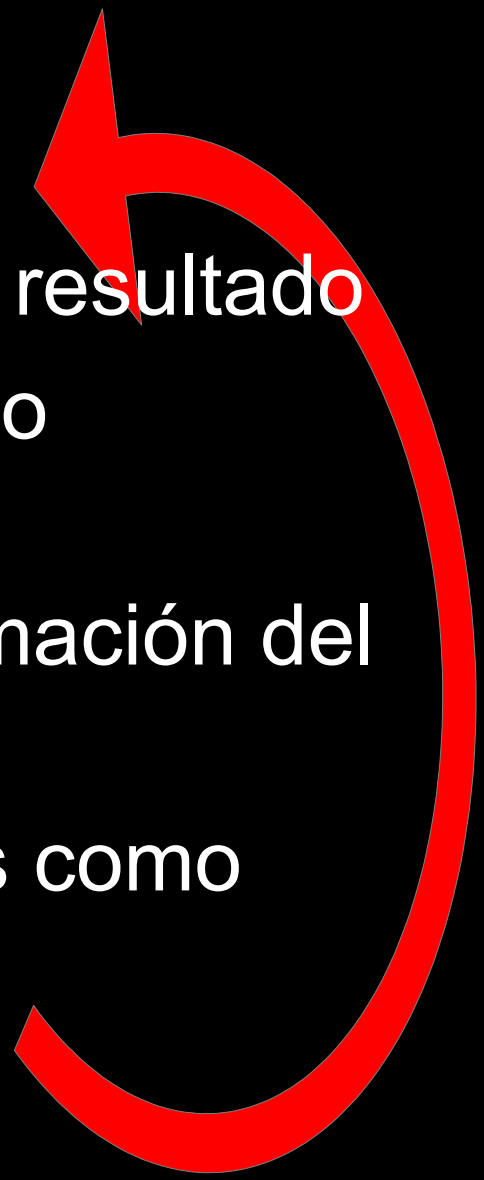
Relevance Feedback – Evaluación

- ¿Precisión y Recall?
  - Con los documentos ya evaluados
  - Sin los documentos ya evaluados
  - Dos colecciones
- ¿Cuánto tiempo le lleva a un usuario encontrar información relevante?
- ¿Cuántos documentos relevantes encuentra un usuario en un tiempo fijo?

Pseudo Relevance Feedback

(1) Consulta del usuario
(2) Conjunto de documentos como resultado
(3) Tomar los *k* primeros docs como relevantes!
(4) Revaluar la necesidad de información del usuario
(5) Nuevo conjunto de documentos como resultado

Eventualmente se repite



Relevance Feedback Indirecto

- Implícito – fuentes de evidencia indirecta
- Más confiable que pseudo RF
- Estadísticas de preferencias de usuarios
  - Consultas y documentos
- Dar mayor puntaje a documentos más populares – elegidos muchas veces
  - Relacionado con la consulta o no
- *clickstream mining*
- Anuncios comerciales

Expansión de Consultas

- Métodos globales
- Los usuarios pueden dar alternativas para cada término de la consulta
- Se sugieren otras consultas relacionadas
- Se usa diccionario de sinónimos
- Se puede combinar con el puntaje dado a los términos
  - Términos modificados – menor importancia

Expansión de Consultas

- Diccionario de sinónimos
  - Lista manual
  - Lista derivada automáticamente – ocurrencias simultaneas – en la colección
  - Minería de logs de consultas – consultas y respuestas

Intención del Usuario

- Modelos para entender a los usuarios
- Broder (2002)
  - Informacional
  - Navegacional
  - Transaccional
- Clasificación automática

Intención del Usuario

- Jansen, Booth y Spink (2007) “Determining the User Intent of Web Search Engine Queries”
- Clasificación automática de ~ un millón de consultas en base a características identificadas en 5 millones de consultas
  - 80% informacionales
    - “How to”, “What is”
    - List, playlist, etc.
    - Cuando el usuario inspecciona varias páginas resultado
  - 10% navegacionales y transaccionales
  - Evaluación de resultados en base a 400 consultas clasificadas manualmente
  - 75% de resultados correctos

Recomendaciones Filtrado Colaborativo

- Grandes cantidades de datos
  - Se espera que los usuarios evalúen los objetos
- Extracción del “conocimiento colectivo” para recomendar objetos a usuarios
  - Búsquedas de usuarios con gustos similares
  - Recomendaciones específicas para cada usuario
  - Usando información de muchos usuarios
  - No es lo mismo que contabilizar el total de accesos a un documento

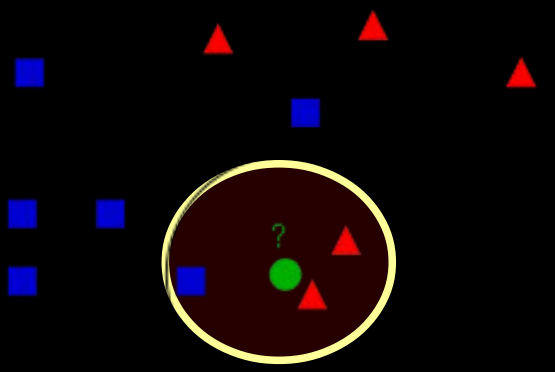
Filtrado Colaborativo

- Similitud entre usuarios
  - Dado un usuario activo representado por el vector de su elecciones previas
  - Buscar usuarios que evalúan de forma similar los objetos
    - Coseno o Pearson (refleja la correlación (-1,1))
  - Encontrar nuevos objetos para recomendar al usuario activo

	d1	d2	d3	d4	...
u1	3	5	4	1	
u2	5	0	2	4	
u3	2	4	?	?	
...					

Filtrado Colaborativo

- Similitud entre usuarios
  - Dado un usuario activo representado por el vector de su elecciones previas
- Algoritmo de agrupamiento
  - “vecino más cercano” - knn clustering



Filtrado Colaborativo

- Similitud entre objetos
  - Buscar objetos similares
    - Evaluados de forma similar por los usuarios
    - Matriz de relaciones entre parejas de objetos
    - Coseno o Pearson (refleja la correlación (-1,1))
  - Encontrar el “gusto” o “perfil” de los usuarios
- Amazon

Minería de Bitácoras de Consultas

- Información de los usuarios
  - Evaluación de los resultados no se puede tomar en una escala
  - Evaluación de los resultados se toma sólo como relevante/no relevante
  - Evaluación de los resultados se toma como el acceso a los documentos

Minería de Bitácoras de Consultas

- Otros métodos
  - Representación mediante grafos
  - Agrupamiento aglomerativo
    - “Aglomerative clustering of search engine query logs” de Beeferman y Berger (2000)

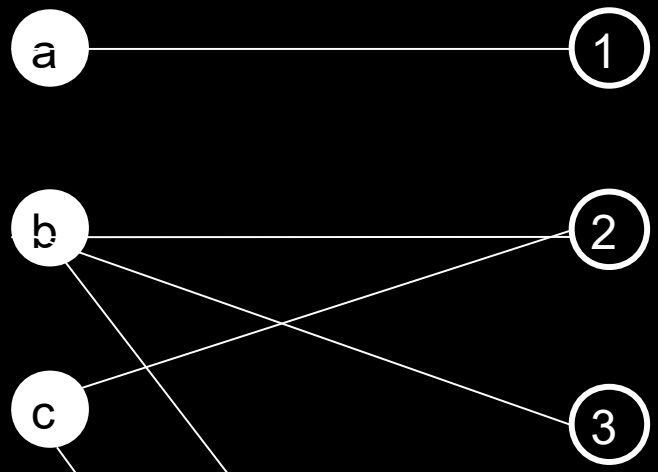
Minería de Bitácoras de Consultas- Agrupamiento Aglomerativo

- Grafo bipartito
- Nodos
  - Consultas
  - Documentos (URLs)

- No importa el contenido
- No influyen diferentes formas de escribir las mismas consultas

Minería de Bitácoras de Consultas- Agrupamiento Aglomerativo

Consultas Documentos (URLs)



Minería de Bitácoras de Consultas- Agrupamiento Aglomerativo

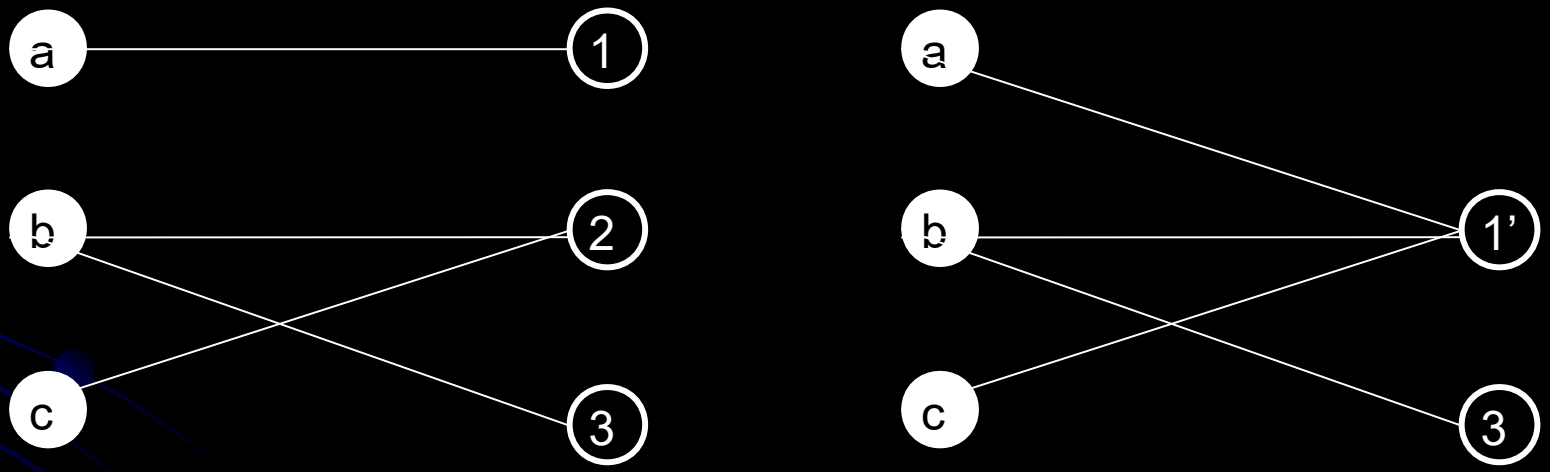
- Si *N(x)* son los vecinos de *x*
- Dos vértices *x* e *y* son similares si tiene una alto grado de coincidencia en *N(x)* y *N(y)*
- *Similitud (Jaccard)*

$$\sigma(x,y) \stackrel{\text{def}}{=} \begin{cases} \dfrac{\mathcal{N}(x) \cap \mathcal{N}(y)}{\mathcal{N}(x) \cup \mathcal{N}(y)}, & \text{if } \mid \mathcal{N}(x) \cup \mathcal{N}(y) \mid > 0 \\ 0, & \text{otherwise} \end{cases}$$

- Esta definición de similitud no diferencia entre vértices que coinciden en 1 sólo vecino o en varios

Minería de Bitácoras de Consultas- Agrupamiento Aglomerativo

- Fusionar los nodos más similares de cada lado
- Alternar



- Hasta condición de parada
  - Reducir a componentes no conexas
    - Hasta que la similitud más pequeña es 0
  - Hasta llegar a un número determinado de grupos de documentos – temas