Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Industrial
IN5526

Repaso y pauta

1. HITS

HITS es un algoritmo que se basa en definir por búsqueda dos puntajes distintos para cada elemento (página web), diferenciando si son un centro o una autoridad. Se basa en que un centro es tan bueno como la suma reputación de autoridades que provee, y las autoridades son tan buenas como la suma de reputación de centros que las proveen. Esto en ecuaciones es:

$$h(v) = \sum_{v \leftarrow y} a(y) \quad (1)$$

$$a(v) = \sum_{y \leftarrow v} h(y) \quad (2)$$

Es importante remarcar que HITS busca primero a partir de una query los posibles documentos relevantes (por ejemplo a través de tf-idf), una vez que ha seleccionado esto busca los valores h y a que satisfacen (1) y 2. Para esto, tomando M como la matriz de adyacencia del digrafo se itera

$$H \leftarrow M \overrightarrow{A} \quad (3)$$

$$A \leftarrow M^T \overrightarrow{H} \quad (4)$$

Aquí además tomando la ec (3) y aplicandola en 2 y recíprocamente se tienen las relaciones

$$H \leftarrow MM^T \overrightarrow{H} \quad (5)$$

$$A \leftarrow M^T M \overrightarrow{A} \quad (6)$$

Dado que MM^T y $M^T M$ no son estocásticas, es necesario normalizar H y A en cada paso. Es importante notar que en esta última formulación el vector H nunca se calcula a partir de A, ni A a partir de H, por lo que el problema se separa.

2. Preguntas y respuestas:

- De dos ventajas y dos desventajas de PageRank y HITS:
  HITS tiene la ventaja de tener dos scores para cada elemento lo que puede ser relevante en cuiertos tipos de escenarios y que los resultados (scores) están especializados en la query. Desventajas son que debe computarse en el tiempo en que se hace la query (más lento que pagerank) y que pueden aparecer centros y autoridades irrelevantes como relevantes debido a la circularidad del algoritmo.
  PageRank tiene la ventada de ser precomputado y por lo tanto eficiente en tiempo de consulta, además el algoritmo tiene asegurada la convergencia a un único valor. Algunas desventajas es que no entrega un valor contextual a la búsqueda y puede tener comportamientos problemáticos con los ciclos.

- En el contexto de WUM comente sobre el filtrado de logs no relevantes o que podrían generar relaciones espurias:
  - Logs generados por Crawlers
  - Logs que no están relacionados con el doc. HTML (gifs, sonido, etc.)
  - Sesioes difíciles de identificar
  - Sesiones cortas
  - Sesiones muy largas con periodos largos de inactividad
- ¿De qué sirve para un negocio identificar autoridades (en sentido amplio) en comunidades de la web? Desde un punto de vista de negocio sirve para reconocer elementos (páginas) en las cuales ciertos tipos de publicidad puedan ser más efectivos o con lo cuales hacer publicidad (personas).
- ¿Qué similitudes tiene la mineria de datos de redes sociales con la mineria de estructura web? En ambos casos la estructura subyacente se puede modelar como un grafo dirigido. Además de compartir estructura en el modelamiento, tienen también problemas análogos que si bien pueden representar distintas situaciones, se resuelven con el mismo enfoque (algoritmo), un ejemplo de esto puede ser PageRank para encontrar usuarios influyentes o páginas relevantes.

3. Para responder:

- ¿Cómo es un filtro colaborativo para sistemas de recomendación?
- ¿En qué ayuda la WUM para el diseño de la estructura de una página web?