Web information retrieval - 2018/2019

Exam - January 23rd, 2019

Time: 60 minutes

Assignment 1

1. Consider tf-idf weights for documents.
    - Write the $tf \times idf$ weighing equation. Explain how each term is defined.
    - Assume we compare three pairs of documents: i) two docs that have *only* frequent words (the, a, an, of, etc.) in common; ii) two docs that have *no* word in common; iii) two docs that have many rare words in common. Rank the above pairs according to their cosine similarity score and explain convincingly the reasoning behind your choice.
2. State at least one convincing reason why IR systems usually use cosine similarity instead of Euclidean distance.

You should clearly motivate your answers

Assignment 2

Consider tf-idf weights for documents.

- Write the $tf \times idf$ weighing equation. Explain how each term is defined.
- Assume we compare three pairs of documents: i) two docs that have *only* frequent words (the, a, an, of, etc.) in common; ii) two docs that have *no* word in common; iii) two docs that have many rare words in common. Rank the above pairs according to their cosine similarity score and explain convincingly the reasoning behind your choice.
- State at least one convincing reason why IR systems usually use cosine similarity instead of Euclidean distance.

Assignment 3

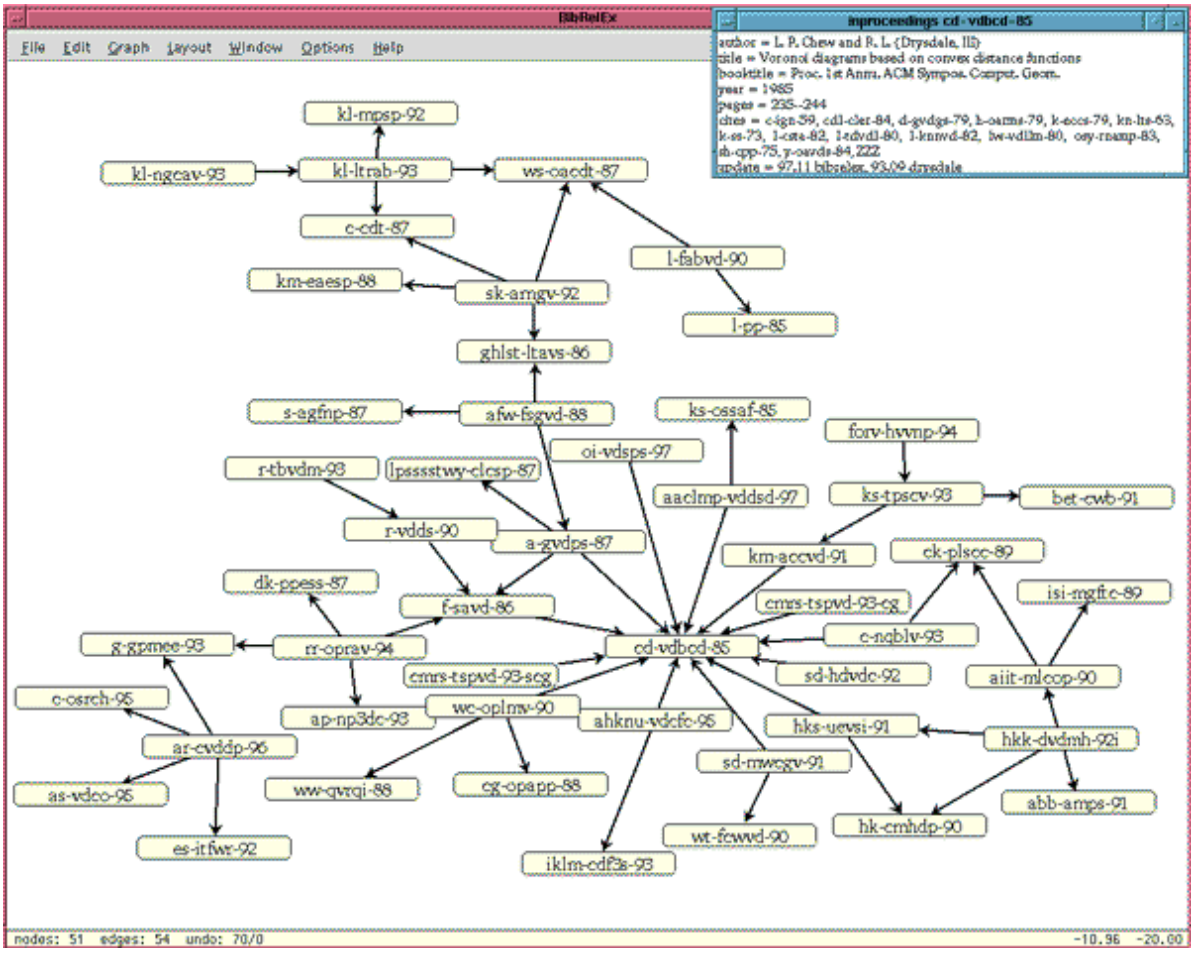
Consider the HITS algorithm applied to a (directed) graph with adjacency matrix $\mathbf{A}$. Denote by $\mathbf{a}(t)$ and $\mathbf{h}(t)$ respectively the t-th values of the authority and hub vectors.

1. Write down one iteration of the algorithm, i.e., show how the hub and authority vectors are updated in each round of the algorithm;
2. Give an example of a network in which some of the vertices have hub score 0.

Introduce whatever notation you think necessary.

Assignment 3

We are a citation graph like the one in the picture below.



Here, vertices represent scientific papers (their label are unique Bibtex identifiers) while, given two vertices u and v, *directed* edge (u, v) exists if and only if paper u cites paper v. A standard measure of a scientific work's important is the *impact factor* which, for a given paper u, is simply the in-degree of the corresponding vertex.

3.1. i) Discuss how Pagerank could be used as an alternative measure of papers' scientific impact; ii) Give an example of a (small) citation network in which the importance of a paper according to the standard impact factor and to Pagerank might be considerably different.

3.2. Assume that, given a citation network G and paper u, you want to rank all other papers in the network with respect to i) their authoritativeness; ii) some notion of "closeness" to u. Give the details of a Pagerank-based method that might achieve this goal.

Introduce whatever notation you think necessary.