

Written exam

Mauro Brunato

Elisa Cilia

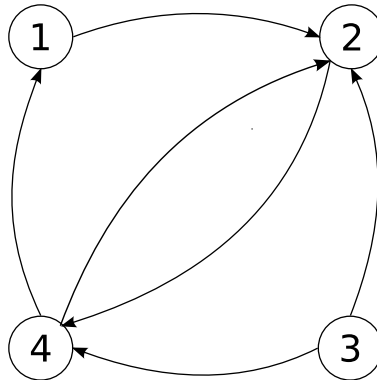
February 9, 2009

Exercise 1

Given the following collection D of documents composed of four pages:

- d1 : The new generation of operating systems currently being developed in Apple and Microsoft will be based on information retrieval.
- d2 : Apple has not officially announced to the market the new operating system based on information retrieval, yet. However, there is evidence to claim that they are operating with a project similar to Microsoft's Longhorn.
- d3 : Microsoft and Apple are currently leaders in operating systems.
- d4 : Microsoft has officially announced that Longhorn will appear on the market on the year 2006. They are trying to fire Apple which is working on a similar project.

with the following hyperlink structure:



1.1) Define the set of (ten) keywords and give documents the corresponding representation in the TF-IDF space (consider $IDF(t) = \frac{|D|}{|D_t|}$), after having defined the stop word elimination and the stemming processes.

1.2) Consider two queries:

q1 : Longhorn

q2 : information retrieval

rank the documents with respect to each one of the queries according to the scalar product and then according to the cosine similarity. Are the results the same?

1.3) Let $\rho(d, q)$ be the cosine similarity and denote by $x(d)$ the PageRank value of d according to the constructed hyperlink structure. Let us define the following integrated notion of similarity which takes into account both the link analysis and the cosine query-document similarity as follows: $\phi(d, q) = \rho(d, q) \cdot x(d)$

Find the new ranking corresponding to queries q1 and q2.

1.4) Assume that the documents —limited to the set of ten keywords— agree with a multinomial model (where term occurrences are considered as independent) and estimate the parameters of the distribution by using the least squares method.