Nome e Cognome:

Matricola:

Ricerca dell'Informazione nel Web

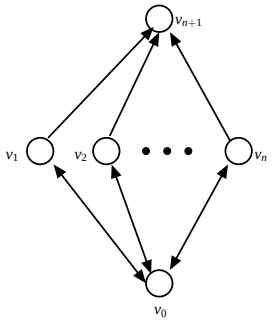
Compito di esame del 2 Febbraio 2012, *tempo a disposizione: 90 minuti*
5 punti/problema

Problema 1

1. For which of the following queries are skip pointers most useful, and for which are completely useless? Briefly explain your answers.
   (a) `x AND y`, where `x` is a frequent term and `y` rare
   (b) `x AND y`, where both `x` and `y` are frequent terms.
   (c) `x OR y`, where `x` is a frequent term and `y` rare
   (d) `x OR y`, where both `x` and `y` are frequent terms.
2. Write a pseudocode of an algorithm for merging two postings lists for a query of the type `term1 AND term2` using skip pointers.

Problema 2

1. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α. Compute the pagerank values for $n = 4$ and $\alpha = 1/2$.
2. Consider the random surfer model. Assume now that in addition to the teleporting, the user has a probability β to go to page v_0. Can this process be modeled as a Markov chain? If so write the necessary equations needed to calculate the stationary distribution.
3. Assume now that in each step, the user in addition can use the *back* button, allowing him to go one step back. Can this process be modeled as a Markov chain? If so write the necessary equations needed to calculate the stationary distribution.



Problema 3

1. Write the *tf*× *idf* weighting equation. Explain what each term represents, and the reasoning about the equation.
2. Consider an IR system where we use the *tf*× *idf* weighting scheme. We compare three pairs of documents:
   (a) Two docs that have only frequent words (the, a, an, of, etc.) in common.

(b) Two docs that have no word in common.

(c) 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.

3. State two reasons that in IR we usually use cosine similarity instead of Euclidean distance.

Problema 4

The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of a collection of 30 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list.

R R N R N N R R N N N R N N N N R N N R N N N N N N R N R N

1. What is the precision of the system on the top 20?
2. What is the recall on the top 20?
3. What is the F_1 measure on the top 20?
4. Draw the precision-recall curve.
5. What is the interpolated precision at 33% recall?