

## 2010 Paper 8 Question 9

### Information Retrieval

- (a) The vector space model of document retrieval uses the notion of an *information space*.
- (i) What are the typical basis vectors for this space? What are the vectors in the space intended to represent? [2 marks]
  - (ii) How is *term weighting* used to position objects in the space? Why is term weighting important for effective document retrieval? [3 marks]
  - (iii) What is the *orthogonality assumption* typically employed in vector space models of document retrieval? Why might this assumption be false, and why might it lead to errors in retrieval? [3 marks]
  - (iv) Suggest one way in which the orthogonality assumption could be relaxed. (Just a short high-level description of a possible method is required.) [2 marks]
- (b) The PageRank algorithm uses a model of a “random surfer” to calculate the validity of a page.
- (i) Explain how the random surfer moves about the web. [2 marks]
  - (ii) Describe how the random surfer can be modelled as an ergodic Markov chain, and how this leads to the PageRank values being calculated as the principal left eigenvector of the transition probability matrix. (You are not required to give a formal definition of an ergodic Markov chain; an informal description will suffice.) [8 marks]