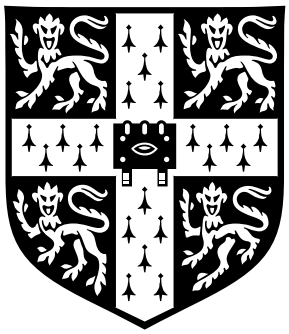
Probabilistic Document Modelling

Philip J. Cowans
Churchill College
Cambridge

A dissertation submitted in candidature for the degree of Doctor of Philosophy,
University of Cambridge

Inference Group
Cavendish Laboratory
University of Cambridge



January 2006

Declaration

I hereby declare that my dissertation entitled "Probabilistic Document Modelling" is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University.

I further state that no part of my dissertation has already been or is being concurrently submitted for any such degree or diploma or other qualification.

Except where explicit reference is made to the work of others, this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration. This dissertation does not exceed sixty thousand words in length.

Date: ..... Signed: .....

Philip J. Cowans
Churchill College
Cambridge
January, 2006

Abstract

In this thesis the development and application of probabilistic models of documents is considered. The initial focus is on language models which provide a way of modelling plain text documents. In particular the hierarchical Dirichlet language model, which is derived from simple Bayesian theory, is investigated and is shown to be well approximated by an existing method known as generalised PPM-A. Using this equivalence, generalised PPM-A is extended to produce a language model which while working on the level of individual letter-like symbols is able to make use of the division of the text stream into words. It is shown that the new model can be used in conjunction with a word list to improve performance when very little information from which to learn the statistics of the language is available.

The hierarchical Dirichlet model is then applied to the task of information retrieval, producing a new retrieval method which naturally includes document frequency information. This information has traditionally been used in retrieval systems, but previously had either been missing or introduced heuristically in language model based approaches to the problem. The hierarchical approach is also extended to the task of retrieval at the passage level where it is shown to give promising results.

Finally, the scope of the investigation is broadened to include documents which contain diagrams as well as plain text. A method is developed to group fragments of digitised ink strokes into perceptually relevant components of a diagram, while at the same time labelling the components with an object class. The approach, which is based on the conditional random field, is shown to work well both in terms of grouping and improving labelling performance when compared to other methods.

Acknowledgements

I would like to thank Ryan Adams, Tom Minka, Iain Murray, Edward Ratzer, Oliver Stegle, David Stern, Martin Szummer, Keith Vertanen, Hanna Wallach, David Ward, Seb Wills and John Winn for ideas, suggestions, discussions and feedback without which the work presented in this thesis would not have been possible.

During my studies I was lucky enough to be able to spend three months working in the Machine Learning and Perception group at Microsoft Research UK, an opportunity for which I am grateful and which produced many of the ideas presented in Chapter 6.

Above all, I would like to thank my supervisor, Professor David MacKay for an unending supply of inspiration, discussion and feedback, and for providing invaluable advice on the draft of this thesis. I would also like to thank my parents, Sue and Trevor Cowans, and my brother, Nick, for their support and encouragement during the last four years.

This work was supported by a grant from Microsoft Corporation.

Contents