A BAYESIAN APPROACH TO INFORMATION RETRIEVAL FROM SETS OF ITEMS

Katherine A. Heller[1], Zoubin Ghahramani[2,3]

(1) Gatsby Computational Neuroscience Unit,
University College London, London WC1N 3AR, UK

(2) Department of Engineering,
University of Cambridge, Cambridge CB2 1PZ, UK
http://learning.eng.cam.ac.uk/zoubin

(3) Machine Learning Department,
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

We consider the problem of retrieving items from a concept or cluster, given a query consisting of a few items from that cluster. We formulate this as a Bayesian inference problem based on models of human categorization and generalization and describe a very simple algorithm for solving it. Our algorithm ends up with a score which can be evaluated exactly using a single sparse matrix multiplication. This makes it possible to apply the method to retrieval from very large datasets (i.e. millions of items). We evaluate our algorithm on three problems: retrieving movies from a database of movie preferences, finding sets of similar authors based on their word usage in a scientific conference, and finding completions of sets of words appearing in encyclopaedia articles. Compared to "Google Sets", we show that our "Bayesian Sets" retrieval method gives very reasonable set completions. Finally, we show how the Bayesian Sets algorithm can form the basis of a Content-Based Image Retrieval (CBIR) system. I will describe and demonstrate this Bayesian CBIR system and mention a range of other applications of our approach.

Key Words: Information retrieval, Vision, Image Retrieval, Google