Proposal for CS4246R Project: Markov Decision Process for Focused Web Crawling

U096883L Shawn Tan

December 22, 2011

1 Introduction

Since its conception, the amount of content on the World Wide Web has been growing exponentially. Systems have been theorised and built to cope with this large amount of data by either categorising them, or indexing them to make search possible.

With the introduction of more user-generated content in recent years, efforts have been made toward understanding the data put up on the web. Data-mining usually involves crawling websites and analysing the data using machine learning techniques.

However, crawling the web takes up a significant amount of resources. In particular, crawlers typically open several data streams to download pages simultaneously. Many of these pages do not have important data, resulting in wastage of bandwith. Focused crawling aims to reduce this redundant IO costs by crawling only pages deemed relevant to the topic requested.

The task of focused crawling requires the crawler to be able to discern wanted pages from unwanted pages, but by immediately discarding unwanted pages, the crawler may miss wanted pages that the discarded page linked to. As a result, some elements of planning have to be included in the design of a focused crawler in order to reduce loss of relevant pages.

The proposed project intends to investigate how Markov Decision Processes can be used for the purpose of focused crawling and its viability compared to other approaches.

2 Related Work

3 Implementation

4 Conclusion