Wikipedia Structure Mining
Proposal

Harrison McKenzie Chapter       Evan Hecht
Aleksandr Movsesyan          Nathan Giusti

October 20, 2009

Abstract

We will be mining a provided dataset from Wikipedia to attempt to identify “communities” of pages which are related. We plan to implement a “fuzzy” clustering algorithm to identify clusters which share tertiary nodes. Further, we hope to identify useful metrics for the “closeness” of two Wikipedia pages using both page linking structure as well as comparing extra, provided mark-up “tag” information on the pages.
1 Introduction

We will be attempting to derive community categorization from a provided, marked-up set of XML documents containing the link structure of a set of Wikipedia pages, as well as some additional “tagging” data which has been added. We will perform an analysis of the link structure of the Wikipedia documents to attempt to find small, coherent sections of the Wikipedia “web”, which represent a particular topic. Recognizing that many pages on Wikipedia do not belong exclusively to one community, we have chosen to pursue a “fuzzy” clustering algorithm which does not rely upon the strict partitioning of segments of the web. We hope that this approach will enable us to produce more meaningful results.

The data for this project is being acquired from the INEX corpus of XML encoded Wikipedia pages. INEX is holding a competition to find effective means of clustering the data to increase the efficiency of searching for pages based on context information. If the data that they have can be effectively clustered into communities surrounding a subject, the search space for finding a related page can be greatly decreased. If our efforts prove fruitful, we may contribute our solution back to their efforts to improve search efficiency for their document collection.

2 Problem Statement

For a set of documents $D$, encoded in XML with tag information, find the communities $C$ which consist of highly mutually-relevant documents.

3 Proposed Solution

We intend to use the algorithm described in [1], which combines two compatible methods to determine clusters and iteratively evaluates their performance to find the right number of clusters. Fuzzy K-means clustering finds local-minimum cluster centers, and is strongly determined by initial positions. Fuzzy maximum-likelihood clustering generally finds better results, but only when starting from ‘good’ positions. The algorithm starts with random positions, and uses K-means to determine cluster centers, which it uses as initial positions for the maximum likelihood method. It then measures the cluster performance, and increases the number of clusters until optimal clusters are found.

4 Dataset

We will be working with the “small” INEX dataset, which consists of about 50,000 Wikipedia pages. The data includes structure, tags, and semantic markup.
We will likely focus on link structure between pages, using semantic tags for illustrative rather than algorithmic purposes. The data is provided pre-processed for various purposes, so we will essentially be dealing with just the link information.

5 Bibliography

References


