Proposal: INEX 2009 XML Mining Track
Competition

James Reed  Jonathan McElroy
jreed@calpoly.edu  jmcelroy@calpoly.edu

Bryan Clevenger
bcleveng@calpoly.edu

Abstract

Data in the world, and more specifically on the Internet is growing to massive sizes. In order to make this information more useful, it must first be more accessible. The INEX Initiative competition is aimed with the goal of identifying and comparing methodologies for categorizing information into clusters. The competition will be run on 60 gigabytes of data from Wikipedia, with the ultimate goal being accurate categorization and clustering in order to reduce search time through this data. As a part of that competition, we propose to run a clustering algorithm on thee data using different evaluation criteria, in an attempt to find the optimal classifications.
1 Introduction

The ultimate goal of the competition here appears to be to optimize search times based off of search keywords by grouping pages into some form of categorical clusters. However, an important side-note is that the contest guidelines also clearly stipulate that in addition to looking for a best-off solution, they’re also interested in a comparison between solutions that are derived from standard evaluation criteria such as:

- Entropy
- F-score
- Normalised Mutual Information
- Etc...

As such, it seems important to attempt to isolate and control standard variables to determine their immediate effect on the results of the algorithms.

2 Problem Statement

Given a known set of XML tagged Wikipedia pages, devise an unsupervised learning strategy to categorize the pages into optimal clusters. Specifically, we intend to focus more strongly on the classification task as outlined by INEX.

3 Proposed Solution

We expect the main focus of our solution to be on the task of providing effective classifications for the given data. At this point, it is our intent to create the different classifications, and run different already-known machine learning algorithms to sort the pages. After this, we hope to be able to measure the effectiveness of different classification categories.

In order to initiate our experimentation, the first step will be getting a better handle on the offered INEX data corpus, as discussed in 4. Next, we expect to be able to devise metrics to determine appropriate initial rough categorization. The end-goal is to be able to refine these metrics to provide the most efficient (or at least relatively decent) categorizations to assist in the ultimate end-goal of creating effective clusters/communities.

4 Data/Datasets

The dataset for our experimentation will consist of the XML tagged Wikipedia pages provided by INEX[1]. Specifically we intend to use the ”small” data set
which consists of roughly 50,000 interlinked pages. This data is provide pre-processed, and comes in the numerical form. At this point, we’re not entirely certain the as to the exact content of the data. The next step in our project is to begin parsing through and determining the best way to use what is available.

5 References

References