

LFY - A Google-based People Search Engine

(1) Project Goal

Given a person's name, we can search for it in GOOGLE and get a huge number of results, but these results are normally not about the same person because so many people have the same name. It is annoying and difficult to identify what web pages are really about the person you are looking for.

The project goal is to classify these results into groups like Athletic, Professor, Business Man, Office Worker, Government Employee, Actor, Dancer, and Singer...

(2) Dataset

Dataset 1:

From Online directory (Yahoo, Looksmart, Google) : fetch all web pages in the target directory and convert them into txt files. For example: singer category, we fetch all web pages in that directory and extract text words

Dataset 2:

From Google Web API: Google is providing a web service interface so that we can get all Google search result through this interface. The same way, given the category, we can use this category as a keyword to fetch all web pages and extract text words.

(3) Vocabulary for each category

Extract vocabulary from Dataset based on the frequency of words, domain independent words like "is", "you" and "and" will be ignored.

(4) ML Techniques we use

Naive Bayes classifier, the nearest classifier, decision trees and subspace method are commonly used. Naive Bayes method is our first choice, and if we have got more time, we use multiple methods as meta-learner and then combine them.

(5) Group Member

Haibo Zhao: zhaohb@uga.edu Department of Computer Science
Haibao Tang: bao@uga.edu Department of Plant Biological

(6) Project Homepage

<http://denali.cs.uga.edu/lfy/>