People Search
Sam Henke, Yandong Liu, Donee McAllister, Dan Muriello, Ashish Panday, Jennifer Sims

The Problem
Information about people on the Internet is both unstructured and disperse.

It is unstructured in the sense that it is represented in a human-readable — rather than a machine-readable — format, and it is disperse in the sense that information about any one person may be spread out across many pages on the Internet.

This presents a problem when people try to use a general-purpose search engine to locate information.

The Goal: SuperPages
What we want is to be able to search once and find everything about a person — his or her contact information, photos, links and snippets about a person, etc.

This should be aggregated and presented to the user on a single page.

Our Solution
MySQL

django

Building Superpage

Identifying Home pages
TF-IDF Model
How much of Bob's PH entry data is in the page?

How much of anyone else's PH entry data is in the page?

If the spread is too even, not a homepage.

Tips:
Remove all stop words
Only look within Bob's department

Learning
Training Data:
People who listed their homepage in the phonebook

Features:
LastName, FirstName, UICID, UIC Phone, Department, Phone, Address

Learn a linear classifier:

Tradeoffs:
A learned classifier is very fast (~25 s/page), but accuracy is upper-bounded near 75% due to inconsistencies in the training data.

TF-IDF is very slow (~60 seconds/page), but accuracy can get as high as 86%.