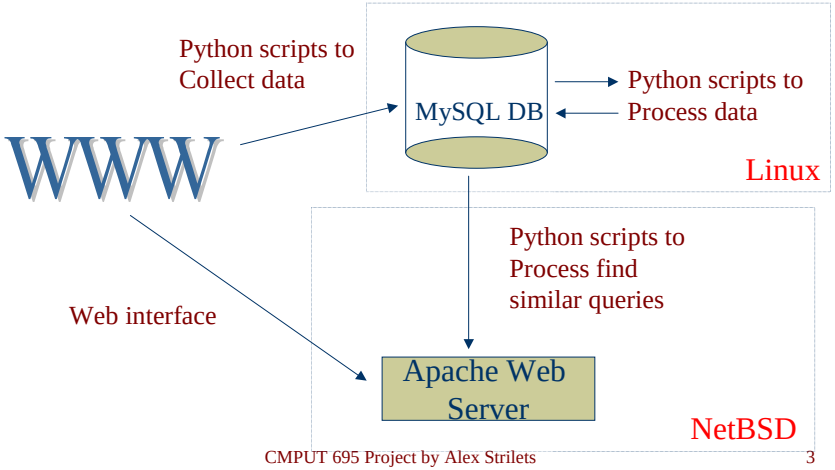
CMPUT 695
Knowledge Discovery in Databases

- Introduction
- Architecture Overview
- Tools Used
- Database design
- Different similarity measurements
- On-line demo
- Questions and Discussion

Introduction

- Find web queries similar to the one user submitted and present user with list of similar queries
  - Collect queries that users submit to the search engine
  - Develop algorithm(s) that will find similar queries
  - Develop web interface

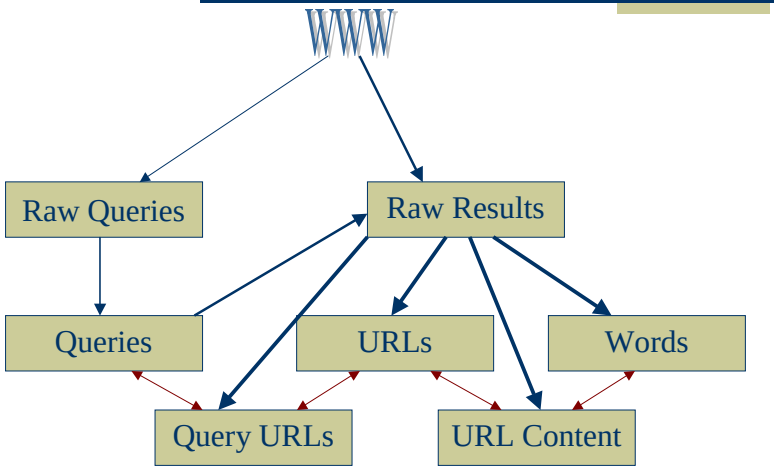
Architecture Overview



Tools Used

- NetBSD/Linux Operating Systems
- MySQL database
- Apache web server
- Python scripting language

Database design



Different similarity measurements

- Popular queries with similar words
  - Given query string, find all other queries that contain at least one word from the query string; sort result by query popularity(occurrence count)
- Popular queries with similar URLs
  - Given query string, submit it to the search engine, retrieve URLs and find all queries from our DB that have at least one similar URL, sort result by query popularity

Different similarity measurements

- Queries with most common URL
  - Submit given query to the search engine, find all URLs returned back, then find all queries that retuned the most common URLs (exclude queries that retuned same URLs back)
  - Possible to specify min and max % threshold for similar URLs (due to the lack of data collected they are set to 0 and 0.95)

Different similarity measurements

- Common title words
  - Submit query to the search engine, find all titles of URLs returned and find all queries that retuned words in title that are most common to the words in titles of URLs of original query
- Common content words
  - Submit query to the search engine, find all snippets of URLs returned and find all queries that retuned words in snippets that are most common to the words in titles of URLs of original query

On-line Demo

- Click here to get to the on-line demo