Search language (SL)

White paper

CS W4115: Programming Languages and Translators

Professor Stephen A. Edwards

Computer Science Department

Fall 2007 Columbia University

Submitted by

Majid Khan

{UID = mk2759,

Email = majidkhan@yahoo.com }

Objective

Search engines are still an emerging markets and lack tools which enable building search applications e.g. Search engines faster and easier. This search language would give capability to users to create search engines or search engine type applications with lesser efforts. The scope of this project would be limited to text search in HTML and plain text files.

Brief overview

Any search language would need to deal with at least 3 items.

1) Able to fetch data from repositories and understand document formats.
2) Able to manipulate data and able to store statistically manipulated data for later use.
3) Able to provide a search capability and provide ranked search results.

Repository – A repository would be a hierarchy of directories containing documents.

Document – HTML and plain texts would be the file types used as documents. Only files with extensions htm, htm, txt and without extensions (would be considered as plain text) would be processed by this language.

Data manipulation – It is a multi step process which is as following

Parse data and get words.

Perform stemming on these words (http://en.wikipedia.org/wiki/Stemming)

Get frequencies of stemmed words and create an inverted vector table using this data.

Calculate term frequency–inverse document frequency (tf-idf) for every stemmed word.

Store this data in for future use.

Search – Search would need two items. Query string and a repository on which we have already performed data manipulation. Query is a series of alphanumeric words separated by white space.

This process would return a list of documents and also a number that represents the closeness of query with document.

A sample program

repository arep is location=”c:\adirectory\anotherDirectory” include=”*.html,*.txt,*.htm,*” type=”cosine”;

process arep with location=”c:\adirectory\anotherDirectory” stemmer =”paice” overrwite=”true” filename=”arep.rep”;

performquery q on arep with query=”search this data please” store=”search.query”

Output

The output for returned by perform query is stored in a file. This file would contain list of files returned by search with the result weight. Higher the weight closer the match is.

pathFromRepositoryRoot\filename1 weight1

pathFromRepositoryRoot\filename2 weight2

pathFromRepositoryRoot\filename3 weight3

Summary

This language is a primitive step towards a language which would enable search engine creation, design and development easier. Above mention language is the initial scope of this project, if time permits (which I really doubt) I would try to add a search for familiar pages and also try to introduce two types i.e. document and document collection.