The Hong Kong University of Science and Technology

Comp 4321 Team Project

A Web Search Engine

Report (Phase 1)

Team members:

Name: Wing San Luk Name: Lee Wasin

ITSC: wslukaa ITSC: wleead

SID: 20193803 SID: 20266284

Name: Gao Xiao Wen Name: Cheung Wai Kwan

ITSC: xgaoah ITSC: wkcheungai

SID: 20264676 SID: 20272910

# The Design of the JBDM Database

Each class contains a hashtable which is used to store data and we can call the data by using the function “hashtable.get(index)”.

## Indexer class

Page **Indexer**

URL <-> PageID

Title Indexer

Title <-> TitleID

Word Indexer

Word <-> WordID

Child Parent

Parent-ID <-> ChildID

## InvertedIndex class

Inverted index

word ID -> {pageID, Freq}

Forward index

Page-ID -> {keywords}

Link-based index

Parent-ID -> {ChildID}

## PageProperty class

Page properties

Page-ID -> title, URL, last-date-of-modification, size

## Explanation

After crawling a website, we will have a page ID from the current website. First, we can save Title, URL, Last-date-of-modification, Size and the Child page in the Index class because we need the index and basic information about the page.

Then, we will save the related Keywords and its Page ID and the Frequency it appeared in the website from the Inverted Index class. We chose to use Inverted index is because we can have a faster retrieval when we want to call the data.

Finally, we generate the spider\_result.txt using this method.