Molly

by

Richard J.I. Drake

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science (MSc)

in

Faculty of Science

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Ken Q. Pu

August 2013

Abstract

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Preface

Background and Motivation

The introduction of keyword search has revolutionized how we find information. Over the years, numerous techniques have been developed which make searching through large amounts of information for one or more keywords extremely fast. With the advent of faster computer hardware, we are able to not only search through small attributes of a document (eg. Title, synopsis, etc.) but rather the entirety of the document itself.

An example of an early system which utilized keyword search would be a library catalogue. Such a system would allow a user to search, by keyword, for the title of a book, manuscript, etc. The results would show item titles matching the keyword(s), as well as other information such as whether or not the item is in circulation, as well as where it is located within the library. This information would come from a relational database.

With the rise of the World Wide Web, much information was placed online. This information would be easy to access if one knew how to locate it. Unfortunately, over time, so much information existed on the World Wide Web that it became difficult to keep track of it all. There was a need to index all of this information and make it accessible. This need was filled by a Web search engine.

Initial Web search engines comprised of simple scripts that gathered listings of files on FTP servers; they were essentially link farms. A few short years later, the first full-text (keyword) search engine, WebCrawler, was released.

While full-text search engines provided an excellent means for locating information, as the

Web grew larger, the volume of noise also grew larger. In addition, every Web page could be structured in a different way; the Web was largely a collection of unstructured documents. That is, there were few obvious links between them.

Search engines such as Google attempted to solve this problem by introducing new algorithms, such as PageRank, to rank Web pages on both the relevance of their content as well as their reputation. The idea was if a page is linked to often, it is considered to be more authoritative on a subject than a page with fewer links. This allowed the relevance of a page to be computed based on not only its contents, but its artificial importance.

Molly attempts to avoid some of the issues plaguing search engines. It deals primarily with structured, filtered data. This allows us to provide results with less noise. In addition, the fact that it deals with structured data means links between documents are explicitly stated. Rather than inferring a link between documents based on hyperlinks, we know when two documents are linked together.

This thesis provides an overview of the Molly system.

Outline

Changed.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

RDBMS to Document Store

```clojure
(ns molly.core
  (:gen-class)
  (:use clojure.pprint
        molly.conf.config
        molly.server.serve
        molly.index.build
        molly.search.lucene
        molly.search.query-builder
        [clojure.tools.cli :only (cli)]
        [molly.algo.bfs-atom :only (bfs-atom)]
        [molly.algo.bfs-ref :only (bfs-ref)]
        [molly.algo.bfs :only (bfs)]))

(defn parse-args
  [args]
  (cli args
       ["-c" "--config" "Path to configuration (properties) file"]
       ["--serve"       "Start the server"
        :default false
        :flag true]
       ["--port"        "Port to run the server on"
        :default 8080
        :parse-fn #(Integer. %)]
       ["--algorithm"   "Algorithm to run"]
       ["-s" "--source" "Source node"]
       ["-t" "--target" "Target node"]
       ["--index"       "Build an index of the database"
        :default false
        :flag true]
       ["--benchmark"   "Run benchmarks"
        :default false
        :flag true]
       ["-h" "--help"   "Show help"
        :default false
        :flag true]))

(def counter (atom 0))

(defn ns-to-ms
  [nanosec]
  (/ nanosec (* 1000.0 1000.0)))

(defn warmup
  [func searcher source target seconds]
  (let [start     (System/nanoTime)
        time-in-s (* seconds 1000 1000 1000)]
    (while (< (- (System/nanoTime) start) time-in-s)
      (func searcher source target)
      (swap! counter inc)))
```

```
(ns molly.conf.mycampus
  (:use molly.conf.config
        molly.datatypes.database
        molly.datatypes.index
        molly.datatypes.schema
        [clojureql.core :only (table project join where rename)])
  (:import (molly.datatypes.database Sqlite)
           (molly.datatypes.index Lucene)
           (molly.datatypes.schema EntitySchema)))

(def schedules-table
  (->
    (join
      (->
        (table :sections)
        (project [[:id :as :sec_id] [:id :as :section_id] :actual
                  :campus :capacity :credits :levels
                  :registration_start :registration_end :semester
                  :sec_code :sec_number :year :course]))
      (->
        (join
          (->
            (table :schedules)
            (project [[:id :as :sch_id] :date_start :date_end :day
                      :schedtype :hour_start :hour_end :min_start
                      :min_end :classtype :location]))
          (->
            (table :teaches)
            (project [:position :teaches.schedule_id
                      :teaches.instructor_id [:id :as :teaches_id]])))
          (where (= :schedules.id :teaches.schedule_id))))
      (where (= :sections.id :section_id)))))

(def schedules-id
  [[:schedules  :schedule_id]
   [:teaches    :teaches_id]
   [:sections   :section_id]])

(def mycampus-schema
  [(EntitySchema.
     {:T      :entity
      :C      :courses
      :sql    (table :courses)
      :ID     :code
      :attrs  [:code :title :description]
      :values [:code :title]})
   (EntitySchema.
     {:T      :entity
      :C      :instructors
      :sql    (table :instructors)
```

```clojure
(ns molly.datatypes.database
  (:require
    [clojure.java.jdbc :as sql]
    [clojureql.core :as cql]))

(defprotocol Database
  (execute-query [this query f]))

(deftype Sqlite [conn]
  Database
  (execute-query
    [this query f]
    (sql/with-connection conn (cql/with-results [rs query]
                                                (doseq [res rs]
                                                  (f res))))))
```

Figure 1.3: datatypes/database.clj

```
(ns molly.datatypes.entity
  (:use molly.util.nlp)
  (:import
    [clojure.lang IPersistentMap IPersistentList]
    [org.apache.lucene.document
     Document Field
     Field$Index Field$Store]))

(defn special?
  [field-name]
  (and (.startsWith field-name "__") (.endsWith field-name "__")))

(defn uid
  "Possible inputs include:
    row :T :ID
    row [[:T :ID] [:T :ID]]
    row [[:T :ID :desc] [:T :ID :desc]]"
  ([row C id]
     (if (nil? (row id))
       (throw
         (Exception.
           (str "ID column " id " does not exist in row " row ".")))
       (str (name C)
            "|"
            (clojure.string/replace (row id) #"\s+" "_"))))
  ([row Tids]
   (clojure.string/join " " (for [[C id] Tids]
                              (uid row C id)))))
(defn field
  [field-name field-value]
  (Field. field-name
          field-value
          Field$Store/YES
          Field$Index/ANALYZED))

(defn document
  [fields]
  (let [doc (Document.)]
    (do
      (doseq [[field-name field-value] fields]
        (.add doc (field (name field-name) (str field-value))))
      doc)))

(defn row->data
  ^{:doc "Transforms a row into the internal representation."}
  [this schema]
  (let [T         (schema :T)
        C         (schema :C)
        attr-cols (schema :attrs)
        attrs     (if (nil? attr-cols)
```

```
(ns molly.datatypes.index
  (:import
    (java.io File)
    (org.apache.lucene.analysis WhitespaceAnalyzer)
    (org.apache.lucene.index IndexReader
                             IndexWriter
                             IndexWriter$MaxFieldLength)
    (org.apache.lucene.queryParser QueryParser)
    (org.apache.lucene.search IndexSearcher)
    (org.apache.lucene.store SimpleFSDirectory)
    (org.apache.lucene.util Version)))

(def version
     Version/LUCENE_35)
(def default-analyzer
  (WhitespaceAnalyzer. version))
(def unlimited-fields IndexWriter$MaxFieldLength/UNLIMITED)

(defprotocol Index
  (path [this])
  (open-searcher [this])
  (open-writer
    [this]
    [this analyzer])
  (close-writer [this idx])
  (search
    [this searcher query]
    [this searcher query analyzer])
  (add-doc [this idx doc]))

(deftype Lucene [idx-path]
  Index
  (path
    [this]
    (-> idx-path File. SimpleFSDirectory.))
  (open-searcher
    [this]
    (-> (IndexReader/open (path this)) IndexSearcher.))
  (open-writer
    [this]
    (open-writer this default-analyzer))
  (open-writer
    [this analyzer]
    (IndexWriter. (path this) analyzer unlimited-fields))
  (close-writer
    [this idx]
    (doto idx
      (.commit)
      (.optimize)
      (.close)))
```

```clojure
(ns molly.datatypes.schema
  (:use molly.datatypes.database
        molly.datatypes.entity
        molly.search.lucene
        molly.util.nlp)
  (:require [clojureql.core :as cql]))

(defprotocol Schema
  (crawl [this db-conn idx-w])
  (klass [this])
  (schema-map [this]))

(deftype EntitySchema [S]
  Schema
  (crawl
    [this db-conn idx-w]
    (let [sql (S :sql)]
      (println sql)
      (execute-query db-conn sql
                     (fn [row]
                       (add-doc idx-w
                                (data->doc (row->data row S)))))

      (if (= (S :T) :entity)
        (doseq [value (S :values)]
          (let [query (->
                       sql
                       (cql/project [value])
                       (cql/grouped [value]))]
            (execute-query db-conn query
                           (fn [row]
                             (add-doc idx-w (data->doc
                                (row->data row
                                           (assoc S
                                                  :T :value
                                                  ))))))))))
  (klass
    [this]
    ((schema-map this) :C))
  (schema-map
    [this]
    S))
```

Figure 1.6: datatypes/schema.clj

```clojure
(ns molly.index.build
  (:use molly.conf.config
        molly.conf.mycampus
        molly.datatypes.database
        molly.datatypes.entity
        molly.datatypes.schema
        molly.search.lucene)
  (:import (molly.conf.mycampus Mycampus)))

(defn build
  [db-path path]
  (let [conf    (Mycampus. db-path path)
        db-conn (connection conf)
        ft-path (idx-path (index conf))
        idx-w   (idx-writer ft-path)
        schemas (schema conf)]
    (doseq [ent-def schemas]
      (println "Indexing " (name (klass ent-def)) "...")
      (crawl ent-def db-conn idx-w))

    (close-idx-writer idx-w)))
```

Figure 1.7: index/build.clj

```clojure
(ns molly.util.nlp)

(defn q-gram
  ([S]
   (q-gram S 3 "$"))
  ([S n]
   (q-gram S n "$"))
  ([S n s]
   (let [padding (clojure.string/join "" (repeat (dec n) "$"))
         padded-S (str padding
                       (clojure.string/replace S " " padding)
                       padding)]
     (clojure.string/join " "
                          (for [i (range
                                   (+ 1 (- (count padded-S) n)))]
                            (. padded-S substring i (+ i n)))))))
```

Figure 1.8: util/nlp.clj

Chapter 2

Search

```clojure
(ns molly.search.lucene
  (:import
    (java.io File)
    (org.apache.lucene.analysis WhitespaceAnalyzer)
    (org.apache.lucene.index IndexReader IndexWriter
                             IndexWriter$MaxFieldLength)
    (org.apache.lucene.queryParser QueryParser)
    (org.apache.lucene.search IndexSearcher)
    (org.apache.lucene.store SimpleFSDirectory)
    (org.apache.lucene.util Version)))

(def version
     Version/LUCENE_35)
(def default-analyzer
  (WhitespaceAnalyzer. version))
(def unlimited-fields IndexWriter$MaxFieldLength/UNLIMITED)

(defn idx-path
  [path]
  (-> path File. SimpleFSDirectory.))

(defn idx-searcher
  [idx-path]
  (-> (IndexReader/open idx-path) IndexSearcher.))

(defn idx-writer
  ([idx-path analyzer]
    (IndexWriter. idx-path analyzer unlimited-fields))
  ([idx-path]
    (idx-writer idx-path default-analyzer)))

(defn close-idx-writer
  [idx-writer]
  (doto idx-writer
    ; Lucene docs say not to use .optimize anymore.
    (.commit)
    (.close)))

(defn idx-search
  [idx-searcher query topk]
  (let [results (. (. idx-searcher search query topk) scoreDocs)]
    (map (fn [result] (.doc idx-searcher (.doc result))) results)))

(defn add-doc
  [idx doc]
  (. idx addDocument doc))
```


Figure 2.1: search/lucene.clj

```clojure
(ns molly.search.query-builder
  (:import (org.apache.lucene.index Term)
           (org.apache.lucene.search BooleanClause$Occur
                                     BooleanQuery
                                     PhraseQuery)))

(defn query
  [kind & args]
  (let [field-name    (condp      = kind
                        :type     "__type__"
                        :class    "__class__"
                        :id       "__id__"
                        :text     "__all__"
                        ; Assume "kind" is an attribute name.
                        (condp = (type kind)
                          clojure.lang.Keyword  (name kind)
                          java.lang.String      kind))
        phrase-query  (PhraseQuery.)]
    (doseq [arg args]
      (. phrase-query add (Term. field-name (name arg))))

    phrase-query))

(defn boolean-query
  [args]
  (let [query (BooleanQuery.)]
    (doseq [[q op] args]
      (. query add q (condp = op
                       :and BooleanClause$Occur/MUST
                       :or  BooleanClause$Occur/SHOULD
                       :not BooleanClause$Occur/MUST_NOT)))

    query))
```

Figure 2.2: search/query_builder.clj

```
(ns molly.server.serve
  (:use noir.core
        molly.datatypes.entity
        molly.search.lucene
        molly.search.query-builder
        molly.util.nlp
        [molly.algo.bfs :only (bfs)]
        [molly.algo.bfs-atom :only (bfs-atom)]
        [molly.algo.bfs-ref :only (bfs-ref)])
  (:require [noir.server :as server]
            [noir.response :as response]))

(def runtime (Runtime/getRuntime))

(defn start!
  [props]
  (let [searcher (idx-searcher (idx-path (props :index)))]
    (defn dox
      [q1 field S op topk]
      (let [bq      (boolean-query
                      (concat [[q1 :and]]
                              (for [s S]
                                [(query field s) op])))
            result  (map doc->data (idx-search searcher bq topk))
            wat     (fn [data] {:meta (meta data) :results data})]
        (map wat result)))

    (defn entities
      [field q topk]
      (dox (query :type :entity)
           field
           (clojure.string/split q #"\s{1}")
           :and
           topk))

    (defpage "/" []
             (response/redirect "/index.html"))

    (defpage "/value" {:keys [q]}
             (response/json
               {:result
                (dox (query :type :value)
                     :text
                     (clojure.string/split (q-gram q) #"\s{1}")
                     :or
                     (props :topk_value))}))
```



```
    (defpage "/entities" {:keys [q]}
             (response/json
               {:result
```

Chapter 3

Concurrency

```clojure
(ns molly.algo.common
  (:use molly.datatypes.entity
        molly.search.lucene
        molly.search.query-builder))

(defn find-entity-by-id
  [G id]
  (let [query (boolean-query [[(query :type :entity) :and]
                              [(query :id id) :and]])]
    (map doc->data (idx-search G query 10))))

(defn find-group-for-id
  [G id]
  (let [query   (boolean-query [[(query :type :group) :and]
                                [(query :entities id) :and]])
        results (map doc->data (idx-search G query 10))
        big-str (clojure.string/join " "
                                     (map #(% :entities) results))]
    (distinct (clojure.string/split big-str #"\s{1}"))))

(defn find-adj
  [G v]
  ;(find-group-for-id G v))
  (remove #{v} (find-group-for-id G v)))
```

Figure 3.1: algo/common.clj

```clojure
(ns molly.algo.bfs
  (use molly.algo.common))

(defn update-adj
  [G marked dist prev u]
  (loop [adj      (find-adj G u)
         marked   marked
         dist     dist
         prev     prev
         frontier []]
    (if (empty? adj)
      [(conj marked u) dist prev frontier]
      (let [v     (first adj)
            adj'  (rest adj)]
        (if (marked v)
          (recur adj' marked dist prev frontier)
          (let [dist' (assoc dist v (inc (dist u)))
                prev' (assoc prev v u)]
            (recur adj' marked dist' prev' (conj frontier v)))))))))

(defn bfs
  [G s t]
  (loop [Q      (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
         marked #{}
         dist   {s 0}
         prev   {s nil}]
    (if (or (empty? Q)
            (some (fn [node] (= node t)) marked))
      [marked dist prev]
      (let [u   (first Q)
            Q'  (rest Q)
            [marked' dist' prev' frontier]
            (update-adj G marked dist prev u)]
        (recur (concat Q' frontier) marked' dist' prev')))))
```

Figure 3.2: algo/bfs.clj

```clojure
(ns molly.algo.bfs-atom
  (use molly.algo.common))

(defn initial-state
  [s]
  (atom {:Q       (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
         :marked  #{s}
         :dist    {s 0}
         :prev    {}
         :done    false}))

(defn update-state
  [state u v]
  (let [Q       (state :Q)
        marked  (state :marked)
        dist    (state :dist)
        prev    (state :prev)]
    (assoc state
           :Q       (conj Q v)
           :marked  (conj marked v)
           :dist    (assoc dist v (inc (dist u)))
           :prev    (assoc prev v u))))

(defn deref-future
  [dfd]
  (if (future? dfd)
    (deref dfd)
    dfd))

(defn update-adj
  [state-ref G u]
  (let [marked?   (@state-ref :marked)
        deferred  (doall
                   (for [v (find-adj G u)]
                     (if (marked? v)
                       nil
                       (future
                         (swap! state-ref update-state u v)))))]
    (doall (map deref-future deferred))))

(defn bfs-atom
  [G s t]
  (let [state-ref (initial-state s)]
    (while (and (not (empty? (@state-ref :Q)))
                (not (@state-ref :done)))
      (let [u     (first (@state-ref :Q))
            Q'    (pop (@state-ref :Q))]
        (swap! state-ref assoc :Q Q')
        (if (some (fn [node] (= node t)) (@state-ref :marked))
          (swap! state-ref assoc :done true)
```

```clojure
(ns molly.algo.bfs-ref
  (use molly.algo.common))

(defn initial-state
  [s]
  (ref {:Q      (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
        :marked #{s}
        :dist   {s 0}
        :prev   {}
        :done   false}))

(defn update-state
  [state u v]
  (let [Q       (state :Q)
        marked  (state :marked)
        dist    (state :dist)
        prev    (state :prev)]
    (assoc state
           :Q       (conj Q v)
           :marked  (conj marked v)
           :dist    (assoc dist v (inc (dist u)))
           :prev    (assoc prev v u))))

(defn deref-future
  [dfd]
  (if (future? dfd)
    (deref dfd)
    dfd))

(defn update-adj
  [state-ref G u]
  (let [marked?   (@state-ref :marked)
        deferred  (doall
                   (for [v (find-adj G u)]
                     (if (marked? v)
                       nil
                       (future (dosync (alter
                                        state-ref
                                        update-state
                                        u
                                        v))))))]
    (doall (map deref-future deferred))))

(defn bfs-ref
  [G s t]
  (let [state-ref (initial-state s)]
    (while (and (not (empty? (@state-ref :Q)))
                (not (@state-ref :done)))
      (let [u   (first (@state-ref :Q))
            Q'  (pop (@state-ref :Q))]
```

Chapter 4

Literature Review

Chapter 5

Theoretical

The problem of efficiently searching through a graph is the subject of countless articles and journal papers. Numerous algorithms have been proposed to perform graph search in an efficient manner. Many of these algorithms build upon their predecessors, making assumptions and changing aspects to better suit a particular problem area.

The applications of efficient graph search algorithms are endless. Algorithms such as Dijkstra's are used heavily in path finding, for example in a portable GPS unit. A* Search is used in the area of computer vision to approximate the best path to take. Companies such as Amazon make use of graph search algorithms in order to find related products for consumers to purchase.

This thesis concentrates on building a system which can be utilized in order to discover related information. The system that was built is capable of finding not only related information from neighbouring nodes, but also the best path between two arbitrary nodes.

Section 5.1 provides an overview of how the data is represented in the system. It also defines notation to represent this data. Section 5.2 outlines how the relational database is related to the data in the system. Section 5.3 provides an example data corpus which is utilized throughout this thesis. It details the "mycampus" dataset. Finally, Section 5.4 provides justification for the algorithm chosen to perform the graph search. It discusses multiple different graph search algorithms, as well as why the one used in this thesis was chosen.

5.1 Data Representation & Notation

The data from the database is represented in various data structures. There are separate representations for each type of data: values, entities, and entity groups.

Value

Definition 1. A **Value** represents a single piece of information. To avoid repetition, each value is unique. That is, $\exists!\ v \in V$, where v is a value in the set V of all values.

Entity

Definition 2. An **Entity** is a collection of attributes, a_n, each mapped to a single value, v_n. An entity also includes additional information such as a unique identifier.

id	$T_n \vert v_{id}$
a_1	v_1
a_2	v_2
$\vdots$	$\vdots$
a_n	v_n

Figure 5.1: The structure of an entity

Entities are analogous to rows in a database table. Thus, the unique identifier is generated based on the table name, T_n, as well as unique key in the table, v_{id}. The unique key identifies the row, and the table name identifies the table. Together they uniquely identify the entity within the entire database. Attributes are analogous to columns in a database table.

$\exists!\ e_{id} \in E$, where E is the set of all entities.

Entity Group

Definition 3. An **Entity Group** joins together two or more entities. These entity groups can also have attributes, a_n, and values, v_n, associated with them much like entities.

$$
\begin{array}{ll}
e_L & [e_1, e_2, \ldots, e_n] \\
a_1 & v_1 \\
a_2 & v_2 \\
\vdots & \vdots \\
a_n & v_n
\end{array}
$$

Figure 5.2: The structure of an entity group

5.2 Relational Database Abstraction

5.3 Data Corpus

For illustrative purposes, the mycampus dataset will be used. This data comes from UOIT's course registration system.

There are several different entities which comprise the mycampus dataset:

- Courses
- Instructors
- Schedules
- Sections
- Teaches

The **Courses** entity represents a course. A course has an individual code, along with a title and a description (See Table 5.1). The code uniquely identifies the course.

Column	Type	Description
code	VARCHAR	Unique course code
title	VARCHAR	Title of the course
description	TEXT	A brief description

Table 5.1: Courses entity schema

The **Instructors** entity represents an individual instructor. Each instructor has a unique identifier, as well as a name (See Table 5.2). An instructor can be a Professor, Lecturer, Sessional Instructor, or Teaching Assistant.

Column	Type	Description
id	INT	Unique identifier
name	VARCHAR	The instructor's name

Table 5.2: Instructors entity schema

The **Sections** entity represents a section of a course. Courses may have many sections. For example, one section could be a lecture, while another is a lab. Some courses may have over a dozen sections, depending on the associated term.

Each section contains a unique identifier, capacity information (students enrolled, spots open, etc.), when registration is open for the section, how many credits it is worth, what level (eg. undergraduate or graduate), and what year the section is offered in (See Table 5.3).

Column	Type	Description
id	INT	Unique identifier
actual	INT	Number of people enrolled in the course
campus	VARCHAR	String uniquely identifying the campus
capacity	INT	Maximum number of people that may be enrolled in the section
credits	FLOAT	Number of credits awarded upon successful completion
levels	VARCHAR	The level of the course (eg. undergraduate, graduate, etc.)
registration_start	DATE	Date registration for the section opens
registration_end	DATE	Date registration for the section ends
semester	VARCHAR	String that uniquely identifies the semester
sec_code	INT	Unique section code (called a CRN)
sec_number	INT	Sequential number identifying the number of the section
year	INT	The year the section is offered in

Table 5.3: Sections entity schema

The **Schedules** entity represents a scheduled meeting of a section. A section may have many schedules. For example, a lecture section may meet twice a week.

Each schedule has a unique identifier, a date range in which the schedule is active, the time of the schedule, the type, location, and the day which the class takes place (See Table 5.4).

Column	Type	Description
id	INT	Unique identifier
date_start	DATE	First day of class
date_end	DATE	Last day of class
day	VARCHAR	Single character representing the day of the week
schedtype	VARCHAR	Lecture, tutorial, lab, etc.
hour_start	INT	Hour the class starts at
hour_end	INT	Hour the class ends at
min_start	INT	Minute the class starts at
min_end	INT	Minute the class ends at
classtype	VARCHAR	The type of class
location	VARCHAR	Unique location name where class is held

Table 5.4: Schedules entity schema

A **Teaches** entity is used to link together an instructor and a schedule. Each link has a unique identifier along with the instructor's position (eg. Teaching Assistant, Lecturer, etc.) (See Table 5.5).

Column	Type	Description
id	INT	Unique identifier
position	VARCHAR	Position of the Instructor with regard to a Schedule

Table 5.5: Teaches entity schema

5.4 Graph Search Algorithms

Careful consideration was given to which graph search algorithm was to be used. Among the choices were:

- Breadth-First
- Bellman-Ford
- Dijkstra
- A*

The first choice, Breadth-First Search, is among the simplest of the algorithms. It has the advantage of being simple to implement. It can only handle fixed costs for travelling between nodes, which can be a disadvantage.

Bellman-Ford is similar to BFS. It has the ability to deal with variable cost. This comes at the cost of increased difficulty in implementation. When the cost between nodes is fixed, Bellman-Ford essentially becomes BFS.

A greedy version of BFS is Dijkstra's Algorithm. It utilizes a priority queue rather than a regular queue, allowing it to be faster. As it is a greedy algorithm, Dijkstra's Algorithm may not return the optimal result.

An extension to Dijkstra's is A* search. Graph search spaces can be rather large. A* attempts to prune the search space based on a heuristic. A* is a natural choice to perform graph search in certain areas such as computer vision where an obvious heuristic exists. It has a disadvantage of consuming large amounts of memory (though IDA* attempts to limit memory consumption).

There are many more graph search algorithms. The above were primarily considered as many of the other algorithms are simple extensions with different data structures.

For this application, there is no obvious heuristic. This eliminates A*. There is also no obvious cost function. This eliminates Dijkstra's Algorithm. Bellman-Ford is very similar to BFS with the added ability to deal with variable cost. As the cost function is not obvious and thus constant, Bellman-Ford essentially reverts back to BFS.

For these reasons, BFS was chosen as the graph search algorithm. While the other candidates and others provide numerous advantages over BFS in many situations, this is not one of them.

Chapter 6

Implementation

6.1 Functional vs. Procedural Programming of Algorithms

Functional and procedural programming languages are entirely different paradigms. In functional programming, data is immutable. Functions must be pure and predictable (free of "side effects"). Loops and control flow are accomplished with recursion and pattern matching.

Procedural languages allow for more flexibility at a cost of unpredictability.

Key differences between FP and procedural (for algorithms)

6.1.1 Functional Data Structures

Clojure immutable data structures, versioning (persistent data structures), STM vs. locking

6.2 Tuneable Parameters

Tuneable parameters

Chapter 7

System Implementation

7.1 Technology Selection

7.1.1 Programming Language

Numerous programming languages were considered for the implementation. Among them were: Ruby, Python, and Clojure. Each of the languages presented both pros and cons. Ruby and Python are rather similar object oriented languages. Clojure is a functional language.

Ruby features a clean syntax and is very object oriented. Its object model was inspired by that of Smalltalk. It has a very active web development community and numerous web frameworks (eg. Rails, Sinatra, Padrino, etc.). In contrast to Python, it features a lean core, instead choosing to depend on third party libraries.

Python also features clean syntax and is object oriented. It embraces a "batteries included" philosophy to its standard library, featuring libraries for everything from serial communications to importing/exporting CSV files. It too has a strong web development community and numerous web frameworks (eg. Django, Flask, web2py, etc.).

Python is used extensively by scientists. As such, it has numerous plotting, computation, and simulation libraries. Examples include Matplotlib, SciPy, NumPy, and NetworkX.

Both of the above languages embrace functional programming elements. Python has filter, reduce, and map. Ruby's collections are also capable of performing filter (select), reduce (inject),

and map. They both lack an important aspect of functional programming: immutable objects.

Clojure is a purely functional language built on top of the JVM. Like Python and Ruby, it is dynamically typed. Unlike Python and Ruby, it is a Lisp dialect and as such features macros and code-as-data.

As it runs on the JVM, Clojure allows us to utilize existing Java languages. Using Java libraries through Python or Ruby requires using their respective native code interfaces to Java's JNI. This process is typically automated through a tool such as SWIG.

Clojure was chosen as the implementation language for a number of reasons. First of all, as it is built on top of the JVM, it can utilize JDBC for database access. It can also make use of several other useful libraries. Secondly, its dynamic Lisp nature is a natural way of processing data. Clojure allows us to model our data structures directly after the data. In a language such as Python, one would need to make use of classes.

7.1.2 Relational Database

Several relational database management engines (RDBMS) were evaluated for this project. Key requirements were SQL support, and JDBC driver availability.

SQL, or Structured Query Language, is a powerful way to make complex queries. It is loosely based on Relational Algebra (RA)[3, p. 243]. In addition to being a data-manipulation language (similar to Relational Algebra), it is also a data-definition language.

JDBC, or Java Database Connectivity, was required in order for our application to access the database. JDBC is an API which allows us to access a database using a standard interface. It is the responsibility of the JDBC Driver to translate the API calls into the correct syntax for the database engine.

Any RDBMS that supports SQL-92 or higher and has a JDBC driver for it should be usable in this system. For testing purposes, we chose to use SQLite. SQLite provided a SQL interface, had a stable JDBC driver, and is ubiquitous. SQLite is an embedded database, allowing for local development without running a separate daemon process.

7.1.3 Full-Text Search Database

Two full-featured full-text search databases (FTSDB) were evaluated. The first, Xapian, is written as a C++ library. It features a rich query language, high performance, and a scalable design[5]. Bindings for Java are made available via a simple JNI library.

The second, Lucene, is a pure-Java library. It too features a rich query language and reasonable performance. Lucene also forms the basis of the Apache Solr search platform. As it is written in Java, Lucene runs natively on the JVM.

Both choices provided plenty of features, Java interfaces, and excellent performance. Ultimately Lucene was chosen. As it ran on the JVM, interfacing with the Lucene library from our Clojure code was simple. The fact that it was written in pure-Java and required no native libraries meant the entire project could be bundled as a Java Archive (JAR).

7.1.4 Web Stack

Clojure is a young language. As such, few web frameworks exist specifically for it (it can, however, interface with any Java web framework).

Clojure web frameworks are built on top of Ring, itself an HTTP abstraction layer with adapters for Java servlets as well as the Jetty web server. Ring is similar to WSGI in Python, or Rack in Ruby.

The web framework chosen for this project was Noir. Noir provided a simple Clojure API for building the JSON API to allow for client-side querying of the project's API.

7.2 System Design

The system is split into numerous components.

A configuration file is used to describe the entities and entity groups. It defines the relationship between entities, as well as their attributes and values. This configuration file is used by the crawler and indexer in order to retrieve data from the relational database and index it in a full-text search database.

Once the index is built, another component provides an internal API for querying the full-text search database. It is a thin Clojure wrapper over the Lucene API which provides the basis for search and retrieval of results. This wrapper is utilized by the system to discover relationships among entities.

Finally, the system can be interfaced with via a JSON API. The JSON API provides a simple HTTP wrapper around the search and discovery features of the system. The use of JSON allows client-side JavaScript to send and receive queries with XMLHttpRequest.

To do (1)

7.2.1 Configuration

Configuration files are simply Clojure code that follow a well-defined protocol. The protocol defines three main components of the system:

`connection` The relational database connection. It must be an instance of the `Database` protocol.

`index` The path to the full-text search database index file.

`schema` A definition of the database schema. The definition includes both the definition of entities, as well as the definition of entity groups. The schema is a list of instances of the `Schema` protocol.

`Database` Protocol

The `Database` protocol is defined as follows.

```
(defprotocol Database
        (execute-query [this query f]))
```

Each database instance must provide a function which accepts a query and function as arguments. It must execute the query provided and call function `f` on each row returned.

`Schema` Protocol

The `Schema` protocol is defined as follows.

```
(defprotocol Schema
  (crawl [this db-conn idx-w])
  (klass [this])
  (schema-map [this]))
```

A schema consists of the following information:

- A symbol (typically the name of the table).
- A hash map containing the structure of the schema.

Schemas are also responsible for populating the full-text search index with data from the relational database.

A concrete implementation for this protocol is provided by the `EntitySchema` type.

7.3 Implementation Issues

Chapter 8

Performance & Evaluation

8.1 Environment

All benchmarks were run on the same machine using the same software versions. The machine has the following specs.

Item	Description
Model	Dell PowerEdge 2950
Processor	2 x Quad-Core Intel® Xeon™ 5300 3.0GHz
Memory	8 GB DDR2 ECC RAM 667 MHz

Along with the following software versions.

Item	Description
Operating System	GNU/Linux Ubuntu 10.04.4 LTS
Kernel	2.6.32-28-generic
Java™ SE Runtime Environment	1.6.0_26-b03
Java HotSpot™ 64-Bit Server VM	20.1b02

8.2 Methodology

Evaluating the performance of code is a difficult problem at best. It is difficult to determine the impact on performance of various uncontrollable factors. Virtual Machines add another layer of abstraction which introduces even more factors.

As this project runs a top of the Java Virtual Machine (JVM), there are many factors to consider with regard to performance. In addition to uncontrollable events such as garbage collection, the JVM's behaviour can differ based on the operating system (OS) it is running on.

8.2.1 Pitfalls of the Java Virtual Machine

Timing Execution

There are two methods provided by the JVM for getting a precise time from the OS; `System.currentTimeMillis()` and `System.nanoTime()`. The former returns the "wall" time. It is possible for time to leap forward or backward using this method. If daylight savings time occurs between calls to `System.currentTimeMillis()`, it can result in a negative time.

The latter uses a variety of methods in order to obtain the precise time. The method it uses depends on both the OS and the hardware itself. Under Windows, `System.nanoTime()` makes use of the `QueryPerformanceCounter(QPC)` call. This call may make use of the "programmable-interval-timer (PIT), or the ACPI power management timer (PMT), or the CPU-level timestamp-counter (TSC)." [4] Accessing the PIT and PMT is a slow operation. Accessing the TSC is a very fast operation, but may result in varying numbers [2].

Linux attempts to use the TSC when possible. If it finds the values to be unreliable (eg. different cores vary too much), it makes use of the High Precision Event Timer (HPET) [2].

Platform	Resolution (ms)
Windows 95/98	55
Windows NT, 2000, XP (single processor)	10
Windows XP (multi processor)	15.625
Linux 2.4 Kernel	10
Linux 2.6 Kernel	1

Table 8.1: Reported resolutions of `currentTimeMillis()` by platform [1].

The JVM's timing functions are one area where the JVM's behaviour differs based on the OS. Different operating systems provide varying degrees of time resolution. The speed at which the OS-level calls to these timing functions return can even differ.

Garbage Collection

Just-in-Time Compilation

A typical JVM only loads classes when they're first used [1]. This task typically involves disk I/O, along with extra processing. As a result, the initial run of a task may only...

8.2.2 Mitigating JVM Pitfalls

JIT Compilation

```
33        ["-h" "--help"   "Show help"
34         :default false
35         :flag true]))
36
37 (def counter (atom 0))
38
39 (defn ns-to-ms
40   [nanosec]
41   (/ nanosec (* 1000.0 1000.0)))
```

Figure 8.1: core.clj

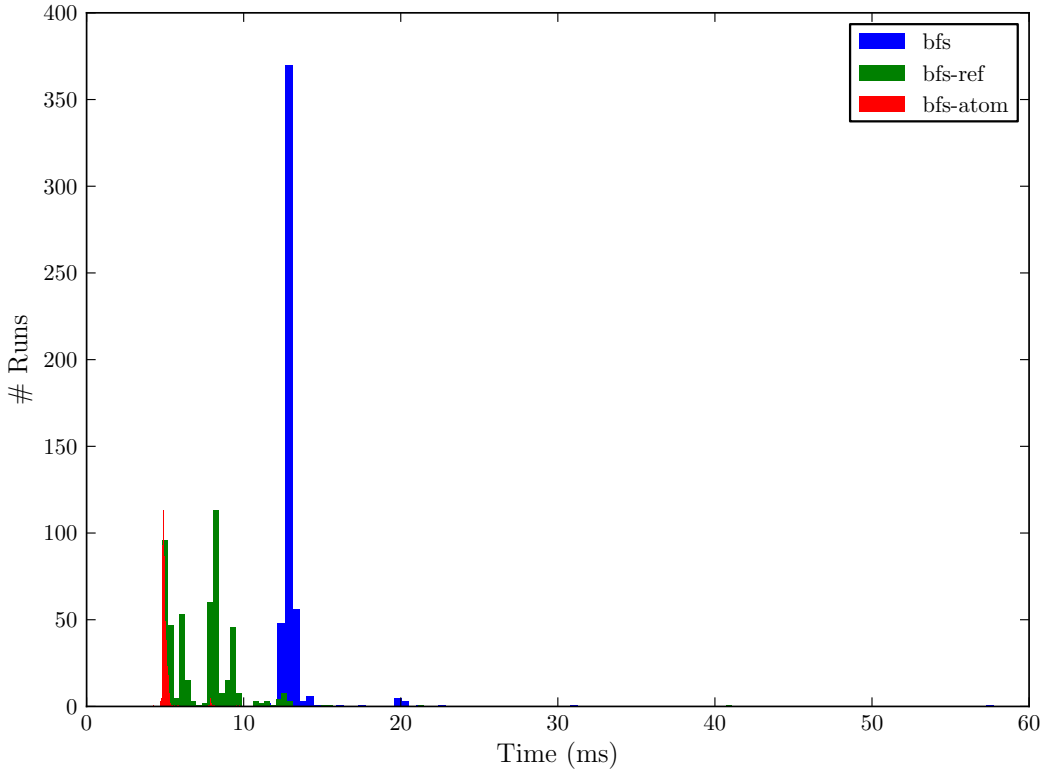

Figure 8.2: 1 Hop

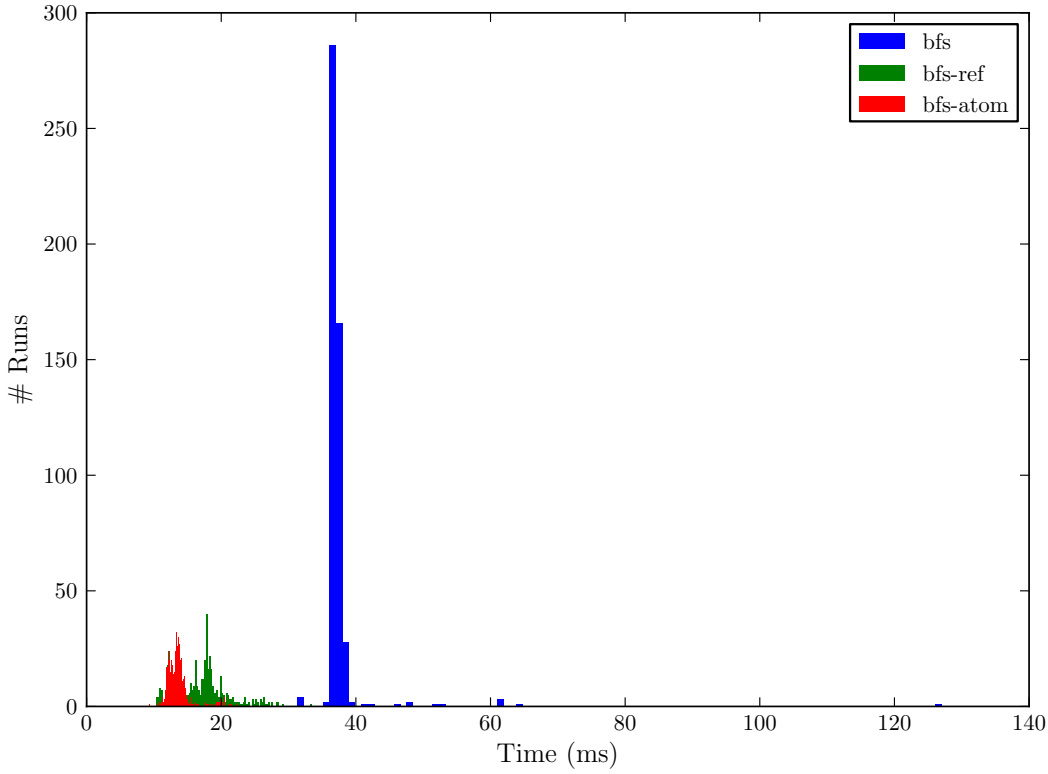

Figure 8.3: 2 Hops

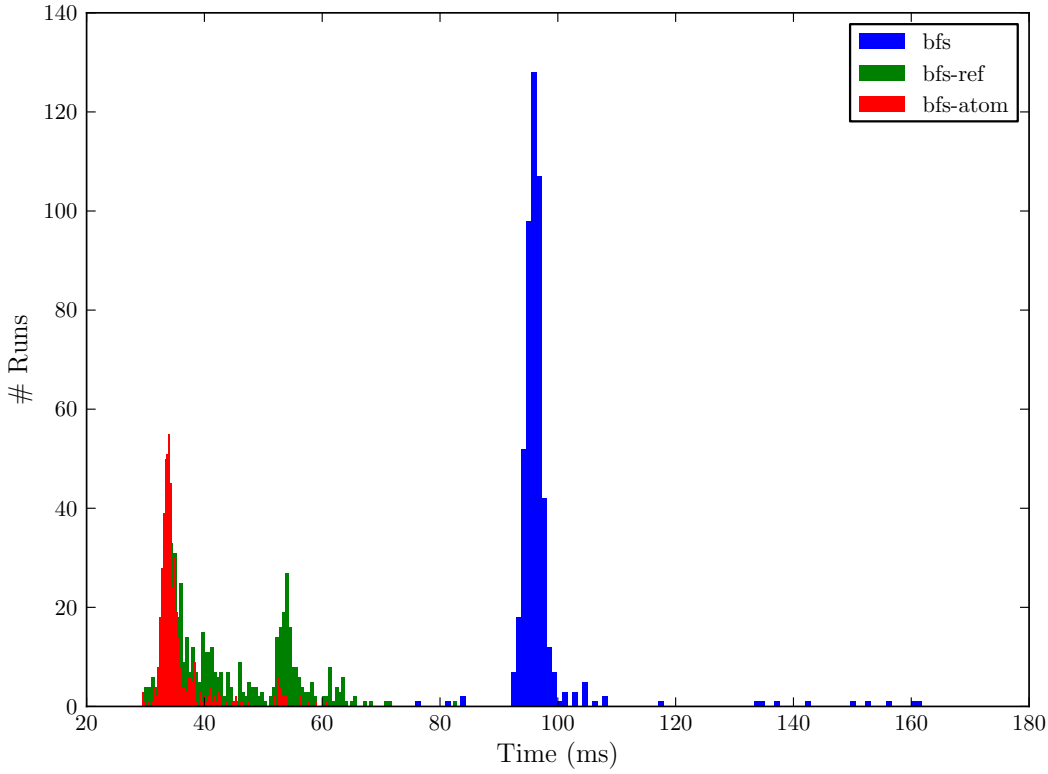

Figure 8.4: 3 Hops

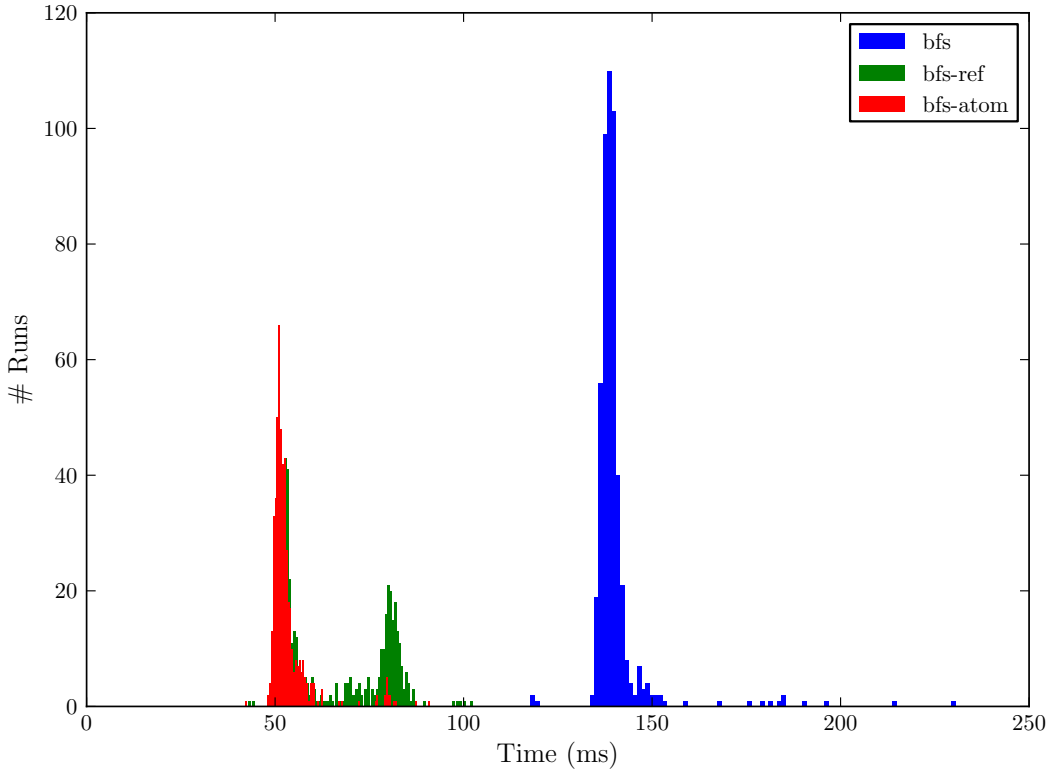

Figure 8.5: 4 Hops

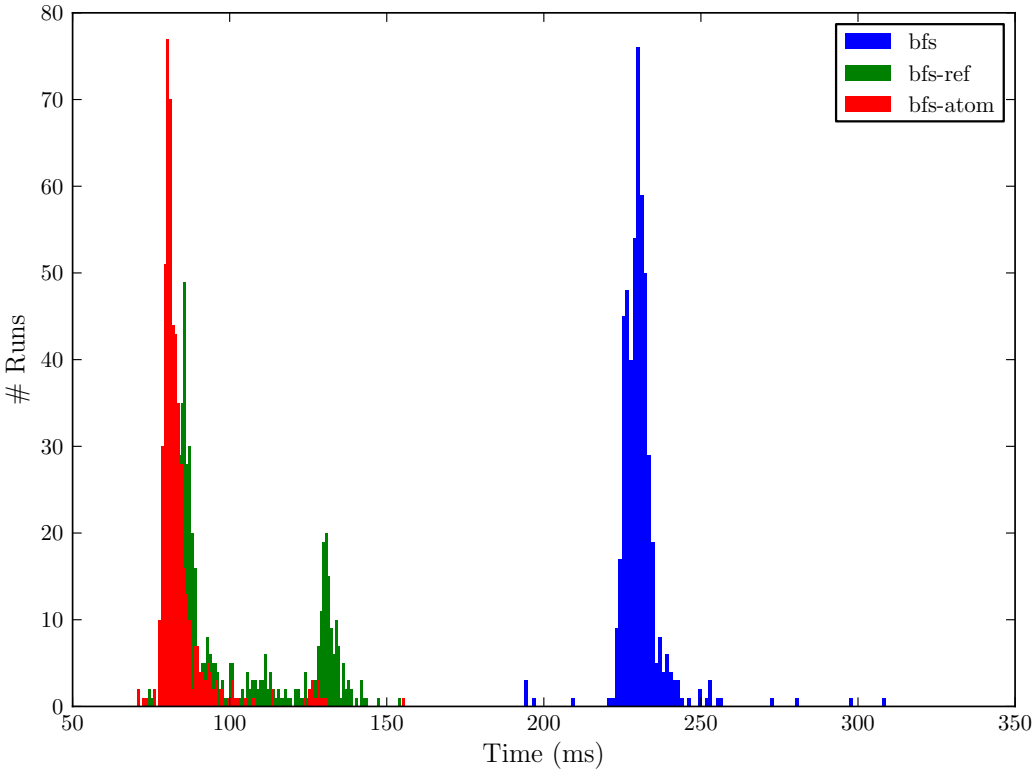

Figure 8.6: 5 Hops

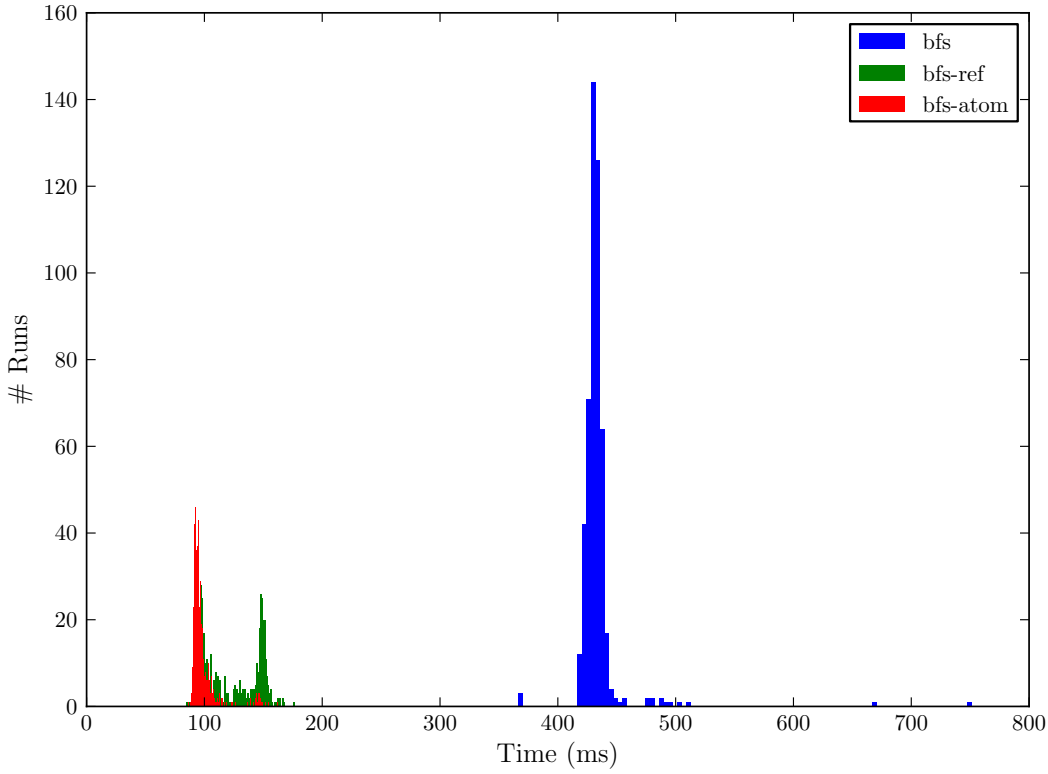

Figure 8.7: 6 Hops

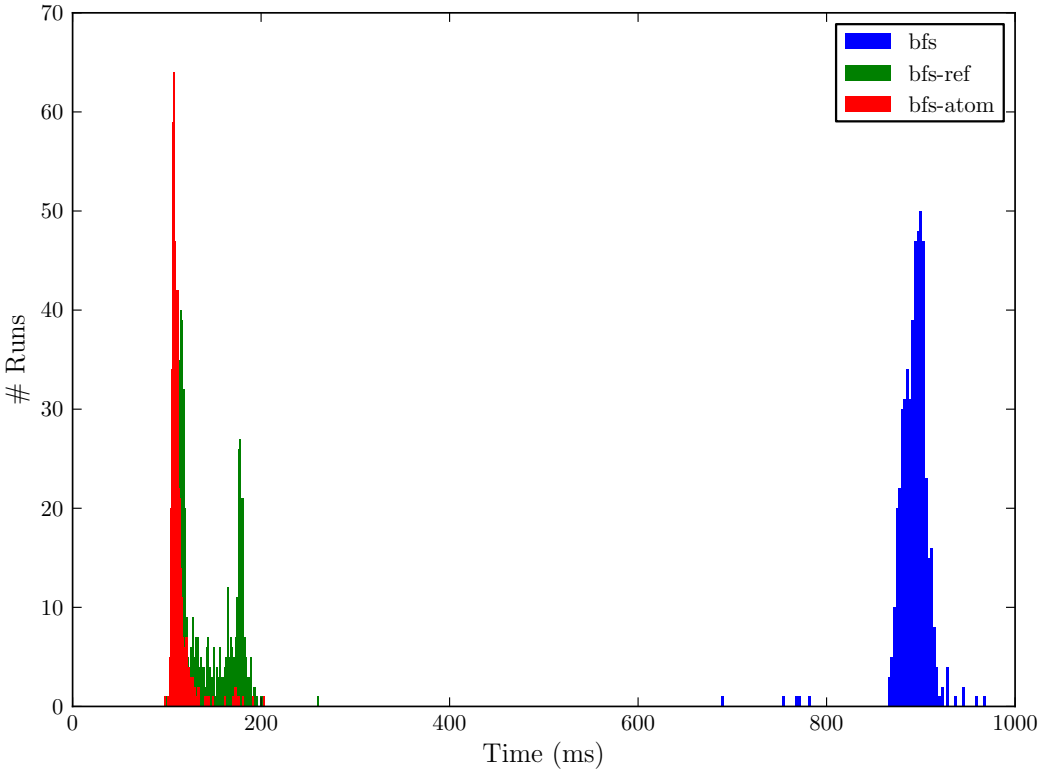

Figure 8.8: 7 Hops

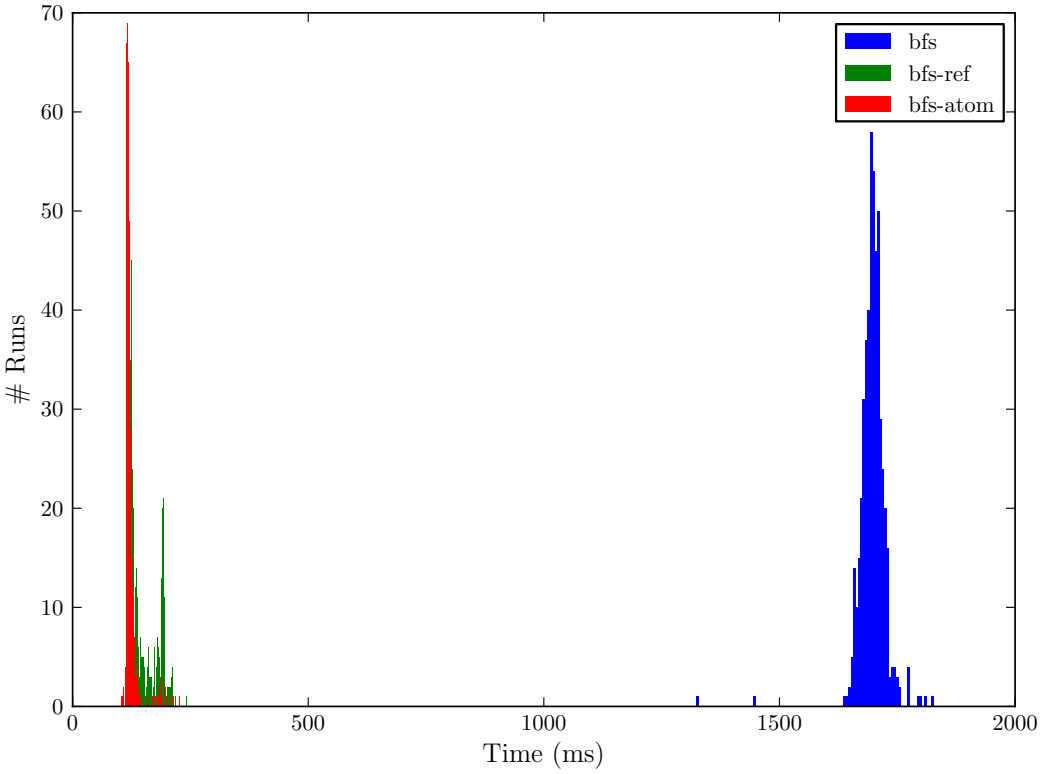

Figure 8.9: 8 Hops

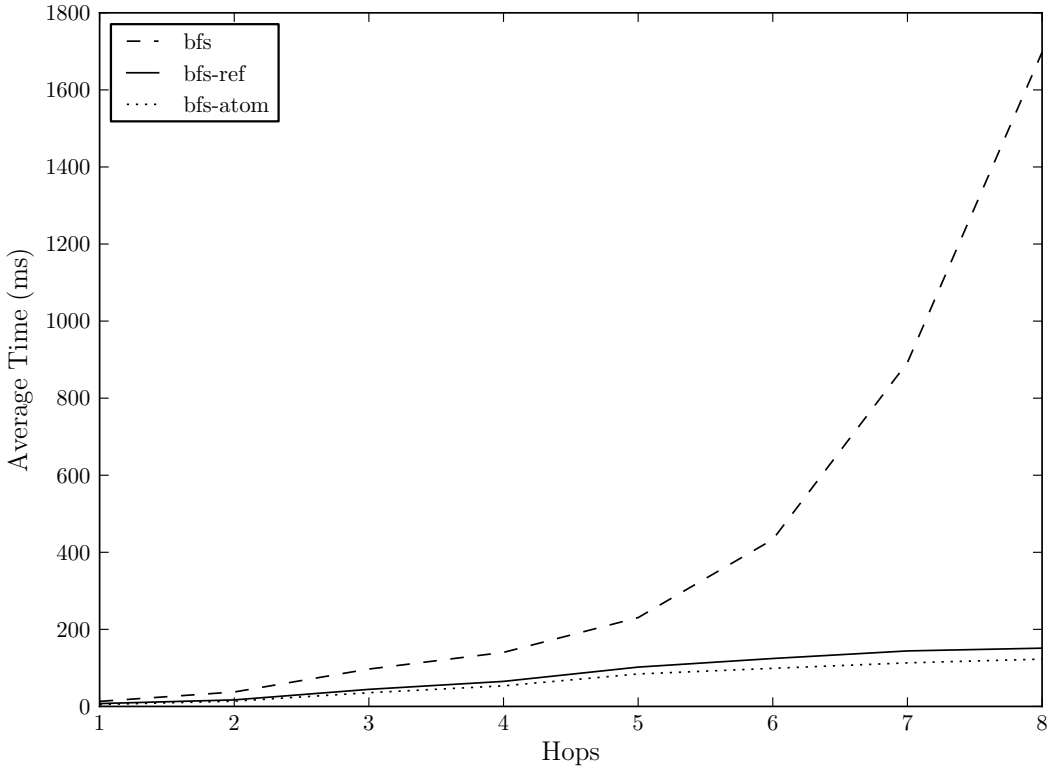

Figure 8.10: Comparison between single threaded and concurrent graph search

Chapter 9

Conclusion

[1]

Bibliography

[1] Brent Boyer. *Robust Java benchmarking, Part 1: Issues*. June 2008. URL: http://www.ibm.com/developerworks/java/library/j-benchmark1/index.html.

[2] Jonathan Corbet. *Counting on the time stamp counter*. Nov. 2006. URL: http://lwn.net/Articles/209101/.

[3] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009.

[4] David Holmes. *Inside the Hotspot VM: Clocks, Timers and Scheduling Events - Part I - Windows*. Oct. 2006. URL: https://blogs.oracle.com/dholmes/entry/inside_the_hotspot_vm_clocks.

[5] *The Xapian Project : Features*. URL: http://xapian.org/features.

To do...

- [ ] 1 (p. 29): Add overview diagram of system design.