

THE CHRONICLES OF 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

Copyright © Richard J.I. Drake 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.

Keywords relational database; full-text search

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

1	Definitions	1
1.1	Data Representation & Notation	1
2	RDBMS to Document Store	3
3	Concurrency	7
4	Theoretical	11
4.1	Data Representation & Notation	12
4.2	Relational Database Abstraction	13
4.3	Data Corpus	13
4.4	Graph Search Algorithms	15
5	Implementation	17
5.1	Functional vs. Procedural Programming of Algorithms	17
5.1.1	Functional Data Structures	17
5.2	Tuneable Parameters	17
6	System Implementation	18
6.1	Technology Selection	18
6.1.1	Programming Language	18
6.1.2	Relational Database	19
6.1.3	Full-Text Search Database	20

6.1.4	Web Stack	20
6.2	System Design	20
6.2.1	Configuration	21
6.3	Implementation Issues	22
7	Performance & Evaluation	23
7.1	Environment	23
7.2	Methodology	23
7.2.1	Pitfalls of the Java Virtual Machine	24
7.2.2	Mitigating JVM Pitfalls	25
8	Conclusion	35
A	Source Code	36
A.1	molly	36
A.1.1	molly.core	36
A.2	molly.conf	39
A.2.1	molly.conf.config	39
A.2.2	molly.conf.mycampus	40
A.3	molly.datatypes	44
A.3.1	molly.datatypes.database	44
A.3.2	molly.datatypes.entity	45
A.3.3	molly.datatypes.schema	48
A.4	molly.index	49
A.4.1	molly.index.build	49
A.5	molly.util	50
A.5.1	molly.util.nlp	50
A.6	molly.search	51
A.6.1	molly.search.lucene	51

A.6.2	molly.search.query_builder	53
A.7	molly.server	54
A.7.1	molly.server.serve	54
A.8	molly.algo	57
A.8.1	molly.algo.common	57
A.8.2	molly.algo.bfs	59
A.8.3	molly.algo.bfs_atom	60
A.8.4	molly.algo.bfs_ref	61

List of Tables

4.1	Courses entity schema	13
4.2	Instructors entity schema	14
4.3	Sections entity schema	14
4.4	Schedules entity schema	15
4.5	Teaches entity schema	15
7.1	Reported resolutions of <code>currentTimeMillis()</code> by platform [1].	24

List of Figures

1.1	The structure of an entity	1
1.2	The structure of an entity group	2
2.1	Determining whether or not a field is classified as “special.”	3
2.2	Unique identifier generation.	3
2.3	Creation of a field.	4
2.4	Creation of a document.	4
2.5	Transformation of a row into the internal representation.	5
2.6	Transformation of a document into the internal representation.	6
2.7	Transformation of the internal representation into a document.	6
3.1	Retrieves the entity with the given ID.	7
3.2	Finds all groups containing an entity with the given ID.	7
3.3	Finds adjacent nodes in the graph (entities linked in groups).	8
3.4	Discovers new nodes along the frontier and updates the state accordingly.	8
3.5	BFS using recursion.	8
3.6	Discovers new nodes along the frontier and updates the state accordingly.	9
3.7	Implements a concurrent version of BFS using atoms.	9
3.8	Discovers new nodes along the frontier and updates the state accordingly.	10
3.9	Implements a concurrent version of BFS using references.	10
4.1	The structure of an entity	12

4.2	The structure of an entity group	13
7.1	core.clj	25
7.2	1 Hop	26
7.3	2 Hops	27
7.4	3 Hops	28
7.5	4 Hops	29
7.6	5 Hops	30
7.7	6 Hops	31
7.8	7 Hops	32
7.9	8 Hops	33
7.10	Comparison between single threaded and concurrent graph search	34

List of Algorithms

Chapter 1

Definitions

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

Figure 1.1: 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.

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

Figure 1.2: The structure of an entity group

Chapter 2

RDBMS to Document Store

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

Figure 2.1: Determining whether or not a field is classified as “special.”

```
13 (defn uid
14   "Possible inputs include:
15   row :T :ID
16   row [[:T :ID] [:T :ID]]
17   row [[:T :ID :desc] [:T :ID :desc]]"
18   ([row C id]
19    (if (nil? (row id))
20        (throw
21          (Exception.
22            (str "ID column " id " does not exist in row " row ".")))
23        (str (name C)
24              "|")
25        (closure.string/replace (row id) #"\\s+" "_"))))
26   ([row Tids]
27    (closure.string/join " " (for [[C id] Tids]
28                               (uid row C id)))))
```

Figure 2.2: Unique identifier generation.

```
30 (defn field
31   [field-name field-value]
32   (Field. field-name
33           field-value
34           Field$Store/YES
35           Field$Index/ANALYZED))
```

Figure 2.3: Creation of a field.

```
37 (defn document
38   [fields]
39   (let [doc (Document.)]
40     (do
41       (doseq [[field-name field-value] fields]
42         (.add doc (field (name field-name) (str field-value))))
43       doc)))
```

Figure 2.4: Creation of a document.

```

45 (defn row->data
46   ^{:doc "Transforms a row into the internal representation."}
47   [this schema]
48   (let [T      (schema :T)
49         C      (schema :C)
50         attr-cols (schema :attrs)
51         attrs    (if (nil? attr-cols)
52                     this
53                     (select-keys this attr-cols))
54         meta-data {:type T :class C}
55         id-col    (schema :ID)]
56     (with-meta (if (= T :group)
57                 (conj attrs {:entities (uid this id-col)})
58                 attrs)
59              (condp = T
60                  :value   (assoc meta-data
61                                   :class
62                                   (clojure.string/join "|"
63                                   (map name
64                                        [C (first attr-cols)])))
65                  :entity (assoc meta-data :id
66                                   (if (coll? id-col)
67                                       (uid this id-col)
68                                       (uid this C id-col)))
69                  :group  (assoc meta-data
70                                   :entities
71                                   (uid this id-col))
72                  (throw
73                   (IllegalArgumentException.
74                    "I only know how to deal with types :value,
75                    :entity, and :group"))))))))

```

Figure 2.5: Transformation of a row into the internal representation.

```

77 (defn doc->data
78   ^{:doc "Transforms a Document into the internal representation."}
79   [this]
80   (let [fields      (.getFields this)
81         extract     (fn [x] [(keyword (clojure.string/replace
82                                     (.name x) "_" ""))
83                               (.stringValue x)])
84         check-special (fn [x] (special? (.name x)))
85         filter-fn     (fn [f] (apply hash-map
86                                   (flatten
87                                    (map extract
84                                     (filter f fields))))))]
88     (with-meta (filter-fn (fn [x] (not (check-special x))))
89               (filter-fn check-special))))
90

```

Figure 2.6: Transformation of a document into the internal representation.

```

92 (defn data->doc
93   ^{:doc "Transforms the internal representation into a Document."}
94   [this]
95   (let [int-meta  (meta this)
96         T         (int-meta :type)
97         all       (clojure.string/join " "
98                                     (if (= T :entity)
99                                       (conj (vals this)
100                                              (name
101                                               (int-meta :class)))
102                                       (vals this)))
103         luc-meta  [[:__type__ (name T)]
104                   [[:__class__ (name (int-meta :class))]]
105                   [[:__all__ (if (= T :value)
106                                (q-gram all)
107                                all))]]
108         raw-doc   (concat luc-meta
109                           this
110                           (condp = (int-meta :type)
111                                :value  [[:value all]]
112                                :entity [[:__id__ (int-meta :id)]]
113                                :group  [[]]))
114   (document raw-doc))

```

Figure 2.7: Transformation of the internal representation into a document.

Chapter 3

Concurrency

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

Figure 3.1: Retrieves the entity with the given ID.

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

Figure 3.2: Finds all groups containing an entity with the given ID.

```

21 (defn find-adj
22   [G v]
23   (remove #{v} (find-group-for-id G v)))

```

Figure 3.3: Finds adjacent nodes in the graph (entities linked in groups).

```

4  (defn update-adj
5    [G marked dist prev u]
6    (loop [adj      (find-adj G u)
7            marked   marked
8            dist     dist
9            prev     prev
10           frontier []]
11      (if (empty? adj)
12          [(conj marked u) dist prev frontier]
13          (let [v      (first adj)
14                adj'   (rest adj)]
15              (if (marked v)
16                  (recur adj' marked dist prev frontier)
17                  (let [dist' (assoc dist v (inc (dist u)))
18                        prev'  (assoc prev v u)]
19                      (recur adj' marked dist' prev' (conj frontier v))))))))

```

Figure 3.4: Discovers new nodes along the frontier and updates the state accordingly.

```

21 (defn bfs
22   [G s t]
23   (loop [Q      (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
24          marked #{ }
25          dist   {s 0}
26          prev   {s nil}]
27       (if (or (empty? Q)
28               (some (fn [node] (= node t)) marked))
29           [marked dist prev]
30           (let [u      (first Q)
31                 Q'     (rest Q)
32                 [marked' dist' prev' frontier]
33                 (update-adj G marked dist prev u)]
34               (recur (concat Q' frontier) marked' dist' prev')))))

```

Figure 3.5: BFS using recursion.

```

4 (defn update-adj
5   [state-ref G u]
6   (let [marked? (@state-ref :marked)
7         deferred (doall
8                     (for [v (find-adj G u)]
9                       (if (marked? v)
10                          nil
11                          (future
12                           (swap! state-ref update-state u v))))))]
13     (doall (map deref-future deferred))))

```

Figure 3.6: Discovers new nodes along the frontier and updates the state accordingly.

```

15 (defn bfs-atom
16   [G s t]
17   (let [state-ref (initial-state s)]
18     (while (and (not (empty? (@state-ref :Q)))
19                (not (@state-ref :done)))
20       (let [u (first (@state-ref :Q))
21             Q' (pop (@state-ref :Q))]
22         (swap! state-ref assoc :Q Q')
23         (if (some (fn [node] (= node t)) (@state-ref :marked))
24             (swap! state-ref assoc :done true)
25             (update-adj state-ref G u))))
26     [(@state-ref :marked) (@state-ref :dist) (@state-ref :prev)]))

```

Figure 3.7: Implements a concurrent version of BFS using atoms.

```

4 (defn update-adj
5   [state-ref G u]
6   (let [marked? (@state-ref :marked)
7         deferred (doall
8                     (for [v (find-adj G u)]
9                       (if (marked? v)
10                          nil
11                          (future (dosync (alter
12                                     state-ref
13                                     update-state
14                                     u
15                                     v))))))]
16     (doall (map deref-future deferred))))

```

Figure 3.8: Discovers new nodes along the frontier and updates the state accordingly.

```

18 (defn bfs-ref
19   [G s t]
20   (let [state-ref (initial-state s)]
21     (while (and (not (empty? (@state-ref :Q)))
22                (not (@state-ref :done)))
23       (let [u (first (@state-ref :Q))
24             Q' (pop (@state-ref :Q))]
25         (dosync (alter state-ref assoc :Q Q'))
26         (if (some (fn [node] (= node t)) (@state-ref :marked))
27             (dosync (alter state-ref assoc :done true))
28             (update-adj state-ref G u)))
29     [(@state-ref :marked) (@state-ref :dist) (@state-ref :prev)]))

```

Figure 3.9: Implements a concurrent version of BFS using references.

Chapter 4

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 4.1 provides an overview of how the data is represented in the system. It also defines notation to represent this data. Section 4.2 outlines how the relational database is related to the data in the system. Section 4.3 provides an example data corpus which is utilized throughout this thesis. It details the "mycampus" dataset. Finally, Section 4.4 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.

4.1 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 4. 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 5. 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	$T_n v_{id}$
a_1	v_1
a_2	v_2
\vdots	\vdots
a_n	v_n

Figure 4.1: 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 6. 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.

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

Figure 4.2: The structure of an entity group

4.2 Relational Database Abstraction

4.3 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 4.1). 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 4.1: Courses entity schema

The **Instructors** entity represents an individual instructor. Each instructor has a unique identifier, as well as a name (See Table 4.2). 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 4.2: 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 4.3).

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 4.3: 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 4.4).

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 4.4: 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 4.5).

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

Table 4.5: Teaches entity schema

4.4 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 5

Implementation

5.1 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)

5.1.1 Functional Data Structures

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

5.2 Tuneable Parameters

Tuneable parameters

Chapter 6

System Implementation

6.1 Technology Selection

6.1.1 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.

6.1.2 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.

6.1.3 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).

6.1.4 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.

6.2 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)

6.2.1 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.

6.3 Implementation Issues

Chapter 7

Performance & Evaluation

7.1 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

7.2 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.

7.2.1 Pitfalls of the Java Virtual Machine

Reliable benchmarking on the JVM is a difficult problem. The standard HotSpot VM utilizes a non-deterministic garbage collector (GC). In addition, the JIT is also non-deterministic. These two problems combine into a larger issue.

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 7.1: 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...

7.2.2 Mitigating JVM Pitfalls

JIT Compilation

```
33  [nanosec]
34  (/ nanosec (* 1000.0 1000.0)))
35
36  (defn warmup
37    [func searcher source target seconds]
38    (let [start      (System/nanoTime)
39          time-in-s  (* seconds 1000 1000 1000)]
40      (while (< (- (System/nanoTime) start) time-in-s)
41        (func searcher source target)))
```

Figure 7.1: core.clj

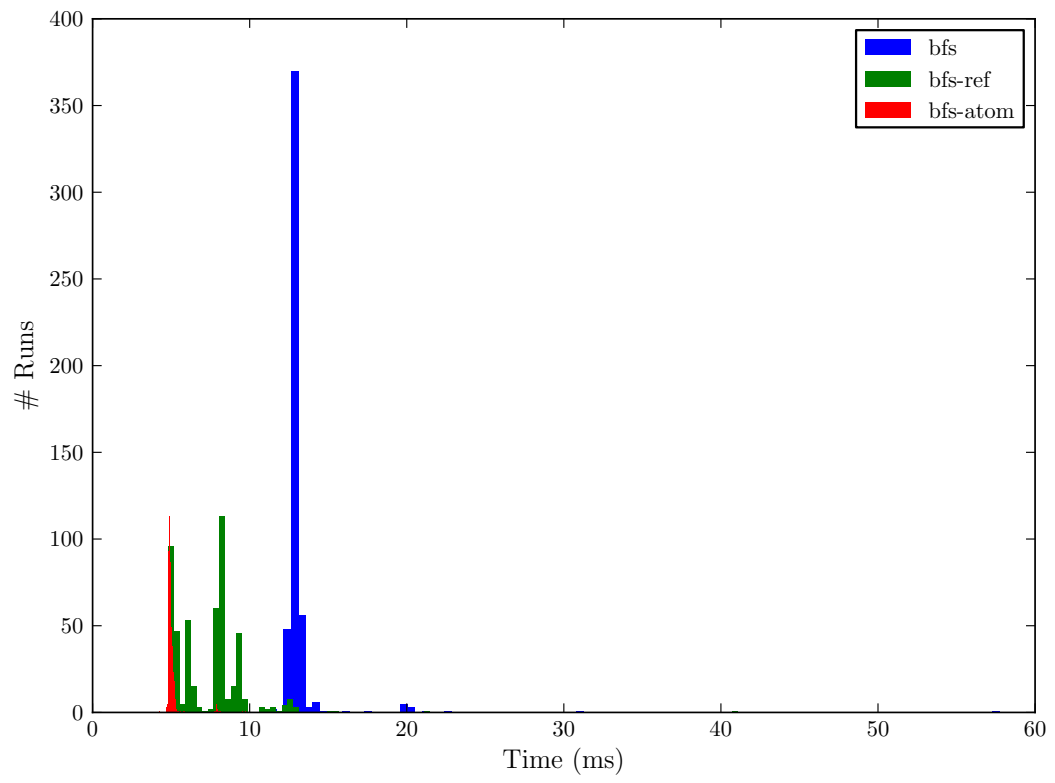


Figure 7.2: 1 Hop

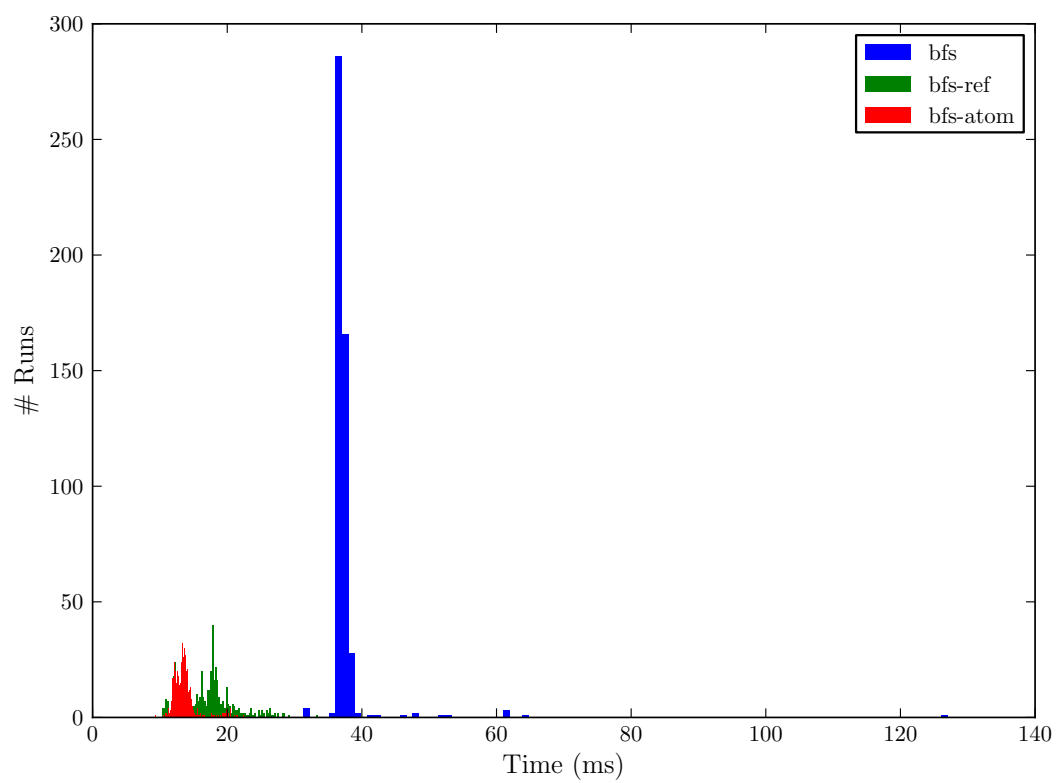


Figure 7.3: 2 Hops

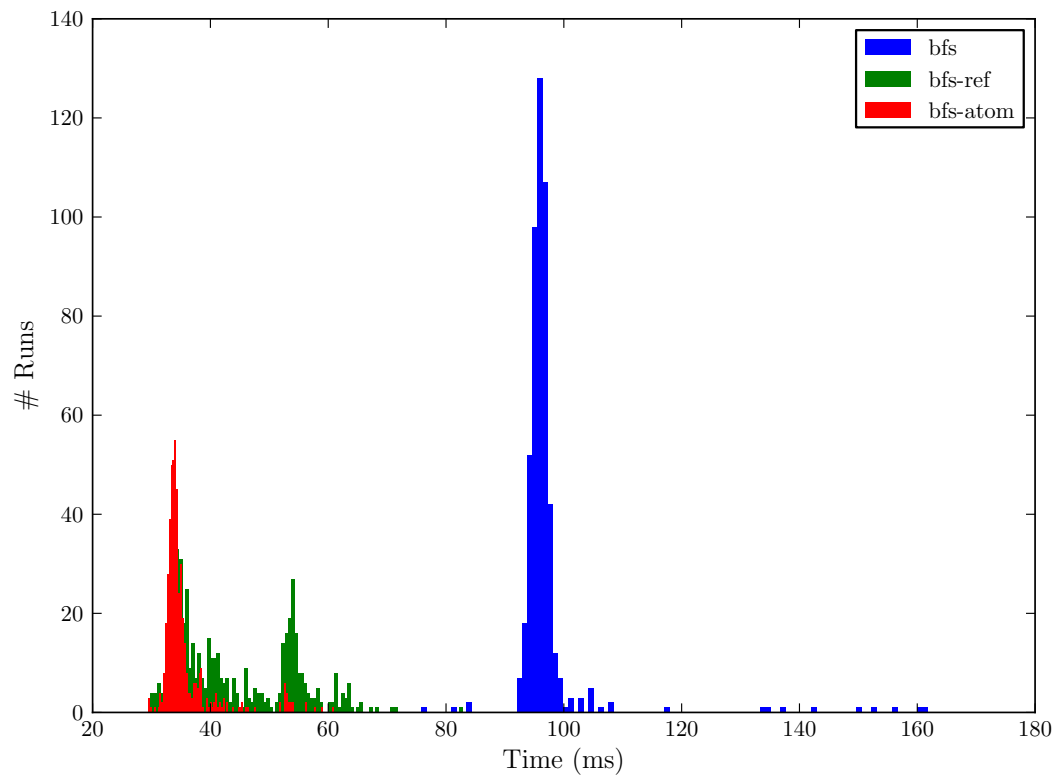


Figure 7.4: 3 Hops

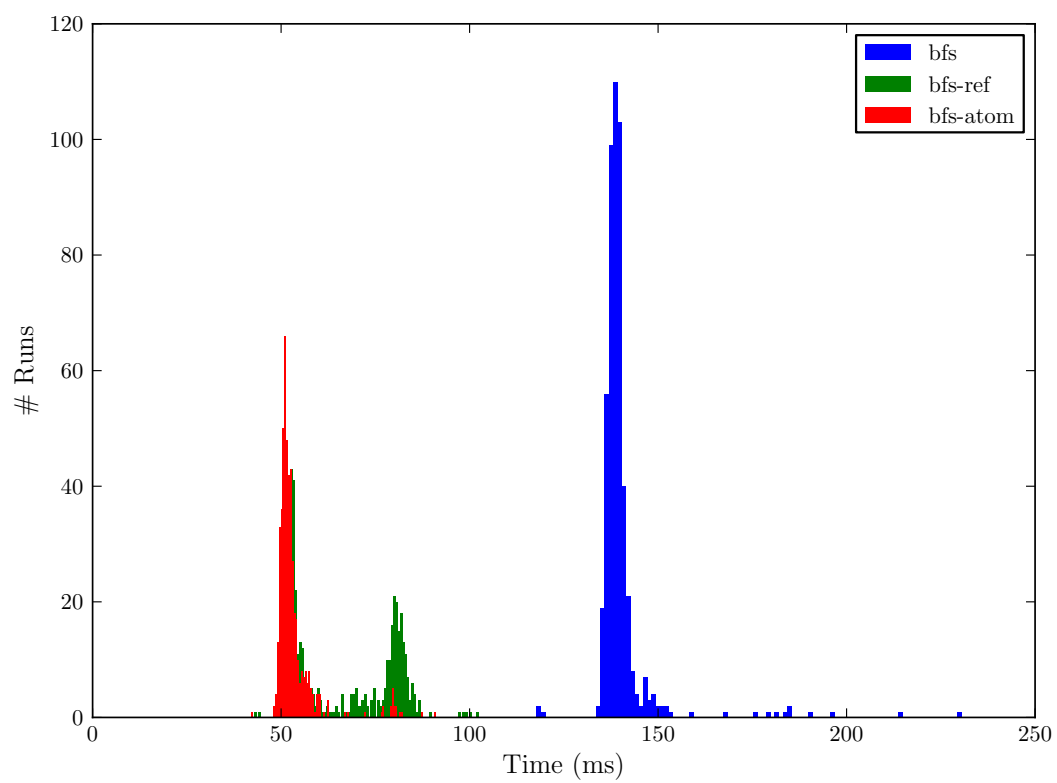


Figure 7.5: 4 Hops

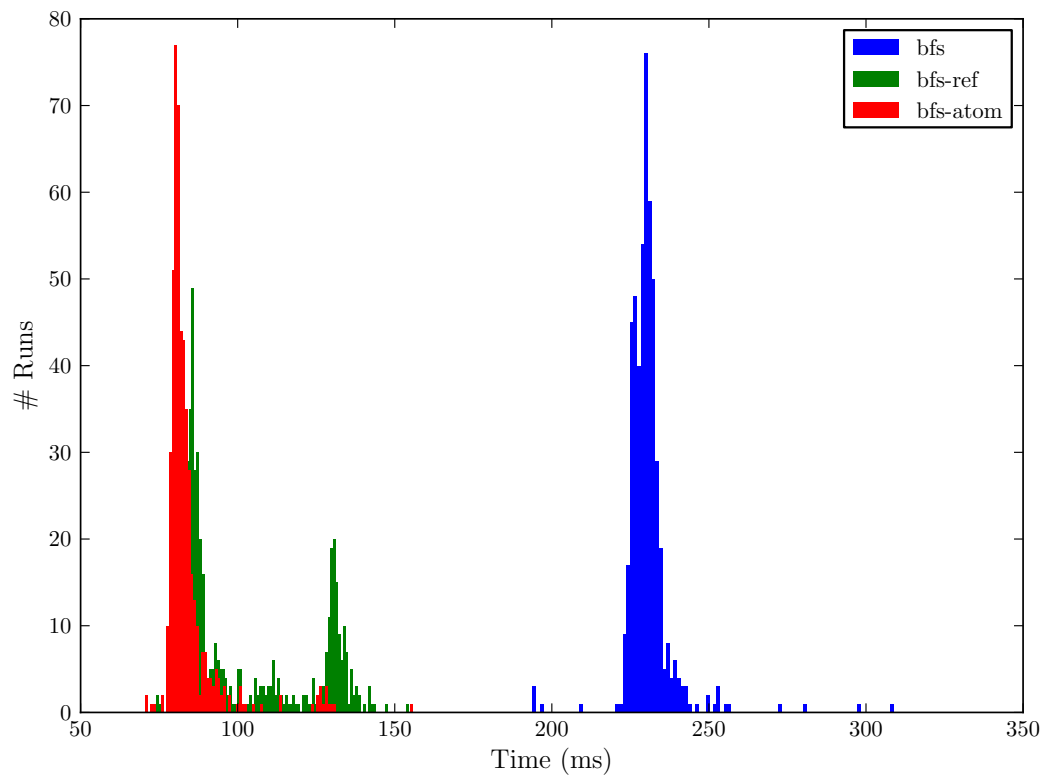


Figure 7.6: 5 Hops

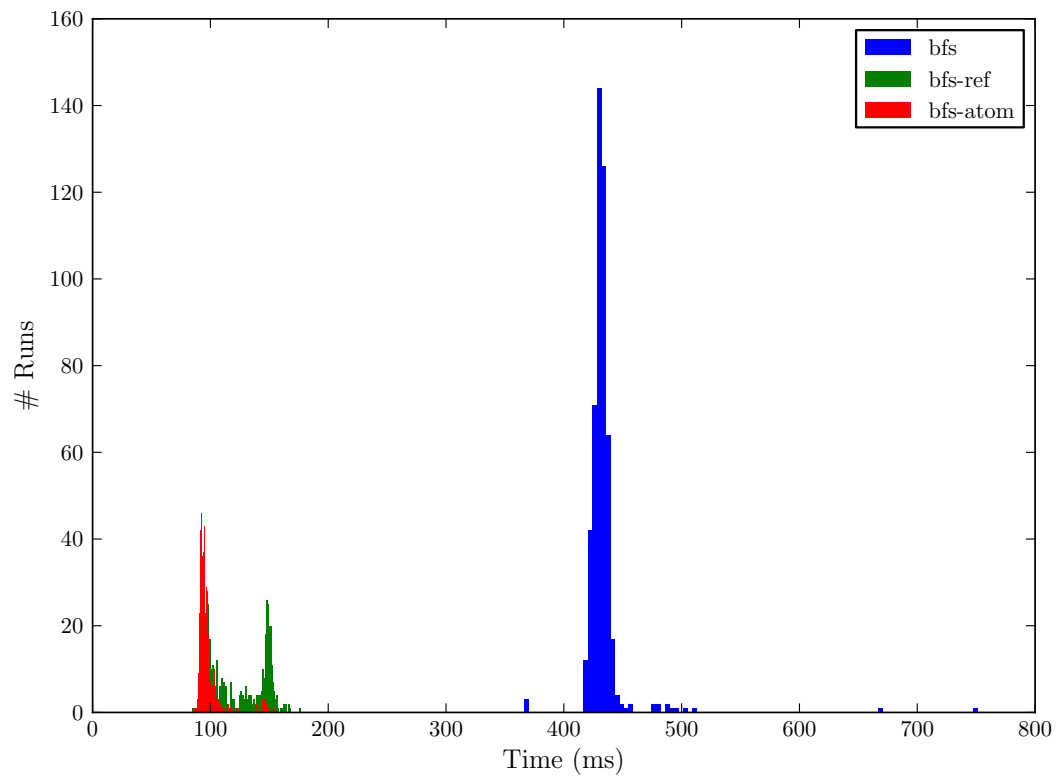


Figure 7.7: 6 Hops

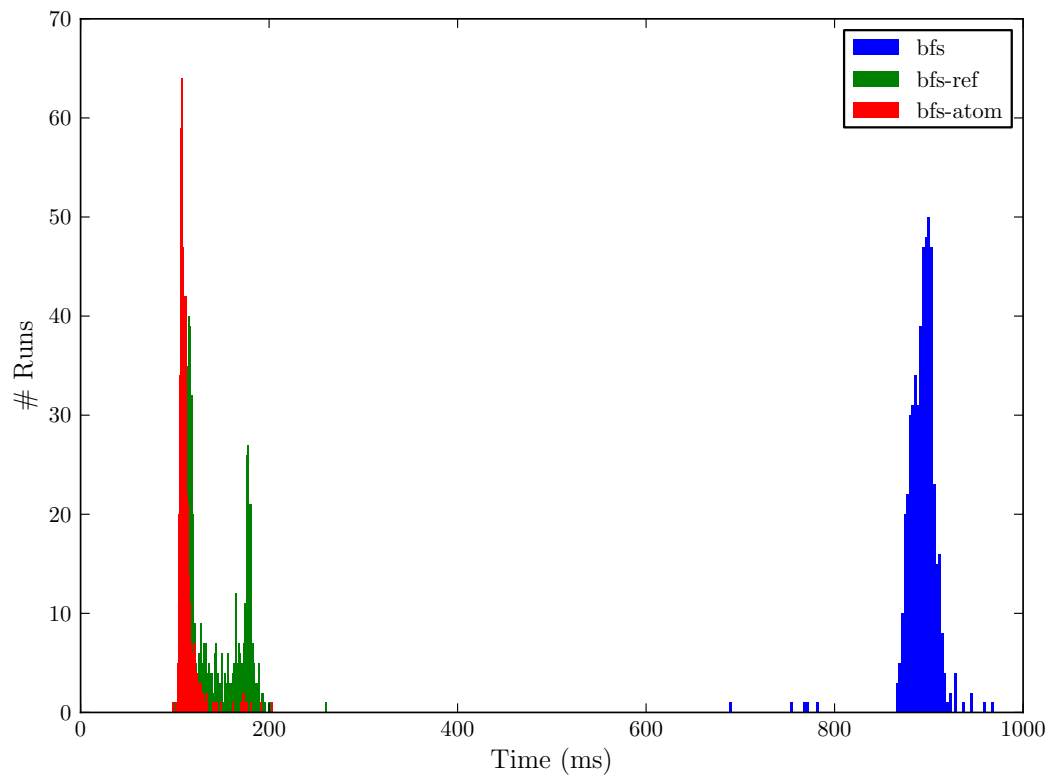


Figure 7.8: 7 Hops

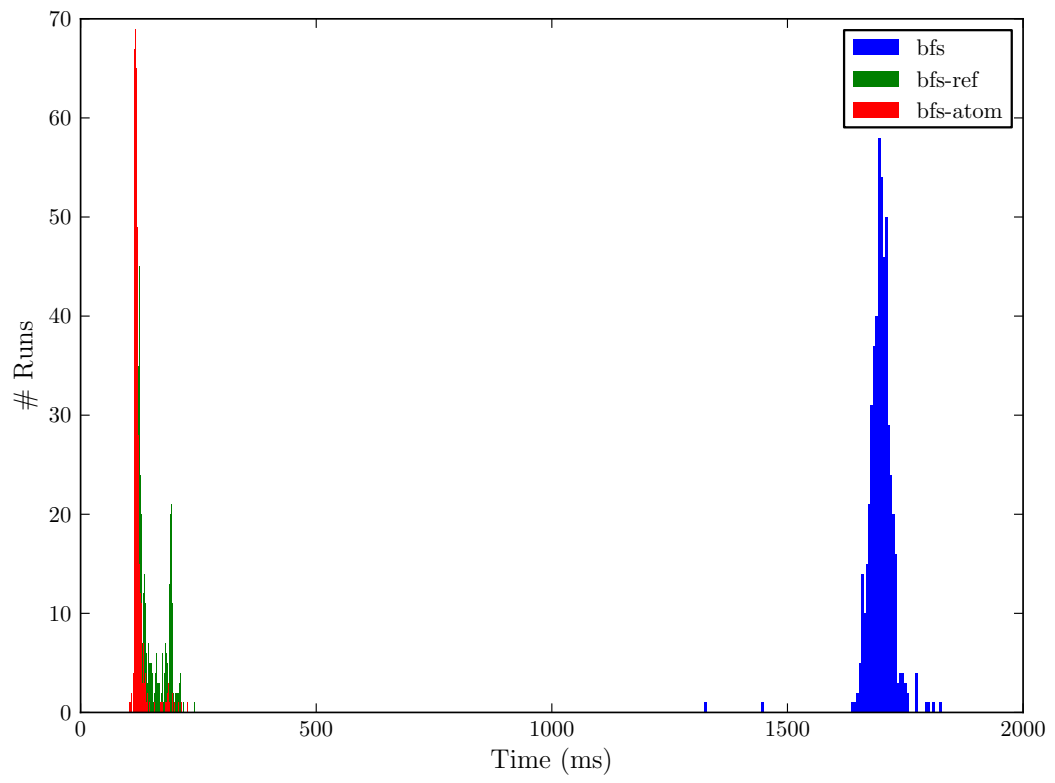


Figure 7.9: 8 Hops

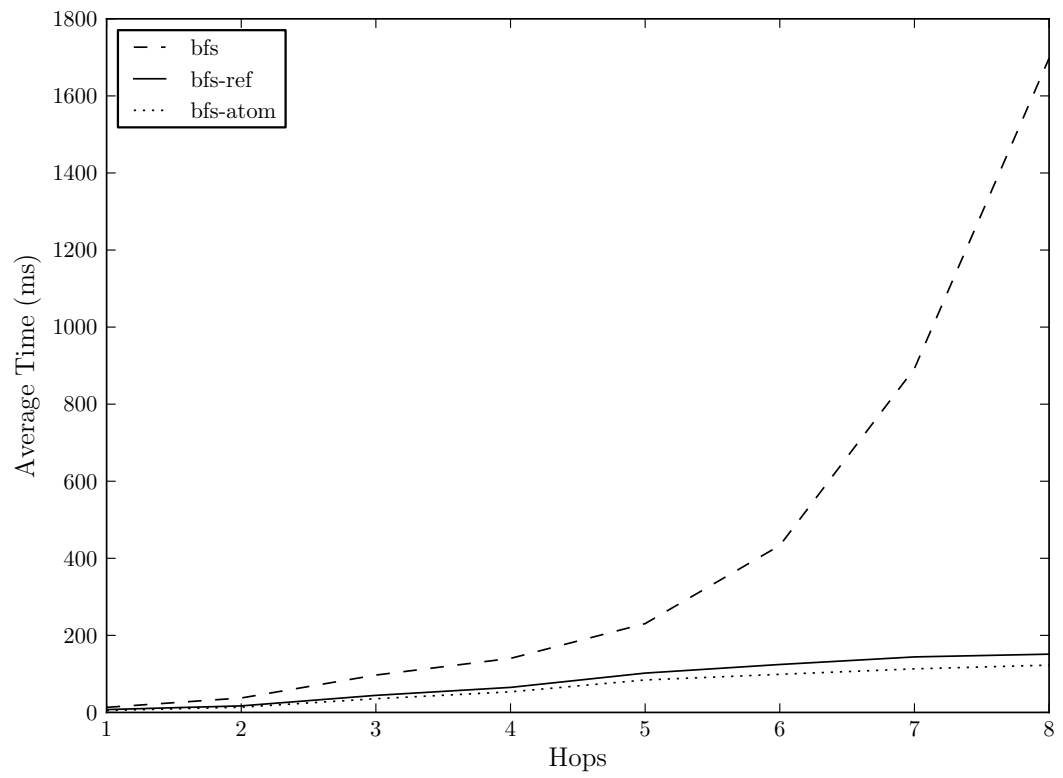


Figure 7.10: Comparison between single threaded and concurrent graph search

Chapter 8

Conclusion

[1]

Appendix A

Source Code

Each namespace in the code is divided into sections in the thesis document.

A.1 molly

A.1.1 molly.core

```
1 (ns molly.core
2   (:gen-class)
3   (:use clojure.pprint
4         molly.conf.config
5         molly.index.build
6         molly.search.lucene
7         molly.search.query-builder
8         [clojure.tools.cli :only (cli)]
9         [molly.algo.bfs-atom :only (bfs-atom)]
10        [molly.algo.bfs-ref :only (bfs-ref)]
11        [molly.algo.bfs :only (bfs)]))
12
13 (defn parse-args
14   [args]
15   (cli args
16     ["-c" "--config" "Path to configuration (properties) file"]
17     ["--algorithm" "Algorithm to run"]
18     ["-s" "--source" "Source node"]
19     ["-t" "--target" "Target node"]
20     ["--index" "Build an index of the database"
21      :default false
22      :flag true]
23     ["--benchmark" "Run benchmarks"
24      :default false
25      :flag true]
26     ["-h" "--help" "Show help"]
```

```

27         :default false
28         :flag true]))
29
30 (def counter (atom 0))
31
32 (defn ns-to-ms
33   [nanosec]
34   (/ nanosec (* 1000.0 1000.0)))
35
36 (defn warmup
37   [func searcher source target seconds]
38   (let [start (System/nanoTime)
39         time-in-s (* seconds 1000 1000 1000)]
40     (while (< (- (System/nanoTime) start) time-in-s)
41       (func searcher source target)
42       (swap! counter inc)))
43
44   @counter)
45
46 (defn benchmark
47   [f searcher source target]
48   (let [start (System/nanoTime)]
49     (f searcher source target)
50     (let [elapsed (- (System/nanoTime) start)]
51       (println
52        (str "Elapsed time: " (ns-to-ms elapsed) " msecs")))))
53
54 (defn bench
55   [f searcher source target]
56   (warmup f searcher "courses|alsu_1101u"
57    "schedules|schedules|3067sections|1246teaches|893" 15)
58   (benchmark f searcher source target))
59
60 (defn -main
61   [& args]
62   (let [[opts arguments banner] (parse-args (flatten args))]
63     (when (and (opts :help) (not (opts :config)))
64       (println banner)
65       (System/exit 0))
66
67     (let [properties (load-props (opts :config))
68           algo (fn [f searcher source target]
69                  (if (opts :benchmark)
70                      (bench f searcher source target)
71                      (f searcher source target)))]

```



```

72 (cond
73   (opts :index)      (let [database (properties :database)
74                           index      (properties :index)]
75                         (build database index))
76   (opts :algorithm) (let [searcher (idx-searcher
77                                   (idx-path
78                                    (properties :index)))
79                           source    (opts :source)
80                           target    (opts :target)]
81     (condp = (opts :algorithm)
82       "bfs" (algo bfs
83                  searcher
84                  source
85                  target)
86       "bfs-atom" (algo bfs-atom
87                        searcher
88                        source
89                        target)
90       "bfs-ref" (algo bfs-ref
91                      searcher
92                      source
93                      target))
94     (shutdown-agents))))))

```

A.2 molly.conf

A.2.1 molly.conf.config

```
1 (ns molly.conf.config
2   (:require clojure.java.io))
3
4 (defn load-props
5   [file-name]
6   (let [res (clojure.java.io/resource file-name)]
7     (if (nil? res)
8       (throw
9         (IllegalArgumentException. "Unable to load properties."))
10      (with-open [^java.io.Reader reader
11                  (clojure.java.io/reader res)]
12        (let [props (java.util.Properties.)]
13          (.load props reader)
14          (into {}
15              (for [[k v] props] [(keyword k)
16                                  (read-string v)])))))))
17
18 (defprotocol IConfig
19   (connection [this])
20   (schema [this])
21   (index [this]))
```

A.2.2 molly.conf.mycampus

```
1 (ns molly.conf.mycampus
2   (:use molly.conf.config
3         molly.datatypes.database
4         molly.datatypes.schema
5         korma.core
6         korma.db)
7   (:import (molly.datatypes.database Sqlite)
8             (molly.datatypes.schema EntitySchema)))
9
10 (declare Campus Course Subject Term Section Schedule
11       Location Instructor db-conn)
12
13 (defentity Campus
14   (has-many Location))
15
16 (defentity Location
17   (belongs-to Campus))
18
19 (defentity Subject
20   (has-many Course))
21
22 (defentity Course
23   (pk :code)
24   (belongs-to Subject)
25   (has-many Section))
26
27 (defentity Instructor
28   (has-many Schedule))
29
30 (defentity Term
31   (has-many Section))
32
33 (defentity Section
34   (pk :crn)
35   (has-many Schedule)
36   (belongs-to Term)
37   (belongs-to Course {:fk :course_code}))
38
39 (defentity Schedule
40   (belongs-to Section)
41   (belongs-to Instructor)
42   (belongs-to Location))
```

```

43
44 (def mycampus-schema
45   [(EntitySchema.
46     {:T      :entity
47      :C      :course
48      :sql    Course
49      :ID     :code
50      :attrs  [:code :title]
51      :values [:code :title]})
52    (EntitySchema.
53      {:T      :entity
54       :C      :instructor
55       :sql    Instructor
56       :ID     :id
57       :attrs  [:name]
58       :values [:name]})
59    (EntitySchema.
60      {:T      :entity
61       :C      :location
62       :sql    Location
63       :ID     :id
64       :attrs  [:name]
65       :values [:name]})
66    (EntitySchema.
67      {:T      :entity
68       :C      :subject
69       :sql    Subject
70       :ID     :id
71       :attrs  [:id :name]
72       :values [:id :name]})
73    (EntitySchema.
74      {:T      :entity
75       :C      :campus
76       :sql    Campus
77       :ID     :id
78       :attrs  [:name]
79       :values [:name]})
80    (EntitySchema.
81      {:T      :entity
82       :C      :term
83       :sql    Term
84       :ID     :id
85       :attrs  [:id :name]
86       :values [:id :name]})
87    (EntitySchema.

```

```

88     {:T      :entity
89      :C      :section
90      :sql     Section
91      :ID      :crn
92      :attrs   [:crn :reg_start :reg_end :credits
93                :section_num :levels]
94      :values  [:crn]})
95 (EntitySchema.
96  {:T      :entity
97   :C      :schedule
98   :sql     Schedule
99   :ID      :id
100  :attrs   [:days :sch_type :date_start :date_end
101            :time_start :time_end :week]
102  :values  []})
103 (EntitySchema.
104  {:T      :group
105   :C      "Instructor schedule"
106   :sql     (->
107             (select* Schedule)
108             (with Instructor))
109   :ID      [[:instructor :instructor_id "Instructor ID"]
110            [:schedule :id "Schedule ID"]]
111   :attrs   []
112   :values  []})
113 (EntitySchema.
114  {:T      :group
115   :C      "Course schedule"
116   :sql     (->
117             (select* Schedule)
118             (with Section
119              (with Course)))
120   :ID      [[:section :crn "CRN"]
121            [:course :code "Code"]
122            [:schedule :id "Schedule ID"]]
123   :attrs   []
124   :values  []})
125 (EntitySchema.
126  {:T      :group
127   :C      "Schedule location"
128   :sql     (->
129             (select* Schedule)
130             (with Location
131              (with Campus)))
132   :ID      [[:campus :campus_id "Campus ID"]

```

```

133         [:location :location_id "Location ID"]
134         [:schedule :id "Schedule ID"]]]
135     :attrs []
136     :values []})
137 (EntitySchema.
138   {:T :group
139    :C "Course subject"
140    :sql (->
141          (select* Course)
142          (with Subject))
143    :ID [[[:course :id "Course"]
144          [:subject :subject_id "Subject"]]]
145    :attrs []
146    :values []})
147 (EntitySchema.
148   {:T :group
149    :C "Section term"
150    :sql (->
151          (select* Section)
152          (with Term))
153    :ID [[[:section :id "Section"]
154          [:term :term_id "Term"]]]})
155   ])
156
157 (deftype Mycampus [db-path idx-path]
158   IConfig
159   (connection
160     [this]
161     (defdb db-conn (sqlite3 {:db db-path}))
162     (Sqlite. db-conn))
163   (schema
164     [this]
165     mycampus-schema)
166   (index
167     [this]
168     idx-path))

```

A.3 molly.datatypes

A.3.1 molly.datatypes.database

```
1 (ns molly.datatypes.database
2   (:use korma.core
3         korma.db))
4
5 (defprotocol Database
6   (execute-query [this query f]))
7
8 (deftype Sqlite [conn]
9   Database
10  (execute-query
11    [this query f]
12    (with-db conn
13      (doseq [result (-> query (select))]
14        (f result)))))
```

A.3.2 molly.datatypes.entity

```
1 (ns molly.datatypes.entity
2   (:use molly.util.nlp)
3   (:import
4     [clojure.lang IPersistentMap IPersistentList]
5     [org.apache.lucene.document
6       Document Field
7       Field$Index Field$Store]))
8
9 (defn special?
10   [field-name]
11   (and (.startsWith field-name "__") (.endsWith field-name "__")))
12
13 (defn uid
14   "Possible inputs include:
15   row :T :ID
16   row [[:T :ID] [:T :ID]]
17   row [[:T :ID :desc] [:T :ID :desc]]"
18   ([row C id]
19     (if (nil? (row id))
20       (throw
21         (Exception.
22           (str "ID column " id " does not exist in row " row ".")))
23       (str (name C)
24            "| "
25            (clojure.string/replace (row id) #"\\s+" "_")))))
26   ([row Tids]
27     (clojure.string/join " " (for [[C id] Tids]
28                                   (uid row C id)))))
29
30 (defn field
31   [field-name field-value]
32   (Field. field-name
33           field-value
34           Field$Store/YES
35           Field$Index/ANALYZED))
36
37 (defn document
38   [fields]
39   (let [doc (Document.)]
40     (do
41       (doseq [[field-name field-value] fields]
42         (.add doc (field (name field-name) (str field-value))))))
```



```

43         doc)))
44
45 (defn row->data
46   ^{:doc "Transforms a row into the internal representation."}
47   [this schema]
48   (let [T      (schema :T)
49         C      (schema :C)
50         attr-cols (schema :attrs)
51         attrs    (if (nil? attr-cols)
52                     this
53                     (select-keys this attr-cols))
54         meta-data {:type T :class C}
55         id-col    (schema :ID)]
56     (with-meta (if (= T :group)
57                  (conj attrs {:entities (uid this id-col)})
58                  attrs)
59              (condp = T
60                  :value   (assoc meta-data
61                                   :class
62                                   (clojure.string/join "|"
63                                   (map name
64                                        [C (first attr-cols)])))
65                  :entity  (assoc meta-data :id
66                                   (if (coll? id-col)
67                                       (uid this id-col)
68                                       (uid this C id-col)))
69                  :group   (assoc meta-data
70                                   :entities
71                                   (uid this id-col)))
72              (throw
73               (IllegalArgumentException.
74                "I only know how to deal with types :value,
75                :entity, and :group"))))))
76
77 (defn doc->data
78   ^{:doc "Transforms a Document into the internal representation."}
79   [this]
80   (let [fields      (.getFields this)
81         extract     (fn [x] [(keyword (clojure.string/replace
82                                         (.name x) "_" ""))
83                               (.stringValue x)])
84         check-special (fn [x] (special? (.name x)))
85         filter-fn     (fn [f] (apply hash-map
86                                       (flatten
87                                        (map extract

```

```

88                                     (filter f fields))))))]]
89     (with-meta (filter-fn (fn [x] (not (check-special x))))
90                 (filter-fn check-special))))
91
92 (defn data->doc
93   ^{:doc "Transforms the internal representation into a Document."}
94   [this]
95   (let [int-meta (meta this)
96         T        (int-meta :type)
97         all       (clojure.string/join " "
98                                         (if (= T :entity)
99                                             (conj (vals this)
100                                                    (name
101                                                       (int-meta :class)))
102                                             (vals this)))
103         luc-meta  [[:__type__ (name T)]
104                   [[:__class__ (name (int-meta :class))]]
105                   [[:__all__ (if (= T :value)
106                                (q-gram all)
107                                all))]]
108         raw-doc   (concat luc-meta
109                           this
110                           (condp = (int-meta :type)
111                                :value  [[:value all]]
112                                :entity [[:__id__ (int-meta :id)]]
113                                :group  []))])
114   (document raw-doc)))

```

A.3.3 molly.datatypes.schema

```
1 (ns molly.datatypes.schema
2   (:use molly.datatypes.database
3         molly.datatypes.entity
4         molly.search.lucene
5         molly.util.nlp
6         korma.core))
7
8 (defprotocol Schema
9   (crawl [this db-conn idx-w])
10  (klass [this])
11  (schema-map [this]))
12
13 (deftype EntitySchema [S]
14   Schema
15   (crawl
16     [this db-conn idx-w]
17     (let [sql (S :sql)]
18       (execute-query db-conn sql
19         (fn [row]
20           (add-doc idx-w
21             (data->doc (row->data row S)))))))
22
23     (if (= (S :T) :entity)
24       (doseq [value (S :values)]
25         (let [query (->
26                   sql
27                   (modifier "DISTINCT")
28                   (fields value)
29                   (group value))]
30           (execute-query db-conn query
31             (fn [row]
32               (add-doc idx-w (data->doc
33                 (row->data row
34                   (assoc S
35                     :T :value)))))))))))
36   (klass
37     [this]
38     ((schema-map this) :C))
39   (schema-map
40     [this]
41     S))
```

A.4 molly.index

A.4.1 molly.index.build

```
1 (ns molly.index.build
2   (:use molly.conf.config
3         molly.conf.mycampus
4         molly.datatypes.database
5         molly.datatypes.entity
6         molly.datatypes.schema
7         molly.search.lucene)
8   (:import (molly.conf.mycampus Mycampus)))
9
10 (defn build
11   [db-path path]
12   (let [conf      (Mycampus. db-path path)
13         db-conn   (connection conf)
14         ft-path   (idx-path (index conf))
15         idx-w     (idx-writer ft-path)
16         schemas   (schema conf)]
17     (doseq [ent-def schemas]
18       (println "Indexing" (name (klass ent-def)) "...")
19       (crawl ent-def db-conn idx-w))
20
21     (close-idx-writer idx-w)))
```

A.5 molly.util

A.5.1 molly.util.nlp

```
1 (ns molly.util.nlp)
2
3 (defn q-gram
4   ([S]
5    (q-gram S 3 "$"))
6   ([S n]
7    (q-gram S n "$"))
8   ([S n s]
9    (let [padding (clojure.string/join "" (repeat (dec n) "$"))
10          padded-S (str padding
11                        (clojure.string/replace S " " padding)
12                        padding)]
13      (clojure.string/join " "
14                            (for [i (range
15                                    (+ 1 (- (count padded-S) n)))]
16                              (. padded-S substring i (+ i n)))))))
```

A.6 molly.search

A.6.1 molly.search.lucene

```
1 (ns molly.search.lucene
2   (:import
3     (java.io File)
4     (org.apache.lucene.analysis.core WhitespaceAnalyzer)
5     (org.apache.lucene.index IndexReader IndexWriter
6       IndexWriterConfig)
7     (org.apache.lucene.search IndexSearcher)
8     (org.apache.lucene.store Directory SimpleFSDirectory)
9     (org.apache.lucene.util Version)))
10
11 (def version
12   Version/LUCENE_44)
13 (def default-analyzer
14   (WhitespaceAnalyzer. version))
15
16 (defn ^Directory idx-path
17   [path]
18   (-> path File. SimpleFSDirectory.))
19
20 (defn idx-searcher
21   [^IndexSearcher idx-path]
22   (-> (IndexReader/open idx-path) IndexSearcher.))
23
24 (defn ^IndexWriter idx-writer
25   ([^Directory idx-path analyzer]
26    (IndexWriter. idx-path (IndexWriterConfig. version analyzer)))
27   ([^Directory idx-path]
28    (idx-writer idx-path default-analyzer)))
29
30 (defn close-idx-writer
31   [^IndexWriter idx-writer]
32   (doto idx-writer
33     ; Lucene docs say not to use .optimize anymore.
34     (.commit)
35     (.close)))
36
37 (defn idx-search
38   [idx-searcher query topk]
39   (let [results (. (. idx-searcher search query topk) scoreDocs)]
40     (map (fn [result] (.doc idx-searcher (.doc result))) results)))
```

```
41  
42 (defn add-doc  
43   [idx doc]  
44   (. idx addDocument doc))
```

A.6.2 molly.search.query_builder

```
1 (ns molly.search.query-builder
2   (:import (org.apache.lucene.index Term)
3             (org.apache.lucene.search BooleanClause$Occur
4                                         BooleanQuery
5                                         PhraseQuery)))
6
7 (defn query
8   [kind & args]
9   (let [field-name (condp = kind
10                      :type    "__type__"
11                      :class    "__class__"
12                      :id       "__id__"
13                      :text     "__all__"
14                      ; Assume "kind" is an attribute name.
15                      (condp = (type kind)
16                            clojure.lang.Keyword (name kind)
17                            java.lang.String      kind))
18       phrase-query (PhraseQuery.)]
19     (doseq [arg args]
20       (. phrase-query add (Term. field-name (name arg)))))
21
22   phrase-query))
23
24 (defn boolean-query
25   [args]
26   (let [query (BooleanQuery.)]
27     (doseq [[q op] args]
28       (. query add q (condp = op
29                        :and BooleanClause$Occur/MUST
30                        :or  BooleanClause$Occur/SHOULD
31                        :not BooleanClause$Occur/MUST_NOT)))
32
33     query))
```

A.7 molly.server

A.7.1 molly.server.serve

```
1 (ns molly.server.serve
2   (:use compojure.core
3         molly.conf.config
4         molly.datatypes.entity
5         molly.search.lucene
6         molly.search.query-builder
7         molly.util.nlp
8         [molly.algo.bfs :only (bfs)]
9         [molly.algo.bfs-atom :only (bfs-atom)]
10        [molly.algo.bfs-ref :only (bfs-ref)])
11   (:require [noir.response :as response]
12             [compojure.handler :as handler]
13             [compojure.route :as route]))
14
15 (def runtime (Runtime/getRuntime))
16 (def props (load-props ".properties"))
17 (def searcher (idx-searcher (idx-path (props :index))))
18
19 (defn dox
20   [q1 field S op topk]
21   (let [bq (boolean-query
22           (concat [[q1 :and]]
23                   (for [s S]
24                       [(query field s) op])))
25         result (map doc->data (idx-search searcher bq topk))
26         fmt (fn [data] {:meta (meta data) :results data})]
27     (map fmt result)))
28
29 (defn entities
30   [field q topk]
31   (dox (query :type :entity)
32        field
33        (clojure.string/split q #"\\s{1}")
34        :and
35        topk))
36
37 (defn home-page []
38   (response/redirect "/index.html"))
39
40 (defn get-value [q]
```

```

41 (response/json
42   {:result
43     (dox (query :type :value)
44           :text
45           (clojure.string/split (q-gram q) #"\\s{1}")
46           :or
47           (props :topk_value)))))
48
49 (defn get-entities [q]
50   (response/json
51     {:result
52       (entities :text q (props :topk_entities))}))
53
54 (defn get-entity [q]
55   (response/json
56     {:result
57       (entities :id q (props :topk_entity))}))
58
59 (defn get-span [e0 eL method]
60   (let [start (System/nanoTime)
61         [visited dist prev]
62         (condp = method
63           "atom" (bfs-atom searcher e0 eL)
64           "ref" (bfs-ref searcher e0 eL)
65           (bfs searcher e0 eL))
66         t (- (System/nanoTime) start)
67         eids (for [[k v] prev] k)
68         get-entities (fn [eid]
69                       {(keyword eid)
70                        (entities :id eid
71                                (props :topk_entity))})]
72     (response/json
73       {:from e0
74        :to eL
75        :prev prev
76        :entities entities
77        :debug {:time t
78                :mem_total (.totalMemory runtime)
79                :mem_free (.freeMemory runtime)
80                :mem_used (- (.totalMemory runtime)
81                             (.freeMemory runtime))
82                :properties props}})))
83
84
85 (defroutes app-routes

```

```
86         (GET "/" [] (home-page))
87         (GET "/value" [q] (get-value q))
88         (GET "/entities" [q] (get-entities q))
89         (GET "/entity" [q] (get-entity q))
90         (GET "/span" [e0 eL method] (get-span e0 eL method))
91         (route/files "/" {:root "resources/public"}))
92
93 (def app (handler/site app-routes))
```

A.8 molly.algo

A.8.1 molly.algo.common

```
1 (ns molly.algo.common
2   (:use molly.datatypes.entity
3         molly.search.lucene
4         molly.search.query-builder))
5
6 (defn find-entity-by-id
7   [G id]
8   (let [query (boolean-query [[(query :type :entity) :and]
9                                [(query :id id) :and]])]
10     (map doc->data (idx-search G query 10))))
11
12 (defn find-group-for-id
13   [G id]
14   (let [query (boolean-query [[(query :type :group) :and]
15                                [(query :entities id) :and]])]
16     results (map doc->data (idx-search G query 10))
17     big-str (clojure.string/join " "
18                                   (map #(% :entities) results)))
19     (distinct (clojure.string/split big-str #"\\s{1}"))))
20
21 (defn find-adj
22   [G v]
23   (remove #{v} (find-group-for-id G v)))
24
25 (defonce initial-state
26   (fn
27     [s]
28     (atom {:Q      (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
29            :marked #{s}
30            :dist   {s 0}
31            :prev   {}
32            :done   false})))
33
34 (defonce update-state
35   (fn
36     [state u v]
37     (let [Q (state :Q)
38           marked (state :marked)
39           dist (state :dist)
40           prev (state :prev)]
```

```
41      (assoc state
42        :Q      (conj Q v)
43        :marked (conj marked v)
44        :dist   (assoc dist v (inc (dist u)))
45        :prev   (assoc prev v u))))
46
47 (defonce deref-future
48   (fn
49     [dfd]
50     (if (future? dfd)
51         (deref dfd)
52         dfd)))
```

A.8.2 molly.algo.bfs

```
1 (ns molly.algo.bfs
2   (use molly.algo.common))
3
4 (defn update-adj
5   [G marked dist prev u]
6   (loop [adj      (find-adj G u)
7          marked    marked
8          dist      dist
9          prev      prev
10         frontier []]
11     (if (empty? adj)
12         [(conj marked u) dist prev frontier]
13         (let [v      (first adj)
14               adj'    (rest adj)]
15             (if (marked v)
16                 (recur adj' marked dist prev frontier)
17                 (let [dist' (assoc dist v (inc (dist u)))
18                       prev' (assoc prev v u)]
19                     (recur adj' marked dist' prev' (conj frontier v))))))))))
20
21 (defn bfs
22   [G s t]
23   (loop [Q      (-> (clojure.lang.PersistentQueue/EMPTY) (conj s))
24          marked #{ }
25          dist   {s 0}
26          prev   {s nil}]
27       (if (or (empty? Q)
28               (some (fn [node] (= node t)) marked))
29           [marked dist prev]
30           (let [u      (first Q)
31                 Q'      (rest Q)
32                 [marked' dist' prev' frontier]
33                 (update-adj G marked dist prev u)]
34             (recur (concat Q' frontier) marked' dist' prev')))))
```

A.8.3 molly.algo.bfs_atom

```
1 (ns molly.algo.bfs-atom
2   (use molly.algo.common))
3
4 (defn update-adj
5   [state-ref G u]
6   (let [marked? (@state-ref :marked)
7         deferred (doall
8                     (for [v (find-adj G u)]
9                       (if (marked? v)
10                          nil
11                          (future
12                           (swap! state-ref update-state u v))))))]
13     (doall (map deref-future deferred))))
14
15 (defn bfs-atom
16   [G s t]
17   (let [state-ref (initial-state s)]
18     (while (and (not (empty? (@state-ref :Q)))
19                (not (@state-ref :done)))
20       (let [u (first (@state-ref :Q))
21             Q' (pop (@state-ref :Q))]
22         (swap! state-ref assoc :Q Q')
23         (if (some (fn [node] (= node t)) (@state-ref :marked))
24             (swap! state-ref assoc :done true)
25             (update-adj state-ref G u))))
26     [(@state-ref :marked) (@state-ref :dist) (@state-ref :prev)]))
```

A.8.4 molly.algo.bfs_ref

```
1 (ns molly.algo.bfs-ref
2   (use molly.algo.common))
3
4 (defn update-adj
5   [state-ref G u]
6   (let [marked? (@state-ref :marked)
7         deferred (doall
8                     (for [v (find-adj G u)]
9                       (if (marked? v)
10                          nil
11                          (future (dosync (alter
12                                      state-ref
13                                      update-state
14                                      u
15                                      v))))))]
16     (doall (map deref-future deferred))))
17
18 (defn bfs-ref
19   [G s t]
20   (let [state-ref (initial-state s)]
21     (while (and (not (empty? (@state-ref :Q)))
22                (not (@state-ref :done)))
23       (let [u (first (@state-ref :Q))
24             Q' (pop (@state-ref :Q))]
25         (dosync (alter state-ref assoc :Q Q'))
26         (if (some (fn [node] (= node t)) (@state-ref :marked))
27             (dosync (alter state-ref assoc :done true))
28             (update-adj state-ref G u))))
29     [(@state-ref :marked) (@state-ref :dist) (@state-ref :prev)]))
```

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...

- ☐ I (p. 21): Add overview diagram of system design.