Towards a Concurrent Implementation of Keyword Search Over Relational Databases

by

Richard J.I. Drake

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

Master of Science (M.Sc.)

in

The Faculty of Science

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Ken Q. Pu

December 2013

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

CSV comma-separated values. 28

FK foreign key. 3, 4, 6

JSON JavaScript object notation. 28

RDBMS relational database management system. 7, 8, 10

SQL structured query language. 5, 18

List of Symbols

N number of documents in collection. 11

M size of T. 11

r set of named tuples. 2, 3, 6, 22

τ unique terms in . 11–14, 18

t named tuple. 2, 3, 6, 21, 22

T all terms in document collection. ix, 11

Chapter 1

Background (2 days)

Literature search on:

- DBExplore
- XRank
- BANKS
- ...

Chapter 2

A Tale of Two Data Models

The term "data model" refers to a notation for describing data and/or information. It consists of the data structure, operations that may be performed on the data, as well as constraints placed on the data [GUW09].

In this chapter we provide a formal definition of the relational data model, discuss its merits, its shortcomings, and contrast it to the document data model. Contrary to the relational model, the document model permits fast and flexible keyword search without requiring explicit domain knowledge of the data. In addition, we demonstrate the feasibility of encoding a relational model into a document model in a lossless manner.

2.1 Relational Model

In its most basic form, the relational data model is built upon sets and tuples. Each of these sets consist of a set of finite possible values. Tuples are constructed from these sets to form relations.

Definition 1 (Named Tuple). A named tuple t is an instance of a relation r, consisting of values corresponding to the attributes of r. For example,

Example 1. Given a tuple $t = \{\text{code} : \text{"CDPS 101"}, \text{title} : \text{"Human-Mutant Relations"}, \text{subject} : \text{"CDPS"}\}$, we denote the attributes of t as $\textsc{Attr}[t] = \{\text{code}, \text{title}, \text{subject}\}$. The values are $t[\text{code}] =$

"CDPS 101", t[title] = "Human-Mutant Relations", and t[subject] = "CDPS".

Definition 2 (Relation). A relation r is a set of named tuples, $r = \{t_1, t_2, \ldots, t_n\}$, such that all the named tuples share the same attributes.

$$\forall t, t' \in r, \textsc{Attr}[t] = \textsc{Attr}[t'] \tag{2.1}$$

Example 2. An example Course relation, r, would be

$$r = \left\{ \begin{array}{llllll} \{\text{code} & : \text{"CDPS 101"}, & \text{title} & : \text{"Human-Mutant Relations"}, & \text{subject} & : \text{"CDPS"}\}, \\ \{\text{code} & : \text{"CDPS 201"}, & \text{title} & : \text{"Humans and You"}, & \text{subject} & : \text{"CDPS"}\}, \\ \{\text{code} & : \text{"MATH 360"}, & \text{title} & : \text{"Complex Analysis"}, & \text{subject} & : \text{"MATH"}\} \end{array} \right\}$$

Relations are typically represented as tables.

code	title	subject
CDPS 101	Human-Mutant Relations	CDPS
CDPS 201	Humans and You	CDPS
MATH 360	Complex Analysis	MATH

Table 2.1: Course relation

Definition 3 (Keys). Keys are constraints imposed on relations. A key constraint K on a relation r is a subset of $\textsc{Attr}[r]$ which may uniquely identify a tuple. Formally, we say r satisfies the key constraint K, denoted as $r \vDash \mathrm{K}$, subject to

$$\forall t, t' \in r, t \neq t' \implies t[\mathrm{K}] \neq t'[\mathrm{K}]$$

For example, in Table 2.1, the relation satisfies the key constraint {code} or {title}, but not {subject}.

Definition 4 (Foreign Keys). A FK constraint applies to two relations, r_1, r_2. It asserts that values of certain attributes of r_1 must appear as values of some corresponding attributes of r_2. A FK constraint is written as

$$\theta = r_1(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k}) \to r_2(\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k})$$

where $\alpha_{1i} \subseteq \text{Attr}[r_1]$ and $\alpha_{2i} \subseteq \text{Attr}[r_2]$. We say (r_1, r_2) satisfies θ, denoted as $(r_1, r_2) \vDash \theta$, if for every tuple $t \in r_1$, there exists a tuple $t' \in r_2$ such that $t[\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k}] = t'[\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k}]$.

We say r_1 is the source, while r_2 is the target.

Example 3. Suppose we have a relation Course(code, title, subject). We impose a FK constraint of

$$\theta = \text{Course(subject)} \rightarrow \text{Subject(id)} \tag{2.2}$$

which asserts (Course, Subject) $\vDash \theta$. Therefore, if

$$t = \{\text{code} : \text{“CDPS 101”}, \text{title} : \text{“Human-Mutant Relations”}, \text{subject} : \text{“CDPS”}\}$$

then $\exists!\, t' \in$ Subject such that $t'[\text{id}] =$ “CDPS”.

Definition 5 (Relational Database). A relational database, $\mathbb{D}$, is a named collection of relations (as defined by Definition 2 on the preceding page), keys (as defined by Definition 3 on the previous page), and foreign key constraints (as defined by Definition 4 on the preceding page).

We use $\text{Name}[\mathbb{D}]$ to denote the name of $\mathbb{D}$, $\text{Rel}[\mathbb{D}]$ the list of relations in $\mathbb{D}$, $\text{Key}[\mathbb{D}]$ the list of key constraints of $\mathbb{D}$, and $\text{FK}[\mathbb{D}]$ the list of foreign key constraints of $\mathbb{D}$.

2.1.1 Schema Group

Definition 6 (Schema Graph). If we view relations as vertices, and foreign key constraints as edges, a database $\mathbb{D}$ can be viewed as a *schema graph* $\boldsymbol{G}$, formally defined as

$$\text{vertices} : \text{V}(\boldsymbol{G}) = \text{Rel}[\mathbb{D}] \tag{2.3}$$

$$\text{edges} : \text{E}(\boldsymbol{G}) = \text{FK}[\mathbb{D}] \tag{2.4}$$

Example 4. Given the schema in Fig. 2.1 on the next page

and the FK constraints in Eq. (2.5) on the following page

Subject(id, name)

Course(code, title, subject)

Term(id, name)

Section(crn, term, course)

Figure 2.1: Subset of mycampus dataset schema

$$\text{Course(subject)} \rightarrow \text{Subject(id)} \tag{2.5}$$

$$\text{Section(term)} \rightarrow \text{Term(id)} \tag{2.6}$$

$$\text{Section(course)} \rightarrow \text{Course(code)} \tag{2.7}$$

we produce the schema graph in Fig. 2.2

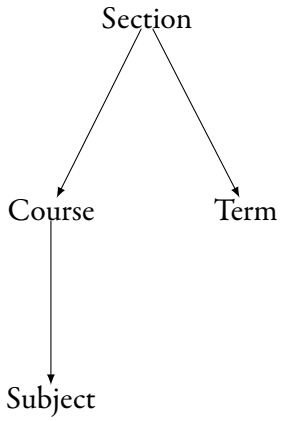

Figure 2.2: Graph representation of relations (Fig. 2.1) and FK (Eq. (2.5)).

The relational data model is particularly powerful for analytic queries. Given the schema graph in Fig. 2.2, one can formulate the following analytic queries in a query language known as structured query language (SQL).

Example 5. Using SQL, find all section CRNs for the subject titled "Community Development & Policy Studies."

The SQL query in Fig. 2.3 on the next page results in Table 2.2 on the following page.

```
SELECT section.crn
FROM   section
       JOIN course
         ON section.course_code = course.code
       JOIN subject
         ON subject.id = course.subject_id
WHERE  subject.name = 'Community Development & Policy Studies';
```

Figure 2.3: Query to find section CRNs for a subject name.

crn
10000
10001
10002

Table 2.2: Results of the query in Fig. 2.3.

2.1.2 Entity Group

Definition 7 (Entity Group)**.** An entity group is a forest, T, of tuples interconnected by join conditions defined by the FK constraints in the schema graph G.

Given two vertices $t, t' \in \mathrm{V}(T)$, $\exists r_1, r_2 \in \textsc{Rel}[\mathbb{D}]$ such that $t \in r_1$, $t' \in r_2$, and $(r_1, r_2) \in G$. That is, t and t' belong to two relations that are connected by the schema graph.

Let $r_1(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k}) \to r_2(\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k})$ be the FK that connects r_1, r_2. We further assert that $t[\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1k}] = t'[\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2k}]$.

Entity groups define complex, structured objects that include more information than individual tuples in the relations.

Example 6. The information in Table 2.3 on the next page all relates to the Course titled Human-Mutant Relations, however no single tuple in the database has all of this information as a result of database normalization.

We require an entity group (Fig. 2.4 on the following page) to join together all pieces of informa-

Attribute	Value
code	CDPS 101
title	Human-Mutant Relations
subject	Community Development & Policy Studies

Table 2.3: Properties of the Course titled Human-Mutant Relations.

tion related to this course.

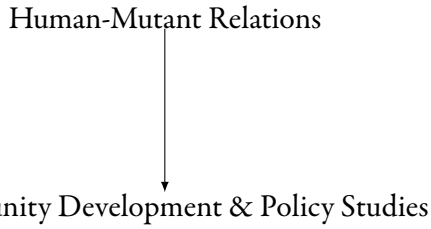

Figure 2.4: Human-Mutant Relations entity group

2.1.3 Pros and Cons of the Relational Model

In order to better understand the motivation behind this work, it is important to examine both the strong and weak points of the relational model.

Pros

The enforcement of constraints is essential to the relational model. There are several types of constraints, including uniqueness and FKs. The first constraint maintains uniqueness.

The Course relation (Table 2.1 on page 3) has the attribute `code` as its primary key. In order for other relations to reference a specific named tuple, the `code` attribute must be unique.

Example 7 (Unique Constraint)**.** Attempt to insert another course with a `code` of "CDPS 101."

```
1 INSERT INTO course
2 VALUES      ('CDPS 101',
```

```
3               'Mutant-Human Relations',
4               'CDPS');
```

The relational database management system (RDBMS) enforces the primary key constraint on the `code` attribute, rejecting the insertion.

```
Error: column code is not unique
```

With the uniqueness of named tuples guaranteed (as demonstrated in Example 7 on the preceding page), we must ensure that any named tuples that are referenced actually exist. If they do not, the database must not permit the operation to continue. Doing so would lead to dangling references.

Example 8 (Referential Integrity). Attempt to insert the tuple ("CHEM 101", "Introductory Chemistry", "'CHEM") in the Course relation.

```
1  INSERT INTO course
2  VALUES      ('CHEM 101',
3               'Introductory Chemistry',
4               'CHEM');
```

Again we see the RDBMS protecting the integrity of the data.

```
Error: foreign key constraint failed
```

In addition to enforcing consistency, the relational model is capable of providing higher-level views of the data through aggregation.

Example 9 (Aggregation). Find the number of sections offered for the subject named "Community Development & Policy Studies."

```
1  SELECT Count(*)
2  FROM   section
3         JOIN course
4           ON section.course = course.code
5         JOIN subject
6           ON subject.id = course.subject
7  WHERE  subject.name = 'Community Development & Policy Studies';
```

Information stored within a properly designed database is normalized. That is, no information is repeated.

Example 10 (Normalization)**.** For example, suppose Emma Frost became headmistress and the subject named "Community Development & Policy Studies" was renamed to "Community Destruction & Policy Studies." If this information were not normalized, each course in this subject would need to be updated. Since this information is normalized, the following query will suffice.

```
1 UPDATE subject
2 SET    name = 'Community Destruction & Policy Studies'
3 WHERE  id = 'CDPS';
```

The above examples are some of the most important reasons for choosing the relational model over others. Unfortunately, the relational model is not without its downsides.

Cons

While the relational model excels at ensuring data consistency, aggregation, and reporting; it is not suitable for every task. In order to issue queries, a user must be familiar with the schema. This requires specific domain knowledge of the data.

An example of a complicated query involving two joins is give in Fig. 2.3 on page 6.

A casual user is unlikely to determine the correct join path, name of the tables, name of the attributes, etc. This is in contrast to the document model, where the data is semi-structured or unstructured, requiring minimal domain knowledge.

The relational model is also rigid in structure. If a relation is modified, every query referencing said relation may require a rewrite. Even a simple attribute being renamed (e.g. $\rho_{\text{name/alias}}(\text{Person})$) is capable of modifying the join paths. This rigidity places additional cognitive burden on users.

In addition to having a rigid structure, most relational database management systems lack flexible string matching options. Assuming basic SQL-92 compliance, a RDBMS only supports the `LIKE` predicate [ISO11].

Example 11 (`LIKE` Predicate). Find all courses with a title that contains "man."

```
SELECT *
FROM   course
WHERE  title LIKE '%man%';
```

There are a couple of limitations to the `LIKE` predicate. First, it only supports basic substring matching. If a user accidentally searches for all courses with a title containing "men," nothing would be found.

Second, unless the predicate is applied to the end of the string and the column is indexed, performance will be poor. The database must scan the entire table in order to answer the query, resulting in performance of $\mathcal{O}(n)$, where n is the number of named tuples in the relation.

2.2 Document Model

In contrast to the relational model, the document model represents semi-structured as well as unstructured data. Examples of information suitable to the document model includes emails, memos, book chapters, etc.

These pieces, or units, of information are broken into documents. Groups of related documents (for example, a library catalogue) are referred to as a document collection.

Definition 8 (Terms and Document). A term, τ, is an indivisible string (e.g. a proper noun, word, or a phrase). A document, d, is a bag of words. Let $\mathrm{freq}(\tau, d)$ be the frequency of terms τ in document d.

Let T denote all possible terms, and Bag[T] be all possible bag of terms.

Remark 1. We use the bag-of-words model for documents. This means that position information of terms in a document is irrelevant, but the frequency of terms are kept in the document. Documents are non-distinct sets.

Definition 9 (Document Collection). A document collection $\mathbb{C}$ is a set of documents, written $D\mathbb{C} = \{d_1, d_2, \ldots, d_k\}$. The size of $\mathbb{C}$ is denoted N.

Example 12. Consider the following short sentences.

1. Superman is strong on Earth and lives on Earth.
2. Batman was born on Earth.
3. Superwoman is fast on Earth.
4. Superman was born on Krypton.

Each sentence represents a document, giving us the following documents.

$$d_1 = \{\text{“and”} : 1, \text{“on”} : 2, \text{“is”} : 1, \text{“lives”} : 1, \text{“earth”} : 2, \text{“strong”} : 1, \text{“superman”} : 1\} \quad (2.8)$$

$$d_2 = \{\text{“batman”} : 1, \text{“on”} : 1, \text{“was”} : 1, \text{“earth”} : 1, \text{“born”} : 1\} \quad (2.9)$$

$$d_3 = \{\text{“on”} : 1, \text{“is”} : 1, \text{“superwoman”} : 1, \text{“fast”} : 1, \text{“earth”} : 1\} \quad (2.10)$$

$$d_4 = \{\text{“krypton”} : 1, \text{“born”} : 1, \text{“on”} : 1, \text{“was”} : 1, \text{“superman”} : 1\} \quad (2.11)$$

$$(2.12)$$

2.2.1 Vectorization of Documents

One of the most fundamental approaches for searching documents is to treat documents as high dimensional vectors, and the document collection as a subset in a vector space. The search query becomes a nearest neighbour query in a vector space equipped with a distance measure.

The first step is to convert a bag of terms into vectors. The standard technique [MRS08] uses a scoring function that measures the relative importance of terms in documents.

Definition 10 (TF-IDF Score). **To do** (1) The term frequency is the number of times a term τ appears in a document d, as given by $\mathrm{freq}(\tau, d)$. The document frequency of a term τ, denoted by $\mathrm{df}(\tau)$, is the number of documents in $\mathbb{C}$ that contains τ. It is defined as

$$\mathrm{df}(\tau) = |\{d \in \mathbb{C} : \tau \in d\}|$$

The combined TF-IDF score of τ in a document d is given by

$$\text{tf-idf}(\mathbb{C}, \tau, d) = \frac{\mathrm{freq}(\tau, d)}{|d|} \cdot \log \frac{N}{\mathrm{df}(\tau)}$$

Remark 2. The first component, $\frac{\mathrm{freq}(\tau,d)}{|d|}$, measures the importance of a term within a document. It is normalized to account for document length. The second component, $\log \frac{N}{\mathrm{df}(\tau)}$, is a measure of the rarity of the term within the document collection $\mathbb{C}$.

Example 13. Using the documents from Example 12 on the preceding page, the TF-IDF scores are as follows.

	d_1	d_2	d_3	d_4
τ_1 : "and"	0.2857	0.0000	0.0000	0.0000
τ_2 : "on"	0.0000	0.0000	0.0000	0.0000
τ_3 : "superwoman"	0.0000	0.0000	0.4000	0.0000
τ_4 : "batman"	0.0000	0.4000	0.0000	0.0000
τ_5 : "is"	0.1429	0.0000	0.2000	0.0000
τ_6 : "fast"	0.0000	0.0000	0.4000	0.0000
τ_7 : "born"	0.0000	0.2000	0.0000	0.2000
τ_8 : "krypton"	0.0000	0.0000	0.0000	0.4000
τ_9 : "earth"	0.1186	0.0830	0.0830	0.0000
τ_{10} : "lives"	0.2857	0.0000	0.0000	0.0000
τ_{11} : "strong"	0.2857	0.0000	0.0000	0.0000
τ_{12} : "was"	0.0000	0.2000	0.0000	0.2000
τ_{13} : "superman"	0.1429	0.0000	0.0000	0.2000

Definition 11 (Document Vector). Given a document collection $\mathbb{C}$ with M unique terms $\mathrm{T} = [\tau_1, \tau_2, \ldots, \tau_n]$, each document d can be represented by an M-dimensional vector.

$$\vec{d} = \begin{bmatrix} \text{tf-idf}(\tau_1, d) \\ \text{tf-idf}(\tau_2, d) \\ \vdots \\ \text{tf-idf}(\tau_n, d) \end{bmatrix}$$

Example 14. The documents in Example 12 on page 11 would produce the following vectors.

$$\vec{d}_n = \begin{bmatrix} \text{tf-idf}(\tau_1, d_n) \\ \text{tf-idf}(\tau_2, d_n) \\ \text{tf-idf}(\tau_3, d_n) \\ \text{tf-idf}(\tau_4, d_n) \\ \text{tf-idf}(\tau_5, d_n) \\ \text{tf-idf}(\tau_6, d_n) \\ \text{tf-idf}(\tau_7, d_n) \\ \text{tf-idf}(\tau_8, d_n) \\ \text{tf-idf}(\tau_9, d_n) \\ \text{tf-idf}(\tau_{10}, d_n) \\ \text{tf-idf}(\tau_{11}, d_n) \\ \text{tf-idf}(\tau_{12}, d_n) \\ \text{tf-idf}(\tau_{13}, d_n) \end{bmatrix}, \vec{d}_1 = \begin{bmatrix} 0.2857 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.1429 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.1186 \\ 0.2857 \\ 0.2857 \\ 0.0000 \\ 0.1429 \end{bmatrix}, \vec{d}_2 = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.4000 \\ 0.0000 \\ 0.0000 \\ 0.2000 \\ 0.0000 \\ 0.0830 \\ 0.0000 \\ 0.0000 \\ 0.2000 \\ 0.0000 \end{bmatrix}, \vec{d}_3 = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.4000 \\ 0.0000 \\ 0.2000 \\ 0.4000 \\ 0.0000 \\ 0.0000 \\ 0.0830 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{bmatrix}, \vec{d}_4 = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.2000 \\ 0.4000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.2000 \\ 0.2000 \end{bmatrix}$$

Definition 12 (Search Query). A search query q is simply a document, namely a bag of terms. **To do** (2)

The top-k answers to q with respect to a collection $\mathbb{C}$ is defined as the k documents, $\{d_1, d_2, \ldots, d_k\}$, in $\mathbb{C}$, such that $\{\vec{d}_i\}$ are the closest vectors to $\vec{q}$ using Euclidean distance measure in $\mathbb{R}^N$.

Example 15. Given the search query $q = \{\text{superwoman, was, born, on, krypton}\}$, compute the vector $\vec{q}$ within the document collection $\mathbb{C}$ (as defined in Example 12 on page 11).

$$\vec{q} = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.1474 \\ 0.2644 \\ 0.0000 \\ 0.2644 \\ 0.1474 \\ 0.0000 \end{bmatrix}$$

In order to determine the top-k documents for search query q, we need a way of measuring the similarity between documents.

Definition 13 (Cosine Similarity). Given two document vectors, $\vec{d}_1$ and $\vec{d}_2$, the cosine similarity is the dot product $\vec{d}_1 \cdot \vec{d}_2$, normalized by the product of the Euclidean distance of $\vec{d}_1$ and $\vec{d}_2$ in $\mathbb{R}^N$. It is denoted as $\text{similarity}(\vec{d}_1, \vec{d}_2)$.

$$\text{similarity}(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{\|\vec{d}_1\| \cdot \|\vec{d}_2\|} \tag{2.13}$$

$$= \frac{\sum_{i=1}^{N} \vec{d}_{1,i} \times \vec{d}_{2,i}}{\sqrt{\sum_{i=1}^{N} (\vec{d}_{1,i})^2} \times \sqrt{\sum_{i=1}^{N} (\vec{d}_{2,i})^2}} \tag{2.14}$$

Recall we may represent search queries as documents and thus document vectors. Therefore we may compute the score of a document d for a search query q as

$$\text{similarity}(\vec{d}, \vec{q})$$

Example 16. Given the document collection $\mathbb{C}$ (from Example 12 on page 11) and search query q, compute the similarity between q and every document $d \in \mathbb{C}$.

$$\text{similarity}(\vec{d}_1, \vec{q}) = 0.000000 \tag{2.15}$$

$$\text{similarity}(\vec{d}_2, \vec{q}) = 0.191533 \tag{2.16}$$

$$\text{similarity}(\vec{d}_3, \vec{q}) = 0.265877 \tag{2.17}$$

$$\text{similarity}(\vec{d}_4, \vec{q}) = 0.618553 \tag{2.18}$$

2.2.2 Extending the Document Model

In the extended document model, documents have fields: Field$[d]$, and each attribute have values (e.g. date, string, integer), or bag of terms. Thus:

$$d \ : \ \text{Field}[d] \rightarrow \text{Bag}[\text{T}]$$

Example 17 (Semi-Structured Document)**.** We see that d_2 is about Batman. The document contents are semi-structured, containing both a name and the name of a planet. By adding fields to the document, we are left with Table 2.4.

Field	Value
name	Batman
birthplace	Earth
body	Batman was born on Earth.

Table 2.4: Person document for Batman

which is similar in structure to the `Person` table. **To do** (3)

2.2.3 Approximate String matching

Definition 14 (N-Gram). An n-gram is a contiguous sequence of substrings of string S of length n. An algorithm for computing the n-gram of S is given in Algorithm 1.

Algorithm 1 N-GRAM(S, n, s)

Require: S is a string, $n \geq 1$, and s is a character

Ensure: the list of n-grams of S

1: $G \leftarrow []$
2: $p \leftarrow \text{REPEAT}(s, n-1)$
3: $S \leftarrow \text{PAD}(S, p)$
4: $S \leftarrow \text{REPLACE}(S, \text{' '}, p)$
5: **for** $i = 0$ **to** $l - n + 1$ **do**
6: append $S[i, i+n]$ to G
7: **end for**
8: **return** G

Where l is the length of S, $\text{REPEAT}(S, n)$ repeats s character n times, $\text{PAD}(S, p)$ prefixes and postfixes S with p, and $\text{REPLACE}(S, s, p)$ replaces character s with p in string S.

Example 18. Given a string $S =$ "superman", compute the trigram of S using Algorithm 1.

$$G = \{\text{“\$\$s”}, \text{“\$su”}, \text{“sup”}, \text{“upe”}, \text{“per”}, \text{“erm”}, \text{“rma”}, \text{“man”}, \text{“an\$”}, \text{“n\$\$”}\}$$

We use n-grams in order to permit approximate string matching.

Example 19. Given a string S (Example 18), let $S' =$ "superwoman". Compute the trigram of S' and compare it to S.

$$G' = \{\text{“\$\$s”}, \text{“\$su”}, \text{“sup”}, \text{“upe”}, \text{“per”}, \text{“erw”}, \text{“rwo”}, \text{“wom”}, \text{“oma”}, \text{“man”}, \text{“an\$”}, \text{“n\$\$”}\}$$

$$
\begin{array}{lcc}
 & G & G' \\
\tau_1 : \text{“an\$”} & 1 & 1 \\
\tau_2 : \text{“oma”} & 0 & 1 \\
\tau_3 : \text{“\$su”} & 1 & 1 \\
\tau_4 : \text{“rwo”} & 0 & 1 \\
\tau_5 : \text{“rma”} & 1 & 0 \\
\tau_6 : \text{“man”} & 1 & 1 \\
\tau_7 : \text{“erw”} & 0 & 1 \\
\tau_8 : \text{“\$\$s”} & 1 & 1 \\
\tau_9 : \text{“upe”} & 1 & 1 \\
\tau_{10} : \text{“n\$\$”} & 1 & 1 \\
\tau_{11} : \text{“per”} & 1 & 1 \\
\tau_{12} : \text{“wom”} & 0 & 1 \\
\tau_{13} : \text{“sup”} & 1 & 1 \\
\tau_{14} : \text{“erm”} & 1 & 0
\end{array}
$$

Figure 2.5: Comparison between n-grams of G and G'.

Comparing G to G' results in the following matrix

As Fig. 2.5 shows, using n-grams yield a similarity of $\frac{8}{14}$.

2.2.4 Pros and Cons of the Document Model

There are numerous reasons to use the document model. It allows users without domain knowledge and working knowledge of a complex query language such as SQL to find information. **To do** (4)

Example 20 (Simple Queries)**.** Find all documents related to “Superman” or “Earth”. This query, if the default operator is `OR`, would simply be `Superman Earth`. The result of the query q would be

$$\text{query}(\text{“superman”}) \cup \text{query}(\text{“earth”}) \rightarrow \{d_1, d_2, d_3, d_4\}$$

Users can also modify queries to require certain terms be present or not present.

Example 21 (`AND` Query). Find all documents containing both “Superman” and “Earth”. This query would return the following set of documents

$$\text{query}(\text{“superman”}) \cap \text{query}(\text{“earth”}) \rightarrow \{d_1\}$$

as only d_1 contains both terms.

Example 22 (`NOT` Query). Find all documents containing “Superman” but not “Earth”. This query would return different results than Example 21.

$$\text{query}(\text{“superman”}) \neg \text{query}(\text{“earth”}) \rightarrow \{d_4\}$$

While none of the above queries required domain knowledge, it is possible to use the extended document model (Section 2.2.2 on page 15) to search specific fields. Doing so allows users to have finer control over what documents are retrieved.

Example 23 (Extended Query). Find all documents with a superhero named “Superman” that contain the term “Earth”.

$$\text{query}(\text{“name”}, \text{“superman”}) \cap \text{query}(\text{“earth”}) \rightarrow \{d_3\}$$

Assuming the first term of every document is also the value of the name attribute.

People utilize keyword query search every day through web search engines such as Google.

Not only does the document model provide a familiar interface to search for information with, it also ranks the results. In the relational model a search for “Superman” would return all named tuples that contained that term. In the document model, documents are ranked against the query q and the top-k documents are returned.

The advantage is that users have the result of q already ranked so only the most relevant documents may be explored. As the number of documents matching q for a large corpus can be high, showing only the top-k relevant documents may save the user a substantial amount of time.

The relational model does not permit approximate string matching. By utilizing the document model with n-grams (Section 2.2.3 on page 16), users who substitute, delete, or insert characters from the desired term may still receive results for their intended term (see Example 19 on page 16 for a demonstration of how n-grams overcome character substitutions).

Unfortunately the document model does not support the concept of foreign keys (Definition 4 on page 3). While information is easily accessible due to flexible search, each document is a discrete unit of information. Aggregate queries are unsupported, as these units are not linked amongst one another.

Chapter 3

Best of Both Worlds

3.1 Encoding Named Tuples into Documents

Recall in the extended document model (Section 2.2.2 on page 15), a document d consists of fields $f_1, f_2, \ldots, f_n$. Using the extended document model, we are left with a straight forward mapping of a tuple t to document d.

For tuple t, every attribute $\alpha \in \textsc{Attr}[t]$ maps to field f in d. Every attribute value must be analyzed into an indexable form in order to store it in a field.

$$\textsc{Attr}[t] \xrightarrow{analyzed} \textsc{Field}[d] \tag{3.1}$$

$$\alpha_1, \alpha_2, \ldots, \alpha_n \xrightarrow{analyzed} f_1, f_2, \ldots, f_n \tag{3.2}$$

We denote the document encoding of t as $\textsc{Doc}[t]$.

3.2 Mapping of Entity Groups to Documents

Recall that an entity group (Definition 7 on page 6) is a forest T of tuples t such that

$$\forall (t, t') \in T, t \neq t' \Rightarrow \textsc{Rel}[t] \neq \textsc{Rel}[t']$$

That is, all tuples are from distinct relations.

Given the restriction

$$\forall (r, r') \in G\mathbb{D}, \exists ! \, (r, r') \vDash \theta$$

we assert that if $(t, t') \in T$, then $(t, t') \in \textsc{Rel}[t] \bowtie \textsc{Rel}[t']$.

Let $\mathrm{V}(T)$ be the vertices of T, $\mathrm{E}(T)$ be the edges of T.

Claim 1. *T can always be reconstructed from $\mathrm{V}(T)$ without loss of information.*

Proof. Given $\mathrm{V}(T)$, we must reconstruct $\mathrm{E}(T)$ in order to complete T.

Choose any $(t, t') \in \mathrm{V}(T)$. If $(\textsc{Rel}[t], \textsc{Rel}[t']) \in G\mathbb{D}$, then (t, t') is an edge in T.

Recall our earlier assertion that $G\mathbb{D}$ is cycle-free and foreign keys must be unique. □

3.3 Encoding an Entity Group as a Document Group

Given a entity group T, we construct two or more documents in order to represent the entity group in the document model.

For every $t \in \mathrm{V}(T)$, we construct a document $\textsc{Doc}[t]$ (Section 3.1 on the preceding page). With each tuple t stored in the document collection $\mathbb{C}$, we construct an additional document which stores the association information.

Let x be the indexing document of T.

$$x[\text{“entities”}] = \bigcup_{t \in \mathrm{V}(T)} \mathrm{UID}[t]$$

Thus, the encoding of T is defined as

$$T \xrightarrow{\text{encode}} \{\textsc{Doc}[t] : t \in V(T)\} \cup \{x\}$$

It's easy to see that from encode(T) we can recover $\mathrm{V}(T)$, the tuples in T.

By Claim 1, this is sufficient to recover T entirely.

3.4 Encoding Attribute Values into Searchable Documents

Each value for user selected attributes are converted into n-grams, and stored in special documents.

3.5 Iterative Search Using Document Encodings

A document database supports fast and flexible keyword search queries. A search query is characterized by $q = (f, w)$, where f is a field name, and w is a search phrase.

Search[q] is the set of documents returned by the text index. The search function allows us to do many things:

1. Suggest values, correcting spelling errors
2. Given attribute values, search all relevant entities
3. Search for all relevant entity groups of one or more entities, using the indexing document
4. We can connect entities via entity groups using (hyper)graph search algorithms.

Chapter 4

Along Came Clojure

Talk about how great Clojure is.

4.1 Basic principles of functional programming (2 days)

- immutable data structures
- persistent data structures using multi-versioning
- functions (and higher order functions) as values

4.1.1 Features of Clojure (2 days)

- Data structures supporting the universal design pattern
- Concurrency + STM
- Interoperability with JVM (including Lucene)

4.2 Search w/ Clojure

4.2.1 Thirdparty libraries (1 day, week 4)

- Lucene

4.2.2 Indexing of relational objects (5 days, week 5)

- Schema definition
- Crawling using SQL
- Indexing using relational objects
- Fuzzy indexing of values (typed by classes)

4.2.3 Keyword Search in document space (5 days, week 6)

- Disambiguate keywords using fuzzy search (suggestion, overloaded terms)
- Flexibility keyword search for documents
- Translate search result back to relational space

4.2.4 Graph Search in document space (5 days, week 7)

- Why we need graph search
- Search in document graph using graph search algorithms with functional implementations: (Ford Fulkerson, BFS)
- Speed up using concurrency
- Clojure specific optimization: ref + atom

Chapter 5

Experimental Evaluation

5.1 Implementation

- Choice of language
- Statistics about the code base: LOC, classes, ?
- Github hosted

5.2 The data set

- Description of the data set
- Statistics of the data set

5.3 Runtime Evaluation

Scripts were written to coordinate the execution, collection, and transformation of the performance data of our implementation.

5.3.1 Methodology

We used Criterium[1] to handle the execution of the benchmarks as it handles unique concerns stemming from benchmarking on the JVM. These include:

- Statistical processing of multiple evaluations
- Inclusion of a warm-up period, designed to allow the JIT compiler to optimize its code
- Purging of the garbage collector before testing, to isolate timings from GC state prior to testing
- A final forced GC after testing to estimate impact of cleanup on the timing results

Unfortunately this requires a much longer runtime as each function must be invoked numerous times. In extreme cases (Ford-Fulkerson, 8 hops) this can take upwards of 4 hours in our test environment.

Data Collection

Criterium provides us with a Clojure map with performance data. It performs analysis, presenting us with outliers, samples, etc. As this data collection process can take several hours or more, this data is collected and stored for offline analysis.

In order to utilize the Clojure output in Python, a data interchange format (JSON) is used. The benchmark function writes the Criterium performance analysis out as a JSON string to stdout and Python captures the output, JSONifies it, and stores it in an array. This array is written to disk in JSON as well so it can be loaded into the data transformation script.

For example:

```
[{"results": {...}, "method": "bfs", "max-hops": 1}, ...]
```

[1] http://hugoduncan.org/criterium/

Data Processing

Several scientific computing libraries are used in the processing and visualization.

There are two forms the data takes:

- comma-separated values (CSV)
- JavaScript object notation (JSON)

The CSV data is generated from the JSON data which is generated as described in Section 5.3.1 on the preceding page.

With the data loaded, we're interested in a handful of pieces of date per each entry.

- Max hops
- Method
- Mean execution time

We can easily load and parse the JSON data.

- Index speed
- Keyword search speed
- Graph search speed:
  - Ford Fulkerson
  - BFS
  - Concurrent BFS using refs
  - Concurrent BFS using atoms

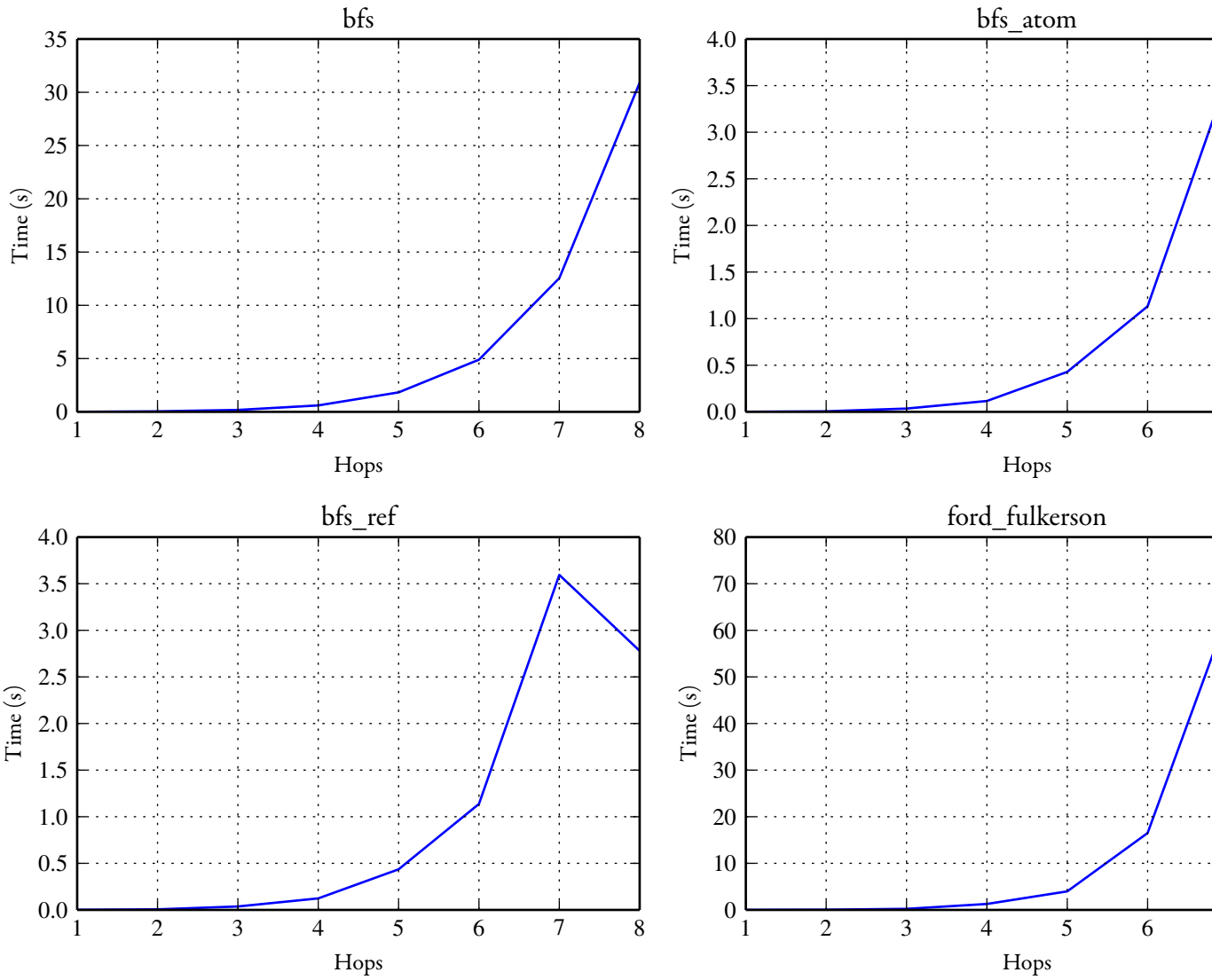

Figure 5.1: Growth of graph search times based on number of hops, plotted separately

5.4 Lessons learned

- Simple algorithms are easier to parallelize
- STM is effective: transactions do not rollback (that much), so we observe impressive speed-up in concurrent versions.
- Fine tuning is beneficial: atom is better than ref.

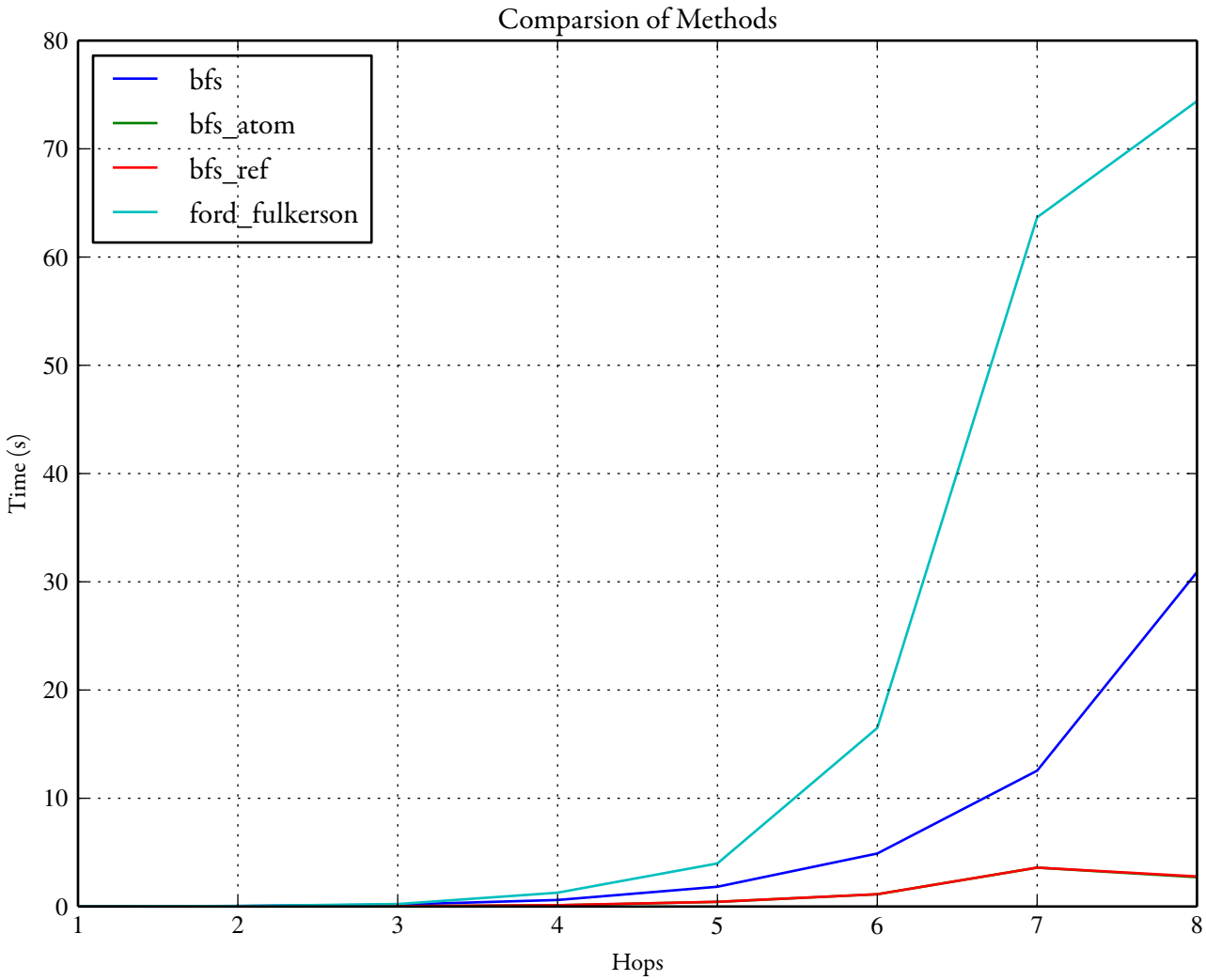

Figure 5.2: Growth of graph search times based on number of hops, combined plot

- The clojure way: correctness first, runtime optimization latter (ref to atom is natural).

Chapter 6

Conclusion (0 days)

Survived Clojure.

Bibliography

[Cod69] E. F. Codd. Derivability, redundancy and consistency of relations stored in large data banks. *Ibm research report, san jose, california*, RJ599, 1969.

[GUW09] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009.

[ISO11] ISO. Information technology – database languages – sql – part 2: foundation (sql/foundation). ISO (ISO/IEC 9075-2:2011). Geneva, Switzerland: International Organization for Standardization, 2011.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN: 0521865719, 9780521865715.

To do...

- ☐ 1 (p. 11): What is a TF-IDF? What does TF-IDF stand for?
- ☐ 2 (p. 13): Awkward wording
- ☐ 3 (p. 15): What table number
- ☐ 4 (p. 17): You state that there are numerous reasons, but only list one