Cloud Computing

Information Retrieval in the Cloud

Dell Zhang
Birkbeck, University of London
2014/15
First, nomenclature...

• Information Retrieval (IR)
  – Focus on textual information (= text/document retrieval)
  – Other possibilities include image, video, music, ...
First, nomenclature...

• What do we search?
  – Generically, “collections”
  – Less-frequently used, “corpora”

• What do we find?
  – Generically, “documents”
  – Even though we may be referring to web pages, PDFs, PowerPoint slides, paragraphs, etc.
Information Retrieval Cycle

- Source Selection
- Query Formulation
- Search
- Selection
- Examination
- Delivery

- Resource
- Query
- Results
- Documents
- Information

- System discovery
- Vocabulary discovery
- Concept discovery
- Document discovery

- source reselection
The Central Problem in Search

Do these represent the same concepts?

Searcher
- Concepts
- Query Terms
  - “tragic love story”

Author
- Concepts
- Document Terms
  - “fateful star-crossed romance”
Abstract IR Architecture

- Query
- Representation Function
- Query Representation
- Comparison Function
- Hits
- Index
- Documents
- Representation Function
- Document Representation

Online vs. Offline

Document acquisition (e.g., web crawling)
How do we represent text?

• “Bag of words”
  – Treat all the words in a document as index terms
  – Assign a “weight” to each term based on “importance” (or, in simplest case, presence/absence of word)
  – Disregard order, structure, meaning, etc. of the words
  – Simple, yet effective!
How do we represent text?

• Assumptions
  – Term occurrence is independent
  – Document relevance is independent
  – “Words” are well-defined
What's a word?

What's a word?
McDonald's slims down spuds
Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

...
Counting Words...

documents → bag of words → inverted index

case folding, tokenization, stopword removal, stemming

syntax, semantics, world knowledge, etc.
Boolean Retrieval

• Users express queries as a Boolean expression
  – AND, OR, NOT
  – Can be arbitrarily nested

• Retrieval is based on the notion of sets
  – Any given query divides the collection into two sets: retrieved, not-retrieved
  – Pure Boolean systems do not define an ordering of the results
Inverted Index: Boolean Retrieval

Doc 1
one fish, two fish

Doc 2
red fish, blue fish

Doc 3
cat in the hat

Doc 4
green eggs and ham

1 2 3 4 blue
1 cat
1 egg
1 fish
1 green
1 ham
1 hat
1 one
1 red
1 two

1 2 3 4 blue
2 cat
3 egg
4 fish
1 green
4 ham
3 hat
1 one
2 red
1 two
Boolean Retrieval

• To execute a Boolean query:
  – Build query syntax tree
    $$\text{( blue AND fish ) OR ham}$$
  – For each clause, look up postings
    $$\begin{array}{c}
    \text{blue} \\
    \text{fish}
    \end{array} \rightarrow \begin{array}{c}
    2 \\
    1 \rightarrow 2
    \end{array}$$
  – Traverse postings and apply Boolean operator
Boolean Retrieval

• Efficiency analysis
  – Postings traversal is linear (assuming sorted postings)
  – Start with shortest posting first
Strengths and Weaknesses

• Strengths
  – Precise
    • If you know the right strategies
    • If you have an idea of what you’re looking for
  – Implementations are fast and efficient
Strengths and Weaknesses

• Weaknesses
  – Users must learn Boolean logic
  – Boolean logic insufficient to capture the richness of language
  – No control over size of result set: either too many hits or none
  – When do you stop reading? All documents in the result set are considered “equally good”
  – What about partial matches? Documents that “don’t quite match” the query may be useful also
Ranked Retrieval

• Order documents by how likely they are to be relevant to the information need
  – Estimate relevance\( (q, d_i) \)
  – Sort documents by relevance
  – Display sorted results

• User model
  – Present hits one screen at a time, best results first
  – At any point, users can decide to stop looking
Ranked Retrieval

• How do we estimate relevance?
  – Assume document is relevant if it has a lot of query terms
  – Replace relevance\((q, d_i)\) with \(\text{sim}(q, d_i)\)
  – Compute similarity of vector representations
Assumption:
Documents that are “close together” in vector space “talk about” the same things
Therefore, retrieve documents based on how close the document is to the query (i.e., similarity ~ “closeness”)
Simularity Metric

• Use “angle” between the vectors:

\[
\cos(\theta) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|d_j\| \|d_k\|}
\]

\[
sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|d_j\| \|d_k\|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}}
\]

• Or, more generally, inner products:

\[
sim(d_j, d_k) = \vec{d}_j \cdot \vec{d}_k = \sum_{i=1}^{n} w_{i,j} w_{i,k}
\]
Term Weighting

• Term weights consist of two components
  – Local: how important is the term in this document?
  – Global: how important is the term in the collection?
Term Weighting

• Here’s the intuition:
  – Local: Terms that appear often in a document should get high weights
  – Global: Terms that appear in many documents should get low weights

• How do we capture this mathematically?
  – Local: Term Frequency (TF)
  – Global: Inverse Document Frequency (IDF)
TF.IDF Term Weighting

\[ w_{i,j} = tf_{i,j} \cdot \log \frac{N}{n_i} \]

\( w_{i,j} \)  \hspace{1cm} \text{weight assigned to term } i \text{ in document } j

\( tf_{i,j} \)  \hspace{1cm} \text{number of occurrence of term } i \text{ in document } j

\( N \)  \hspace{1cm} \text{number of documents in entire collection}

\( n_i \)  \hspace{1cm} \text{number of documents with term } i
Inverted Index: TF.IDF

Doc 1
one fish, two fish

Doc 2
red fish, blue fish

Doc 3
cat in the hat

Doc 4
green eggs and ham

<table>
<thead>
<tr>
<th>Term</th>
<th>Doc 1</th>
<th>Doc 2</th>
<th>Doc 3</th>
<th>Doc 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>egg</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ham</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hat</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>two</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

tf

<table>
<thead>
<tr>
<th>Term</th>
<th>tf 1</th>
<th>tf 2</th>
<th>tf 3</th>
<th>tf 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>egg</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ham</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hat</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>two</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

df

<table>
<thead>
<tr>
<th>Term</th>
<th>df 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
</tr>
<tr>
<td>egg</td>
<td>1</td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
</tr>
<tr>
<td>green</td>
<td>1</td>
</tr>
<tr>
<td>ham</td>
<td>1</td>
</tr>
<tr>
<td>hat</td>
<td>1</td>
</tr>
<tr>
<td>one</td>
<td>1</td>
</tr>
<tr>
<td>red</td>
<td>1</td>
</tr>
<tr>
<td>two</td>
<td>1</td>
</tr>
</tbody>
</table>
Positional Indexes

- Store term position in postings
- Supports richer queries (e.g., proximity)
- Naturally, leads to larger indexes...
Inverted Index: Positional Information

**Doc 1**
one fish, two fish

**Doc 2**
red fish, blue fish

**Doc 3**
cat in the hat

**Doc 4**
green eggs and ham

```
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>egg</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>green</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ham</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>hat</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>one</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>red</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>two</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td></td>
<td></td>
<td>1 [3]</td>
</tr>
<tr>
<td>cat</td>
<td></td>
<td>1</td>
<td>1 [1]</td>
</tr>
<tr>
<td>egg</td>
<td></td>
<td>4</td>
<td>1 [2]</td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
<td>1</td>
<td>2 [4]</td>
</tr>
<tr>
<td>green</td>
<td>1</td>
<td>4</td>
<td>1 [1]</td>
</tr>
<tr>
<td>ham</td>
<td></td>
<td>4</td>
<td>1 [3]</td>
</tr>
<tr>
<td>hat</td>
<td></td>
<td>3</td>
<td>1 [2]</td>
</tr>
<tr>
<td>one</td>
<td>1</td>
<td>1</td>
<td>1 [1]</td>
</tr>
<tr>
<td>red</td>
<td>1</td>
<td>2</td>
<td>1 [1]</td>
</tr>
<tr>
<td>two</td>
<td>1</td>
<td>1</td>
<td>1 [3]</td>
</tr>
</tbody>
</table>
```
Retrieval in a Nutshell

• Look up postings lists corresponding to query terms
• Traverse postings for each query term
• Store partial query-document scores in accumulators
• Select top $k$ results to return
Retrieval: Document-at-a-Time

- Evaluate documents one at a time (score all query terms)

<table>
<thead>
<tr>
<th>blue</th>
<th>9</th>
<th>2</th>
<th>21</th>
<th>1</th>
<th>35</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>fish</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>1</td>
<td>35</td>
<td>2</td>
<td>80</td>
<td>3</td>
</tr>
</tbody>
</table>

Accumulators (e.g. priority queue)

Document score in top k?
- Yes: Insert document score, extract-min if queue too large
- No: Do nothing
Retrieval: Document-at-a-Time

• Tradeoffs
  – Small memory footprint (good)
  – Must read through all postings (bad), but skipping possible
  – More disk seeks (bad), but blocking possible
Retrieval: Query-At-A-Time

- Evaluate documents one query term at a time
  - Usually, starting from most rare term (often with tf-sorted postings)

\[
\text{Score}_{q=x}(\text{doc } n) = s
\]

Accumulators (e.g., hash)

```
blue
9 2 21 1 35 1 ...

fish
1 2 9 1 21 3 34 1 35 2 80 3 ...
```
Retrieval: Query-At-A-Time

• Tradeoffs
  – Early termination heuristics (good)
  – Large memory footprint (bad), but filtering heuristics possible
MapReduce it?

• The indexing problem
  – Scalability is critical
  – Must be relatively fast, but need not be real time
  – Fundamentally a batch operation
  – Incremental updates may or may not be important
  – For the web, crawling is a challenge in itself

Perfect for MapReduce!
MapReduce it?

• The retrieval problem
  – Must have sub-second response time
  – For the web, only need relatively few results

Uh... not so good...
MapReduce: Index Construction

• Map over all documents
  – Emit *term* as key, *(docno, tf)* as value
  – Emit other info as necessary (e.g., term position)
• Sort/shuffle: group postings by term
• Reduce
  – Gather and sort the postings (e.g., by *docno* or *tf*)
  – Write postings to disk
• MapReduce does all the heavy lifting!
Inverted Indexing with MapReduce

Map

<table>
<thead>
<tr>
<th>Key</th>
<th>Doc 1</th>
<th>Doc 2</th>
<th>Doc 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>two</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>fish</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Reduce

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>1</td>
</tr>
<tr>
<td>two</td>
<td>1</td>
</tr>
<tr>
<td>fish</td>
<td>1</td>
</tr>
<tr>
<td>red</td>
<td>1</td>
</tr>
<tr>
<td>blue</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
</tr>
<tr>
<td>hat</td>
<td>1</td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
</tr>
<tr>
<td>red</td>
<td>2</td>
</tr>
<tr>
<td>cat</td>
<td>3</td>
</tr>
<tr>
<td>hat</td>
<td>3</td>
</tr>
</tbody>
</table>

Shuffle and Sort: aggregate values by keys
1: class MAPPER
2:   procedure MAP(docid n, doc d)
3:     \( H \leftarrow \text{new ASSOCIATIVE ARRAY} \)
4:     for all term \( t \in \text{doc d} \) do
5:       \( H\{t\} \leftarrow H\{t\} + 1 \)
6:     for all term \( t \in H \) do
7:       EMIT(term \( t \), posting \( \langle n, H\{t\} \rangle \))

1: class REDUCER
2:   procedure REDUCE(term \( t \), postings \[\langle n_1, f_1 \rangle, \langle n_2, f_2 \rangle \ldots \])
3:     \( P \leftarrow \text{new LIST} \)
4:     for all posting \( \langle a, f \rangle \in \text{postings} \[\langle n_1, f_1 \rangle, \langle n_2, f_2 \rangle \ldots \] \) do
5:       APPEND\( (P, \langle a, f \rangle) \)
6:     \text{SORT}(P)
7:     EMIT(term \( t \), postings \( P \))
Positional Indexes

Map

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Doc 2</th>
<th>Doc 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>one fish, two fish</td>
<td>red fish, blue fish</td>
<td>cat in the hat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>one</th>
<th>red</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>[1]</td>
<td>[1]</td>
<td>[1]</td>
</tr>
<tr>
<td>two</td>
<td>blue</td>
<td>hat</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>[3]</td>
<td>[3]</td>
<td>[2]</td>
</tr>
<tr>
<td>fish</td>
<td>fish</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>[2,4]</td>
<td>[2,4]</td>
<td></td>
</tr>
</tbody>
</table>

Reduce

Shuffle and Sort: aggregate values by keys

<table>
<thead>
<tr>
<th>cat</th>
<th>blue</th>
<th>hat</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>[1]</td>
<td>[3]</td>
<td>[2]</td>
</tr>
<tr>
<td>fish</td>
<td>fish</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>[2,4]</td>
<td>[2,4]</td>
<td></td>
</tr>
<tr>
<td>one</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>[1]</td>
</tr>
<tr>
<td>red</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>[1]</td>
</tr>
</tbody>
</table>
Inverted Indexing: Pseudo-Code

```java
1: class MAPPER
2:     procedure MAP(docid n, doc d)
3:         H ← new ASSOCIATIVEARRAY
4:         for all term t ∈ doc d do
5:             H{t} ← H{t} + 1
6:         for all term t ∈ H do
7:             EMIT(term t, posting ⟨n, H{t}⟩)

1: class REDUCER
2:     procedure REDUCE(term t, postings [⟨n₁, f₁⟩, ⟨n₂, f₂⟩ ...])
3:         P ← new LIST
4:         for all posting ⟨a, f⟩ ∈ postings [⟨n₁, f₁⟩, ⟨n₂, f₂⟩ ...] do
5:             APPEND(P, ⟨a, f⟩)
6:     SORT(P)
7:     EMIT(term t, postings P)
```

What’s the problem?
Scalability Bottleneck

• Initial implementation: terms as keys, postings as values
  – Reducers must buffer all postings associated with key (to sort)
  – What if we run out of memory to buffer postings?
• Uh oh!
Another Try…

How is this different?

• Let the framework do the sorting
• Term frequency implicitly stored
• Directly write postings to disk!

Where have we seen this before?
Inverted Indexing: Pseudo-Code

1: class Mapper
2:   method MAP(docid n, doc d)
3:     $H \leftarrow$ new AssociativeArray
4:     for all term $t \in$ doc $d$ do
5:         $H\{t\} \leftarrow H\{t\} + 1$
6:     for all term $t \in H$ do
7:         EMIT(tuple $\langle t, n\rangle$, tf $H\{t\}$)

1: class Reducer
2:   method INITIALIZE
3:     $t_{prev} \leftarrow \emptyset$
4:     $P \leftarrow$ new PostingsList
5:   method REDUCE(tuple $\langle t, n\rangle$, tf $[f]$)
6:     if $t \neq t_{prev}$ or $t_{prev} \neq \emptyset$ then
7:         EMIT(term $t_{prev}$, postings $P$
8:         $P$.Reset()
9:         $P$.Add($\langle n, f\rangle$)
10:        $t_{prev} \leftarrow t$
11:   method CLOSE
12:     EMIT(term $t$, postings $P$)
Retrieval with MapReduce?

- MapReduce is fundamentally batch-oriented
  - Optimized for throughput, not latency
  - Startup of mappers and reducers is expensive
- MapReduce is not suitable for *real-time* queries!
  - Use separate infrastructure for retrieval...
Important Ideas

• Partitioning (for scalability)
• Replication (for redundancy)
• Caching (for speed)
• Routing (for load balancing)

The rest is just details!
Term vs. Document Partitioning

Term Partitioning

Document Partitioning

D

T

D

T

T_1

T_2

T_3

D_1

D_2

D_3

...
Distributed Lucene

http://katta.sourceforge.net/
Take Home Messages

• Introduction to Information Retrieval
• Basics of indexing and retrieval
• Inverted indexing in MapReduce
• Retrieval at scale