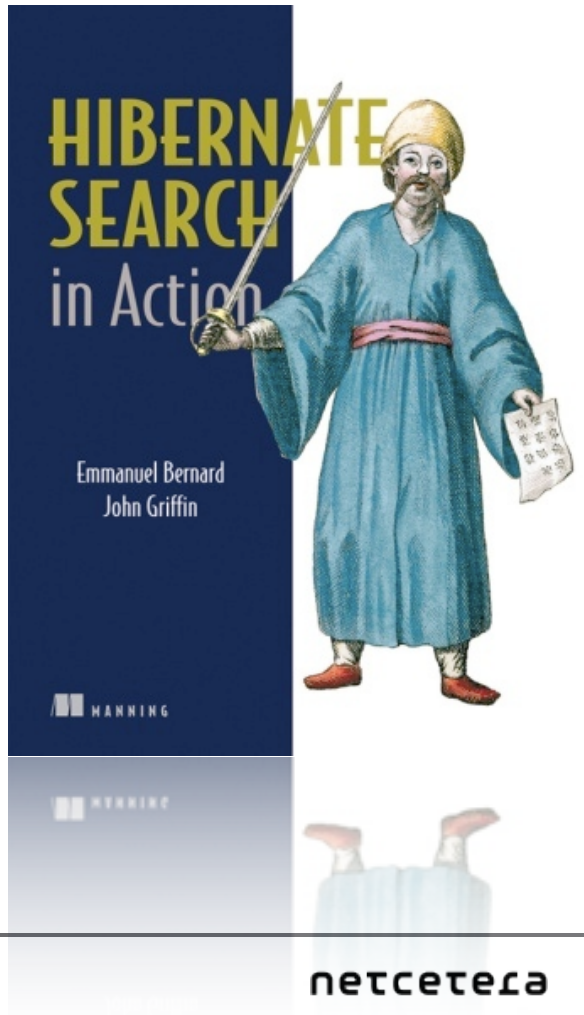
Hibernate Search

Finding data: you deserve better

Emmanuel Bernard
JBoss by Red Hat

Emmanuel Bernard

- Hibernate Search in Action
- blog.emmanuelbernard.com
- twitter.com/emmanuelbernard



Understand what full-text search does for you

Understand the magic sauce: analyzers

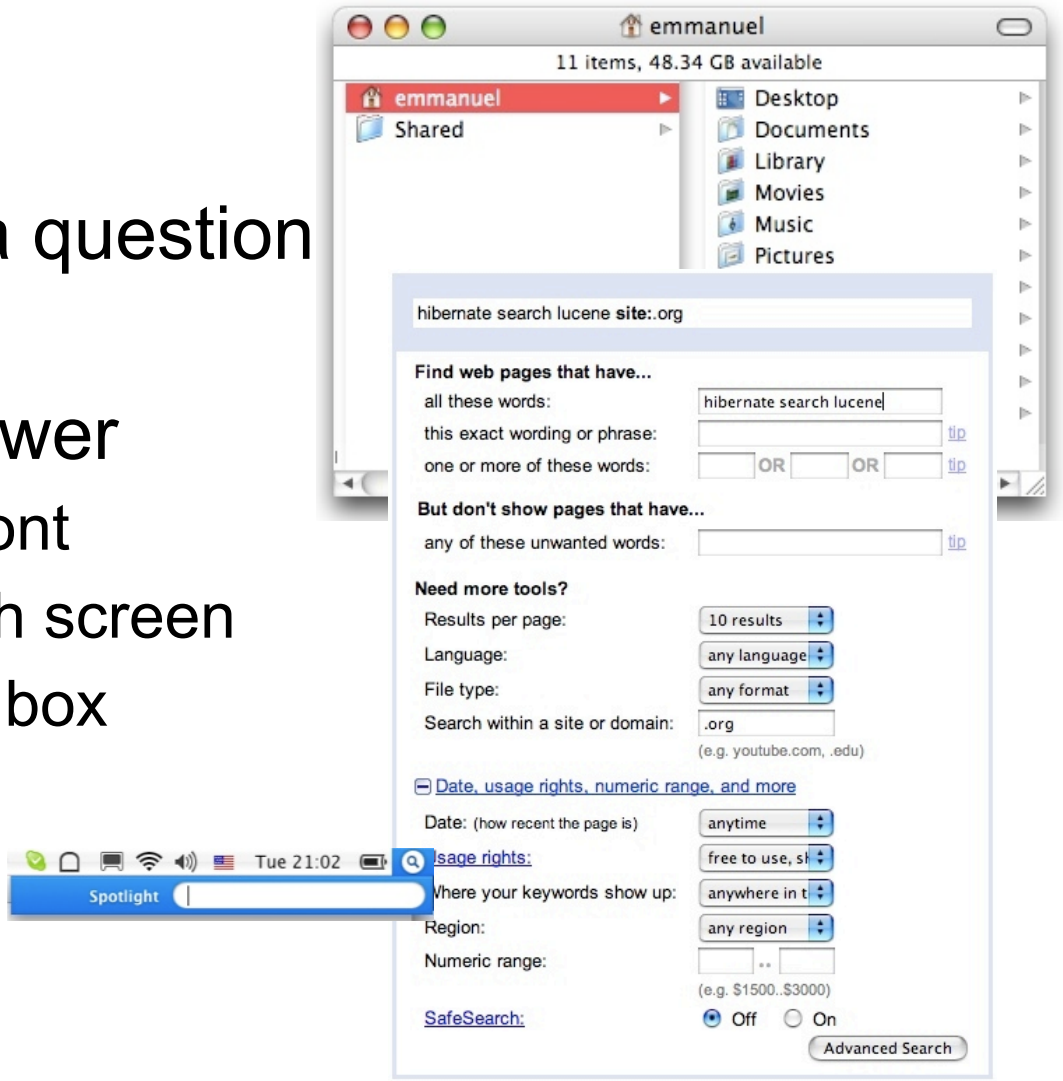
Full-text search and applications: how does it fit?

Bring the *Wow!* effect to existing applications

What is searching?

> Searching is asking a question

> Different ways to answer
  - Categorize data up-front
  - Offer a detailed search screen
  - Offer a simple search box



Human search in a relational world

- where?
  - which columns, which tables
- column != word
  - wildcard queries?
- did you say “car” or “vehicle”?
- cympausium or simposyum?
- Order results by relevance

- How to do that in SQL?

Full Text Search

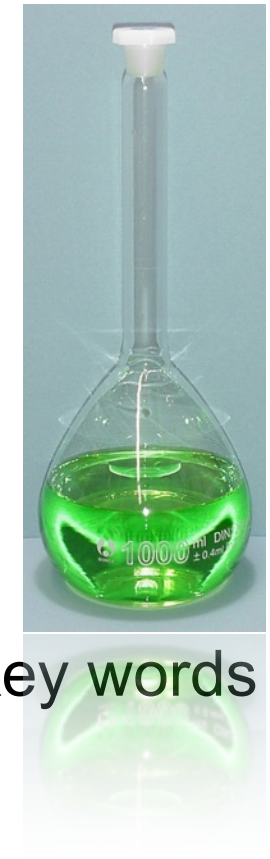
- Search by word
- Dedicated index
  - inverted indices (word frequency, position)
- Very efficient

- Full text products:
  - embedded in the database engine
  - black box / appliance
  - library embeddable like Lucene

Some of the interesting problems

- bring the “best” document first
- recover from typos
- recover from faulty orthography
- find from words with the same meaning
- find words from the same family
- find an exact phrase
- find similar documents

Find by relevance



- Best results first
  - very human sensitive
- Prioritize some fields over others
- The more matches, the better
  - for a given key word per document
  - for a given document the amount of matching key words
- Similarity algorithm

Extracting the quintessence

- Word: Atomic information

- Analyzer
  - Chunk / tokenize the text into individual words
  - Apply filters

    remove common words

    lower case
- One tokenizer
- Some filters

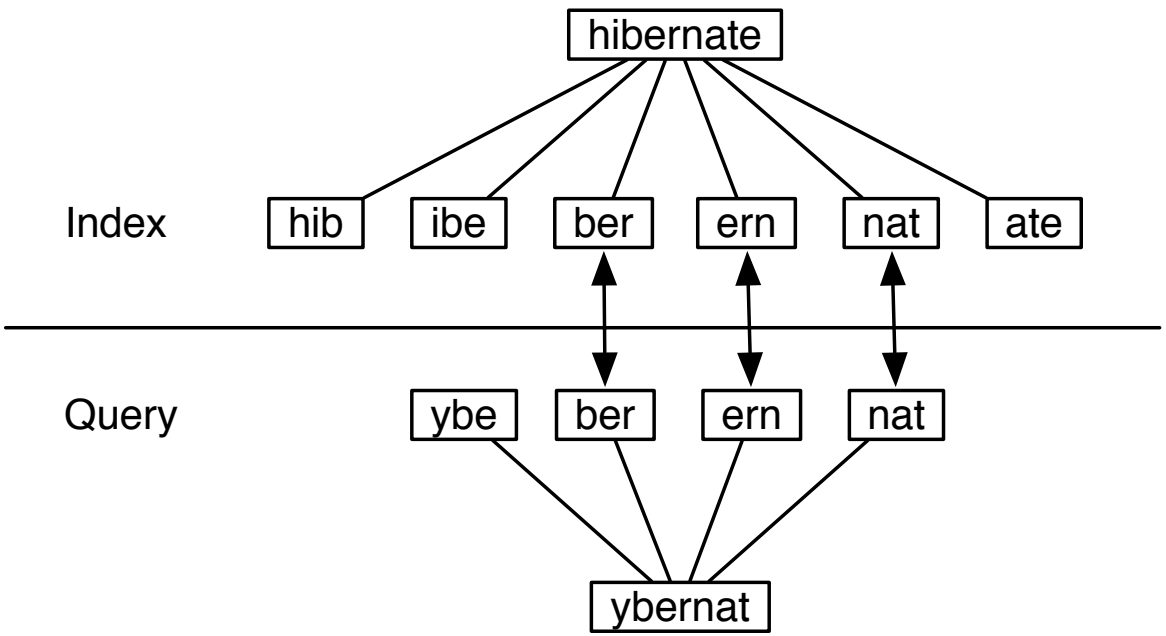
Approximation

- Recover from typos and other approximations
- Fuzzy search
  - query time operation
  - Levenshtein distance (edit distance)

```
Hibernate

Hibrenate
```

- n-gram
  - cut the word in parts of n characters
  - index each piece



- Indexing + quer
  - use a TokenFilt

Demo

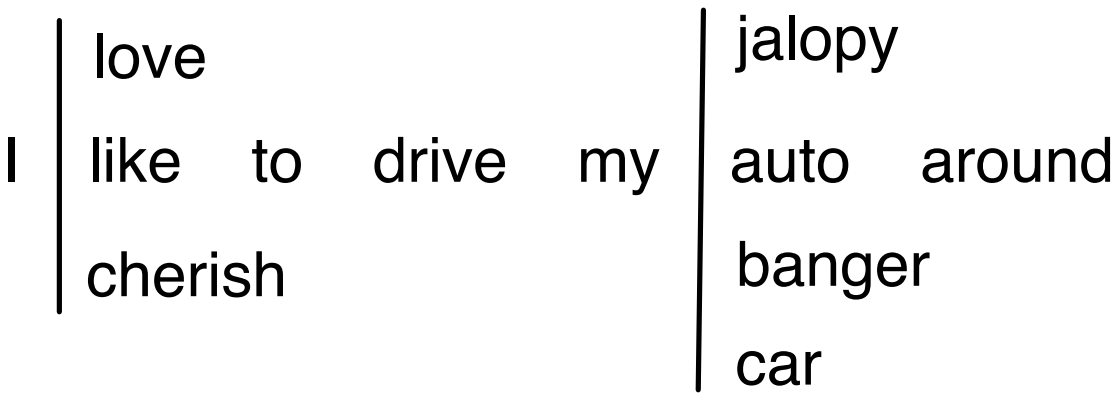
Phonetic search

- Is it “jiroscop” or “gyroscope”
  - not so useful in daily life
- Several phonetic algorithms
  - Soundex
  - Metaphone (JRSKP)
  - mostly for latin languages
- index the phonetic equivalent of a word

- Indexing + query time strategy
  - use a TokenFilter

Synonyms

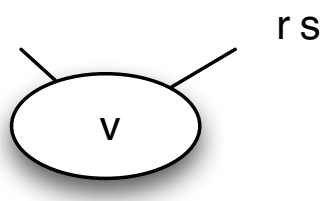
- Based on a synonym dictionary
- index all synonyms of a word in the index

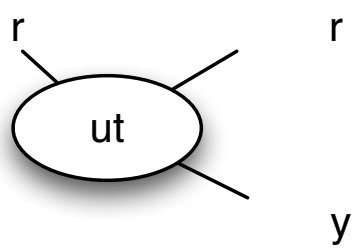


- Indexing time strategy
  - use a TokenFilter

Synonyms

- Based on a synonym dictionary
- index a reference word in the index





t r v y ut r u

v t r v y r r u

r s t r v y r r u

v t r v y ut r u

Words from the same family

- compute, computed, computerized
- Brutal force
  - index all variations of a word
- Stemming
  - Porter algorithm for English
  - Snowball Stemmer for most Indo-European languages

- Indexing + query time strategy
  - use a TokenFilter

Demo

What's the catch

- Lucene is quite low level
- Integration into an application model
- Index synchronization
- Object model conversion
- Programmatic mismatch

Integration into a Java SE / EE app

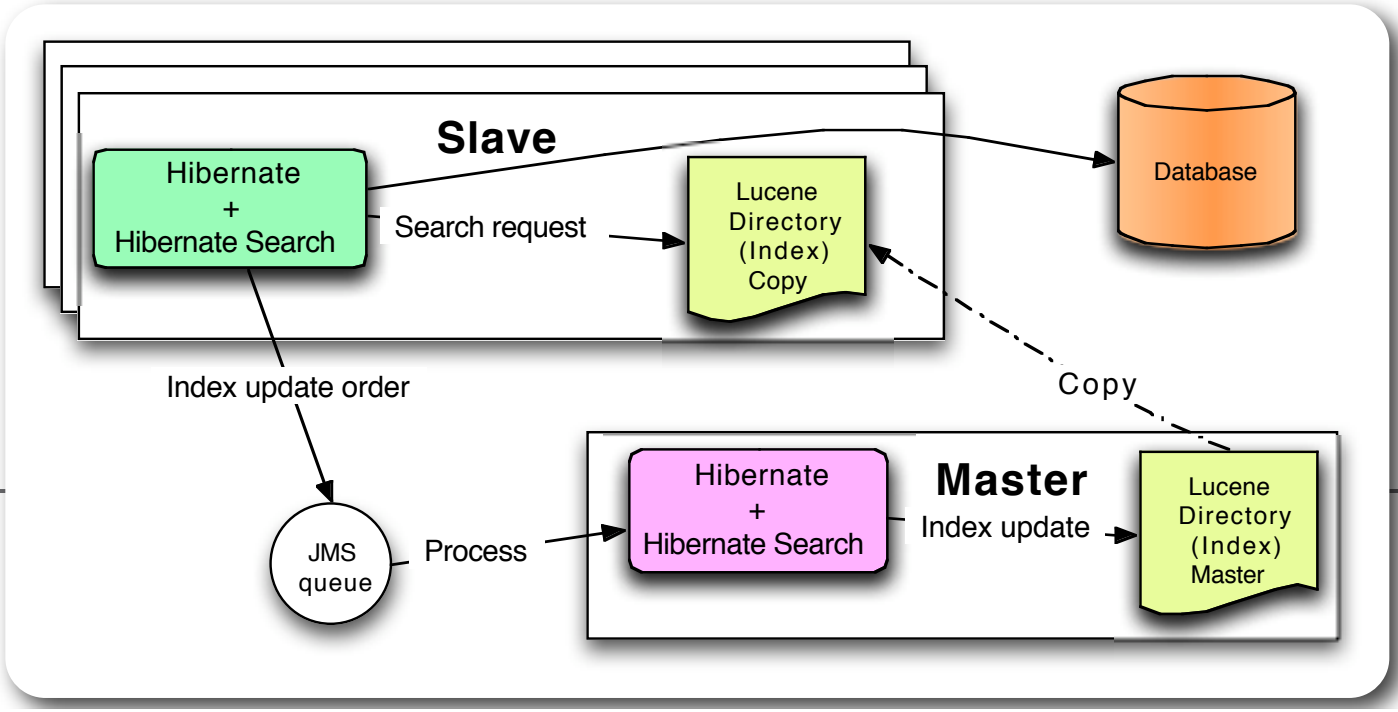
- Hibernate Search bridges
  - Hibernate Core and Java Persistence
  - Apache Lucene
- Transparent index synchronization
  - event based
- Metadata driven conversion
  - annotation based
- Unified programmatic model
  - API
  - semantic

More on Hibernate Search

- Asynchronous clustering
- Projection
- Filters
- Index sharding
- Custom DirectoryProvider (eg. JBoss Cache based)
- JBoss Cache is full text searchable
- Native Lucene access

Asynchronous cluster

- Search local / change sent to master
- Asynchronous indexing (delay)
- No front end extra cost / good scalability



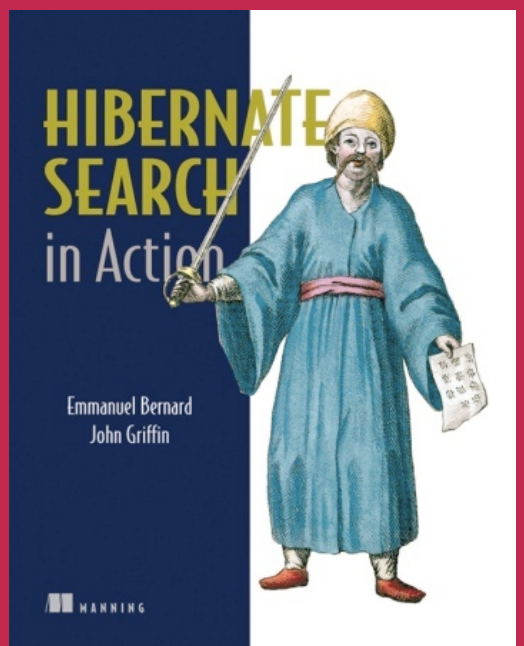
Summary

- Search for humans
- Full text tackles those problems
  - relevance
  - (human) fault tolerance
  - stemming and synonyms
  - incremental search
- Barrier of entry has lowered: Go for it!
  - POJO based approach
  - infrastructural code tackled by frameworks
  - unified programmatic model

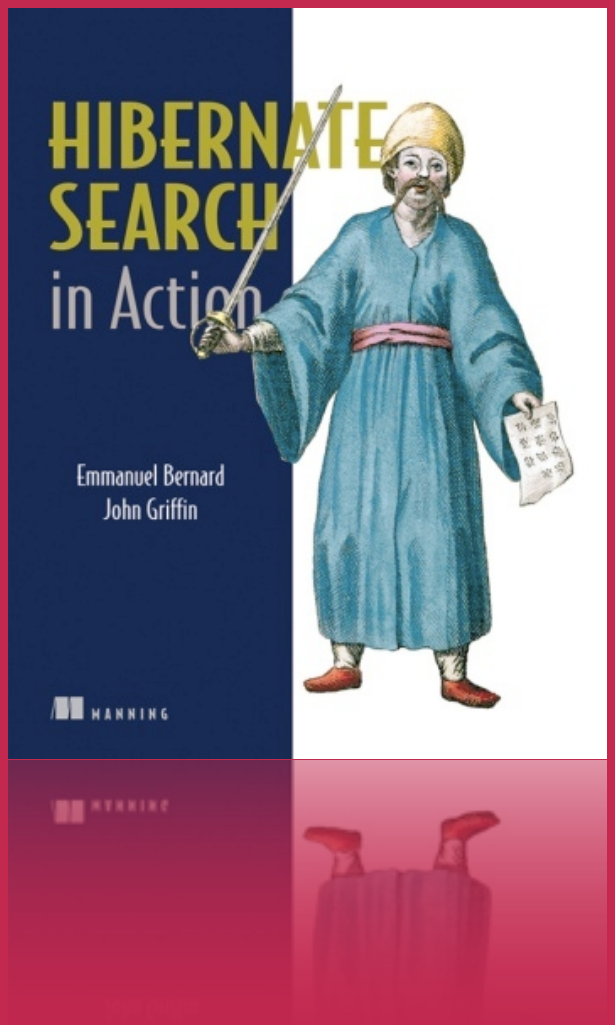
For More Information

- http://search.hibernate.org
- http://lucene.apache.org
- Hibernate Search in Action
  - Manning

- http://in.relation.to
- http://blog.emmanuelbernard.com



Questions


HIBERNATE
SEARCH
in Action
Emmanuel Bernard
John Griffin
MANNING

Emmanuel Bernard
JBoss by Red Hat

http://hibernate.org
emmanuel@hibernate.org

JAZOON09
THE INTERNATIONAL CONFERENCE ON JAVA TECHNOLOGY
JUNE 22–25, 2009 ZURICH

JBoss by Red Hat

netcetera

Sun microsystems