A successful search, a happy user: Make it happen!

Emmanuel Bernard Doer
JBoss, a Division of Red Hat

http://in.relation.to
http://twitter.com/emmanuelbernard
Overall Presentation Goal

Understand full-text search
Knowing how to index a domain model
Understand analyzers: the magic sauce
Speaker’s qualifications

- Lead developer of Hibernate Search
- Hibernate team member
- JCP (JPA 2, Bean Validation JSR-303)

"Hibernate Search in Action from Manning"

http://blog.emmanuelbernard.com

http://twitter.com/emmanuelbernard
Be better than Google Search
Human search: word

- Where?
- Which columns, which tables
- Column != word
- Wildcard queries?
- Did you say “Devoxx” or “Javapolis”?
- Devox or Daivoxx?
- 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

- Hibernate Search, full-text search your domain model
- For Hibernate based apps. No programmatic mismatch
- Index synchronization
- Object model conversion
Interesting problems

- recover from typos
- recover from faulty orthography
- find from words with the same meaning
- find words from the same family
Extracting the quintessence

- Word: Atomic information
- Analyzer
  - Chunk / tokenize the text into individual words
- Apply filters
  - remove common words
- lower case
Approximation

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

Diagram:

```
Index: hib ibe ber ern nat ate
Query: ybe ber ern nat
```

Diagram shows hibernate as an example word, with its n-grams (hib, ibe, ber, ern, nat, ate) indexed and the corresponding query n-grams (ybe, ber, ern, nat) highlighted.
n-gram 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
Synonyms

Based on a synonym dictionary

index the reference synonym of a word in the index

I like to drive my auto around
I love to drive my banger around
I cherish to drive my car around
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
DEMO

Stem your data
Summary

- Full-text search, search for humans
  - relevance
  - (human) fault tolerance
  - stemming and synonyms
  - many more

- Hibernate Search
  - full-text search for Hibernate applications
  - infrastructural code tackled by framework
  - unified programmatic model
Thanks for your attention!

- http://search.hibernate.org
- Hibernate Search in Action - Manning
- http://in.relation.to
- http://blog.emmanuelbernard.com