Snippet Generation for WikiSearch

Felix Geller,
Robert Krahn,
Alexander Krasnogolowy

July 21, 2009
Architecture

Documents

Pre-Processor

Meta Information

Sentences

Snippet Engine

Query Engine

Ranking Engine

User Interface
**Hoover, Alabama**

The city is divided into two battalions. There are eight *engine* companies, three quints, three ALS rescue/ambulances, and two battalion chiefs [4.08]... All *engine* companies are staffed with a minimum of three, with at least two being firefighter/paramedics [4.0]... BE&K Engineering – 302 [3.88]...

<table>
<thead>
<tr>
<th>Official Name:</th>
<th>Hoover, Alabama</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settlement Type:</td>
<td>City</td>
</tr>
</tbody>
</table>

Suckheed Vega, a single *engine* plane that carried four passengers, another system, as well as competitors flights, to assure tight quality adopted to Continentals network [4.08]... Continental Airlines, Inc.

Firefighters. The CFD utilizes three *Engines*, a Ladder truck, a Heavy truck in Bergen County, New Jersey, United States [2.5]... As of the election was 5,917 [2.0]...

Bathyscaphe Trieste, oceanographic ship Mizar and other ships, 8,400 feet (2560 m) below the surface, in six the sail, sonar dome, bow section, *engineering* spaces section, [4.08]... In later life, McCool was sure he would have delayed
Features

- WikiMedia markup parser
- Sentence normalization separated from query processing
- Flexible ranking algorithm [Turpin et al., 2007]
- Account for WikiMedia markup specific features
- 42 unit tests
- LOC: 575 (parser) + 392 (front-end) = 967
  (cat *java | grep \; | grep -v // | wc -l)
Lessons learned

- Wikimedia markup is not *that* simple
- Best to process it in multiple passes
- Sentence normalization should be handled earlier in search engine architecture
- Unicode is fun
if (isHeading)
    rank += HEADING_WEIGHT;
else
    rank += (1.0F / sentence.getInParagraphPosition());

if (isInFirstParagraph)
    rank += FIRST_PARAGRAPH_WEIGHT;

rank += queryTermsCount * REPEATED_QT_WEIGHT;
rank += distinctQueryTermsCount * DISTINCT_QT_WEIGHT;
rank += highestConsecutiveCount * CONSECUTIVE_QT_WEIGHT;
Sentence: "Sun Microsystems has released the Niagara and Niagara 2 chips, both of which feature an eight-core design."

```java
static final float HEADING_WEIGHT = 0.5F; // added
static final float FIRST_PARAGRAPH_WEIGHT = 1.5F; // higher than heading!
static final float REPEATED_QUERY_TERMS_WEIGHT = 1.25F; // multiplied
static final float DISTINCT_QUERY_TERMS_WEIGHT = 1.5F; // multiplied
static final float CONSECUTIVE_QUERY_TERMS_WEIGHT = 1.0F; // multiplied
static final float RANK_THRESHOLD = 1.5F + (1.0F * DISTINCT_QUERY_TERMS_WEIGHT);
static final int SENTENCES_COUNT_THRESHOLD = 3;

float rank = 0.0F;
if (isHeading)
    rank += HEADING_WEIGHT;
else
    rank += (1.0F / sentence.getInParagraphPosition());
if (isInFirstParagraph)
    rank += FIRST_PARAGRAPH_WEIGHT;
// rank = 1
rank += matchedQueryTermsCount * REPEATED_QUERY_TERMS_WEIGHT;
// rank = 1 + 2*1.25 = 3.5
rank += matchedDistinctQueryTermsCount * DISTINCT_QUERY_TERMS_WEIGHT;
// rank = 3.5 + 2*1.5 = 6.5
rank += highestConsecutiveCount * CONSECUTIVE_QUERY_TERMS_WEIGHT;
// rank = 6.5 + 2*1 = 8.5
```
while (!atEnd()) {
    step();
    if (parseHeading()) continue;
    if (parseProcessingAction()) continue;
    if (parseInfobox()) continue;
    if (parseTemplate()) continue;
    if (parseTag()) continue;
    if (parseEscapedTag()) continue;
    if (parseFileLink()) continue;
    if (parseEmbeddedLink()) continue;
    if (parseLink()) continue;
    if (parseList()) continue;
    if (parseDots()) continue;
    if (parseParagraphEnd()) continue;
    if (parseSentenceEnd(WikiPart.SENTENCE)) continue;
    if (parseUnimportantChars()) continue;
    if (parseSentenceChar()) continue;
    throw new Error("Not recognized input");
}