Fall 2012 CSCI-590 (30162)
Directed Research under Prof. Dennis McLeod
Friday, Nov 30, 2012

Jacob Kalakal Joseph
USC ID: 3827-3286-08
jacobkaj@usc.edu

Context-Sensitive Lucene

(Final Report)

Introduction

This project prototypes how to implement context-sensitive search using an open source search engine (Apache Lucene) and an ontology based web-service (Altervista Thesaurus).

Apache Lucene

Apache Lucene libraries include indexing and searching APIs. The 3.0.1 version also comes with a demo web-search application. I leveraged this demo app's front end to construct the Context-Sensitive-Lucene. I added a text-box where the user could mention a context. I also added a related-terms word-cloud feature where the user could view the terms that were used to refine the user's query according to the desired context.

Altervista Thesaurus

Altervista Thesaurus provides a web-service which returns a JSON list of semantically related terms for a given word. I wrote a Java-Servlet that retrieves the related words and recursively queries the Altervista web-service for related words of related words till a desired depth of recursion. This in effect works as my virtual ontology graph of semantically related words and phrases which can be used for directing the search results in the direction the user decides to.

Fall 2012 CSCI-590 (30162)
Directed Research under Prof. Dennis McLeod
Friday, Nov 30, 2012

Jacob Kalakal Joseph
USC ID: 3827-3286-08
jacobkaj@usc.edu

Experiments and Results

I downloaded and indexed ~18 GB of Wikipedia HTML static dump. Next, I connected this web-index to my customized-Lucene project. Finally, I ran search queries for various types, especially the following:

1. Context Sensitive Search

The user specifies a query term and a context term.

Java Query	Software Context

When the user clicks on the search button, the application queries the Altervista web-service using the context term. The web-service returns a JSON object which is parsed by the application.

software	program	code

For each of the related terms the application queries the web-service and parses the resultant JSON object. Finally, the application aggregates all the terms, and assigns weights according to the term-frequency.

software (5)	program (4)	code (3)	computer software (2)	software package (2)	application (2)	coding system (1)

Once the list of related terms and the associated weights is ready, the application appends the query term to this string to form the appended query term.

```
(Java)^1.0
(software)^0.5 (program)^0.4 (code)^0.3 (computer software)^0.2
(software package)^0.2 (application)^0.2 (coding system)^0.1
```

Now, the Lucene searcher is invoked on this appended query term and the results section is updated. The user is also shown a word-cloud of the related terms. If the user choses to click on one of these terms, a new search is executed with the context being the selected related term.

2. Homonym Search

Words like Crane, Washington, Index, Date, and Point have multiple meanings. A homonym search can be used to display the various possible connotations in the word cloud UI.

For example: the word 'left' would result the following word-cloud list: left side, hard to left, larboard, near, nigh side, port, portside, sinister, sinistral, south, abandoned, continuing, departed, extra, forsaken, gone out, leftover, marooned, over, remaining, residual, split, staying, depart, abscond, beat it, break away, clear out, come away, cut out, decamp, defect, desert, disappear, elope, embark, emigrate, escape, exit, flee, flit, fly, forsake, give the slip, go, go away, go forth, head out, issue, migrate, move, move out, part, pull out, push off, quit, relinquish, remove oneself, retire, ride off, run along, sally, say goodbye, scram, set out, slip out, start, step down, take a hike, take leave, take off, vacate, vamoose, vanish, walk out, withdraw, leave, back out, cease, cede, desert, desist, drop, drop out, evacuate, forbear, forsake, give notice, give up, hand over, knock off, maroon, quit, refrain, relinquish, resign, stop, surrender, terminate, waive, yield, forget, allow, drop, have, lay down, leave behind, let, let be, let continue, let go, let stay, mislay, omit, permit.

3. Regular Search

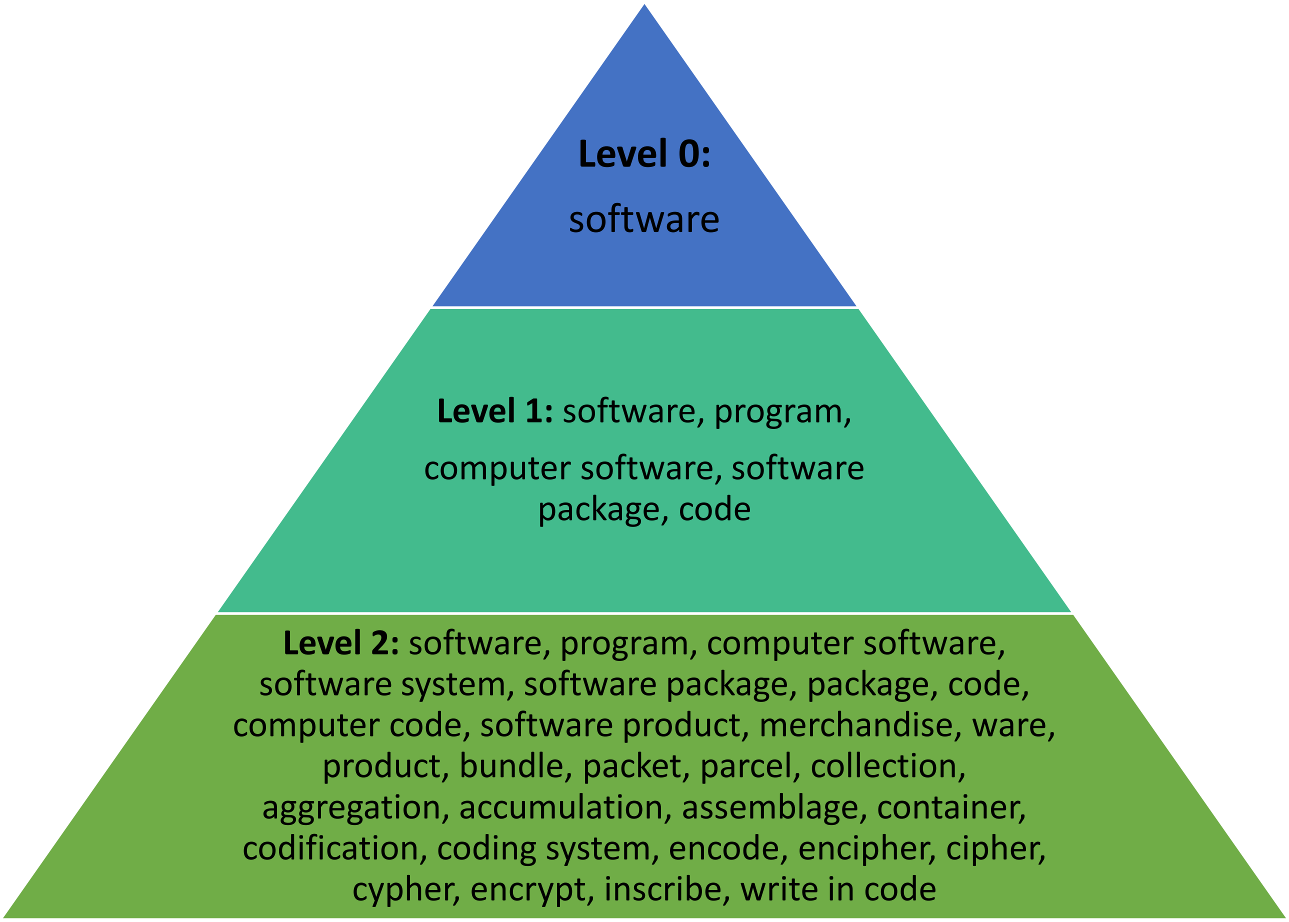
The default Lucene HTML searcher feature is also available for compatibility reasons.

Future work

This research project successfully prototypes how ontologies can be used to narrow down search results and arrive at more useful results. In addition to UI improvements here are some of the possible extensions and feature additions:

1. Custom Depth

It would be nice if the user could decide how deep a particular context term is searched. For example, for the context term 'software', the following would be the resultant related terms for the corresponding depths:



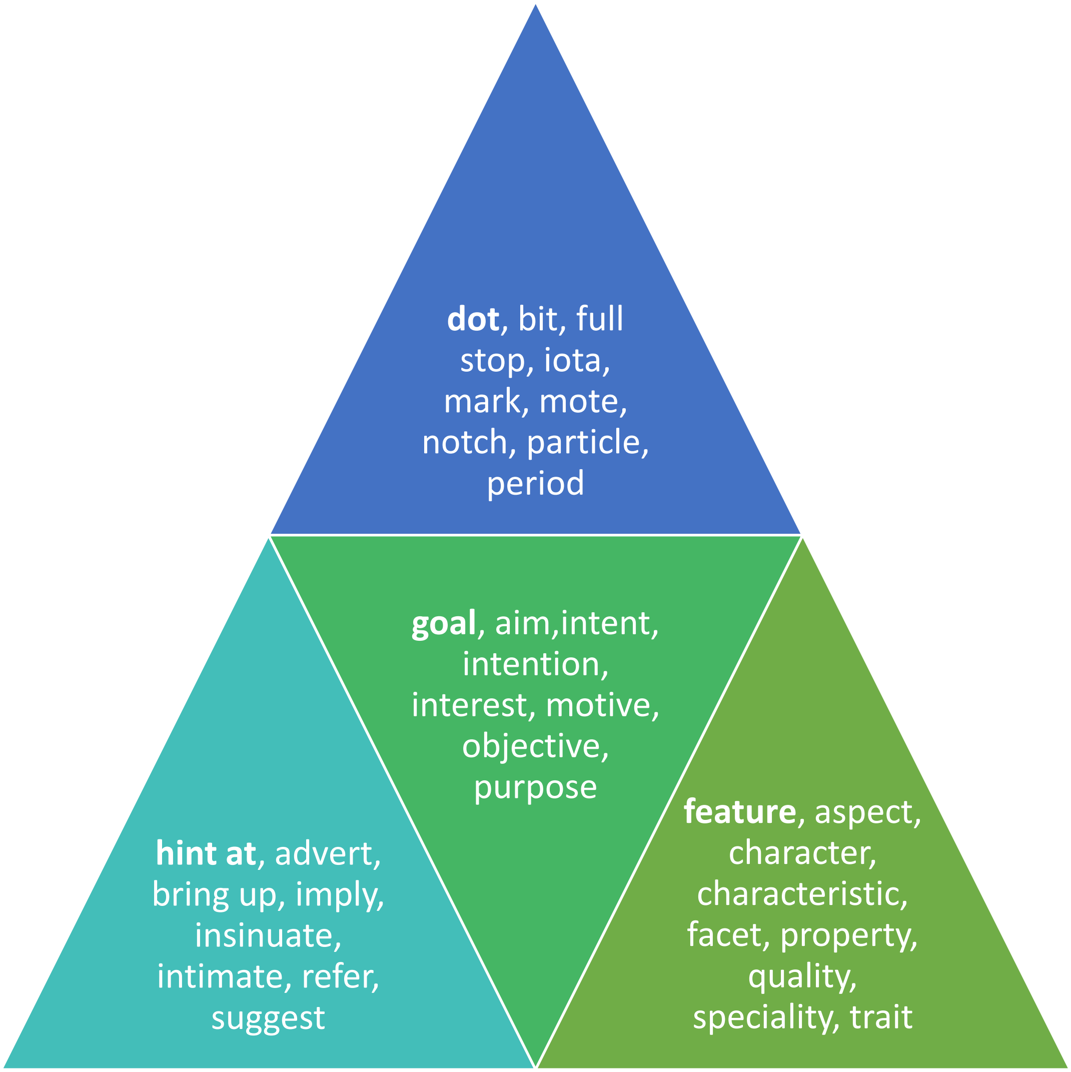
Level 0: software

Level 1: software, program, computer software, software package, code

Level 2: software, program, computer software, software system, software package, package, code, computer code, software product, merchandise, ware, product, bundle, packet, parcel, collection, aggregation, accumulation, assemblage, container, codification, coding system, encode, encipher, cipher, cypher, encrypt, inscribe, write in code

2. Clustering of related related-words

One future feature could be clusters of related "related words" instead of listing out everything in a series in the word-cloud. This is best explained using an example – 'Point' could be clustered as:



dot, bit, full stop, iota, mark, mote, notch, particle, period

goal, aim,intent, intention, interest, motive, objective, purpose

hint at, advert, bring up, imply, insinuate, intimate, refer, suggest

feature, aspect, character, characteristic, facet, property, quality, speciality, trait

Fall 2012 CSCI-590 (30162)
Directed Research under Prof. Dennis McLeod
Friday, Nov 30, 2012

Jacob Kalakal Joseph
USC ID: 3827-3286-08
jacobkaj@usc.edu

Demo Video:

1. YouTube link: http://www.youtube.com/v/qBU6eZWp8VE

References

1. Apache Lucene: http://lucene.apache.org/
2. Altervista Thesaurus: http://thesaurus.altervista.org/
3. Wikipedia HTML static dump: http://dumps.wikimedia.org/other/static_html_dumps/
4. Thesaurus: http://thesaurus.com