Xpantrac connection with IDEAL

Sloane Neidig, Samantha Johnson, David Cabrera, Erika Hoffman

CS 4624
3/6/2014
Project Specification

→ Short Description
  ◆ Integrating Xpantrac into the IDEAL software suite
  ◆ Applying Xpantrac to identify topics for IDEAL webpages

→ Primary Contact
  ◆ Seungwon Yang

→ Deliverable
  ◆ Xpantrac tailored for IDEAL
What is Xpantrac?

➔ Seungwon Yang’s dissertation topic
➔ Based on Expansion-Extraction approach
➔ Algorithm to identify topics in a given webpage
➔ Purpose: Tag topics to easily understand document
➔ Combines Cognitive Informatics with Information Retrieval
➔ Currently, only running on Seungwon’s personal data set
What is Xpantrac?
A single input document is expanded to multiple expanded XML format documents

Ultimately, search API will be replaced with Apache Solr (Different project for IDEAL)

Uses Vector Space Model approach for topic identification -
http://en.wikipedia.org/wiki/Vector_space_model (black box - no need to know to implement)
Example Input/Output

Input
Knife-wielding mob kills 27 at China train station
By CNN Staff

Output
-------- m39 10 topics --------
station, people, attach, news, train, china, xinhua, railway, group, knife
Quick python script written to get html out of .warc (web archive) files

Xpantrac currently runs on local data set (50 .txt files) and gives correct output, using Yahoo Web API

- Needs to eventually use Apache Solr since that is what IDEAL uses

Started indexing .html CNN news files to SOLR/Got familiar with Solr

- Stores information like title, first 30 words, last 30 words, etc. to easily use as input for Xpantrac
Use case

→ Librarian wants to be able to search 100 short stories for a contest by topic
→ He/She will upload the 100 stories using the Xpantrac UI
  ◆ Topic tags can currently be extracted from UI
→ Topics will then be added in the meta data of the pages on the Library website for easy searching
→ User of Library website can now look up a short story by topic name
Figure 36. The overall user interface of the Xpantrac system: (a) The Collection Pane, (b) The Document Pane, and (c) The Topics Pane.
Solr is a popular enterprise search platform that allows you to index and search documents.

Allows for an easy creation of a search engine for websites, databases, and user provided files.

Offers features such as:
- Full Text Searching
- Rich Document Parsing and Indexing (such as Word, PDF, etc)
set theCounter to 0
tell application "Terminal"
  (* The first window starts up the server *)
tell window 1
    set currentTab to do script ("cd Desktop/solr-4.6.1/example")
    do script ("java -jar start.jar") in currentTab
  end tell
delay 10
  (* The second window uploads the html files *)
tell window 2
  set currentTab to do script ("cd Desktop/html")
  repeat 51 times
    delay 5
    do script ("curl \"http://localhost:8983/solr/update/extract? literal.id=doc" & theCounter & "&commit=true\" -F \"myfile" & theCounter & "=@" & theCounter & ".html \"") in currentTab
    set theCounter to theCounter + 1
  end repeat
end tell
A SOLR API query to retrieve a text document in JSON format

- Can also use the web interface
Querying Results

```
{
    "responseHeader":{
        "status":0,
        "QTime":2,
        "params":{
            "indent":"true",
            "q":"web_doc_id:ctr_1 AND web_doc_content:1",
            "wt":"json"},
    "response":{"numFound":1,"start":0,"docs":[
    {
        "id":"4603",
        "web_doc_id":"ctr_1",
        "web_title":"Haiti braces for Isaac’s deluge",
        "web_text":"Haiti braces for Isaac’s deluge

The threat of a direct hit by Isaac on Southeast Florida might be declining but the Keys remain in the firing line. Rain and gusts could affect much of the state.

Related Content

An early look: What’s open and closed as Isaac approaches

Track Tropical Storm Isaac

BY CURTIS MORGAN

MORGAN@MIAMIHerald.COM

PORT-AU-
```
How Solr Fits Into The Puzzle

A diagram showing the indexing of documents and expanded documents using SOLR to emulate a search engine API.
In order to use Xpantrac with Apache Solr, two constraints must be followed:

- It should index a massive number of documents, so that any documents from users could be expanded based on indexed information.
- It should return the most relevant portion of the matching documents, for a query.
Deadlines

➔ February

- 1st: Determine topic, Get client approval
- 15th: Discuss roles & timelines; Get instructor approval; Try out Apache Solr

➔ March

- 1st: Decompress files (with IDEALpages); Begin working with Xpantrac
- 15th: Xpantrac - Begin Batch & Interactive Modes
- 29th: Xpantrac - Continue Batch & Interactive Modes
Deadlines

→ April
   ◆ 12th: Stretch Goal: Tweets and Archive Files
   ◆ 26th: Finalize project & prototype; Final Report

→ May
   ◆ 1st – 6th: Final Presentation
   ◆ 6th: Final Paper
References

- Automatic Identification of Topic Tags from Texts Based on Expansion-Extraction Approach by Yang, Seungwon, Virginia Polytechnic and State University, 2013, 230 pages. (Seungwon’s dissertation)

- SOLR tutorial: https://lucene.apache.org/solr/tutorial.html
Questions?