IPRO 327

Spring 2009

A Geographically Searchable Local News Aggregator

Ori Rawlings (CS)
Dan Copeland (PS)
Evan Estola (CS)
Jay Mundrawala (CPE)
Daniel Price (CS+CPE)
Daniel Sirotzke (CS)
Max Kaim (CS/Psych)
Laura Rodriguez (CPE)
Steven Peterson (ITM)
Wai Gen Yee (Faculty)
Ophir Frieder (Faculty)

I. Abstract

IPRO 327 aims to solve a problem that has long plagued search engines: the system does not actually know or understand what the user is searching for. It simply interprets the user's query as a string of letters and/or numbers, then attempts to match that string with Web pages it has indexed. Our team attempts to solve this problem by creating a search engine that "knows" what the user is searching for, and then returns requested results.

II. Background

A. Our project is faculty initiated with the goal of exploring the impact of emerging semantic technologies on search engines. The expected user base for this system includes marketers, intelligence agents and casual users.

B. Search engines today look at keywords that the user has chosen and compare them to documents that the they have gathered over time. Search engines, however, have very limited ability to discern the deeper meaning of what the user is searching for. Subtle variations in meaning are lost to it. Our project attempts to overcome these problems by using semantic Web technology to ensure quality results while Web searching.

C. Our project takes advantage of emerging semantic Web technology. This technology associates the terms in documents with their real-world meaning, instead of treating them as merely strings of letters and numbers. This allows software to "understand" the meaning that a user is attempting to convey in a query and then match this query with the appropriate documents. Semantic Web technology has, so far, seen little implementation in the modern world. Although modern search engines implement some semantic features (e.g., recognizing a query as an address), few fully semantic search engines exist, and those that do are not well known or used often.

D. Last semester's IPRO built a basic prototype search engine, whose functionality and design we extended. It was only able to search canned data (as opposed to live news feeds), and it had limited support for semantic data. It did, however, support date-based and geography-based search: it returned articles with times and locations specified by the user.

E. One ethical issue that arises is hosting news articles that we crawl on our Web site. These articles are copyrighted and may have restrictions on how they may be used. Since this is currently a research project for an educational institution, we see that we are using the data under the "fair use" conventions (Teska, 2008). If this were to become a commercial news search engine, we would need purchase the rights to host the news articles.

III. Objectives

A. According to our project plan, our team had a variety of objectives at the beginning of the semester. These were as follows:

i. Technical Objectives

1. Support for a greater number of live news feeds
2. More comprehensive entity search functionality for Swyne Project

3. Better organized, more user-friendly interface

ii. Broader Impact Objectives

1. Better accountability for our continuity
2. Stricter use of document sharing & record keeping
3. More clearly defined tracking of progress

iii. Soft Skill Objectives – Improvement of the following:

1. Public communication
2. Interpersonal relations
3. Effective presentations
4. Management of personnel
5. Time management

IV. Methodology

A. Our group decided that splitting the task into 3 parts would be the most effective way to tackle our problem. Consequently, the decision was made to split the group into 3 teams: **crawler**, which would focus on getting news articles from the Web (Technical Objective 1); **server**, which would focus on storing and analyzing those articles (Technical Objective 2); and **usability**, which would ensure that the Web site was as user-friendly as possible (Technical Objective 3). This reflects no change from the initial Project Plan.

B. Approach: We split the Swyne Project members into 3 teams:

i. Crawler team: 'Listen' to various news Web feeds. As articles are published on the Internet, our crawler downloads them and extracts the article body from the Web page (i.e., removes extraneous information, such as ads). The article is then handed off to the Server Indexer.

ii. Server team: The Server Indexer extracts entities out of news article text and stores them in a database. The server sub-team also provides an interface to access the stored data for the usability team.

iii. Usability team: Create a front end for accessing stored news data. The usability sub-team ensures the user-friendliness of the Web site and online interface.

V. Team Structure and Assignments

A. The team structure has remained unchanged from the Project Plan.

B. Teams had the following structures:

i. Crawler team:

1. Ori Rawlings & Dan Copeland:
    a. Programmed a "crawler" system to listen to Web feeds from various news sites, which downloads article web pages and extracts the article text from the page. The system then sends the article text along with article title, publication date, and source to the Swyne server indexer. Also built was a heuristically based method for extracting article text from web pages based on the Text-to-Tag Ratio paper (Weninger and Hsu, 2008). RSS news feeds from 21 different newspapers across the country were examined by the team, representing a huge area of coverage.

ii. Server team:

1. Jay Mundrawala:
    a. Team leader. Created an entity indexer with functionality that allows the disambiguation of entities that share the same names from (Milne, Witten and Nichols, 2007).
2. Evan Estola:
    a. Worked with the user interface team to make sure the server team's functionality could be integrated into the user interface.
    b. Built an entity browser for Swyne.
    c. Ranked search results.
3. Dan Sirotzke:
    a. Implemented the server interface that allows server functionality to be easily accessed by the user interface team.
4. Dan Price:

a. Researched spatial indexing and tested effectiveness of using a spatial index.
b. Created Geocoder, which identifies a location with latitude-longitude coordinates.

iii. Usability team:

1. Laura Rodriguez
a. Coordinated efforts among the sub-team.
b. Communicated with the other sub-teams to request functionality.
c. Worked collaboratively on deciding team goals, motivation and implementation.
d. Designed and administered various surveys.
e. Recorded and analyzed feedback from surveys and assigned tasks accordingly.
f. Troubleshot coding problems.
g. Came up with different layout designs for the text boxes.

2. Max Kaim:
a. Compiled an initial list of team member’s reactions toward the Web site.
b. Worked collaboratively on deciding team goals, motivation and implementation.
c. Implemented auto-update for the radius of the circle in the map.
d. Cleaned up code.

3. Steven Peterson:
a. Created a Wiki for feedback and e-mailed IPRO members regarding ideas for the website implementation.
b. Worked collaboratively on deciding team goals, motivation and implementation.
c. Fixed initial issues with the front page.
d. Came up with several possible headers and implemented the best.
e. Designed a new entities page.
f. Came up with different color schemes and sought feedback from the class.

All teams were coordinated by an overall Project Leader, Evan Estola. The Project Leader managed communications among teams (via Team Leaders) and ensured that subgoals were being met. He also mediated communications between the teams and the faculty advisors.

VI. Budget

A. The only expenses incurred were those stated in the Project Plan. These came from the Server team, which would amounted to one hard drive and associated cabling. This equipment cost approximately $160.

VII. Code of Ethics

A. Our overarching standard for the Code of Ethics is mutual respect. Team members will treat each other with the utmost respect at all times. Team leaders will utilize the five principles of effectively fostering mutual respect: expectations, skills, feedback, consequences, and growth. Team leaders will first lay down their expectations to their teams. The team leaders will then provide their teams with the skills necessary to complete their tasks. Team leaders will provide regular feedback to team members on their performance. Team leaders will also distribute consequences, whether positive or negative, in accordance with the performance of their team members. Finally, the team itself and its members will grow from the experience and achieve a higher state of performance.

B. Our seven canons shall be:
   - i. Team members must show up for class sessions.
     1. Pressure: Not showing up for class.
     2. Risk: Team suffers from lack of input.
     3. Risk: Negative impact on student's grade.
   - ii. Team members must anticipate deadlines.
     1. Pressure: Not being proactive with respect to deadlines.
     2. Risk: Deadlines can surprise team member.
     3. Risk: Workload becomes backed up.
   - iii. Team members must complete assigned portions of task on time.
     1. Pressure: Not completing work.
     2. Risk: Work does not get completed on time.
     3. Risk: Holds up entire team's progress.
   - iv. Team members shall communicate freely and often with other team members.
     1. Pressure: Not being communicative.
     2. Risk: Lack of sharing of information.
     3. Risk: Team's progress is impeded.
   - v. Team members will contact team in advance if deadline cannot be met.
     1. Pressure: Not sharing this info with team.
     2. Risk: Work does not get completed on time.
     3. Risk: Team does not know that work was not completed on time.
   - vi. Team members shall provide input into any/all aspects of project whenever beneficial.
     1. Pressure: Taking a passive role with respect to work.
     2. Risk: Work will not be as high-quality as it could be.
     3. Risk: Team member will be an overall lesser contributor.

vii. Team members shall conduct themselves in a proactive manner, anticipating problems and taking preventative measures to ensure they do not affect the project.

1. Pressure: Taking a passive role in the team itself.
2. Risk: Team as a whole suffers from lack of initiative.
3. Risk: Individual member becomes disillusioned with team’s ability.

VIII. Results

A. The project was a success for the entire team. All objectives we set out to complete were accomplished on time. Results for each sub-team were as follows:

i. Crawler team: We have accomplished the goals set out at the beginning of the semester to a T. We did not encounter any unpredicted obstacles during the course of the semester.

ii. Server team: We have an entity index that stores a vast amount of information about real world entities. The speed of the system was drastically improved through the use of a spatial index. We have an entity browser that makes browsing through entities easy and intuitive.

iii. Usability team:

1. The auto-update function of the circle in the map is working properly: users can change the radius by keyboard entry and see the difference in the map.
2. The header and layout of the buttons have been improved to reflect feedback and for a more user-friendly view.
3. The color scheme was updated for a more current design.
4. The design of the entities page was completed.
5. Three surveys were designed and administered with IPRO class members, and one survey was conducted with non IPRO team members. Results were recorded and analyzed.
6. The number of results is now displayed in the results page.
7. The calendar was updated for a more user-friendly interaction.

IX. Obstacles

A. The team as a whole had to overcome several obstacles. Breaking these down by sub-team, they are:

i. Crawler team:

1. The challenges that the team encountered were: general purpose article extraction and identifying work that the different team members could attend to.

ii. Server team:

1. Not everyone was familiar with IR/Semantic technologies.
2. Sheer amount of work required coupled with a lack of time to complete that work in.

iii. Usability team:

1. The major issue that the team had to deal with was that we could not start working until later because we needed requirements from the different sub-teams. Also, the Usability team worked on some problems with communication at the beginning of the semester, but those issues were resolved in a timely fashion.
2. In terms of coding, the main issue was that none of the team members had prior experience with HTML or CSS, so we had to adopt a “learn on the go” attitude, which was key to our success.
3. Finally, another important problem that the team had was regarding outside testing. The sub-team sought help from the entire IPRO team to come up with sufficient outside feedback, but it was hard to get.

B. Resolution of obstacles:

i. Crawler team:

1. Articles were extracted from each newspaper individually. A separate crawler was necessary for each unique website. All programming duty was handled by Ori Rawlings, while the rest of the work concerning deliverables was delegated to Dan Copeland.

ii. Server team:

1. Each member of the team was polled concerning their individual skill in the area. Adequate time was then partitioned to bring all members of the team up to speed in the necessary areas. The work for the project was then divvied up according to skill level and to ensure timely completion of the project.

iii. Usability team:

1. Increased communication with the other teams helped to alleviate the first obstacle (reliance on progress of other teams). ‘Icebreaking’ activities such as organizing meets outside of IPRO class also helped to develop communication within the team.
2. Team members had to learn what was necessary to catch up, as well as learn while doing the job itself.

3. This obstacle was never quite overcome; sufficient numbers of testers were never procured.

X. Recommendations

A. Our team has several recommendations. By team, they are:

i. Crawler team:

1. Make a very serious effort towards accomplishing something useful for the project. Do not over-estimate or under-estimate how much you can accomplish. Try to predict challenges ahead of time. Schedule work so that it is due shortly after mid-term reviews. You are bound to over-shoot this deadline, but you will still finish in time to have a comfortable amount of time to prepare the final report and IPRO deliverables.

ii. Server team:

1. Better integration of entities; currently, entities get assigned URLs within the swyne.homelinux.org domain. It would be nice if pages were created for these links, or pages would be dynamically created when someone visited one of those links.

iii. Usability team:

1. The team feels that having one member proficient in HTML or CSS would have helped, at least initially, in getting things done. However, the team members were very quick in catching up with the language, so this did not turn out to be such a big drawback.
2. Another recommendation would be to get a larger outside crowd for proper usability testing. The sub-team struggled to get enough people for feedback, so maybe contacting other sources to arrange for a more structured feedback session would be useful.

XI. Resources

A. Broken down by sub-team, our resources were as follows:

i. Crawler team:

No money spent
Man hours:
Ori: 95 hours spent
Dan C.: 88 hours spent

ii. Server team:

$150 spent on new hard drive & cable

Man hours:

Jay: 101
Evan: 115
Dan P.: 102
Dan S. 105

iii. Usability team:

No money spent
Man hours:

Laura: 112.5 hours spent
Max: 83 hours spent
Steven: 123 hours spent

In addition, students used their personal computer for their design, implementation and coding and the final system was hosted on a machine in the Information Retrieval Lab.

B. Outside documents used by team:

1. Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web". *Scientific American Magazine*. http://www.sciam.com/article.cfm?id=the-semantic-web&print=true. Retrieved on 26 March 2008
2. Wiki Page on Semantic Web. http://en.wikipedia.org/wiki/Semantic_Web
3. R. Guha, Rob McCool, Eric Miller, Semantic Search. WWW, 2003. http://portal.acm.org/citation.cfm?id=775152.775250&coll=ACM&dl=ACM&CFID=1639369&CFTOKEN=35105784
4. Wiki Page on Resource Description Framework (RDF). http://en.wikipedia.org/wiki/Resource_Description_Framework
5. Wiki Page on RDF Schema. http://en.wikipedia.org/wiki/RDF_schema
6. W3 Schools tutorials about RDF and RDFS. http://www.w3schools.com/rdf/default.asp
7. Making a PowerPoint Presentation. http://radiographics.rsnajnls.org/cgi/content/full/radiographics;24/4/1177
8. HTML/CSS Crash Primer http://www.communitymx.com/content/article.cfm?page=1&cid=695E2
9. Reference: CSS Property Index http://www.blooberry.com/indexdot/css/propindex/all.htm
10. T. Weninger and W. H. Hsu. Text Extraction from the Web via Text-to-Tag Ratio. In the Proceedings of the Database and Expert Systems Application (DEXA) Conference, 2008.
11. K. Teska. What Can You (Legally) Take from the Web. IEEE Spectrum, April, 2008.

12. D. Milne, I. H. Witten and D. M. Nichols. A Knowledge-Based Search Engine Powered by Wikipedia. In the Proceedings of the ACM Conference on Information and Knowledge Management, 2007.

C. Web sources used by team:

Hyperlocal Web Sites Deliver News without Newspapers

C. C. MIller and B. Stone, NY Times, April 12, 2009

Quintura

General-Purpose Computing on a Semantic Network Substrate

Marko A. Rodriguez

Originally found at http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.3395v3.pdf.

ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web

Kemafor Anyanwu, Amit Sheth

SemRank: Ranking Complex Relationship Search Results on the Semantic Web

Kemafor Anyanwu, Angela Maduko, Amit Sheth

Supporting Complex Thematic, Spatial and Temporal Queries over Semantic Web Data

Matthew Perry , Amit Sheth , Farshad Hakimpour , Prateek Jain

Analyzing Theme, Space, and Time: An Ontology-based Approach

Farshad Hakimpour, Matthew Perry, Amit Sheth

Data Processing in Space, Time and Semantics Dimensions

Farshad Hakimpour, Boanerges Aleman-Meza, Matthew Perry, Amit Sheth

Falcons: Searching and Browsing Entities on the Semantic Web

Gong Cheng, Weiyi Ge, Yuzhong Qu ***(Southeast University)***

Semantic Web

- http://en.wikipedia.org/wiki/Semantic_Web
- http://www.altova.com/semantic_web.html
- Video tutorial - Y. Sure. A Short Tutorial on Semantic Web. Citation Info? - http://videolectures.net/training06_sure_stsw/ (WGY)

Semantic Search

- http://en.wikipedia.org/wiki/Semantic_search

RDF (Resource Description Framework)

- http://en.wikipedia.org/wiki/Resource_Description_Framework
- http://en.wikipedia.org/wiki/RDF_Schema
- http://en.wikipedia.org/wiki/SPARQL
- http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
- http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

RDF Visualizers

- http://simile.mit.edu/welkin/
- http://semweb.salzburgresearch.at/apps/rdf-gravity/

RDF Browsers

- http://simile.mit.edu/wiki/Longwell

RDF Schema

- http://www.w3.org/TR/rdf-schema/
- http://en.wikipedia.org/wiki/RDF_Schema

SPARQL (RDF Query Language)

- http://www.w3.org/TR/rdf-sparql-query/
- http://jena.sourceforge.net/ARQ/Tutorial/

OWL (Web Ontology Language)

- http://en.wikipedia.org/wiki/Web_Ontology_Language
- http://www.w3.org/TR/owl-features/

Wikipedia

- http://wikixmldb.dyndns.org/

Surveys

- Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman, An analysis of search-based user interaction on the Semantic Web, Centrum voor Wiskundeen Informatica, Information Systems, Tech Report INS-E0706 MAY 2007. http://ftp.cwi.nl/CWIreports/INS/INS-E0706.pdf - This paper describes the user interface design of several search systems. Lots of good links to systems in here. (WGY, 2/10/09)

Other

- http://www.csszengarden.com/

XII. Acknowledgements

A. Usability survey participants – We would like to thank the volunteers who helped us identify interface design issues.