A Geographically Searchable Local News Aggregator

IPRO 327
Fall, 2008
Advisors: Frieder, Yee
Goal

- Create a news search engine that understands the contents of articles
  - Allows search by “meaning” versus by “string matching”
  - Example: What happened in Chicago in 1996?
Current Practice – Google
Why News Articles?

- Contains reasonably factual content
- Articles printed regularly
- Articles have relatively standard format
Target Audience

- Researchers (e.g., reporters)
- Intelligence agents
- Real estate agents
- Marketers
- ...

Semester Goals

- Design the system
- Build a working prototype
- Lay a foundation for future work

- Focus – search based on time and location
Approach

- Consider functional units of system
- Consider IPRO goals

- Make a team for each functional unit
  - Consider strengths/interests of students
Other Tasks

- Required everyone to do research and present on related topics, such as the semantic web
  - Developed soft skills (reading, writing, presenting)
  - Learned about the problems and benefits of building such a system
Students

- Nick Bathum (CS, 4\textsuperscript{th} yr)
- Evan Estola (CS, 3\textsuperscript{rd} yr)
- Jaeyeon Kihm (CS, 4\textsuperscript{th} yr)
- Jay Mundrawala (ECE, 2\textsuperscript{nd} yr)
- Yacin Nadji (CS, 4\textsuperscript{th} yr)
- Chris Osswald (BME, 3\textsuperscript{rd} yr)
- Pete Schmitz (CS, 4\textsuperscript{th} yr)
- Cameron Zangenehzadeh (PS, 4\textsuperscript{th} yr)
Organization

- Developed Plan, Timeline
- Used a free Wiki service for discussion, file-sharing and status reporting
  - Centralized information
  - Avoid mass emails and information loss
Functional Units

- Design (User Interface)
  - Usability and presentability

- Web Application Architecture
  - Build infrastructure, pull together work of others

- Entity Database
  - Pull “machine readable” information from articles

- Web Crawling
  - Collection of news articles from the Web
Functional Unit Relationship

- User
- User Interface
- Web Application
- Database
- Crawler
- Web
Architecture Group

Evan Estola
Yacin Nadji
Architecture Group

- Responsible for building Web infrastructure
  - Interconnect other functional units
- Combine work of other groups
Architecture Approach

- Use leading open source software
  - Sesame semantic database for entities
  - MySQL database for crawled article data
  - Java Server Pages, Java Beans and Tomcat
Architecture Challenges

- Little to no Web development experience
- Working with other groups
  - Dependency on other groups
  - Communication
Architecture Outcome

- Success!
  - Working prototype
  - Dynamically builds interface (from Design team) to display article data (from Crawling team) and entities (from Entity Extraction team)
  - Learned to learn
Architecture Future Work

- Optimization (speed)
- Build entity display system
Entity Extraction
Team
Jaeyeon Kihm
Jay Mundrawala
Entity Extraction - Goals

- Given news articles, identify
  - Persons
  - Locations
  - Organizations
  - Dates
- Build database for extracted data
Dr. Wai Gen Yee is a professor in Department of Computer Science at Illinois Institute of Technology in Chicago.
Dr. Wai Gen Yee is a professor in the Department of Computer Science at the Illinois Institute of Technology in Chicago.

- Implemented leading open source software (Stanford NER API)
- Person, Location, Organization, Date, Title entities are extracted successfully
Lat/Lon Coordinate Point

- Generate latitude / longitude coordinate points from location entity.

Diagram:

Address → Google map → Lat/Lon Database
Entity Extraction - Accomplishments

Accomplishment:
- Extract entities from articles
- Build database to store entities and news articles
- Generate Lat/Lon coordinate points from address
Challenges

- Stanford NER does not recognize addresses
  - Wrote custom module
- Poor database performance
  - Added a second custom database
Entity Extraction – Future works

- Future work
  - Unification and disambiguation of entities
    - Ex: “George Bush” vs. “President Bush”
- Ranking of retrieved results
- Support for different types of entities (e.g., sports teams)
- Support for relationships among entities
Web Crawling Team

Nick Bathum
Pete Schmitz
Web Crawling Goal

- Retrieve news articles from the Web
- Submit news articles to Entity Extraction system
Web Crawling Approach

- Select correct software for crawling
  - Nutch selected for its support and scalability
- Determine which news source and how to acquire correct article data
Web Crawling Approach

- Get article links
- Use Nutch software to fetch the articles
- Parse relevant content from articles
Web Crawling Challenges

- Automated extraction of relevant content from Web pages filled with extraneous info
  - Need to identify Web page structure
Web Crawling Outcome

- Functioning article retrieval system
- Retrieve articles from
  - Reuters
  - LA Times
Web Crawling Future Work

- Support for other news sources
Design Team

Chris Osswald (BME)
Cam Zangenehzadeh (PS)
What do we want to do?

- Show off implemented functionality
- Design a simple, user-friendly interface
Approach

- Looked at other systems (EveryBlock.com, PowerSet.com, etc.)
- Identify core functionality
- Developed appropriate interface
  - Shows core function of system
Flow of Information

Homepage

Results Page

Article Page

Tutorial, IPRO 327 Website

Article Source

Navigate Workbench®

Entity Page
The Home Page

- Simple, user-friendly interface for entering query parameters
The Results Page

- Articles listed in familiar format

SwyneProject

I is munchin' all your news articles...

Results

USA: CME nearby hogs and pork bellies close limit down.
   Reuters 1996-01-20
   Chicago, CHICAGO

USA: CCC seeks 33,000 T U.S. white wheat for Bangladesh.
   Reuters 1996-01-20
   U.S., CHICAGO, Bangladesh

USA: Widespread aflatoxin found in south Texas corn.
   Reuters 1996-01-20
   CHICAGO, Texas, TX, College Station

USA: Marcus Theatres expands theater network.
   Reuters
   United States, Chicago, Lake Geneva, Detroit, Wisconsin, St. Louis, Minnesota, Colorado, Wis., Illinois, Kansas City, MILWAUKEE

USA: Corn progress slow, conditions slip in key states.
   Reuters 1996-01-20
   U.S., CHICAGO, Illinois, Indiana, Iowa, Ohio, Nebraska

USA: CCC seeks U.S. wheat for donation to Tajikistan.
USA: CME nearby hogs and pork bellies close limit down.

August and October CME hog futures and August pork belly futures all finished down their respective daily trading limits on pressure from increased hog runs and slaughter and lower cash pork and hog prices that touched off sell stops, traders said.

August and October hogs ended down the 1,500 cent per lb limit with August at 66.025 cents and October at 53.550. Other hog contracts closed mostly down 1.400 to 0.250 cent.

Declining cash prices on larger hog runs pressured hog futures from the start and sell stops were touched off around $4.600 cents in October and December.

"The fact that we are seeing a few more hogs show up has people worried about the supply side but more important than this, gives the packers a little more freedom to improve their margins," said Doug Harper, Brock Associates Inc senior livestock analyst.

Harper noted that packer buying interest will decline despite the large kill capacity the industry has because negative profit margins are starting to take a toll. Packers have been losing money the past two quarters and "they are starting to feel pressure from stockholders. They have to lower their procurement costs."

Traders cited indications for another 50.50 to $1.00 per cwt loss in cash hog Wednesday which added to weakness and knocked hog futures into technical sell stops. Futures went limit down shortly after penetrating moving averages and falling below $4,000 in October and December, traders said.

August pork bellies settled down the limit at 91.375 cents per lb and other months ended 1.250 to 0.400 lower.

Bellies also declined on concern over the rate of hog slaughter, traders said.

"Hog slaughter has continued to weigh on bellies," said John Fisher, partner in Skyline Trading and belly trader.
Exploring Entities

- Allows user to explore relationships
Accomplishments

- Created simple interface
  - Learned Web design
- Access to all implemented functionality
- Created IPRO Web site
Issues Encountered

- Scoping: What is considered “window dressing?”
- Coding of the website itself

- Soft skills
  - Communication
  - Patience
  - Delegating tasks
Future Work

- Support for additional functionality
  - Categorized results
  - Add search support for additional entity types
Ethical Issues

- Professional Issues
  - Software Copyright
  - Content Copyright
  - Terms of Use

- Internal Issues
  - Scheduling Conflicts
  - Interpersonal Conflicts
Group Accomplishments

- Developed working prototype
- Published deliverables on the Web
- Outlined future work
Societal Impact

- Inspire future search engines to use semantic data to provide users with specific answers
  - Improve access to information
  - Published our design documents online
Future Work

- Improve overall speed of system
- Improve accuracy of entity extraction
- Support new the types of queries
- Allow crawling of more types of data sources
Questions & Answers