Internet-based Information Extraction Technologies

Fang Li (李芳)

Dept. of Computer Science & Engineering
Li-fang@cs.sjtu.edu.cn
Contents

- Motivation & Aim
- Related Technologies
- Applications
- Course Introduction & Syllabus
Motivation

- Internet: widely used, easily accessed
- Internet: a great information resource
- Internet: shared by the whole world

- Too much information
- Less time in modern life
- Little powerful search engines
截至2015年6月，我国网民规模达6.68亿，半年共计新增网民1894万人。互联网普及率为48.8%，较2014年底提升了0.9个百分点，整体网民规模增速继续放缓。
## Internet and User (June, 2015)

<table>
<thead>
<tr>
<th>应用</th>
<th>2015 年 6 月</th>
<th>网民使用率</th>
<th>2014 年 12 月</th>
<th>网民使用率</th>
<th>半年增长率</th>
</tr>
</thead>
<tbody>
<tr>
<td>即时通信</td>
<td>60626</td>
<td>90.8%</td>
<td>58776</td>
<td>90.6%</td>
<td>3.1%</td>
</tr>
<tr>
<td>网络新闻</td>
<td>55467</td>
<td>83.1%</td>
<td>51894</td>
<td>80.0%</td>
<td>6.9%</td>
</tr>
<tr>
<td>搜索引擎</td>
<td>53615</td>
<td>80.3%</td>
<td>52223</td>
<td>80.5%</td>
<td>2.7%</td>
</tr>
<tr>
<td>网络音乐</td>
<td>48046</td>
<td>72.0%</td>
<td>47807</td>
<td>73.7%</td>
<td>0.5%</td>
</tr>
<tr>
<td>博客/个人空间</td>
<td>47457</td>
<td>71.1%</td>
<td>46679</td>
<td>72.0%</td>
<td>1.7%</td>
</tr>
<tr>
<td>网络视频</td>
<td>46121</td>
<td>69.1%</td>
<td>43298</td>
<td>66.7%</td>
<td>6.5%</td>
</tr>
<tr>
<td>网络游戏</td>
<td>38021</td>
<td>56.9%</td>
<td>36585</td>
<td>56.4%</td>
<td>3.9%</td>
</tr>
<tr>
<td>网络购物</td>
<td>37391</td>
<td>56.0%</td>
<td>36142</td>
<td>55.7%</td>
<td>3.5%</td>
</tr>
<tr>
<td>微博客</td>
<td>20432</td>
<td>30.6%</td>
<td>24884</td>
<td>38.4%</td>
<td>-17.9%</td>
</tr>
<tr>
<td>网络文学</td>
<td>28467</td>
<td>42.6%</td>
<td>29385</td>
<td>45.3%</td>
<td>-3.1%</td>
</tr>
</tbody>
</table>
Internet in China (cont.)

The number of web pages is dramatically rising.

截至2015年6月，我国IPv6地址数量为19,338块/32，半年增长2.9%。

来源：CNNIC 中国互联网络发展状况统计调查 2015.6
Four kinds of Data on the Internet (from William W. Cohe)

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Non-grammatical snippets, rich formatting & links

<table>
<thead>
<tr>
<th>Lecturer</th>
<th>Contact Information</th>
<th>Position</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barto, Andrew G.</td>
<td>(413) 545-2109</td>
<td>Professor.</td>
<td>Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.</td>
</tr>
<tr>
<td>Berger, Emery D.</td>
<td>(413) 577-4211</td>
<td>Assistant Professor.</td>
<td>Emilio Remolina and Benjamin Kuipers.</td>
</tr>
<tr>
<td>Brock, Oliver</td>
<td>(413) 577-0334</td>
<td>Assistant Professor.</td>
<td>Computer Science.</td>
</tr>
<tr>
<td>Clarke, Lori A.</td>
<td>(413) 545-1328</td>
<td>Professor.</td>
<td>Software verification, testing, and analysis; software architecture and design.</td>
</tr>
<tr>
<td>Cohen, Paul R.</td>
<td>(413) 545-3638</td>
<td>Professor.</td>
<td>Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.</td>
</tr>
</tbody>
</table>

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Grammatical sentences and some formatting & links

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 9:30 AM</td>
<td>Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty</td>
</tr>
<tr>
<td>Joseph Y. Halpern, Cornell University</td>
<td></td>
</tr>
<tr>
<td>9:30 - 10:00 AM</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:00 - 11:30 AM</td>
<td>Technical Paper Sessions:</td>
</tr>
<tr>
<td>Cognitive Robotics</td>
<td>Logic Programming</td>
</tr>
<tr>
<td>Natural Language Generation</td>
<td>Complexity Analysis</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>Games</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>739</td>
<td>A Logical Account of Causal and Topological Maps</td>
<td>Emilio Remolina and Benjamin Kuipers</td>
</tr>
<tr>
<td>116</td>
<td>A System of Problem Solving through Abduction</td>
<td>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</td>
</tr>
<tr>
<td>758</td>
<td>Title Generation for Machine-Translated Documents</td>
<td>Rong-Jin and Alexander G. Hauptmann</td>
</tr>
<tr>
<td>417</td>
<td>Let's go Nats: Complexity of Named Circumscriptive and Abnormality Theories</td>
<td>Marco Cadoli, Thomas Eiter, and Georg Gottlob</td>
</tr>
<tr>
<td>179</td>
<td>Knowledge Extraction and Comparison from Local Function Networks</td>
<td>Kenneth McCurry, Stefan Wermter, and John Machajdik</td>
</tr>
<tr>
<td>540</td>
<td>Online-Execution of cColog Plans</td>
<td>Henrik Grosskreutz and Gerhard Lakemeyer</td>
</tr>
<tr>
<td>131</td>
<td>A Comparative Study of Logic Programs with Preference</td>
<td>Torsten Schaub and Kevin</td>
</tr>
<tr>
<td>246</td>
<td>Dealing with Dependencies between Content Planning and Surface Realization in a Pipeline Generation</td>
<td></td>
</tr>
<tr>
<td>470</td>
<td>A Perspective on Knowledge Compilation</td>
<td>Adriano D’Avanzo and Pierre Marquis</td>
</tr>
<tr>
<td>258</td>
<td>Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series</td>
<td></td>
</tr>
<tr>
<td>355</td>
<td>Temporal Difference Learning Applied to a High Performance Game-Playing</td>
<td></td>
</tr>
</tbody>
</table>
Aim of the Internet-based Information Extraction

Web as an information resource

- Web Crawler
- Categorize, Clustering, Information extraction
- Real Applications
- Acquire
- Process
- Analysis

Heterogeneous
Specific
Structure
Knowledge
Data – Knowledge – Information

- **Data**: recoded facts or figures
- **Knowledge**: the understanding required to convert data into information and to apply it to real-world situations
- **Information**: the value derived from data through the application of knowledge

**Information** = **data** + **knowledge**
Data vs. Knowledge

Knowledge are data with meaning, e.g., a property (or feature) of an object (size of a human, name of a company). Note that the same data element might have several possible interpretations.
For example (data & knowledge)

<table>
<thead>
<tr>
<th>Year</th>
<th>Make</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Ford</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Thunderbird</td>
</tr>
<tr>
<td>Mileage</td>
<td>5,500 miles</td>
</tr>
<tr>
<td>Features</td>
<td>Red, ABS, 6 CD changer</td>
</tr>
<tr>
<td>Price</td>
<td>$33,000</td>
</tr>
<tr>
<td>Phone</td>
<td>(916) 972-9117</td>
</tr>
</tbody>
</table>

Car Advertisement

- Year: 2002
- Make: Ford
- Model: Thunderbird
- Mileage: 5,500 miles
- Features: Red, ABS, 6 CD changer
- Price: $33,000
- Phone: (916) 972-9117
Knowledge vs. Information

**Knowledge**: a model of the world (structural and functional properties of the real world)

**Information**:
- is that part of knowledge which is used to solve a certain problem *(Information System view)*
- Information only exists in concrete problem situations

> Information systems extract that knowledge just in time, a user needs in context of a given situation.
Information System

Main tasks of information systems:

- **Maintain knowledge in digital form as data.**
- **Provide knowledge as useful information to a user.**
  - Information only exists relative to an information consumer/request
  - Information must be interpreted relative to already existing information.
  - There is no communication without information.
Information Extraction (IE): the subdiscipline of Artificial Intelligence.

19 March — A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb — allegedly detonated by urban guerrilla commandos — blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Text → structural data

<table>
<thead>
<tr>
<th>INCIDENT TYPE</th>
<th>bombing</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE</td>
<td>March 19</td>
</tr>
<tr>
<td>LOCATION</td>
<td>El Salvador: San Salvador (city)</td>
</tr>
<tr>
<td>PERPETRATOR</td>
<td>urban guerrilla commandos</td>
</tr>
<tr>
<td>PHYSICAL TARGET</td>
<td>power tower</td>
</tr>
<tr>
<td>HUMAN TARGET</td>
<td>-</td>
</tr>
<tr>
<td>EFFECT ON PHYSICAL TARGET</td>
<td>destroyed</td>
</tr>
<tr>
<td>EFFECT ON HUMAN TARGET</td>
<td>no injury or death</td>
</tr>
<tr>
<td>INSTRUMENT</td>
<td>bomb</td>
</tr>
</tbody>
</table>
Internet-based Information Extraction from William W. Cohe

- **Web site specific**
  - Formatting
  - Amazon.com Book Pages

- **Genre specific**
  - Layout
  - Resumes

- **Wide, non-specific**
  - Language
  - University Names

---

**Jason D. M. Rennie**

Massachusetts Institute of Technology
MIT AI Lab B143-733
209 Technology Sq
Cambridge, MA 02139
http://www.ai.mit.edu/people/jrennie
(617) 253-5339

Research Interests

My main interests lie in the automated analysis and representation of data for the purposes of classification, estimation and the acquiring of new knowledge. I have both interests in applying such techniques to a wide, non-specific domain and to the analysis of web sites and genre specific web sites.

---

**L. Douglas Baker**

Home Address: available upon request
Office Address: Wean Hall, 8102
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
(412) 685-4906
http://www.cs.cmu.edu/~ldbaker

Objective

A position in a dynamic, highly-skilled applied research and development team using statistical machine learning to solve large-scale, real-world tasks such as Information Retrieval and Text Classification.

Education

- Carnegie Mellon University, Pittsburgh, PA
  - Ph.D., Computer Science, in progress
  - M.S., Computer Science, 1999
- Technical University of Berlin
- Humboldt University of Berlin
- Exchange Fellow, 1992-1993
- University of Michigan
- Ann Arbor, MI
- M.S.E., Computer Science and Engineering, 1994 B.S.E., Computer Engineering, Georgia Tech, 1992

Research Experience

- Carnegie Mellon University
  - 1994-present

I am currently pursuing my dissertation research: a hierarchical probabilistic model for novelty detection in text. This work is being done as part of the Topic Detection and Tracking Task under the Defense Advanced Research Projects Agency.
Different tasks of IE

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity
- Person: Jack Welch
- Person: Jeffrey Immelt
- Location: Connecticut

Binary relationship
- Relation: Person-Title
  - Person: Jack Welch
  - Title: CEO
- Relation: Company-Location
  - Company: General Electric
  - Location: Connecticut

N-ary record
- Relation: Succession
  - Company: General Electric
  - Title: CEO
  - Out: Jack Welch
  - In: Jeffrey Immelt

Named entity extraction
- Event extraction
- Lecture of Internet-based IE technologies
Related Technologies

Web-based IE

ML: Machine Learning

IR: Information Retrieval

LT: Language Technology

WM: Web Mining

lecture of Internet-based IE technologies
Language Technology: as normalization

IE vs. LT

Natural Language Processing includes:
- Tokenization
- Morphological analysis
- Special phrases: Date and time, Proper names, Number expressions
- General Phrases: nominal phrases, prepositional phrases, ...
- Structural analysis: complex flat structure
- Semantic analysis: agent, role, target
Natural Language Processing

- **Issues in tokenization:**
  - Hewlett-Packard → Hewlett Packard?
  - Lowercase → lower-case, lower case?
  - San Francisco → one token or two?

- **Issues in morphological analysis**
  - Windows, window → U.S.A, USA
  - Need to **normalize terms**
  - Fed vs. fed, US vs. us → case is helpful

- **Period “.” is quite ambiguous**
  1. Sentence boundary
  2. Abbreviations like Inc. or Dr.
  3. Numbers like .02% or 4.3
Machine Learning: as instruments

- Use machine learning algorithms to automatically draw extraction patterns or rules for information extraction.
- Use machine learning methods to create IE systems for new domains.

Aim: to learn from a training corpus or seed examples and predict the new one.
Machine Learning includes:

**Supervised learning**

Given $D = \{X_i, Y_i\}$, learn $f(\cdot): Y_i = f(X_i)$, s.t. $D^{\text{new}} = \{X_j\} \Rightarrow \{Y_j\}$

we know the values of $f$ for the m samples in the training set $D$. We assume that if we can find a hypothesis, $h$, that closely agrees with $f$ for the members of $D$, then this hypothesis will be a good guess for $f$ especially if $D$ is large.

**Unsupervised learning**

Given $D = \{X_i\}$, learn $f(\cdot): Y_i = f(X_i)$, s.t. $D^{\text{new}} = \{X_j\} \Rightarrow \{Y_j\}$

we simply have a training set of vectors without function value for them, we regard the problem as one of the learning a function.
Machine Learning includes:

Semi-supervised learning

- make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.
- many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.
Interacts with Web Mining  IE vs. WM

- **Data Mining** (from database): information extraction and discovering of relational links
- **Text Mining** (from text documents): data mining using domain-independent shallow text processing
- **Web Mining** (from the Internet): The use of data mining techniques to uncover hidden patterns or relationships among available Web data.
  - Data on the WWW: all the content information available on-line
  - Web log data: users’ on-line activities (Cookies)
  - Web structure data: web linkage information
Web Mining (cont.) IE vs. WM

IE focus on content data

Web Data
- Free Texts
- HTML Files
- XML Files
- Dynamic Content
- Multimedia
- Static Link
- Dynamic Link

Content Data
- Structure Data
- Usage Data
- User Profile Data

data from Web log data – IP addresses, date & time access

lecture of Internet-based IE technologies

data that is user specific – registration and customer profile
Information Retrieval: as a precondition

- Information Retrieval
  - user
  - dynamic, open domain

- Information Extraction
  - system
  - static, predefined -> open domain

IR vs. IE
IE vs. IR: What IR System can do

Find documents relevant to "terrorist attack"
IE vs. IR: What IR System can not do

Given documents describing terrorist attacks, identify the involved criminal names

Sort the terrorist attacks based on when they occurred

IE technologies are required!

lecture of Internet-based IE technologies
IE has a high Application Impact & interacts with

1. Information Retrieval
2. Question Answering
3. Text Classification and Abstractive Summarization
4. Database Systems
5. Knowledge-based Systems

Information extraction is a helpful step in these processes because the data become structured and semantically enriched.
IE for Question-Answering

- **Question-Answering (QA):** query texts using natural language just like querying a database using SQL.
- QA supported by IE:

  - **Question Parse:**
    - *Who kidnapped Anderson?*

  - **Event Query Template:**
    - **Captor:** ?
    - **Captee:** Anderson

  - **Kidnapping Events:**
    - | Captor  | Captee    | Where            | When   |
    - |--------|----------|------------------|--------|
    - | NULL   | Mayors   | northeast Col... |
    - | NULL   | reporters| northeast Col... |
    - | Islamic| Anderson | NULL            | NULL   |
    - | Islamic| Thomas...| NULL            | NULL   |
A QA Example from AskJeeves

List U.S. presidents

Temporal information

Alphabetical Order: Adams to Jackson
- Adams, John
  - 1797-1801
- Adams, John
  - 1825-29
- Arthur, Chester
  - 1881-85
- Buchanan, James
  - 1857-61
- Bush, George H.W.
  - 1989-93
- Bush, George W.
  - 2001-present

Alphabetical Order: Jefferson to Wilson
- Jefferson, Thomas
  - 1801-09
- Johnson, Andrew
  - 1865-69
- Johnson, Lyndon
  - 1963-69
- Kennedy, John F.
  - 1961-63
- Lincoln, Abraham
  - 1861-65
- Madison, James
  - 1809-17
Sample Job Posting:
Job Title: Senior DBMS Consultant
Location: Dallas, TX
Responsibilities:
DBMS Applications consultant works with project teams to define DBMS based solutions that support the enterprise deployment of Electronic Commerce, Sales Force Automation, and Customer Service applications.
Desired Requirements:
3-5 years exp. developing Oracle or SQL Server apps using Visual Basic, C/C++, Powerbuilder, Progress, or similar. Recent experience related to installing and configuring Oracle or SQL Server in both dev. and deployment environments.
Desired Skills:
Understanding of UNIX or NT, scripting language. Know principles of structured software engineering and project management.

Filled Job Template:
title: Senior DBMS Consultant
state: TX
city: Dallas
country: US
language: Powerbuilder, Progress, C, C++, Visual Basic
platform: UNIX, NT
application: SQL Server, Oracle
area: Electronic Commerce, Customer Service
required years of experience: 3
desired years of experience: 5
Discussion: Food Events

Food-Bot is a website designed to help **college students** find **free food events** on campus. The website seeks to serve two groups: **students** looking for free food, and **student groups** looking for some free publicity for their events. [http://food-bot.com](http://food-bot.com)

**How it works**

有否免费的午餐？
Food-Bot: Problem Statement

Given an arbitrary text $d$, determine whether $d$ contains information about a free food event, and if so, return an array of correctly-associated information about each event (date/time, location, and food type).

Question: How it works?

Group Discussion
Food-Bot: methods

1) Naïve Bayes Classification to find whether the d contains a free food event.

2) Event Parsing Theory
   - Identify time, place, food name.
   - Merge into an event.
Mistakes made by Food-bot

Date:
Tuesday, February 15, 2011 - 12:00pm
Location:
12:00 pm - 1:00 pm Stanley Hall
Food Type:
lunch
Organization:
unspecified
Food Quality:
7/10
Food Quantity:
5/10
Time Commitment:
5/10
Awkwardness:
5/10

Speaker/Performer: Erica Whitney QB3-Berkeley is offering a series of workshops for graduate students and postdoctoral researchers in QB3-affiliated labs on understanding the research grant process and how to write fundable proposals.

Each workshop is offered twice per week once on Tuesdays and again on Wednesday. They are held in 221 Stanley Hall from 12-1pm. Please feel free to bring a bag lunch.

lectures of Internet-based IE technologies
王石
标签:企业家,探险运动家,知名...

性别: 男
生日: 1960年以前
出生地点: 广西壮族自治区柳州市
身高: 176cm
标签: 商业,企业家,探险运动家,商业人员,中国

来源: 百度百科


机构关系

万科集团
...月16日，住房和城乡建设部副部长黄卫，四川省建设厅厅长杨洪波、副厅长何建等领导在四川省建设厅抗旱应旱指挥部会见万科集团董事长王石一行，商讨灾后重建工作。黄卫副部长对灾后重建工作提出了具体要求，杨厅长、何副厅长就灾后重建工作与王石进行了交流。...

兰州铁道学院
...们走上工作岗位的学历，并非其人生的最高学历，最高学历这东西，特别是近几年，唉......我就不说了。王石，先当兵，后推荐上了兰州铁道学院的，那学校，我当年常去溜达，在西北也算好学校了。孙宏斌，清华大学的研究生，这所学校我就不必说了。冯...

http://www.tianya.cn/publicforum/content/HELP/1/197465.shtml

http://bbs.city.tianyacity.com/tianyacity/Content/45/1/777287.shtml
Course Introduction

Textbook is not the only reference
→ Research papers are needed to read

Different forms of exercises
→ Read articles, Discussion, Programming and presentation

Many models and technologies are involved, not detailed.
→ Research oriented course
（泛学与精学）
Text Book and References


Ralph Grishman, Information Extraction: Capabilities and challenges
Course Web Sites

My personal web sites:
http://www.cs.sjtu.edu.cn/~li-fang/

This course is in:
Course Contents (overview)

- chapter 1: Overview IE technologies.
- chapter 2: Named entity and relations extraction.
- chapter 3: Event information extraction.
- chapter 4 opinion extraction and sentimental analysis
- chapter 5: Web-based information extraction
- chapter 6: IE systems
- student workshop
The course organized as following:

**Information Extraction**

from:

- Free text
- Web pages

**Information extraction tasks include**

- Named Entity extraction
- Relation extraction
- Event extraction

**Information extraction Research**

- IE systems
- Sentimental analysis
- Student workshop

lecture of Internet-based IE technologies
Updates of the Course

- Design more classroom exercises and discussion topics.
- Sentimental analysis consists of opinion extraction and opinion words extraction.
Course Project: Algorithm implementation

Task 1: relation extraction
- Corpus: news reports
- Extract:
  Employment (person-organization)
  Employment (person-organization-time)
  Employment (person-organization-position)

Choose one of the 3 relations.
Course Project: Algorithm implementation

Task 2: sentiment analysis
Corpus: TV series reviews
Classify:
Positive reviews
Negative reviews
Neutral reviews
Course project: free topic

- Task 3: Exploring a new extraction for some particular application. (system implementation)
Student Workshop

Each group should present their work in the following points:

- Task Requirements
- Group Organization
- Project Introduction and demo:
  1. Algorithm for relation extraction or
  2. Sentimental Classification or
  3. System Design for some applications
Grading (考核标准)

- Attendence & Classroom discussion (40%)
- Workshop Presentation (20%)
- Course Project (40%)

There is no final examination.
What are your expectations from this course? Feedbacks are welcome!