

# Information Extraction

- **Definition** (after Grishman 1997, Eikvil 1999):  
*"The identificiation and extraction of instances of a particular class of events or relationships in a natural language text and their transformation into a structured representation (e.g. a database)."*
  - *IR* retrieves *relevant documents* from collections
  - *IE* retrieves *relevant information* from documents
- **Example: AutoSlog** (Riloff)
  - input:
    - general syntactic patterns
    - annotated (marked-up) training documents
  - output:
    - instantiated patterns that extract particular information
  - **Autoslog-TS**: Extension that replaces need for annotated corpus with manual post-processing of sorted pattern list
- **On the Web**: natural language text → (semi-)structured text

# Extracting Job Openings from the Web

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites

Back Forward Stop

Address <http://www.foodsci>

Links AMEX Rewards

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS  
INTERNATIONAL  
INC.

About | Staff | Job

OPUS: Job Listings - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address [http://www.foodscience.com/jobs\\_midwest.html#top](http://www.foodscience.com/jobs_midwest.html#top)

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

**Job Listings**

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

Test Kitchen-  
Consumer Food Relations

Major food manufacturer in Chicago area seeks a consumer food professional to write and test recipes. Will make presentations; will be a key player in a cross-functional team. Requires a BS in human ecology, nutrition, Food Science, or related field and a minimum three years' professional experience.  
Contact Moira: [e-mail](#)  
1-800-488-2611

**Ice Cream Guru**

If you dream of cold creamy chocolate or coochy coochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.  
Contact Susan: [e-mail](#)  
1-800-488-2611

foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: [www.foodscience.com/jobs\\_midwest.html](http://www.foodscience.com/jobs_midwest.html)


OtherCompanyJobs: foodscience.com-Job1



# Example: A Solution

job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer

Address <http://www.flipdog.com/home.html> Go File Edit View Favorites Tools Help Links >>

 **FlipDog**.com

[Home](#) [Find Jobs](#) [Your Account](#) [Resource Center](#) • [Support](#) • [Employers](#)

Job Search at FlipDog.com: Employment & Career Management




**647,514**  
Job Opportunities  
from **53,641** Employers

[▶ Find a Job!](#)

[▶ Post Your Resume](#)

[▶ Employers](#)  
click here for  
Products & Services



**Job Seekers: Find your dream job!**

- ▶ Check our 'Best Places to Find a Job' [January report](#).
- ▶ Open your [FREE account](#) and put your [resume online](#).
- ▶ Search 24x7 with our FREE automatic [JobHunters™](#).
- ▶ Research our database of over [50,000 employers](#).
- ▶ Get [expert advice](#) at our new [Resource Center](#).
- ▶ Access [salary surveys/calculators](#), [relocation tools](#), [networking opportunities](#), & [training/testing](#) tools.
- ▶ Use FlipDog.com to search jobs right from your desktop! Download [Snippets](#) today!

**Pigskin Places**

- Health Care in NY [2,770](#)
- Health Care in MD [1,262](#)
- Sales in NY [3,751](#)
- Sales in MD [958](#)
- Computing in NY [8,050](#)
- Computing in MD [4,114](#)


**Jobs for Sports Fans**

- [Head Football Coach](#)
- [Football Coach](#)
- [Asst. Football Coach](#)
- [High School Football Coach](#)
- [Univ. Asst. Football Coach](#)


**Job Seeker Newsletter**

Enter your e-mail address:


[Sign Me Up!](#)

 "Top 100 Web Sites"  
PC Magazine, Nov. 2000

 "Top 10 Career Web Site"  
Media Matrix, Sept. 2000

 "Top 10 Job Site"


**Showcase Jobs**



We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.

[Learn More](#)



---



Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs.

[Learn More](#)

powered by  
**WhizBang!**

Start |  |  Microsoft PowerPoint - [sta... |  job search find employem... | 12:12 AM

# Job Openings:

Category = Food Services

Keyword = Baker

Location = Continental U.S.

Slide taken from William Cohen

**FlipDog**  
Fetch Your Next Job Here™

Home Find Jobs Your Account Resource Center

Return to Results | Modify Search | New Search

**The University Alliance**  
A BISK EDUCATION NETWORK  
Degrees Online

Learn While You Earn  
**MBA, BA, AA** Degrees  
Online & Project Mgt.

[Click here to e-mail your resume to 1000's of Head Hunters with ResumeZapper.com](#)

**how to easily DOUBLE your chances when applying FOR JOBS!**

Breakthrough ebook shows why most people are WRONG about how to apply for jobs.

1 - 25 of 47 jobs shown below 1 2 Next >

Search these results for:   [Search tips](#) Show Jobs Posted:

View: [Brief](#) | [Detailed](#)

**Web Jobs:** FlipDog technology has found these jobs on thousands of employer Web sites.

<a href="#">Food Pantry Workers</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Cooks</a> at <a href="#">Lutheran Social Services</a>	October 11, 2002	<a href="#">Archbold, OH</a>
<a href="#">Bakers Assistants</a> at <a href="#">Fine Catering by Russell Morin</a>	October 11, 2002	<a href="#">Attleboro, MA</a>
<a href="#">Baker's Helper</a> at <a href="#">Bird-in-Hand</a>	October 11, 2002	United States
<a href="#">Assistant Baker</a> at <a href="#">Gourmet To Go</a>	October 11, 2002	<a href="#">Maryland Heights, MO</a>
<a href="#">Host/Hostess</a> at <a href="#">Sharis Restaurants</a>	October 10, 2002	<a href="#">Beaverton, OR</a>
<a href="#">Cooks</a> at <a href="#">Alta's Rustler Lodge</a>	October 10, 2002	<a href="#">Alta, UT</a>
<a href="#">Line Attendant</a> at <a href="#">Sun Valley Coporation</a>	October 10, 2002	<a href="#">Huntsville, UT</a>
<a href="#">Food Service Worker II</a> at <a href="#">Garden Grove Unified School District</a>	October 10, 2002	<a href="#">Garden Grove, CA</a>
<a href="#">Night Cook / Baker</a> at <a href="#">SONOCO</a>	October 10, 2002	<a href="#">Houma, LA</a>
<a href="#">Cooks/Prep Cooks</a> at <a href="#">GrandView Lodge</a>	October 10, 2002	<a href="#">Nisswa, MN</a>
<a href="#">Line Cook</a> at <a href="#">Lone Mountain Ranch</a>	October 10, 2002	<a href="#">Big Sky, MT</a>
<a href="#">Production Baker</a> at <a href="#">Whole Foods Market</a>	October 08, 2002	<a href="#">Willowbrook, IL</a>
<a href="#">Cake Decorator/Baker</a> at <a href="#">Mandalay Bay Hotel and Casino</a>	October 08, 2002	<a href="#">Las Vegas, NV</a>
<a href="#">Shift Supervisors</a> at <a href="#">Brueggers Bagels</a>	October 08, 2002	<a href="#">Minneapolis, MN</a>

# IE from Research Papers

**A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p**

File Edit View Favorites Tools Help

← Back → Stop Home Search Favorites History Print Copy Paste

Address <http://citeseer.nj.nec.com/peter90critical.html> Links >>

**A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990) (Correct) (5 citations)**

Peter Norvig Robert Wilensky University of California, Berkeley Computer...  
Thirteenth International Conference on Computational Linguistics, Volume 3

Download: [norvig.com/coling.ps](http://norvig.com/coling.ps)  
Cached: [PS.gz](#) [PS](#) [PDF](#) [DjVu](#) [Image](#) [Update](#) [Help](#)

From: [norvig.com/resume](http://norvig.com/resume) (more)  
Home: [R.Wilensky](#) [HPSearch](#) (Correct)

**NEC ResearchIndex** [Bookmark](#) [Context](#) [Related](#)

[\(Enter summary\)](#)

Rate this article: 1 2 3 4 5 (best)  
[Comment on this article](#)

**Abstract:** this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. [\(Update\)](#)

**Context of citations to this paper:** [More](#)

.... (break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths. Table 1: Empirical Results Comparing Coherence and...

.... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in [Norvig and Wilensky \(1990\)](#). The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

**Cited by:** [More](#)

[Translation Mismatch in a Hybrid MT System - Gawron \(1999\) \(Correct\)](#)

[Abduction and Mismatch in Machine Translation - Gawron \(1999\) \(Correct\)](#)

[Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin \(1990\) \(Correct\)](#)

**Active bibliography (related documents):** [More](#) [All](#)

0.1: [Critiquing Effective Decision Support in Time-Critical Domains - Gertner \(1995\) \(Correct\)](#)

0.1: [Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson \(1995\) \(Correct\)](#)

0.1: [A Probabilistic Network of Predicates - Dekora Liu \(1992\) \(Correct\)](#)

Internet

# What is “Information Extraction”

As a task:

Filling slots in a database from sub-segments of text.

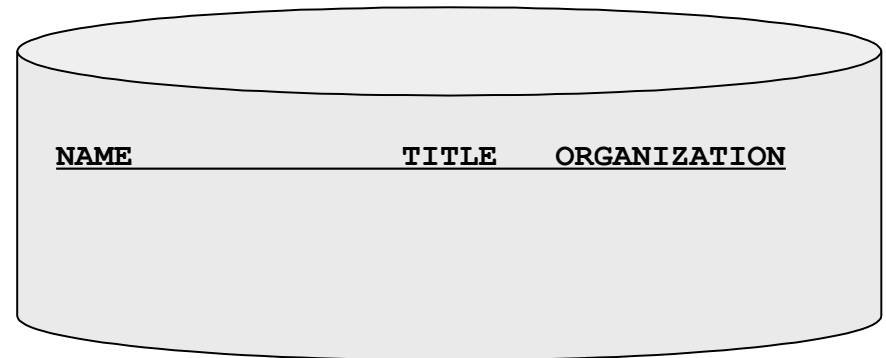
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



# What is “Information Extraction”

As a task:

Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



<u>NAME</u>	<u>TITLE</u>	<u>ORGANIZATION</u>
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

# Landscape of IE Tasks (1/4): Degree of Formatting

## 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.











## Grammatical sentences and some formatting & links

**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.

- Press
- **Contact**
- General information
- Directions maps

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

## Non-grammatical snippets, rich formatting & links

<b>Barto, Andrew G.</b> Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.	(413) 545-2109	<a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a>	CS276	 
<b>Berger, Emery D.</b> Assistant Professor.	(413) 577-4211	<a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a>	CS344	 
<b>Brock, Oliver</b> Assistant Professor.	(413) 577-0334	<a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a>	CS246	 
<b>Clarke, Lori A.</b> Professor. Software verification, testing, and analysis; software architecture and design.	(413) 545-1328	<a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a>	CS304	 
<b>Cohen, Paul R.</b> Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.	(413) 545-3638	<a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a>	CS278	 

## Tables

8:30 - 9:30 AM	<b>Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty</b> <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
<b>Cognitive Robotics</b>	<b>Logic Programming</b>	<b>Natural Language Generation</b>	<b>Complexity Analysis</b>	<b>Neural Networks</b>	<b>Games</b>
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Networks <i>Kenneth McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz and Gerhard Lakemeyer</i>	131: A Comparative Study of Logic Programs with Preference <i>Torsten Schaub and Kewen</i>	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation <i>Adnan Darwiche and Pierre Marquis</i>	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing



# Landscape of IE Tasks (2/4): Intended Breadth of Coverage

## Web site specific

### Formatting

#### Amazon.com Book Pages

## Genre specific

### Layout

#### Resumes

## Wide, non-specific

### Language

#### University Names

The screenshot shows the Amazon.com product page for the book "Machine Learning in Graphical Models" edited by Michael Irwin Jordan. The page features the Amazon logo, navigation tabs for "WELCOME", "YOUR STORE", "BOOKS", "ELECTRONICS", "DVD", and "TOYS & GAMES". A search bar is visible at the top. The book cover is displayed with a "LOOK INSIDE!" feature. The price is listed as \$60.00, with a "NEW Super Saver Shipping FREE" offer. A "Great Buy" banner is at the bottom, suggesting a bundle with "Probabilistic Reasoning in Intelligent Systems" for a total price of \$128.95.

The screenshot shows two resumes. The top one is for Jason D. M. Rennie, listing his affiliation with MIT AI Lab and his research interests in automated data analysis. The bottom one is for L. Douglas Baker, listing his education at Carnegie Mellon University and the University of Michigan, along with his research objective in dynamic machine learning.

The screenshot shows a university website page. At the top, it features a title "Talk: Plausibility Measures: A General Approach for Representing Uncertainty" by Y. Halpern, Cornell University. Below this is a table of "Academic Paper Sessions" with columns for "Natural Language Generation", "Complexity Analysis", "Neural Networks", and "Games". The bottom section features a "Contact" box for Dr. Steven Minton, Founder/CTO, and Frank Huybrechts, COO, providing their roles and a link to "General information" and "Directions maps".

# Landscape of IE Tasks (3/4): Complexity

E.g. word patterns:

## Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

## Complex pattern

U.S. postal addresses

University of Arkansas  
P.O. Box 140  
Hope, AR 71802

Headquarters:  
1128 Main Street, 4th Floor  
Cincinnati, Ohio 45210

## Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

## Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

# Landscape of IE Tasks (4/4): Single Field/Record

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*

# Recognizers

- Simple procedures to find pieces of information based on its appearance
  - e-mail addresses (easy)
  - telephone numbers (tricky)
  - street addresses (difficult)
- Examples:
  - Simple Web Crawlers can (and do) collect huge databases of e-mail addresses
  - Recognizers can also be used to automatically generate training examples for wrapper induction (Kushmerick, 2000)
  - A Firefox plugin can recognize phone numbers on pages and replace them with a link to the Skype dialer

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Hinweis: Ein Klick auf einen E-Mail-Link funktioniert nur, wenn Sie Javascript in Ihrem Browser aktiviert haben.

---

## A

---

### **Abbing**, Jana (Mgr.)

Telekooperation, Raum S2 02 | A013

E-Mail: [jana\(a-t\)informatik.tu-darmstadt.de](mailto:jana(a-t)informatik.tu-darmstadt.de)

Tel:  +49 6151 - 16-5245 , Fax: +49 6151 - 16-3052

### **Achenbach**, Michael

Aspektorientierte Programmierung, Raum S2 02 | A205

Tel:  +49 6151 - 16-4216 , Fax: +49 6151 - 16-5410

### **Adamson**, Anders (Dipl.-Inform.)

Graphisch-Interaktive Systeme, Raum S3 05 | 316

E-Mail: [anders.adamson\(a-t\)gris.informatik.tu-darmstadt.de](mailto:anders.adamson(a-t)gris.informatik.tu-darmstadt.de)

Tel:  +49 6151 - 155-673 , Fax: +49 6151 - 155-430

### **Aderhold**, Markus

Programmiermethodik, Raum S2 02 | A312

E-Mail: [aderhold\(a-t\)informatik.tu-darmstadt.de](mailto:aderhold(a-t)informatik.tu-darmstadt.de)

Tel:  +49 6151 - 16-5668 , Fax: +49 6151 - 16-6241

### **Aitenbichler**, Erwin (Dr.-Ing.)

Telekooperation, Raum S2 02 | A121

E-Mail: [erwin\(a-t\)informatik.tu-darmstadt.de](mailto:erwin(a-t)informatik.tu-darmstadt.de)

Tel:  +49 6151 - 16-2259 , Fax: +49 6151 - 16-3052

### **Andriluka**, Mykhaylo

for wrapper induction (Kushnirenko, 2000)

- A Firefox plugin can recognize phone numbers on pages and replace them with a link to the Skype dialer

# Recognizers

- example for an incorrect extraction

Christine Langhammer  
für den Vorsitzenden der Berufungskommission  
O.Univ.-Prof.Dr. Peter Zinterhof

## Examples.

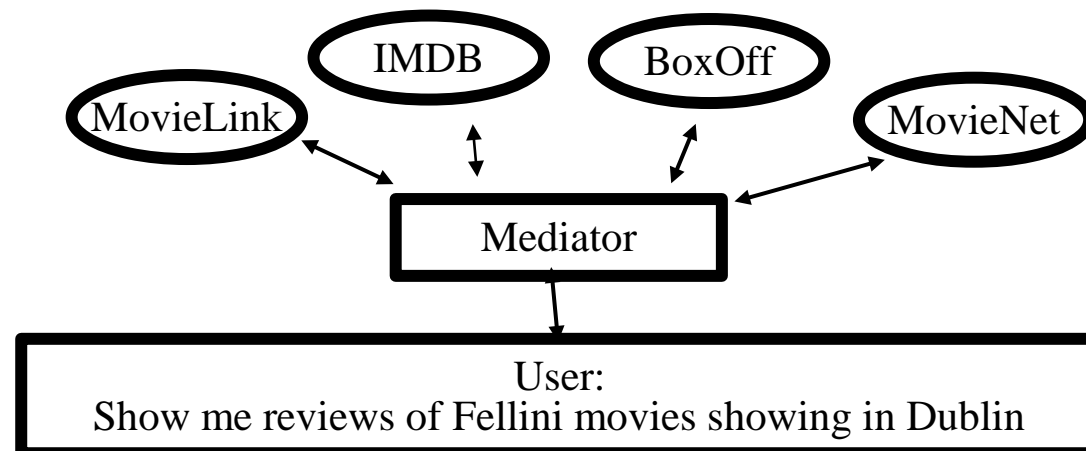
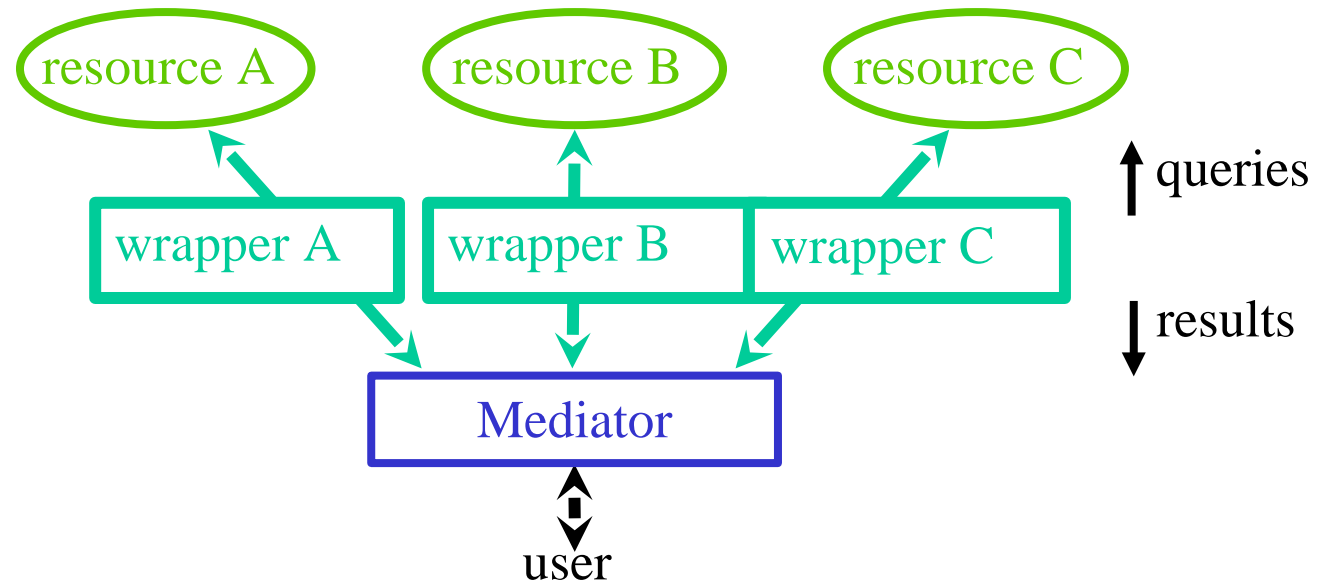
- Simple Web Crawlers can (and do) collect huge databases of e-mail addresses
- Recognizers can also be used to automatically generate training examples for wrapper induction (Kushmerick, 2000)
- A Firefox plugin can recognize phone numbers on pages and replace them with a link to the Skype dialer
- Google-Mail replaces in-line URLs with links to the site

# Wrappers

- Wrapper: (in an Information Extraction context)
  - A procedure that extracts certain pieces of information from (semi-)structured text (HTML)
- Examples:
  - Comparison Shoppers (Junglee, Shopbot/Jango, mySimon)
  - Meta-Search engines (citeseer, metacrawler)
  - News Agents (google news)
- Building Wrappers by hand:
  - time-consuming and error-prone (=> expensive)
  - Web-sites change frequently
    - mean-time to failure of wrappers: 1 month (Weld, 1998)
    - monthly failure rates of wrappers: 8% (Norvig, 1998)

# Wrapper Induction: Motivation

- Wrappers
  - parse the contents of several sites
- Mediators
  - integrate the extracted information
- Example:





# Wrapper Induction

- Automatic generation of wrappers from a few (annotated) sample pages
- Assumptions:
  - regularity in presentation of information
  - often machine-generated answers to queries
    - same header
    - same tail
    - inbetween a table/list of items that constitute the answer to the query
- Learn the delimiters between items of interest

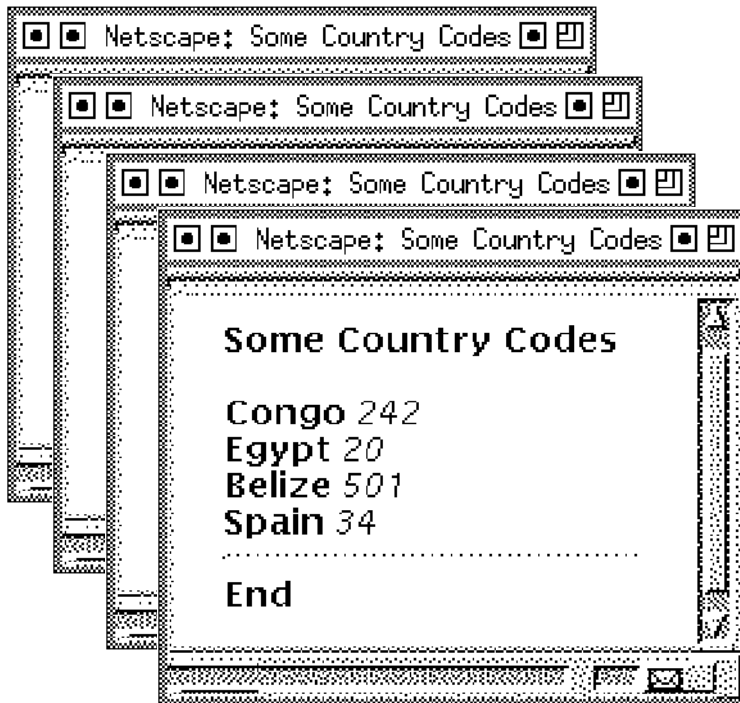
# LR Wrappers (Kushmerick 2000)

- Very simple but nevertheless powerful wrapper class
- Assume that
  - only one "database" per page
  - information can be separated into tuples (records)
  - each tuple contains exactly  $k$  items (attributes)
- Wrapper consists of  $k$  delimiter pairs  $\langle l_i, r_i \rangle$ ,
  - $l_i$  and  $r_i$  are patterns that have to be matched in the text

```
repeat
  foreach  $\langle l_i, r_i \rangle \in \{ \langle l_1, r_1 \rangle, \dots, \langle l_k, r_k \rangle \}$ 
    find next occurrence of  $l_i$ 
    find next occurrence of  $r_i$ 
    extract text inbetween and store as the  $i$ -th value for this tuple
until no more occurrences of  $l_1$ 
```

# Induction of LR Wrappers

Web Pages



Web Pages Labeled for Extraction

```
<HTML><HEAD>Some Country Codes</HEAD>
<HTML><HEAD>Some Country Codes</HEAD>
<HTML><HEAD>Some Country Codes</HEAD>
<HTML><HEAD>Some Country Codes</HEAD>
<B>Congo</B> <I>242</I><BR>
<B>Egypt</B> <I>20</I><BR>
<B>Belize</B> <I>501</I><BR>
<B>Spain</B> <I>34</I><BR>
</BODY></HTML>
```

Extracted Wrapper

$\langle \langle B \rangle, \langle /B \rangle, \langle I \rangle, \langle /I \rangle \rangle$

$\langle l_1, r_1, l_2, r_2 \rangle$

# Induction of LR Wrappers

- **Heads:** text before first tuple for each page
- **Tails:** text after last tuple for each page
- **Separators:** text between subsequent attributes
- Candidate delimiters:
  - **Left:** suffixes of the shortest of all separators to the left (including heads for  $i = 1$ )
  - **Right:** prefixes of the shortest of all separators to the right (including tails for  $i = k$ )
- Among the candidate delimiters, any one that satisfies a set of constraints can be selected
  - Constraints must ensure that the wrapper does not try to extract irrelevant parts of text (false positives)

# Constraints for Delimiters

- the left delimiter  $l_i$ 
  - must be a proper suffix of the text before each instance of the target
    - a proper suffix of a string means that
      - it is a suffix of the string
      - and it does not occur in any other place of the string (so that extraction does not start too early)
    - Example:
      - cde is a proper suffix of deabcde, de is a suffix but not proper
  - $l_i$  must not be part of any pages tail
    - otherwise extraction of a new tuple will be started at the end
- the right delimiter  $r_i$ 
  - must be a prefix of the text after each instance of the target
  - must not be part of any value for attribute  $i$ 
    - otherwise extraction will terminate prematurely

# A Problem with LR-Wrappers

- Distracting text in Head or Tail

$l_1$  fires

```
<HTML><TITLE>Some Country Codes</TITLE>  
<BODY><B>Some Country Codes</B><P>  
<B>Congo</B> <I>242</I><BR>  
<B>Egypt</B> <I>20</I><BR>  
<B>Belize</B> <I>501</I><BR>  
<B>Spain</B> <I>34</I><BR>  
<HR><B>End</B></BODY></HTML>
```



- an LR-Wrapper cannot learn an extractor for this case
  - every candidate delimiter for  $l_1$  occurs in the head
  - every candidate delimiter for  $l_1$  occurs in the tail

# HLRT-Wrappers

- Head-Tail-Left-Right Wrappers:
  - learn a separate delimiter for identifying head and tail

Ignore page's *head* and *tail*

```
<HTML><TITLE>Some Country Codes</TITLE>
<BODY><B>Some Country Codes</B><P>
<B>Congo</B> <I>242</I><BR>
<B>Egypt</B> <I>20</I><BR>
<B>Belize</B> <I>501</I><BR>
<B>Spain</B> <I>34</I><BR>
<HR><B>End</B></BODY></HTML>
```

end of head

} head

} body

} tail

start of tail

# More Expressive Wrapper Classes

- HLRT Wrappers:
  - learn 2 additional delimiters to separate the head and the tail
  - ignores occurrence of  $l_i$  and  $r_i$  before  $h$  and after  $t$
  - allows to process multiple "databases" in one document
- OCLR and HOCLRT Wrapper:
  - for each tuple: learn an (O)pening and (C)losing delimiter
- N-LR and N-HLRT:
  - allows multi-valued attributes
  - allows optional attributes
    - RESTRICTION: if a value is specified, all previous values (of this tuple) must also be specified.



# Evaluation

- Study on 30 randomly selected Web-sites from [www.search.com](http://www.search.com) (at that time a catalogue of hubs for various topics)
  - LR Wrapper was able to wrap 53%
  - LR + HLRT wrapped 60%
  - Addition of OC wrapping did not bring improvements
  - Addition of N-HLRT improved to 70%
- LR Wrappers are not limited to HTML-documents
  - any string can be extracted for delimiters, not just HTML tags
- All wrapper classes are PAC learnable
- Constraints become hard to handle

# SoftMealy (Hsu & Dung, 1998)

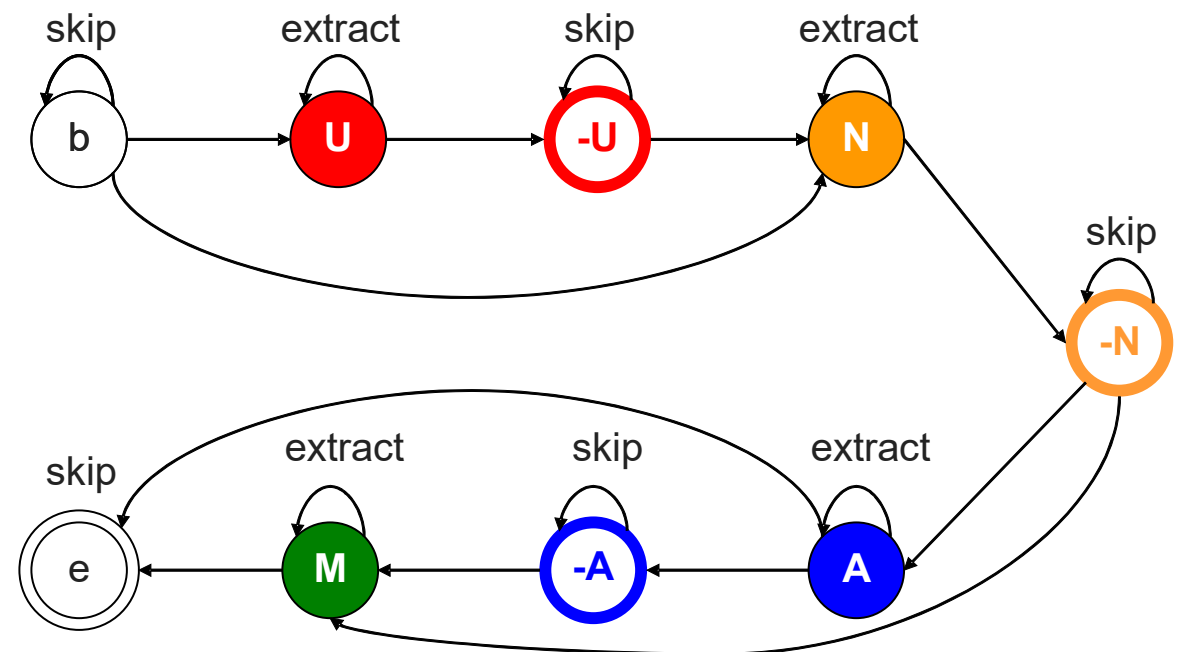
- Problems with LR-Wrappers:
  - no permutations of attributes allowed
  - delimiters may not be sufficient to identify texts
- SoftMealy provides a general solution to problems with
  - missing attributes
  - attributes with multiple values
  - variable order of attributes
- Approach:
  - learn a **finite-state transducer (FST)** that encodes all possible sequences of attributes
  - each state represents a fact to be extracted
  - dummy states are used to skip parts of text
  - use *separators* ("invisible" borders) instead of delimiters
  - learn to recognize separators by defining their left and right context with **contextual rules** (state transitions)

# Labelled Web Page

U (URL)  
 <LI><A HREF="mani.html">  
 N (Name) A (Academic title)  
 Mani Chandy</A>, <I>Professor of Computer Science</I> and  
 M (Admin title)  
 <I>Executive Officer for Computer Science</I>

U (URL)  
 <LI><A HREF="david.html">  
 N (Name) M (Admin title)  
 David E. Breen</A>, <I>Assistant Director of Computer Graphics  
 Laboratory</I>

# Sample FST



v Contextual rule looks like:  
**TRANSFER FROM state N TO state -N IF**  
 left context = capitalized string  
 right context = HTML tag "</A>"

# Wrapper Induction by Inductive Rule Learning

- Training Examples:
  - treat each slot independently (single slot extraction)
  - generate training example that represent the context of the slot (tokens before, after, and in the slot)
- Features are extracted from the context of a slot:
  - *token type*: word, number, punctuation, html-tag, ...
  - *formatting*: capitalized, italics, bold, font, ...
  - *location*: after/before line break, paragraph, ...
  - *html structure*: h1, a, href, table, td, center, ...
  - *relative position*: previous token, next token
- Learn Rules:
  - evaluate rules by counting correct matches as positive, wrong matches as negative (e.g., Laplace heuristic)

# Example Systems

- RAPIER (Califf & Mooney, 1997):
  - based in a logic framework (ILP)
  - integrates some NLP (part-of-speech tags)
  - bottom-up learning with *lgg*: select two examples and compute the minimal generalization that covers both
- SRV (Freitag, 1998):
  - uses a large variety of features both for structured and unstructured text
  - top-down rule learning (Ripper-like)
- Expressive, general rule learning systems (e.g., ILP) could be used as well, but would lack domain-specific optimizations

# WHISK (Soderland, 1999)

- multi-slot extraction
- rules represented as perl-like regular expressions
- can handle (semi-)structured and unstructured text
- top-down rule learning with seed instance (AQ-like)
  - choose a random training example
  - start with the most general rule
  - refine the rule using heuristics as in RIPPER-like algorithms (e.g., Laplace accuracy)
  - but only with conditions that appear in the training example
- use of user-specified semantic classes
  - e.g. BEDROOM = {brs|br|bds|bdrm|bd|bedroom|bedrooms|bed}
- integrated with interactive training based on a simple form of active learning

# Example - WHISK

Training example:

<B>Capitol Hill -</B> 1 bedroom twnhme. fplc D/W  
W/D. Undergrnd pkg incl. \$675. 3 BR, 2<sup>nd</sup> flr of  
turn of ctry HOME. incl. gar, grt N. Hill loc  
\$995. (206) 999-9999 <br>

Label:

- Rental:
  - area: Capitol Hill
  - bedrooms: 1
  - price: 675
- Rental:
  - area: Capitol Hill
  - bedrooms: 3
  - price 995

Starting Rule:

\* ( \* ) \* ( \* ) \* ( \* ) \*

Final Rule:

(after seeing several examples):

START<B> ( \* ) ' - ' \* ( DIGIT )  
BEDROOM \* '\$' ( NUMBER ) \*

# Example - WHISK

Training example:

<B>Capitol Hill -</B> 1 bedroom twnhme. fplc D/W  
W/D. Undergrnd pkg incl. \$675. 3 BR, 2<sup>nd</sup> flr of  
turn of ctry HOME. incl. gar, grt N. Hill loc  
\$995. (206) 999-9999 <br>

---

START<B> ( \* ) ' - ' \* ( DIGIT ) BEDROOM \* '\$' ( NUMBER ) \*

BEDROOM = {brs | br | bds | bdrm | bd | bedroom | bedrooms | bed}



# Example - WHISK

Training example:

<B>Capitol Hill -</B> 1 bedroom twnhme. fplc D/W  
W/D. Undergrnd pkg incl. \$675. 3 BR, 2<sup>nd</sup> flr of  
turn of ctry HOME. incl. gar, grt N. Hill loc  
\$995. (206) 999-9999 <br>

START<B> ( \* ) ' - ' \* ( DIGIT ) BEDROOM \* '\$' ( NUMBER ) \*

BEDROOM = {brs | br | bds | bdrm | bd | bedroom | bedrooms | bed}

# Information Extraction as a Classification Problem

- treat each text position (token boundary / token) as a classification example
  - classification is “beginning” or “ending” of annotation
  - features of examples are extracted from the context
  - similar as in inductive rule learning approach
- advantages in comparison to wrappers
  - use of powerful state-of-the-art classification algorithms
  - concentration on the actual task: extraction of useful information (feature generation)
    - no development of specialized algorithms needed

# Problem Transformation: Boundaries

token		The	quick	brown	fox	jumps	over	the	lazy	dog
position		1	2	3	4	5	6	7	8	9
class		NEG	START	NEG	END	NEG	NEG	NEG	NEG	NEG

- boundary classification patterns
  - INSIDE/OUTSIDE
  - BEGIN/END
  - BEGIN/CONTINUE/END
  - BEGIN/CONTINUE/OUTSIDE
- the right choice depends mainly on the type of information
  - length of the annotations, partial results acceptable etc.

# Problem Transformation: Feature Generation

token	The	quick	brown	fox	jumps	over	the	lazy	dog
position	1	2	3	4	5	6	7	8	9
class	NEG	START	NEG	END	NEG	NEG	NEG	NEG	NEG
token features	the=1 +1.quick=1	quick=1 +1.brown=1 -1.the=1	brown=1 +1.fox=1 -1.quick=1	fox=1 +1.jumps=1 -1.brown=1	jumps=1 +1.over=1 -1.fox=1	over=1 +1.the=1 -1.jumps=1	the=1 +1.lazy=1 -1.over=1	lazy=1 +1.dog=1 -1.the=1	dog=1 -1.lazy=1
character patterns	Xxx=1 X+x+=1	xxxxx=1 x+=1	xxxxx=1 x+=1	xxx=1 x+=1	xxxxx=1 x+=1	xxxx=1 x+=1	xxx=1 x+=1	xxxx=1 x+=1	xxx=1 x+=1
history features		-1.NEG=1	-1.START=1 -2.NEG=1	-1.NEG=1 -2.START=1	-1.END=1 -2.NEG=1	-1.NEG=1 -2.END=1	-1.NEG=1 -2.NEG=1	-1.NEG=1 -2.NEG=1	-1.NEG=1 -2.NEG=1
	DT=1	JJ=1	JJ=1	NN=1	NNS=1	IN=1	DT=1	JJ=1	NN=1

- representing the context
  - set-of-words, word patterns (capitalization etc.)
  - presence of formatting, location, html structure
  - part-of-speech, syntactic parsing
- windowing
  - extend of context usually given in number of words
- classification history
  - include preceding classification as feature

# Information Extraction as a Classification Problem

- unbalanced number of pos. and neg. examples
  - specialized algorithms, e.g. perceptrons with uneven margins (Li et al. 2001)
  - two stage process (Finn and Kushmerick 2004)
    - second classifier is trained on the neighborhood of boundaries
    - validates and corrects decisions of first classifier
- large (and sparse) feature space
  - usually SVM are used, which deal very well with large but sparse feature vectors
- state-of-the-art for standard IE tasks
  - e.g. around 90% on common experiments tasks

	ELIE <sub>L1</sub>			ELIE <sub>L2</sub>			BWI			LP <sup>2</sup>			RAPIER			SNoW-IE		
field	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
speaker	95.8	76.2	84.9	91.0	86.0	88.5	79.1	59.2	67.7	87.0	70.0	77.6	80.9	39.4	53.0	83.3	63.3	73.8
location	96.1	75.9	84.8	93.1	80.7	86.5	85.4	69.6	76.7	87.0	66.0	75.1	91.0	60.5	72.7	90.9	64.1	75.2
stime	99.0	94.4	96.6	98.6	98.5	98.5	99.6	99.6	99.6	99.0	99.0	99.0	96.5	95.3	95.9	99.6	99.6	99.6
etime	99.0	77.8	87.0	95.7	97.3	96.4	94.4	94.4	94.4	94.0	97.0	95.5	95.8	96.6	96.2	97.6	95.0	96.3

SVM

37 boosted LR

rule learner

# Information Integration

- Data Integration (Data Warehousing):
  - Join different databases into a single view
  - Problem: Information may be encoded in different ways
- Information Integration:
  - Join information originating from different wrappers
  - Problem: extracted information is still free text
- Example:
  - *Data source 1*: Wrapper for Movie database
  - *Data source 2*: Wrapper Local movie show times
  - *Task*: Generate a page that integrates reviews into the local show times
  - *Problem*: Key relation (movie titles) will not match exactly

# WHIRL (Cohen 1998)

- extension of DATALOG (or SQL) database queries that allows to deal with free text
  - models the information extracted by a wrapper as a relational table
- addresses the problem that
  - wrappers may not be able to extract the exact text
    - e.g., irrelevant information (directors, ratings, actors, etc.) might be extracted with title
  - text may be formulated differently on different Web-Sites
    - e.g., order and/or abbreviations of first, middle and last names
- Approach:
  - uses vector space model to represents textual fields
  - uses *similarity literals* to specify approximate matches

# DATALOG vs. WHIRL

- Hard Queries:
  - items in a join must match exactly
- Items match or do not match
- Return all matches satisfying the query
- Soft Queries:
  - items in a join need only be "similar"
- Use cosine similarity to compute the degree of match  $[0, 1]$
- Return the best matches according to similarity
  - Use efficient A\*-like search to find the  $r$  best matches according to similarity score ( $r$ -materialization)



# WHIRL - Example

- Given two wrapped relations:

- `review(Movie, Review)`
- `showtime(Cinema, Movie, Time)`

- Sample Queries:

- Hard Query (DATALOG):

`showtime(C, M, T) & review(M, R)`

- Soft Query:

`showtime(C, M1, T) & review(M2, R) & M1 ~ M2`

- If the titles of the reviews could not be wrapped:

`showtime(C, M, T) & review(R) & M ~ R`

- Free text queries:

`showtime(C, M1, T) & review(M2, R) & M1 ~ M2 &  
R ~ "excellent comedy with Bruce Willis"`

M1 is similar to M2



# WHIRL - Scoring

- Possible answers  $\Theta$  to queries  $Q$  are scored, i.e., a function  $SCORE(Q, \Theta)$  is computed

- For a regular literal:  $SCORE(B, \Theta) = s$   
if  $B\Theta$  is a ground fact, 0 otherwise  
(usually  $s = 1$ , "degree of belief in the proposition")

- For a similarity literal  $X \sim Y$ :

$$SCORE(X \sim Y, \Theta) = sim(X\Theta, Y\Theta)$$

- Conjunctive Query  $Q = B_1 \& \dots \& B_n$

$$SCORE(Q, \Theta) = \prod_i SCORE(B_i)$$

- A definite clause  $Head :- B_1, B_2, \dots, B_n.$

$$SCORE(Head) = 1 - \prod_i (1 - SCORE(B_i))$$

# Using WHIRL as Text Classifier

- represent labelled training documents in relation  
`train(Document, Class)`
- The following clause returns labels `C` ordered by similarity score of `D` to `D1`  
`classify(D, C) :- train(D1, C), D ~ D1.`
  - NOTE: multiple ground instantiations of the head (i.e, multiple bindings to the head) are combined using the definite clause similarity score!
- very similar to nearest neighbor classification
  - minor differences in combining evidence (similarity score)
- experimentally very competitive to conventional approaches