

---

# Web Content Mining and NLP

---

Bing Liu

Department of Computer Science

University of Illinois at Chicago

liub@cs.uic.edu

<http://www.cs.uic.edu/~liub>

# Introduction

- The Web is perhaps the **single largest and distributed data source** in the world that is easily accessible.
- **Web mining**
  - **Web usage mining**: mine usage logs, web traffics
  - **Web structure mining**: mine hyperlinks and communities.
  - **Web content mining**: mine page contents.
- We focus on ***Web content mining***.
  - **Still a very large topic**. We will **not** discuss traditional tasks: Web page classification, clustering, etc

# Different types of data

- Structured data
  - The data are usually retrieved from backend databases, and
  - displayed in Web pages following some fixed templates.
- Semi-structured data
  - Each page is organized in some way to some extent, usually as a hierarchy of blocks.
- Unstructured data:
  - natural language text

# Roadmap

- Introduction

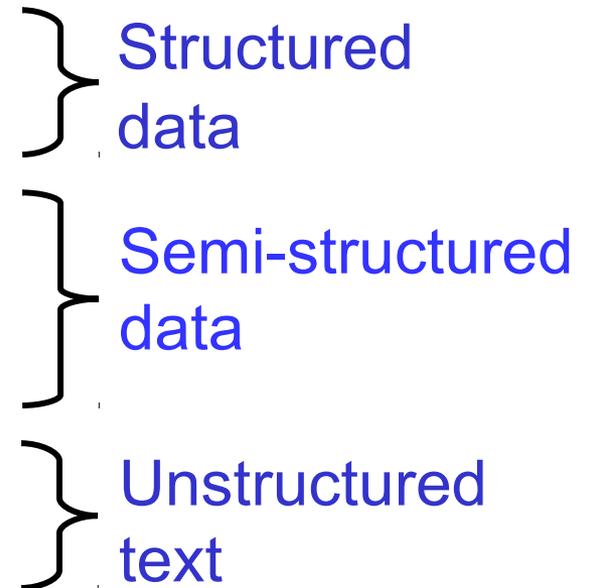
- 1. Structured data extraction**

2. Information integration

3. Information synthesis

4. Opinion mining

- Conclusions

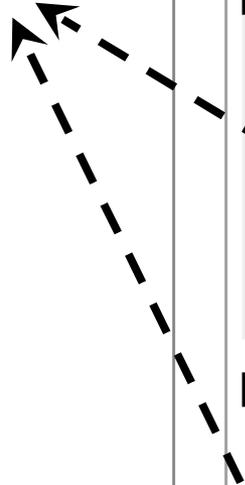


# Structured Data Extraction

- A large amount of information on the Web is contained in **regularly structured data objects**.
  - often data records retrieved from databases.
  - Important: **lists of products and services**.
- **Applications**: Gather data to provide value-added services
  - comparative shopping, object search, opinion mining, etc.
- **Two types of pages with structured data**:
  - **List pages**, and **detail pages**

# List Page – two lists of products

Two lists



CompUSA.com - Product Results - Microsoft Internet Explorer

Address: http://www.compUSA.com/products/products.asp?N=200049&cm\_re=A-\_-HPF-\_-Flat+Panel+%2BLCD%29

Search the Web

### Top Sellers

 <b>EN7410 17-inch LCD Monitor, Black/Dark Charcoal</b> \$299.99 <a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a> Compare » <input type="checkbox"/> <	 <b>17-inch LCD Monitor</b> \$249.99 <a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a> Compare » <input type="checkbox"/> <	 <b>AL1714cb 17-inch LCD Monitor, Black</b> \$269.99 <a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a> Compare » <input type="checkbox"/> <	 <b>SyncMaster 712n 17-inch LCD Monitor, Black</b> Was: \$369.99 <b>\$299.99</b> SAVE \$70 after: \$70.00 mail-in rebate(s) <a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a> Compare » <input type="checkbox"/> <
--	--	---	--

Page 1 of 6: 1 2 3 4 5 6 Next >>

Sort by: Popularity

 <b>EN7410 17-inch LCD Monitor, Black/Dark Charcoal</b> Product Number: 318020 Mfr. Part #: EN7410 Brand: <a href="#">Envision</a>	<b>\$299.99</b>	<a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a>	Compare » <input type="checkbox"/> <
 <b>17-inch LCD Monitor</b> Product Number: 316328 Mfr. Part #: 130611 Brand: <a href="#">Norwood Micro</a>	<b>\$249.99</b>	<a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a>	Compare » <input type="checkbox"/> <
 <b>AL1714cb 17-inch LCD Monitor, Black</b> Product Number: 317993 Mfr. Part #: ET.L1809.031 Brand: <a href="#">Acer</a>	<b>\$269.99</b>	<a href="#">Add To Cart</a> (Delivery / Pick-Up) <a href="#">Penny Shipping</a>	Compare » <input type="checkbox"/> <
 <b>SyncMaster 712n 17-inch LCD Monitor, Black</b> Was: \$369.99			

start | 4 Outlook Express | WWW-05 tutorial | 3 Microsoft PowerP... | CompUSA.com - Prod... | 11:57 AM

# Detail Page - detailed

View Cart | My Account | Order Status | Help

**COMPUSA**

Consumer | **Business** | Services | Auctions | Locations | Gift Cards | Free Shipping\*

Computers & Peripherals | Upgrades | Software | Accessories | Electronics | Games & Movies | Office Supplies | See All »

Search:  **GO!** [Check Out Our Interactive Ad](#)

[CompUSA.com](#) » [Categories](#) » [Electronics](#) » [Digital Photography](#) » [Digital Cameras](#)

## Kodak EasyShare Z730 Digital Camera, 5.0 Megapixels



Brand: Kodak [« Visit their Showcase](#)  
Mfg Part #: 8857963  
SKU: 336442

**Delivery**  Was: \$249.99  
**FREE shipping** **\$179.99** (28% Off)  
SAVE \$70 after: \$70.00 instant savings  
Usually Ships In: 2 - 4 Weeks  
[Estimate Arrival Time](#)

**In-Store**  **\$249.99**  
Ready for Pick-Up In: 15 Minutes  
[Check Store Availability](#)

**Add to Cart**  
[Protect this product \(learn how\)](#)

[▶ Delivery Only Special - pricing not available in store](#)  
[See product info from Kodak](#)

Customer Rating: ★★★★★ 4.6 out of 5  
[Read all 57 reviews](#) [Rate this product](#)

[Overview](#) | [Tech Specs](#) | [Add-Ons](#) | [Rebate Info](#) | [Ratings / Reviews](#) | [Add to Wishlist](#) | [Print](#) | [E-Mail](#) | [Compare](#)

Overview for Kodak EasyShare Z730 Digital Camera, 5.0 Megapixels  
(Based on manufacturer's information)

Imagine. Invent. Inspire.

# Extraction Task: an illustr



## Cabinet Organizers by Copco

9-in. [Round Turntable: White](#) ★★★★★

\$4.95 [BUY](#)

12-in. [Round Turntable: White](#) ★★★★★

\$7.95 [BUY](#)



## Cabinet Organizers

14.75x9 [Cabinet Organizer \(Non-skid\): White](#) ★★★★★

\$7.95 [BUY](#)



## Cabinet Organizers

22x6 [Cookware Lid Rack](#) ★★★★★

\$19.95 [BUY](#)

nesting

image 1	Cabinet Organizers by Copco	9-in.	Round Turntable: White	*****	\$4.95
image 1	Cabinet Organizers by Copco	12-in.	Round Turntable: White	*****	\$7.95
image 2	Cabinet Organizers	14.75x9	Cabinet Organizer (Non-skid): White	*****	\$7.95
image 2	Cabinet Organizers	22x6	Cookware Lid Rack	*****	\$19.95

# Data Model and Solution

## Web data model: Nested relations

- See formal definitions in (Grumbach and Mecca, ICDT-99; Liu, Web Data Mining 2006)

## Solve the problem

- Two main types of techniques
  - Wrapper induction – supervised
  - Automatic extraction – unsupervised
- Information that can be exploited
  - Source files (e.g., Web pages in HTML)
    - Represented as strings or trees
  - Visual information (e.g., rendering information)

# Tree and Visual information

1.  [Apple iBook Notebook M8600LL/A \(600-MHz PowerPC G3, 128 MB RAM, 20 GB hard drive\)](#)

Buy new: **\$1,194.00**  
Usually ships in 1 to 2 days

Customer Rating:  
★★★★☆

Best use: ( <a href="#">what's this?</a> )	Business: ●●●●○	Portability: ●●●●○	Desktop Replacement: ●●●●○	Entertainment: ●●●●○
--	-----------------	--------------------	----------------------------	----------------------

600 MHz PowerPC G3, 128 MB SRAM, 20 GB Hard Disk, 24x CD-ROM, AirPort ready, and Mac OS X, Mac OS X, Mac OS 9.2, Quick Time, iPhoto, iTunes 2, iMovie 2, AppleWorks, Microsoft IE

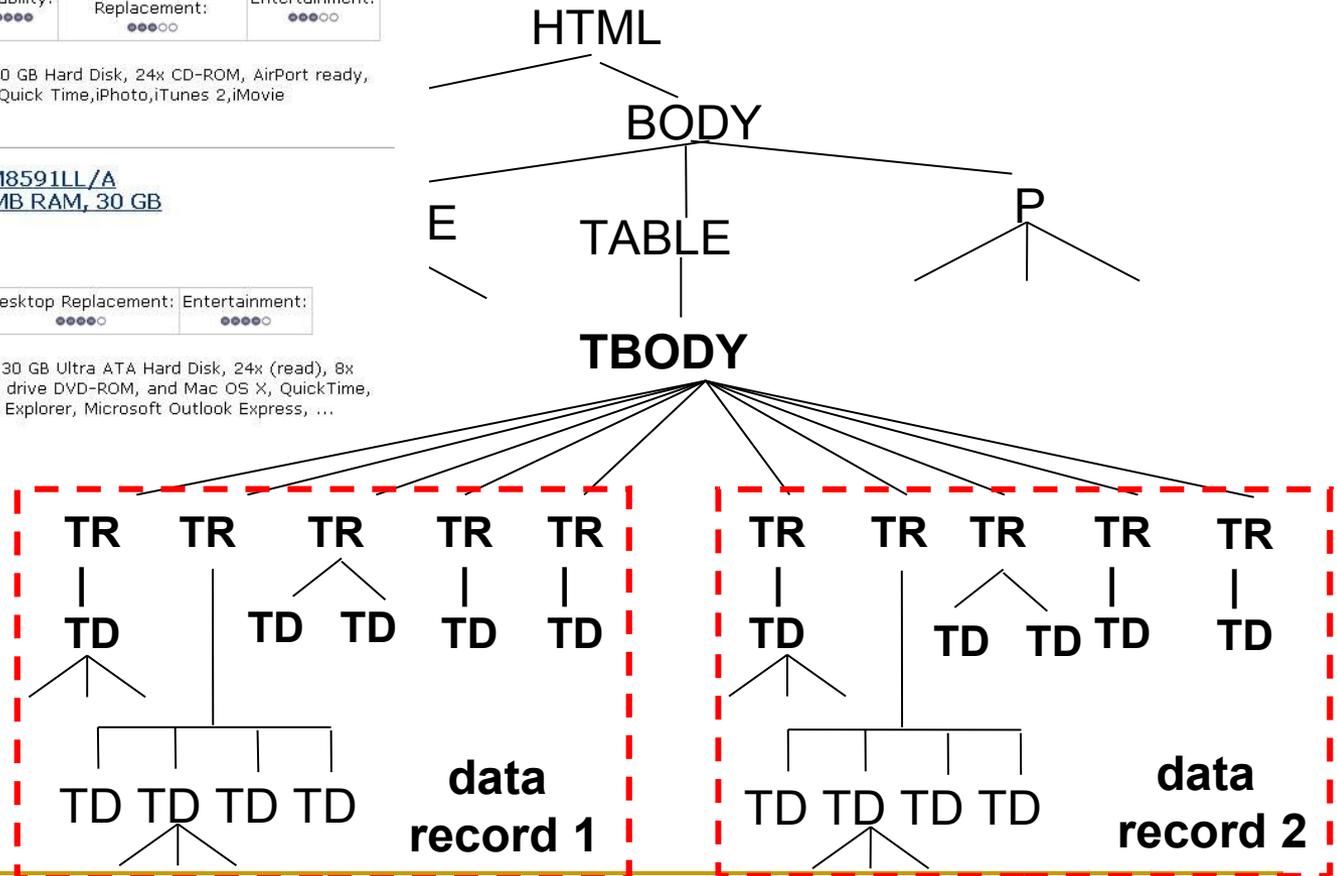
2.  [Apple Powerbook Notebook M8591LL/A \(667-MHz PowerPC G4, 256 MB RAM, 30 GB hard drive\)](#)

Buy new: **\$2,399.99**

Customer Rating:  
★★★★☆

Best use: ( <a href="#">what's this?</a> )	Portability: ●●●●○	Desktop Replacement: ●●●●○	Entertainment: ●●●●○
--	--------------------	----------------------------	----------------------

667 MHz PowerPC G4, 256 MB SDRAM, 30 GB Ultra ATA Hard Disk, 24x (read), 8x (write) CD-RW, 8x; included via combo drive DVD-ROM, and Mac OS X, QuickTime, iMovie 2, iTunes(6), Microsoft Internet Explorer, Microsoft Outlook Express, ...



# Wrapper Induction (Muslea et al., Agents-99)

- Using machine learning to generate extraction rules.
  - The user marks the target items in a few training pages.
  - The system learns extraction rules from these pages.
  - The rules are applied to extract items from other pages.

## Training Examples

E1: 513 Pico, <b>Venice</b>, Phone 1-<b>800</b>-555-1515

E2: 90 Colfax, <b>Palms</b>, Phone (800) 508-1570

E3: 523 1<sup>st</sup> St., <b>LA</b>, Phone 1-<b>800</b>-578-2293

E4: 403 La Tijera, <b>Watts</b>, Phone: (310) 798-0008

## Output Extraction Rules

- Start rules: End rules:
  - R1: SkipTo(()) SkipTo()
  - R2: SkipTo(-<b>) SkipTo(</b>)

# Automated extraction

**There are two main problem formulations:**

**Problem 1:** Extraction based on a single list page (Liu et al., KDD-03; Liu, 2006)

**Problem 2:** Extraction based on multiple input pages of the same type (list pages or detail pages) (Grumbach and Mecca, ICDT-99).

- Problem 1 is more general: Algorithms for solving Problem 1 can solve Problem 2.
  - Thus, we only discuss Problem 1.

# Automatic Extraction: Product

Data region1

Data records

Data region2

CompUSA.com - Product Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address [http://www.compusa.com/products/products.asp?N=200049&cm\\_re=A-\\_-HPF-\\_-Flat+Panel+%28LCD%29](http://www.compusa.com/products/products.asp?N=200049&cm_re=A-_-HPF-_-Flat+Panel+%28LCD%29)

Search the Web

**Top Sellers**

Product Name	Price	Action
EN7410 17-inch LCD Monitor, Black/Dark Charcoal	\$299.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
17-inch LCD Monitor	\$249.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
AL1714cb 17-inch LCD Monitor, Black	\$269.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
SyncMaster 712n 17-inch LCD Monitor, Black	Was: \$369.99 \$299.99 SAVE \$70 after: \$70.00 mail-in rebate(s)	Add To Cart (Delivery / Pick-Up) Penny Shipping

Page 1 of 6: 1 2 3 4 5 6 Next >>

Sort by: Popularity Compare

Product Name	Price	Action
EN7410 17-inch LCD Monitor, Black/Dark Charcoal Product Number: 318020 Mfr. Part #: EN7410 Brand: <u>Erison</u>	\$299.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
17-inch LCD Monitor Product Number: 316328 Mfr. Part #: 13061 Brand: <u>Norwood Micro</u>	\$249.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
AL1714cb 17-inch LCD Monitor, Black Product Number: 317993 Mfr. Part #: FT.L1809.031 Brand: <u>Acer</u>	\$269.99	Add To Cart (Delivery / Pick-Up) Penny Shipping
SyncMaster 712n 17-inch LCD Monitor, Black Was: \$369.99		

start | 4 Outlook Express | WWW-05 tutorial | 3 Microsoft PowerP... | CompUSA.com - Prod... | 11:57 AM

# Solution Techniques (Liu et al. KDD-2003)

- Identify data regions and data records: **by finding repeated patterns**
  - string matching
    - treat HTML source as a string
  - tree matching
    - treat HTML source as a tree
- Align data items: Multiple alignment
  - Align items in more than two data records

# String edit distance (definition)

Assume we are given two strings  $s_1$  and  $s_2$ . The following recurrence relations define the edit distance,  $d(s_1, s_2)$ , of two strings  $s_1$  and  $s_2$ :

$$d(\varepsilon, \varepsilon) = 0 \quad // \varepsilon \text{ represents an empty string}$$

$$d(s, \varepsilon) = d(\varepsilon, s) = |s| \quad // |s| \text{ is the length of string } s$$

$$d(s_1+ch_1, s_2+ch_2) = \min(d(s_1, s_2) + r(ch_1, ch_2), d(s_1+ch_1, s_2) + 1, d(s_1, s_2+ch_2) + 1)$$

where  $ch_1$  and  $ch_2$  are the last characters of  $s_1$  and  $s_2$  respectively, and  $r(ch_1, ch_2) = 0$  if  $ch_1 = ch_2$ ;  $r(ch_1, ch_2) = 1$ , otherwise.

# An example

**Example 1:** We want to compute the edit distance and find the alignment of the following two strings:

$s_1$ : X G Y X Y X Y X  
 $s_2$ : X Y X Y X Y T X

- The edit distance matrix and back trace path

- alignment

$s_1$ : X G Y X Y X Y - X  
 $s_2$ : X - Y X Y X Y T X

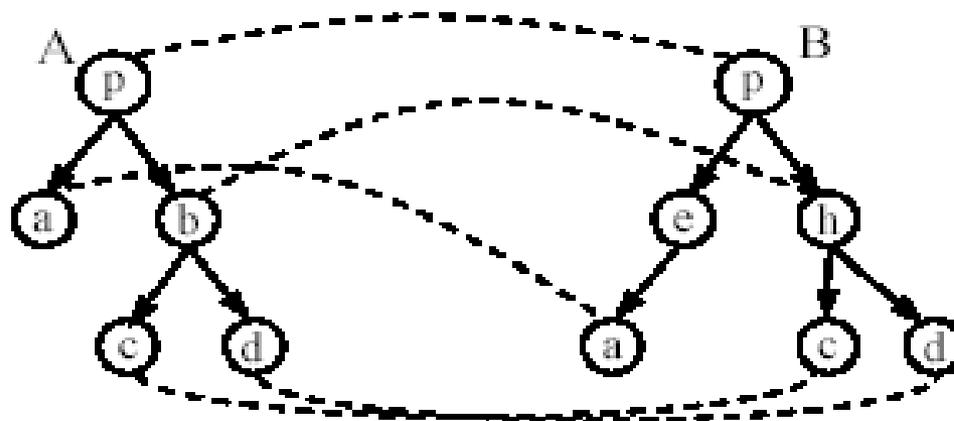
	$s_1$	X	G	Y	X	Y	X	Y	X
$s_2$	0	1	2	3	4	5	6	7	8
X	1	0	1	2	3	4	5	6	7
Y	2	1	1	1	2	3	4	5	6
X	3	2	2	2	1	2	3	4	5
Y	4	3	3	2	2	1	2	3	4
X	5	4	4	3	2	2	1	2	3
Y	6	5	5	4	3	2	2	1	2
T	7	6	6	5	4	3	3	2	2
X	8	7	7	6	5	4	3	3	2

# Tree edit distance or tree

Let  $X$  be a tree and let  $X[i]$  be the  $i$ th node of tree  $X$  in a preorder walk of the tree. A *mapping*  $M$  between a tree  $A$  of size  $n_1$  and a tree  $B$  of size  $n_2$  is a set of ordered pairs  $(i, j)$ , one from each tree, satisfying the following conditions for all  $(i_1, j_1), (i_2, j_2) \in M$ :

- (1)  $i_1 = i_2$  iff  $j_1 = j_2$ ;
- (2)  $A[i_1]$  is on the left of  $A[i_2]$  iff  $B[j_1]$  is on the left of  $B[j_2]$ ;
- (3)  $A[i_1]$  is an ancestor of  $A[i_2]$  iff  $B[j_1]$  is an ancestor of  $B[j_2]$ .

Intuitively, the definition requires that each node appears no more than once in a mapping and the order among siblings and the hierarchical relation among nodes are both preserved. Fig. 16 shows a mapping example.



# Simple Tree Matching (Liu, Web Data Mining 2006)

- Let  $A = R_A: \langle A_1, \dots, A_k \rangle$  and  $B = R_B: \langle B_1, \dots, B_n \rangle$  be two trees, where  $R_A$  and  $R_B$  are their roots

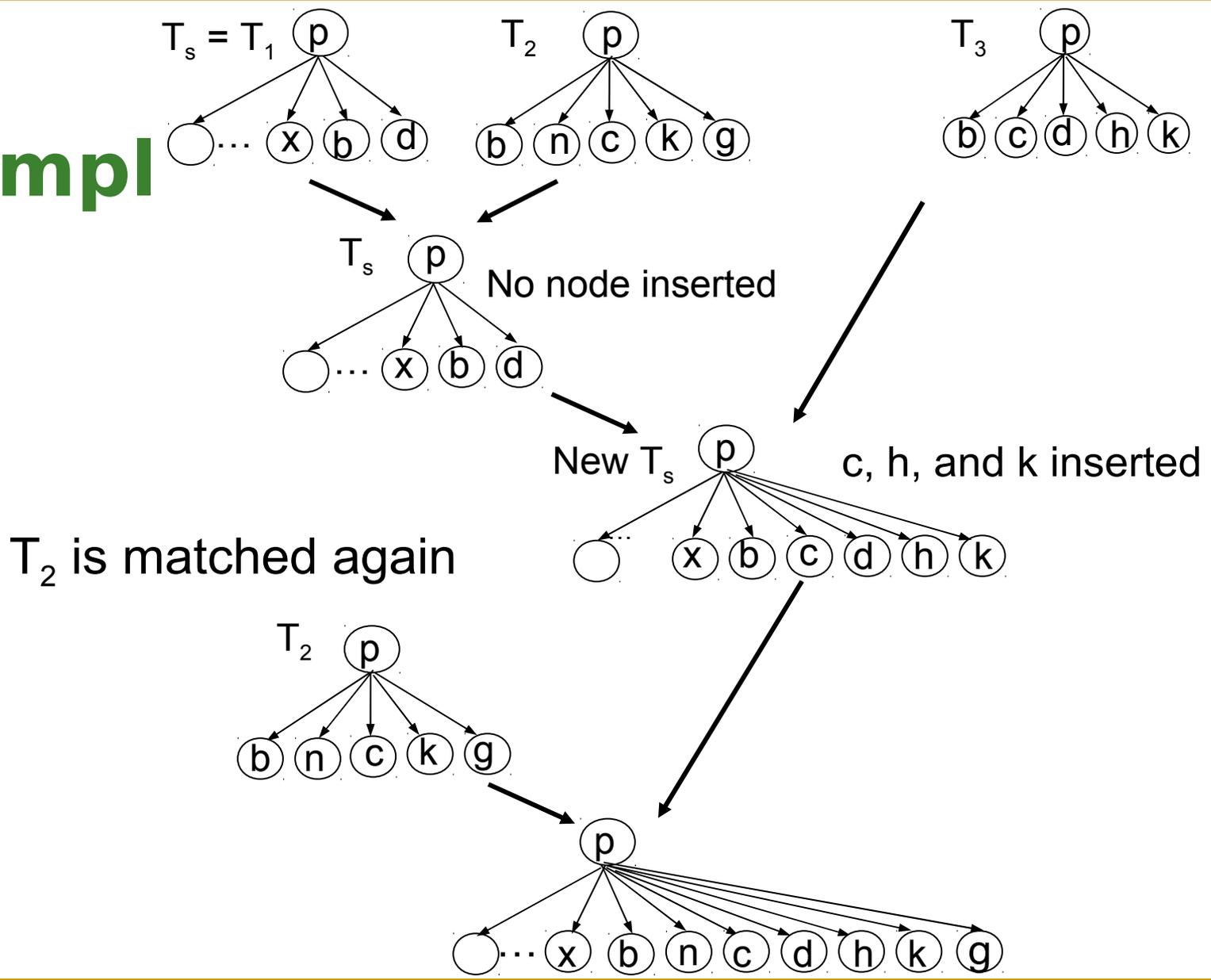
$$W(A, B) = \begin{cases} 0 & \text{if } R_A \neq R_B \\ m(\langle A_1, \dots, A_k \rangle, \langle B_1, \dots, B_n \rangle) + 1 & \text{otherwise} \end{cases}$$

$$\begin{aligned} m(\langle \rangle, \langle \rangle) &= 0 && // \langle \rangle \text{ represents an empty sub-tree list.} \\ m(s, \langle \rangle) = m(\langle \rangle, s) &= 0 && // s \text{ matches any non-empty sub-tree list} \\ m(\langle A_1, \dots, A_k \rangle, \langle B_1, \dots, B_n \rangle) &= \max(m(\langle A_1, \dots, A_{k-1} \rangle, \langle B_1, \dots, B_{n-1} \rangle) + W(A_k, B_n), \\ & \quad m(\langle A_1, \dots, A_k \rangle, \langle B_1, \dots, B_{n-1} \rangle), \\ & \quad m(\langle A_1, \dots, A_{k-1} \rangle, \langle B_1, \dots, B_n \rangle)). \end{aligned}$$

# Multiple alignment

- Pairwise alignment is not sufficient because a web page usually contain more than two data records.
- We need multiple alignment.
- There are many existing techniques, e.g.,
  - **Partial tree alignment**. It iteratively match all trees. In each pairwise matching, only match those nodes that can be matched (Zhai and Liu WWW-05).
  - It is a least commitment approach

# An example



# Roadmap

- Introduction

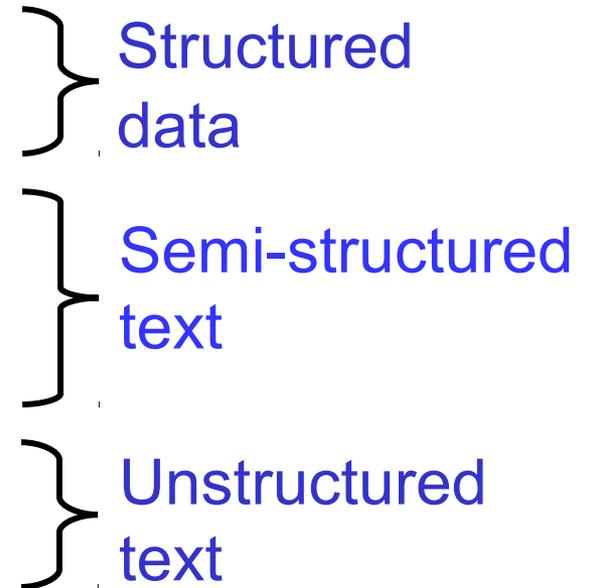
1. Structured data extraction

- 2. Information integration**

3. Information synthesis

4. Opinion mining

- Conclusions



# Information Integration

- The extracted data from different sites need to be **integrated** to produce a consistent database.
- **Integration** means:
  - **Schema match**: match columns in different data tables (e.g., product names).
  - **Data instance match**: match values, e.g., “Coke” = “Coca Cola”?
- Unfortunately, not much research has been done so far in this extraction context.
- Much of the research has been focused on the integration of **Web query interfaces**

# Web Query Interface Integration

(Wu et al., SIGM)



04: Discount at a Global Query Interface

**1 Where and when do you want to travel?**

Leaving from:

Going to:

Departing: (MMDDYY)     Anytime

Returning: (MMDDYY)     Anytime

**2 Who is going on this trip?**

1  Adults (age 19 to 64)

0  Seniors (age 65 and over)

0  Children (age 18 and under)

**3 Do you have any preferences?**

Airline:  No Preference

Class:  Economy / Coach



From:  To:

Round trip  One way  Multi city

Departure date:  Feb  19  Morning

Return:  Feb  19  Morning

Search by  Schedule  Price

Passengers:  1   Check

Depart City:

Destination City:

Depart Date:  May  15  2004  SAT

Return Date:  May  17  2004  MON

Passengers:  1  Adult  0  Child (Age 2 to 11)

Class:  Economy

Leaving from:  Departure date:  Feb  26  Time:  10am

Going to:  Return date:  Mar  05  Time:  10am

Passengers:  1  Preferred cabin:  Economy/Coach

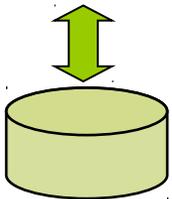
From:  find a city name To:  find a city name

enter airport or city  enter airport or city

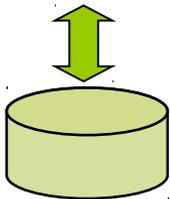
Departure Date:  Feb  27  Return Date:  Mar  2

No same-day returns.

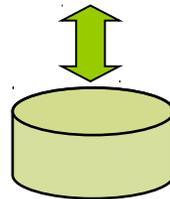
Number of Tickets:  1   START



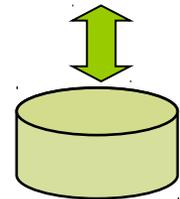
united.com



airtravel.com



delta.com



hotwire.com

# Constructing global query

- A unified query interface:
  - **Conciseness** - Combine semantically similar fields over source interfaces
  - **Completeness** - Retain source-specific fields
  - **User-friendliness** – Highly related fields are close together

1 Where and when do you want to travel?

Leaving from:

Going to:

Departing: (MMDD/YY)   Anytime

Returning: (MMDD/YY)   Anytime

2 Who is going on this trip?

1  Adults (age 19 to 64)

0  Seniors (age 65 and over)

0  Children (age 18 and under)

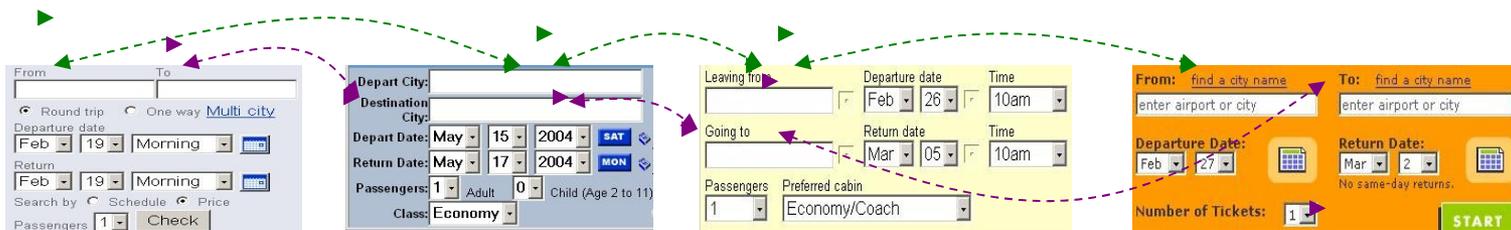
3 Do you have any preferences?

Airline  Class

No Preference  Economy / Coach

- Two-phrased integration

- **Interface Matching** – Identify semantically similar fields

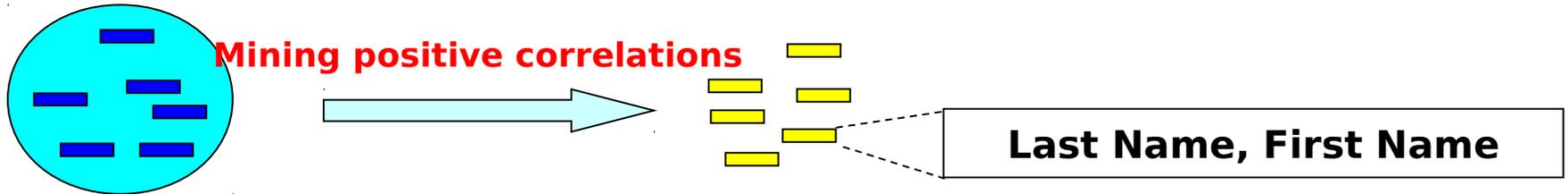


- **Interface Integration** – Merge the source query interfaces

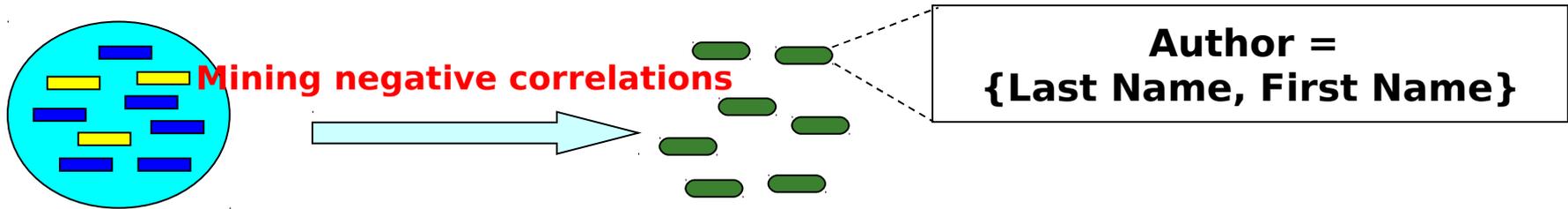
# Schema Matching as Correlation Mining (He and Chang, KDD-04)

- This technique needs a large number of input query interfaces.
- Synonym attributes are **negatively correlated**
  - they are alternatives, *rarely co-occur*.
  - *e.g.*, Author = writer
- Group attributes have **positive correlation**
  - they *often co-occur* in query interfaces
  - *e.g.*, {Last Name, First Name}

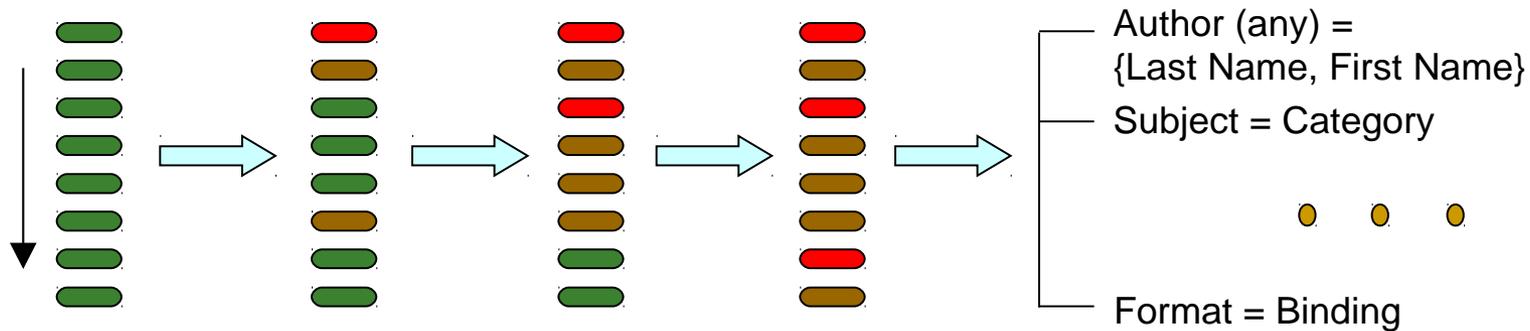
## Positive correlation mining as potential groups



## Negative correlation mining as potential matchings



## Matching selection as model construction



# A clustering approach to schema matching (Wu et al. SIGMOD 2004)

- Hierarchical modeling
- Bridging effect
  - “a2” and “c2” might not look similar themselves but they might both be similar to “b3”
- 1:m mappings
  - Aggregate and is-a types
- User interaction helps in:
  - learning of matching thresholds
  - resolution of uncertain mappings

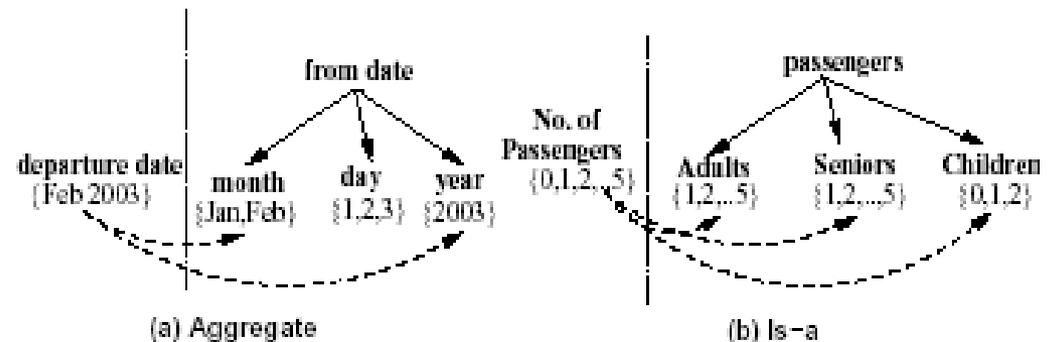
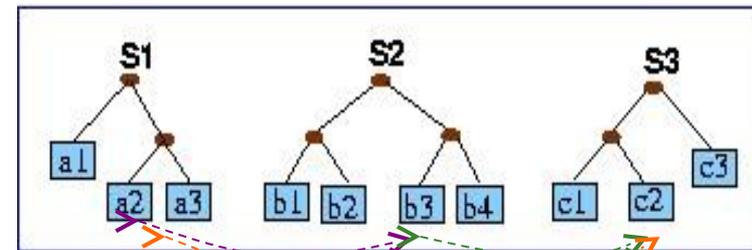
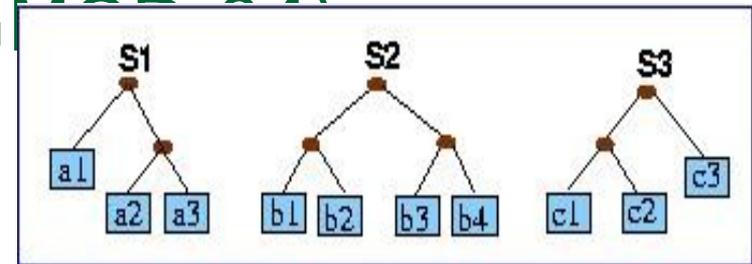
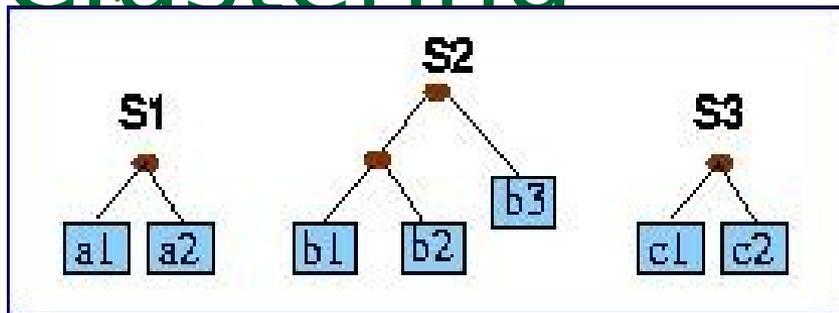


Figure 3: 1:m mappings

# Find 1:1 Mappings via

## Clustering

Interfaces



Initial similarity matrix:

	a1	a2	b1	b2	b3	c1	c2
a1	0	0	.9	0	.85	0	0
a2	0	0	.15	0	0	0	0
b1	.9	.15	0	0	0	.8	0
b2	0	0	0	0	0	.1	.6
b3	.85	0	0	0	0	0	0
c1	0	0	0	.1	0	0	0
c2	0	0	.8	.6	0	0	0

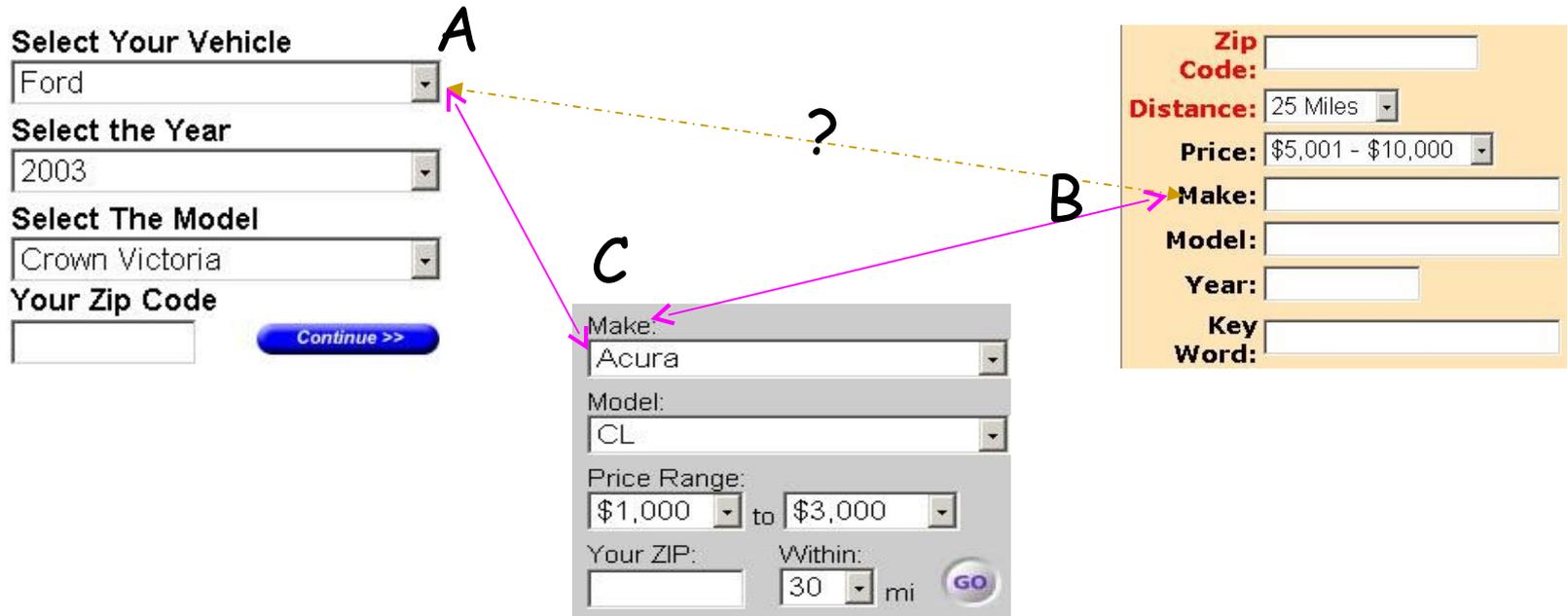
After one merge:

	a2	b2	b3	c1	c2	{a1, b1}
a2	0	0	0	0	0	0
b2	0	0	0	.1	.6	0
b3	0	0	0	0	0	0
c1	0	.1	0	0	0	.8
c2	0	.6	0	0	0	0
{a1, b1}	0	0	0	.8	0	0

- Similarity functions
  - linguistic similarity
  - domain similarity

..., final clusters:  $\{\{a1, b1, c1\}, \{b2, c2\}, \{a2\}, \{b3\}\}$

# “Bridging” Effect



## Observations:

- It is difficult to match “vehicle” field, A, with “make” field, B
- But A’s instances are similar to C’s, and C’s label is similar to B’s
- Thus, C might serve as a “bridge” to connect A and B!

Note: Connections might also be made via labels

# Complex Mappings

**From:\***

**To:\***

**Depart:** 12 March 2003 Morning

**Return:** 26 March 2003 Morning

**Class:** Economy

**Flight type:** Roundtrip

**Travellers:** 1 adult 0 child

**Depart City:**

**Destination City:**

**Depart Date:** May 15 2004 SAT

**Return Date:** May 17 2004 MON

**Passengers:** 1 Adult 0 Child (Age 2 to 11)

**Class:** Economy

**Aggregate type** – contents of fields on the **many** side are part of the content of field on the **one** side

**Commonalities** – (1) field proximity, (2) parent label similarity, and (3) value characteristics

# Complex Mappings (Cont'd)

A screenshot of a flight search form with a yellow background. The form contains the following fields: 'Leaving from' (text input), 'Departure date' (month: Feb, day: 26), 'Time' (10am), 'Going to' (text input), 'Return date' (month: Mar, day: 05), 'Time' (10am), 'Passengers' (dropdown: 1), and 'Preferred cabin' (dropdown: Economy/Coach). A red arrow points from the 'Passengers' dropdown to the 'Passengers' field in the second screenshot.

A screenshot of a flight search form with a blue background. The form contains the following fields: 'Depart City:' (text input), 'Destination City:' (text input), 'Depart Date:' (month: May, day: 15, year: 2004, day of week: SAT), 'Return Date:' (month: May, day: 17, year: 2004, day of week: MON), 'Passengers:' (dropdown: 1, type: Adult), and 'Class:' (dropdown: Economy). A red arrow points from the 'Passengers' dropdown in the first screenshot to the 'Passengers' field in this screenshot. Another red arrow points from the 'Preferred cabin' dropdown in the first screenshot to the 'Class:' dropdown in this screenshot.

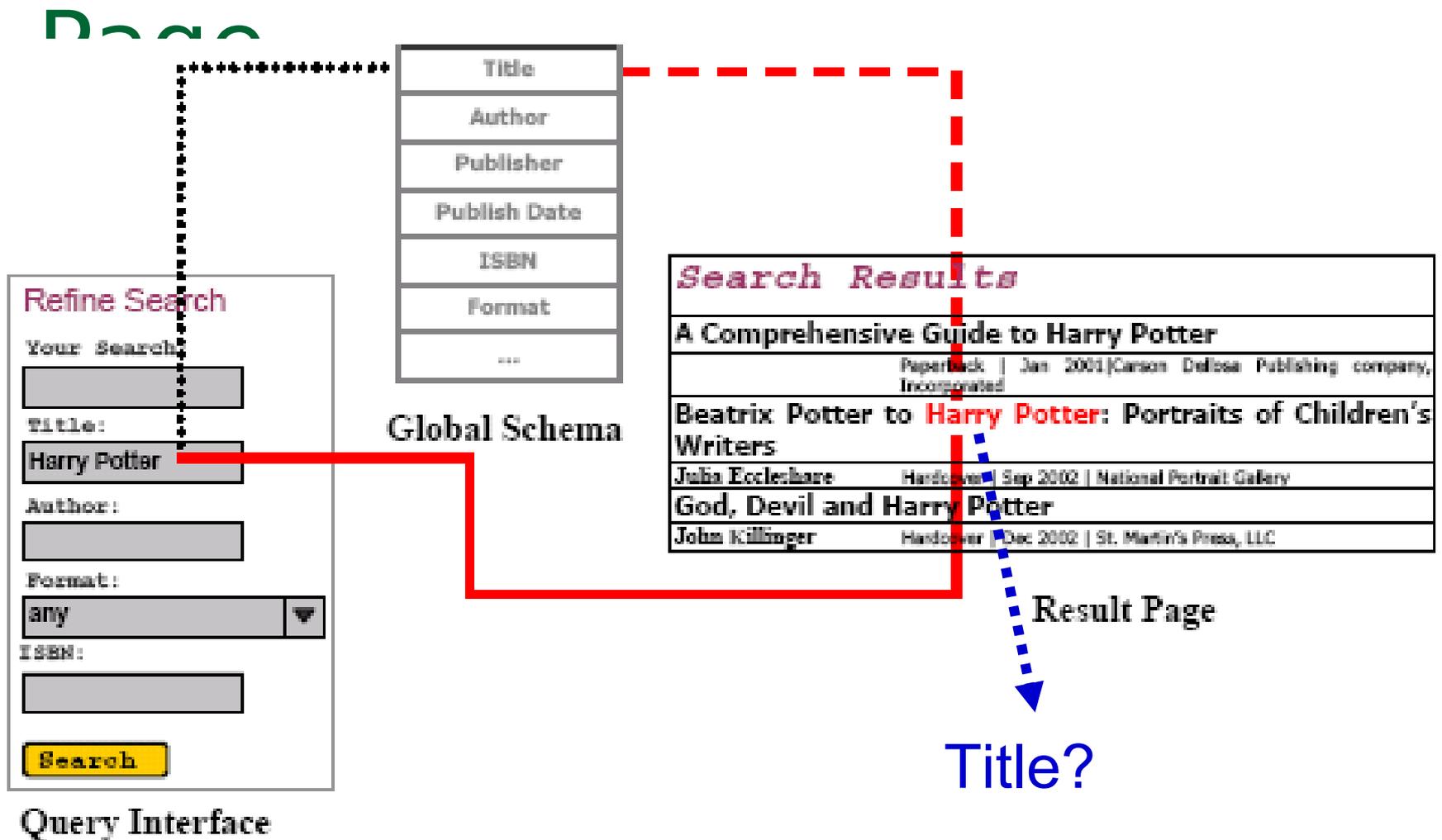
**Is-a type** – contents of fields on the **many** side are sum/union of the content of field on the **one** side

**Commonalities** – (1) field proximity, (2) parent label similarity, and (3) value characteristics

# Instance-Based Matching via Query Probing (Wang et al., VLDB-04)

- Both query interfaces and returned results (instances) are considered in matching.
  - Assumption: A global schema (GS) and a set of instances are given.
  - The method uses each instance value (IV) of every attribute in GS to probe the underlying database to obtain the count of IV appeared in the returned results.
  - These counts are used to help matching.

# Query Interface and Result Page



# The core problem

- Recognizing domain specific synonyms
  - Words
  - Phrases
  - Other general expressions
- An NLP problem!
- Existing methods exploited both linguistic and semi-structured information in Web pages.

# Roadmap

- Introduction

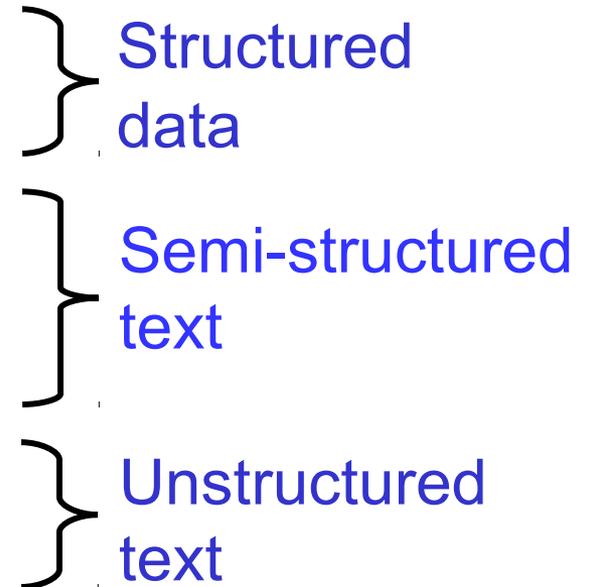
1. Structured data extraction

2. Information integration

- 3. Information synthesis**

4. Opinion mining

- Conclusions



# Information/knowledge synthesis

- **Web search paradigm:**
  - Given a query, a few words
  - A search engine returns a ranked list of pages.
  - The user then browses and reads the top-ranked pages to find what s/he wants.
- Sufficient for **navigational queries**
  - if one is looking for a specific piece of information, e.g., homepage of a person, a paper.
- Not sufficient for **informational queries**
  - open-ended research or exploration

# Information synthesis: a growing trend

- **Problems with individual pages:**
  - Bias
  - incompleteness
- **A growing trend among web search engines:** go beyond the traditional paradigm of presenting a list of ranked pages to provide **more varied**, and **comprehensive** information about a search topic.
- To provide **unbiased** and more **complete** info:
  - Find and integrate related bits and pieces:
  - **Information synthesis!**

# Bing search of “cell phone”

The image shows a screenshot of a Bing search results page for the query "cell phone". At the top, there are navigation links for Web, Images, Videos, Shopping, News, Maps, More, MSN, and Hotmail. The Bing logo is on the left, and the search bar contains "cell phone" with a magnifying glass icon. Below the search bar, the page is divided into a left sidebar and a main content area. The sidebar has a "CELL PHONE" header and a list of categories: Shopping, Cell Phone Brands, Cell Phone Buying Guide, Cell Phone Providers, Cell Phone Plans, Cell Phone Accessories, Reference, News, Videos, and Local Listings. Below these are "RELATED SEARCHES" for Verizon Cell Phones, Cingular Cell Phones, Motorola Cell Phone, and Sprint Cell Phones. The main content area shows "ALL RESULTS" for "1-20 of 149,000,000 results" with an "Advanced" link. The first result is a sponsored site for "Cell phone" from eBay.com, offering 8% off with PayPal. Other results include Nextel Phones from Sprint.com, Free Cell Phones from Wirefly.com, and the AT&T Official Site. A "Shop for cell phone" section lists top brands (Motorola, LG Electronics, Nokia, Aftermarket Group, Samsung) and price ranges. Below that, there are links to Cnet.com, Howstuffworks.com, and Consumerreports.org. The final result is a link to "Cell Phones, Cellular Phone Plans, Prepaid Cell Phones, Free Cell Phones ..." from T-Mobile.com, with a "Cached page" link. At the bottom, there is a link to "Mobile phone - Wikipedia, the free encyclopedia" and a snippet from "The Cell Phone Reader: Essays in Social Transformation, 2006; Kopomaa, Timo. The City in Your".

Web Images Videos Shopping News Maps More MSN Hotmail

bing cell phone

CELL PHONE

ALL RESULTS 1-20 of 149,000,000 results - [Advanced](#)

Shopping

Cell Phone Brands

Cell Phone Buying Guide

Cell Phone Providers

Cell Phone Plans

Cell Phone Accessories

Reference

News

Videos

Local Listings

RELATED SEARCHES

[Verizon Cell Phones](#)

[Cingular Cell Phones](#)

[Motorola Cell Phone](#)

[Sprint Cell Phones](#)

[Cell phone](#) - [www.eBay.com](#) Bing cashback Sponsored sites  
Buy **Cell phone**. You may get 8% off with PayPal if eligible.

[Nextel Phones](#) - [www.sprint.com](#)  
See How Work Gets Done. Now. Only W/ The Sprint Nextel Network!

[Free Cell Phones](#) - [www.Wirefly.com](#)  
The Leading Site For **Phones & Plans** - AT&T, Sprint, Verizon, T-Mobile

[AT&T Official Site](#) - [att.com/wireless](#)  
Huge Selection Of Free **Cell Phones** w/ Plans At \$39.99. Visit AT&T Now.

[Shop for cell phone](#)

**Top brands** - [Motorola](#) - [LG Electronics](#) - [Nokia](#) - [Aftermarket Group](#) - [Samsung](#) - [More...](#)  
**Price** - [below \\$20](#) - [\\$20-\\$150](#) - [above \\$150](#) - Bing cashback  
**Guides** - [Cnet.com](#) - [Howstuffworks.com](#) - [Consumerreports.org](#)

[Cell Phones, Cellular Phone Plans, Prepaid Cell Phones, Free Cell Phones ...](#)  
T-Mobile: **Cell phones** you love, plans you want. Offering the best deals on cellular **phone** service, prepaid **cell phones**, **cell phone** accessories, free **cell phones & more!**  
[www.t-mobile.com](#) - [Cached page](#)

[Mobile phone - Wikipedia, the free encyclopedia](#)  
The **Cell Phone** Reader: Essays in Social Transformation, 2006; Kopomaa, Timo. The City in Your

# Mining a book (Liu et al WWW-2003, Nitin et al, coming)

- Traditionally, when one wants to learn about a topic,
  - one reads a book or a survey paper.
- Learning in-depth knowledge of a topic from the Web is becoming increasingly popular.
  - Web's convenience,
  - richness of information and diversity
  - For emerging topics, it may be essential - no book.
- Can we help such learning by mining “a book” from the Web given a topic?
  - Knowledge in a book is well organized:
    - Table of Contents
    - Detailed description pages

# An example

- Given the topic “data mining”, can the system produce the following, a concept hierarchy?
  - Classification
    - Decision trees
      - ... (Web pages containing the descriptions of the topic)
    - Naïve Bayes
      - ...
    - ...
  - Clustering
    - Hierarchical
    - Partitioning
    - K-means
    - ....
  - Association rules
  - Sequential patterns
  - ...

# Exploiting information redundancy

- **Web information redundancy**: many Web pages contain similar information.
- **Observation 1**: If some phrases are mentioned in a number of pages, they are likely to be important concepts or sub-topics of the given topic.
- This means that we can use data mining to find concepts and sub-topics:
  - What are candidate words or phrases that may represent concepts of sub-topics?

# Each Web page is already organized

- **Observation 2:** The contents of most Web pages are already organized.
  - Different levels of headings
  - Emphasized words and phrases
- They are indicated by various HTML emphasizing tags, e.g., <H1>, <H2>, <H3>, <B>, <I>, etc.
- We utilize existing page organizations to find a global organization of the topic.
  - Cannot rely on only one page because it is often incomplete, and mainly focus on what the page authors are familiar with or are working on.

# Using language patterns to find sub-topics

- Certain syntactic language patterns express some relationship of concepts.
- The following patterns represent hierarchical relationships, concepts and sub-concepts:
  - *Such as*
  - *For example (e.g.,)*
  - *Including*
- E.g., “*There are many clustering techniques (e.g., hierarchical, partitioning, k-means, k-medoids).*”

# PANKOW (Cimiano, et al WWW-04) and KnowItAll (Etzioni et al WWW-04)

- Linguistic patterns, first 4 from (Hearst SIGIR-92):

1: <concept>s such as <instance>

2: such <concepts>s as <instance>

3: <concepts>s, (especially | including)<instance>

4: <instance> (and | or) other <concept>s

5: the <instance> <concept>

6: the <concept> <instance>

7: <instance>, a <concept>

8: <instance> is a <concept>

.....

# Put them together

1. Crawl the set of pages (a set of given documents)
2. Identify important phrases using
  1. HTML emphasizing tags, e.g., `<h1>`, ..., `<h4>`, `<b>`, `<strong>`, `<big>`, `<i>`, `<em>`, `<u>`, `<li>`, `<dt>`.
  2. Language patterns.
3. Perform data mining (frequent itemset mining) to find frequent itemsets (**candidate concepts**)
  - Data mining can weed out peculiarities of individual pages to find the essentials.
1. Eliminate unlikely itemsets (using heuristic rules).
2. Rank the remaining itemsets, which are main concepts.

# Additional techniques

- Segment a page into different sections.
  - Find sub-topics/concepts only in the appropriate sections.
- Mutual reinforcements:
  - Using sub-concepts search to help each other
- ...
- Finding definition of each concept using **syntactic patterns (again)**
  - {is | are} [*adverb*] {called | known as | defined as} {*concept*}
  - {*concept*} {refer(s) to | satisfy(ies)} ...
  - {*concept*} {is | are} [*determiner*] ...
  - {*concept*} {is | are} [*adverb*] {being used to | used to | referred to | employed to | defined as | formalized as | described as | concerned with | called} ...

## Data Mining

Clustering  
Classification  
Data Warehouses  
Databases  
Knowledge Discovery  
Web Mining  
Information Discovery  
Association Rules  
Machine Learning  
Sequential Patterns

## Web Mining

Web Usage Mining  
Web Content Mining  
Data Mining  
Webminers  
Text Mining  
Personalization  
Information Extraction  
Semantic Web Mining  
XML  
Mining Web Data

# Some concepts extraction results

## Classification

Neural networks  
Trees  
Naive bayes  
Decision trees  
K nearest neighbor  
Regression  
Neural net  
Sliq algorithm  
Parallel algorithms  
Classification rule learning  
ID3 algorithm  
C4.5 algorithm  
Probabilistic models

## Clustering

Hierarchical  
K means  
Density based  
Partitioning  
K medoids  
Distance based methods  
Mixture models  
Graphical techniques  
Intelligent miner  
Agglomerative  
Graph based algorithms

---

# The core problems

- Recognize key concepts in a domain
- Discover their relationships
  - Manly hierarchical relations
- Recognize domain specific synonyms
  
- Existing methods exploit structures or organizations in a page and language patterns.

# Roadmap

- Introduction

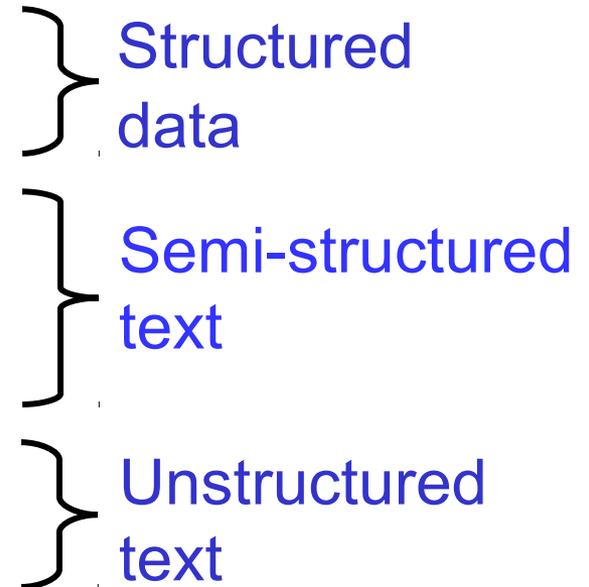
1. Structured data extraction

2. Information integration

3. Information synthesis

- 4. Opinion mining**

- Conclusions



# Opinion mining

- We now move to unstructured text on the Web.
- A major Web content mining research is to extract specific types of information from text in Web pages.
  - Factual information, e.g.,
    - Extract unreported side effects of drugs from Web pages.
    - Extract infectious diseases from online news.
    - Extract economic data from reports of different countries.
  - Opinions
    - We focus on this topic as the Web has enabled the task. There is also a growing interest in this topic.
    - It is useful to everyone: individuals and organizations.

# Word-of-Mouth on the Web

- The Web has dramatically changed the way that people express their opinions. One can
  - post reviews of products at merchant sites, and
  - express opinions on almost anything in forums, discussion groups, and blogs, which are collectively called **the user generated content**.
- **Opinion mining or sentiment analysis aims to extract and summarize opinions**
- **Benefits:**
  - **Potential Customer:** No need to read many reviews, etc.
  - **Product manufacturer:** market intelligence, product benchmarking.

# Sentiment Classification of Reviews

- (Tracy, ACL 02, Pang et al., EMNLP 02, etc.)  
Classify reviews, based on the overall sentiment, expressed by authors, i.e.,
  - Positive or negative
- Related to but different from traditional topic-based text classification.
  - Here the opinion words (e.g., great, beautiful, bad, etc) are important, not topic words.
- Some representative techniques
  - Use opinion phrases
  - Use traditional text classification method
  - Use a custom-designed score function

# Feature-Based Opinion Summarization

- (Sentiment Classification) does not find what exactly consumers **liked** or **disliked**.
- You may say that people can read reviews, **but**



- In online shopping, a lot of people write reviews

- Time consuming and boring to read all the reviews



- How?



- **Opinion summarization is a natural solution**
  - What is an effective summary?

# An Review Example and a Summary

**GREAT Camera.**, Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The **pictures** coming out of this camera are amazing. The '**auto**' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. ...

....

## Summary:

### Feature1: **picture**

Positive: 12

- The **pictures** coming out of this camera are amazing.
- Overall this is a good camera with a really good **picture** clarity.

...

Negative: 2

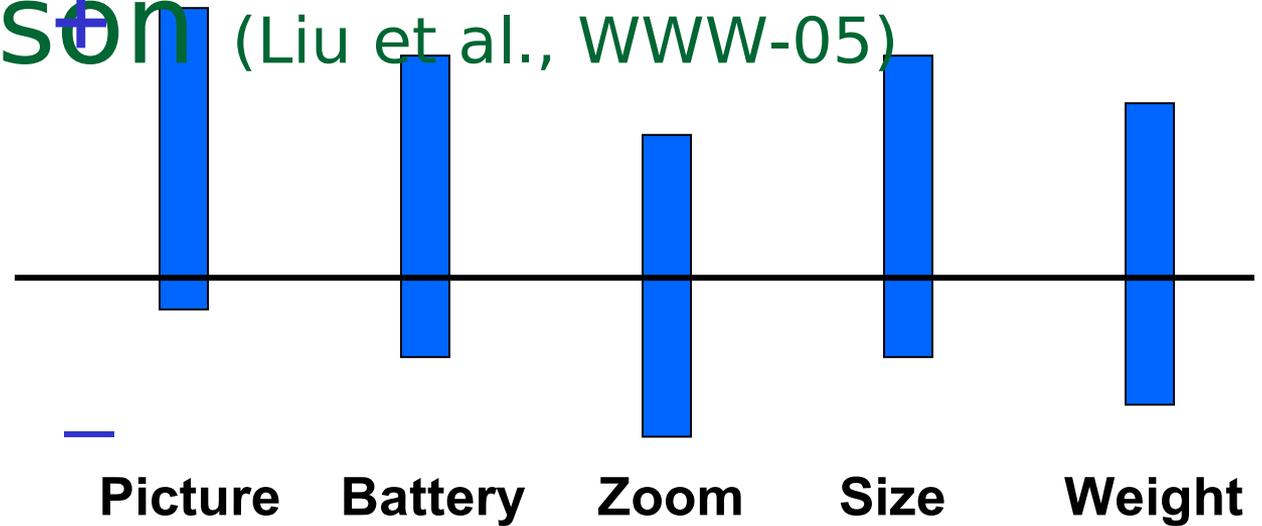
- The **pictures** come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, **pictures** produced by this camera were blurry and in a shade of orange.

### Feature2: **battery life**

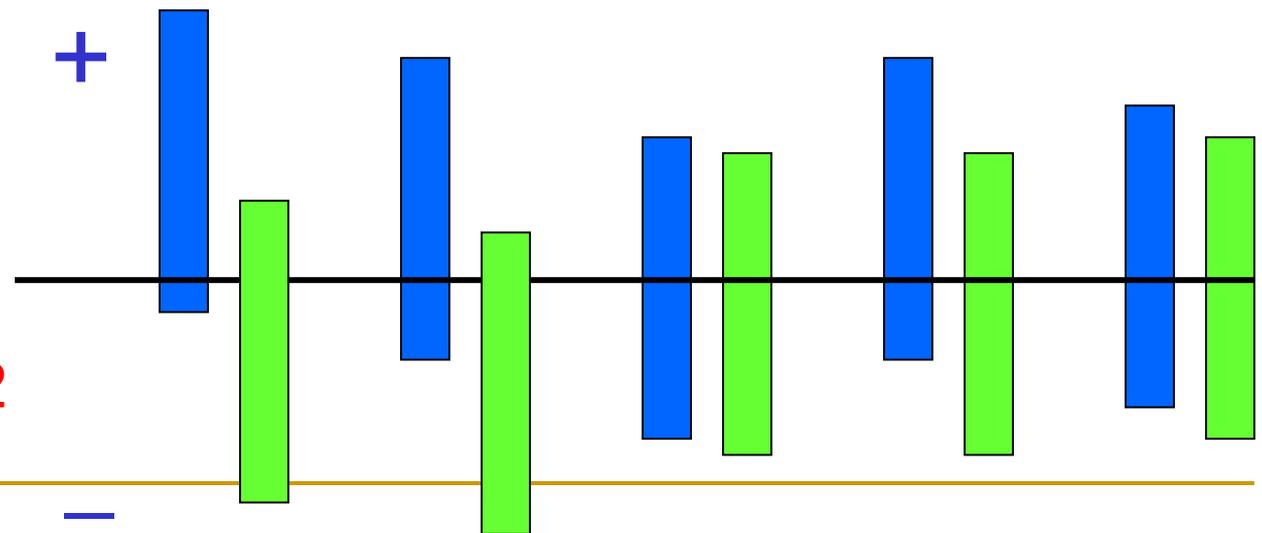
...

# Visual Summarization & Comparison (Liu et al., WWW-05)

- Summary of reviews of **Digital camera 1**



- Comparison of reviews of **Digital camera 1** and **Digital camera 2**



# Mining Tasks

(Hu and Liu, KDD-04; Liu, Web Data Mining book 2006)

**Task 1:** Identifying and extracting object features that have been commented on in each review.

**Task 2:** Determining whether the opinions on the features are positive, negative or neutral.

**Task 3:** Grouping synonym features.

- Produce a feature-based opinion summary.
  - A structured and quantitative summary.

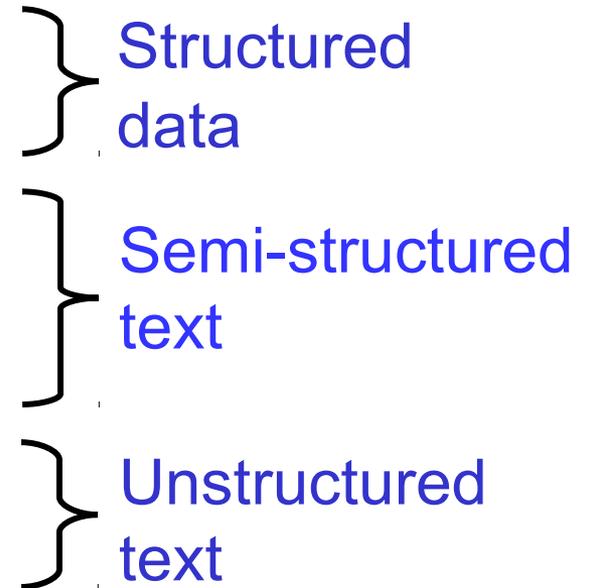
# Existing Research

- Current algorithms are combinations of
  - Natural language processing (NLP) methods, and
    - Part-of-speech tagging, parsing, etc.
    - Pre-compiled opinion words and phrases.
  - Data mining or machine learning techniques.
- **Opinion mining is a fascinating problem**
  - Technically very challenging. It is NLP!
    - It touches every aspect of NLP, yet it is confined/targeted
  - 20-60 companies working on it in USA alone.
- **We will discuss it in more detail tomorrow.**

# Roadmap

## ■ Introduction

1. Structured data extraction
2. Information integration
3. Information synthesis
4. Opinion mining



## ■ **Conclusions**

# Conclusions

- We briefly:
  - Structured data extraction
  - Information integration
  - Information synthesis
  - Opinion mining
- The tasks look different, but there is a common theme:
  - **Extraction and integration**
- All are related to and need some level of NLP.
- Integration has been regarded as the most difficult task by database researchers.
  - **Core problem**: recognizing domain “synonym”: words, phrases and expressions