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 someway 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

Semi-structured data

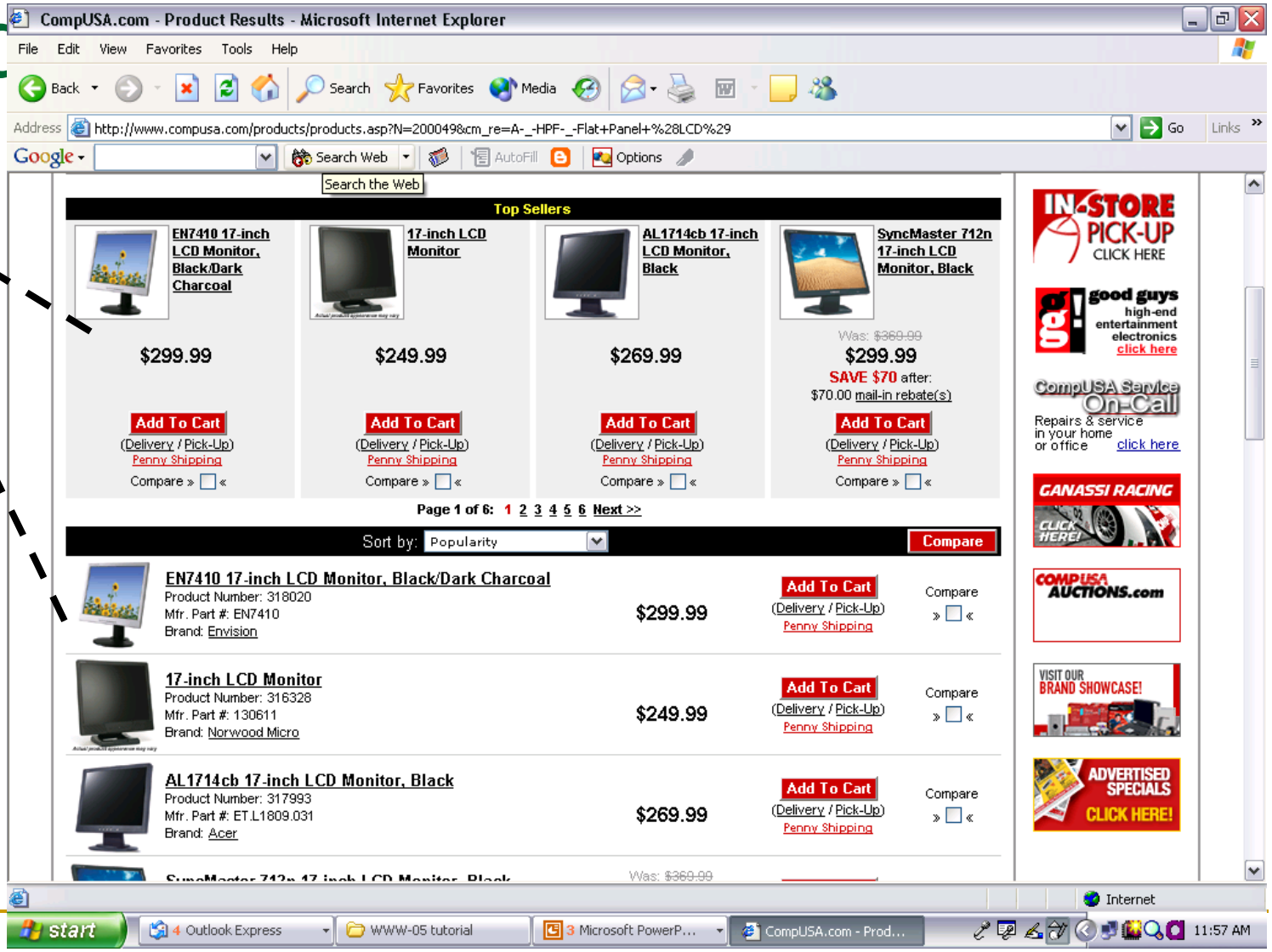
Unstructured text

Structured Data Extraction

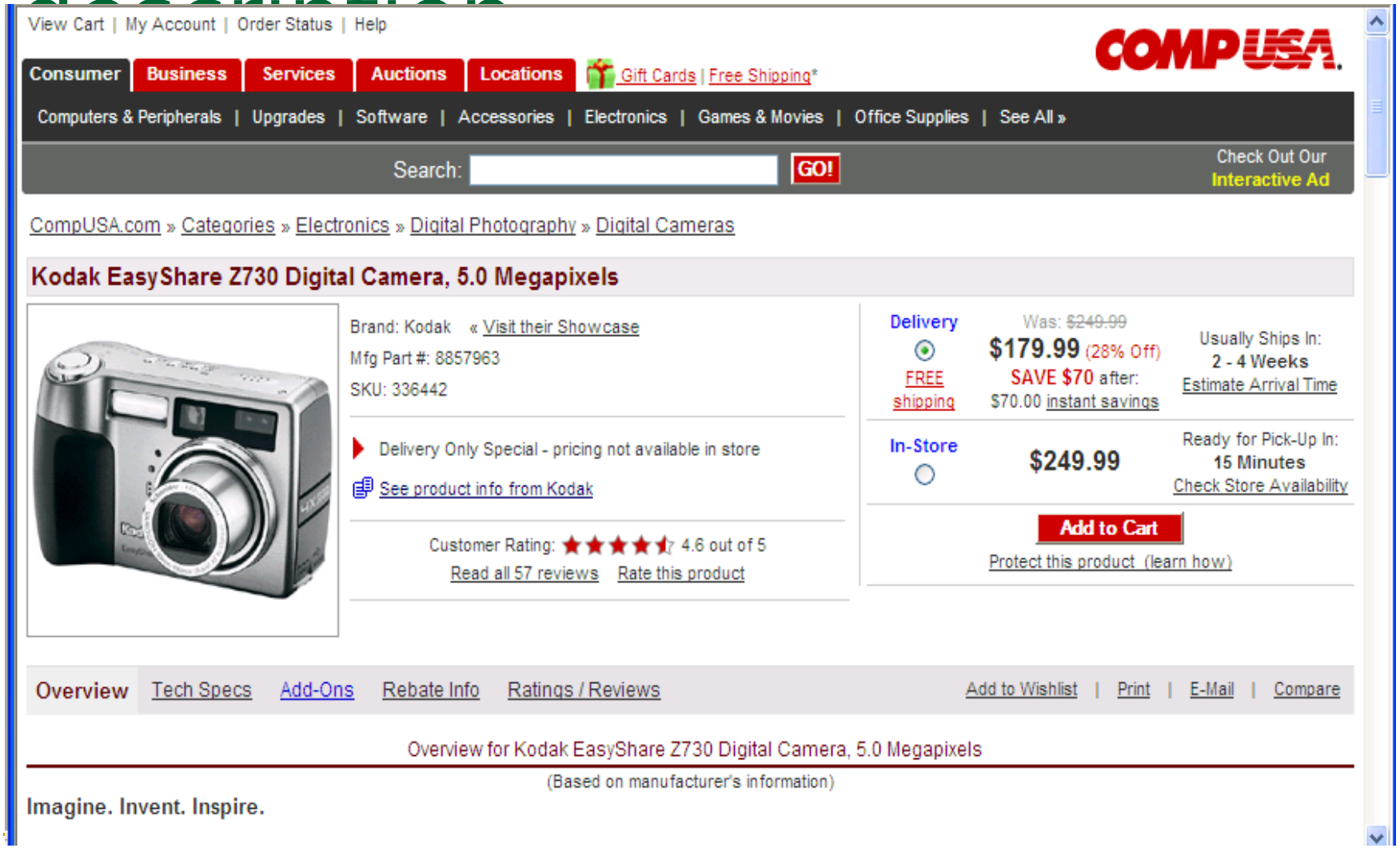
- A large amount of information on the Web is contained in regularly structured data objects.
  - often data records retrieved from databases.
  - Important: lists of products and services.
- Applications: Gather data to provide value-added services
  - comparative shopping, object search, opinion mining,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

List Page – two lists of proc

Two lists



Detail Page – detailed description



Extraction Task: an illustr

Cabinet Organizers by Copco

9-in. Round Turntable: White ★★★★½ $4.95 BUY

12-in. Round Turntable: White ★★★★½ $7.95 BUY

Cabinet Organizers

14.75x9 Cabinet Organizer (Non-skid): White ★★★★½ $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

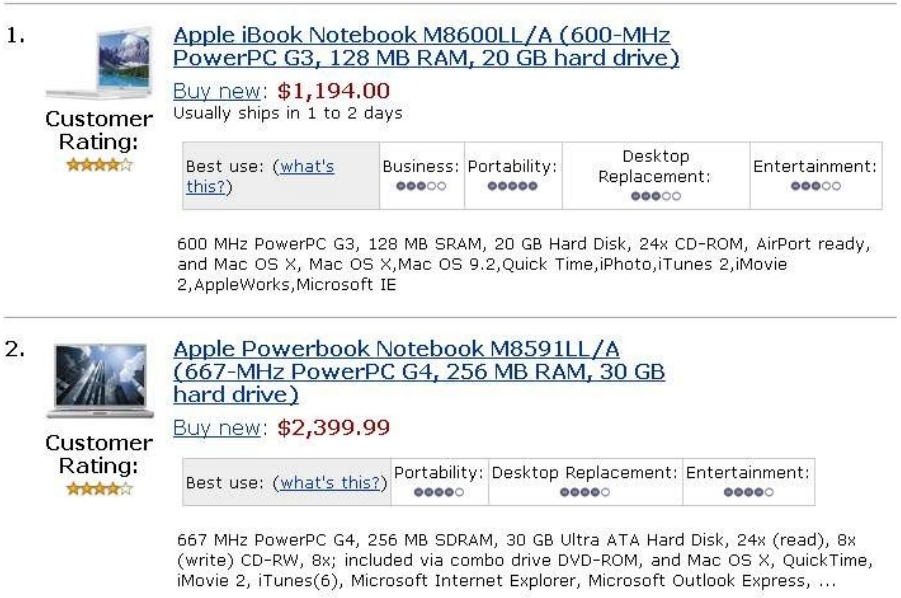
Web data model: Nested relations

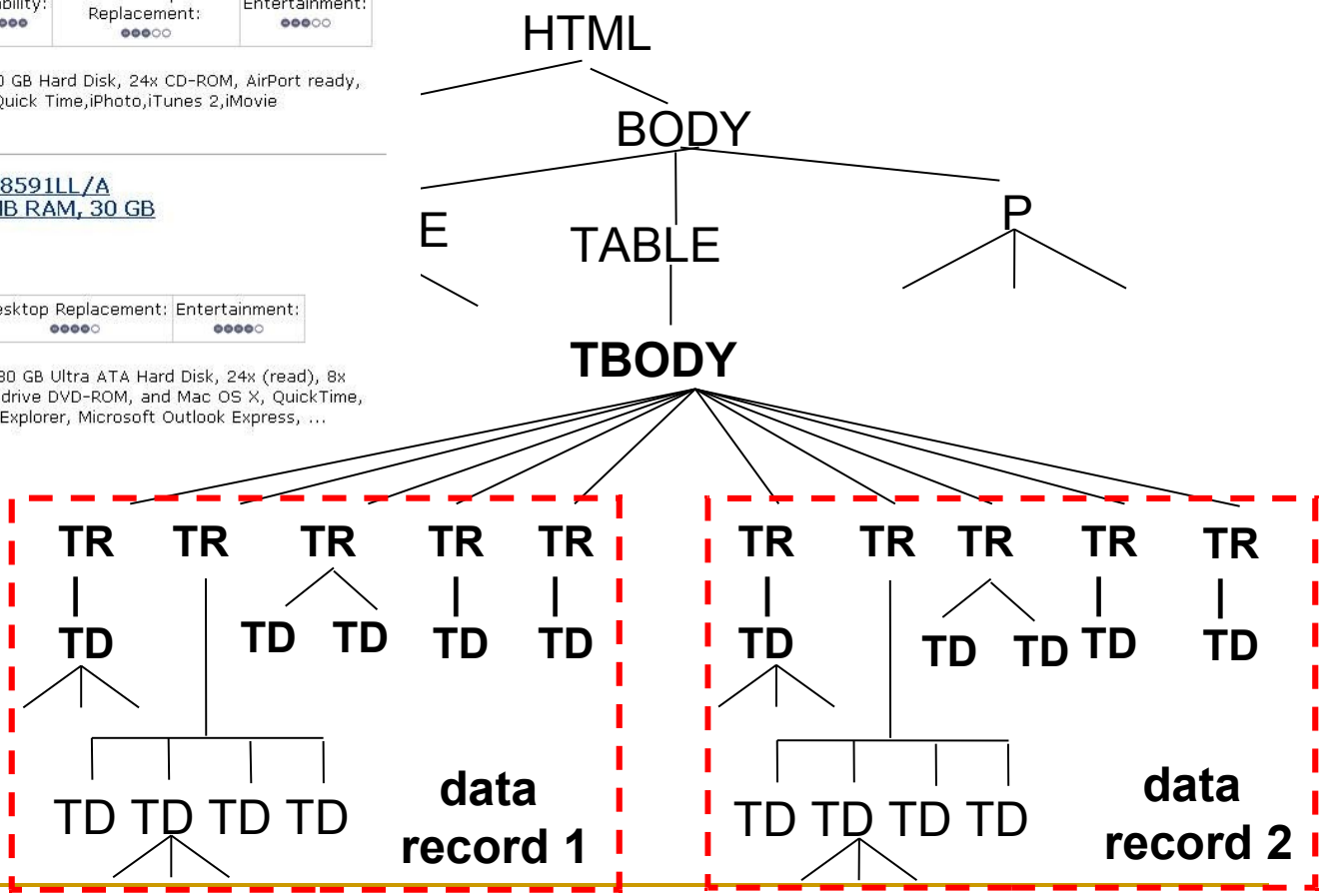
- See formal definitions in (Grumbach and Mecca, ICDT-99; Liu, Web Data Mining 2006)

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





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, Venice, Phone 1-**800**-555-1515
E2: 90 Colfax, Palms, Phone (**800**) 508-1570
E3: 523 1st St., LA, Phone 1-**800**-578-2293
E4: 403 La Tijera, Watts, Phone: (**310**) 798-0008

Output Extraction Rules

- Start rules: End rules:

R1: SkipTo(**(**) SkipTo(**)**)
R2: SkipTo(-) SkipTo()

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).

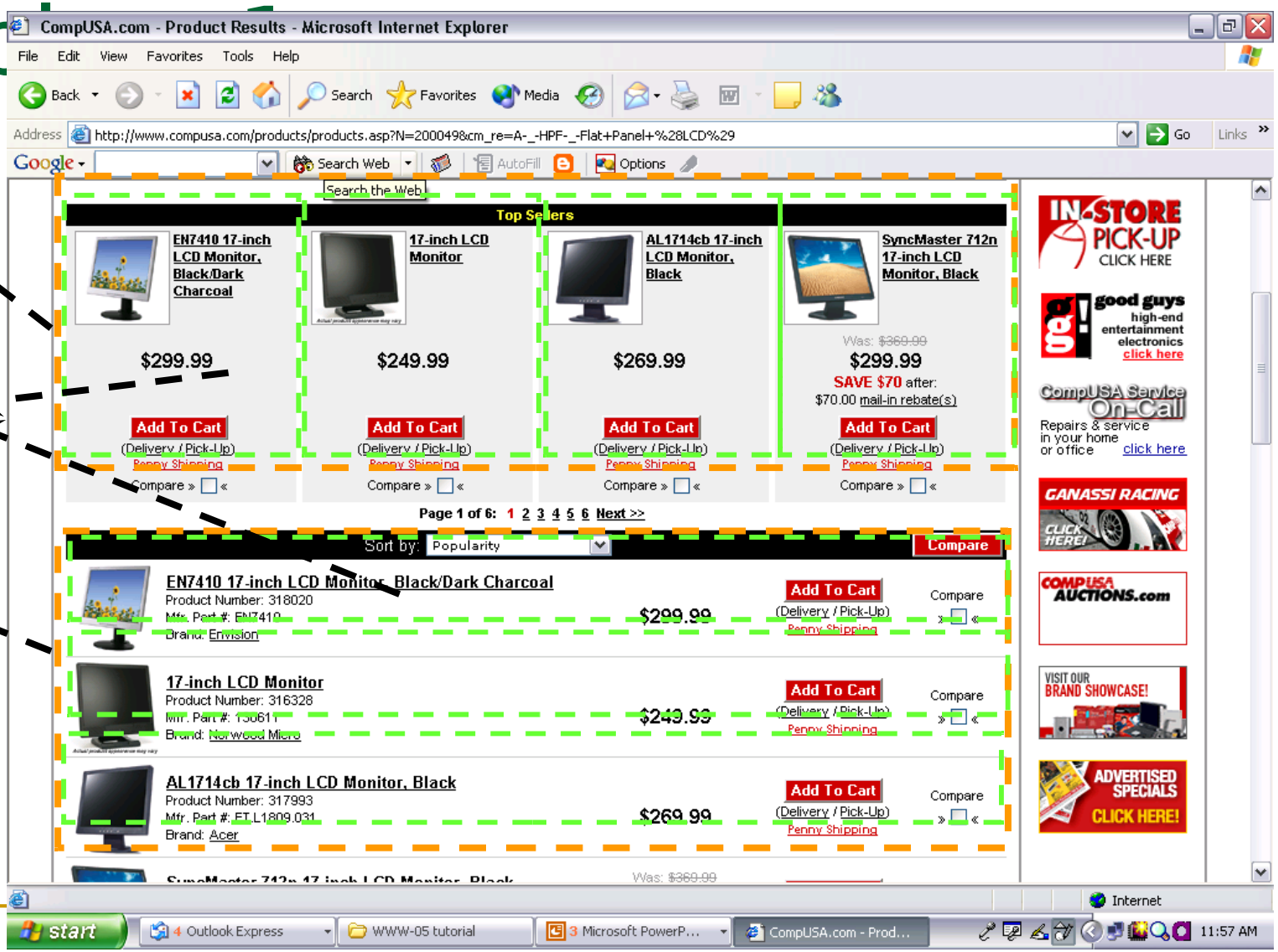
- Problem 1 is more general: Algorithms for solving Problem 1 can solve Problem 2.
  - Thus, we only discuss Problem 1.

Automatic Extraction: Prob

Data region1

Data records

Data region2



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:

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

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

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

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

- alignment

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

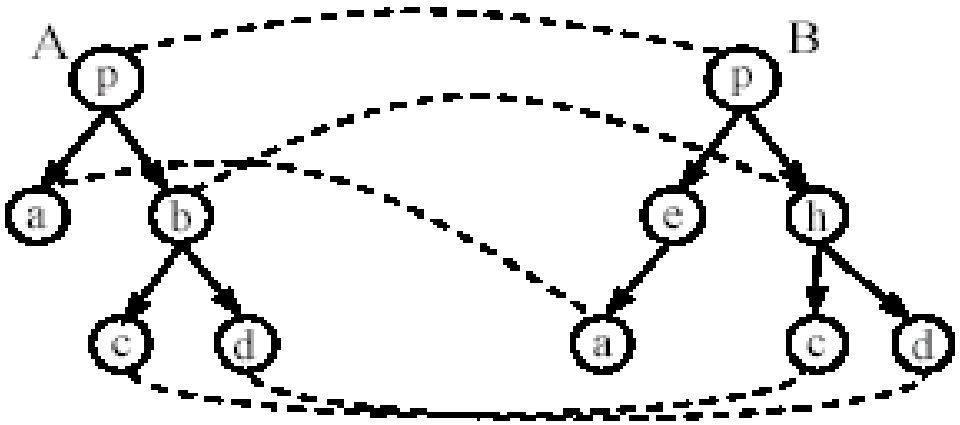
	s_1	X	G	Y	X	Y	X	Y	X
s_2	0	1	2	3	4	5	6	7	8
X	1	0	1	2	3	4	5	6	7
Y	2	1	1	1	2	3	4	5	6
X	3	2	2	2	1	2	3	4	5
Y	4	3	3	2	2	1	2	3	4
X	5	4	4	3	2	2	1	2	3
Y	6	5	5	4	3	2	2	1	2
T	7	6	6	5	4	3	3	2	2
X	8	7	7	6	5	4	3	3	2

Tree edit distance or tree

Let X be a tree and let $X[i]$ be the ith 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 $B[j_2]$;
(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 $A = R_A:(A_1, \ldots, A_k)$ and $B = R_B:(B_1, \ldots, B_n)$ *be two trees, where* R_A *and* R_B *are their roots*

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

$m(\langle\rangle, \langle\rangle) = 0$ // $\langle\rangle$ represents an empty sub-tree list.

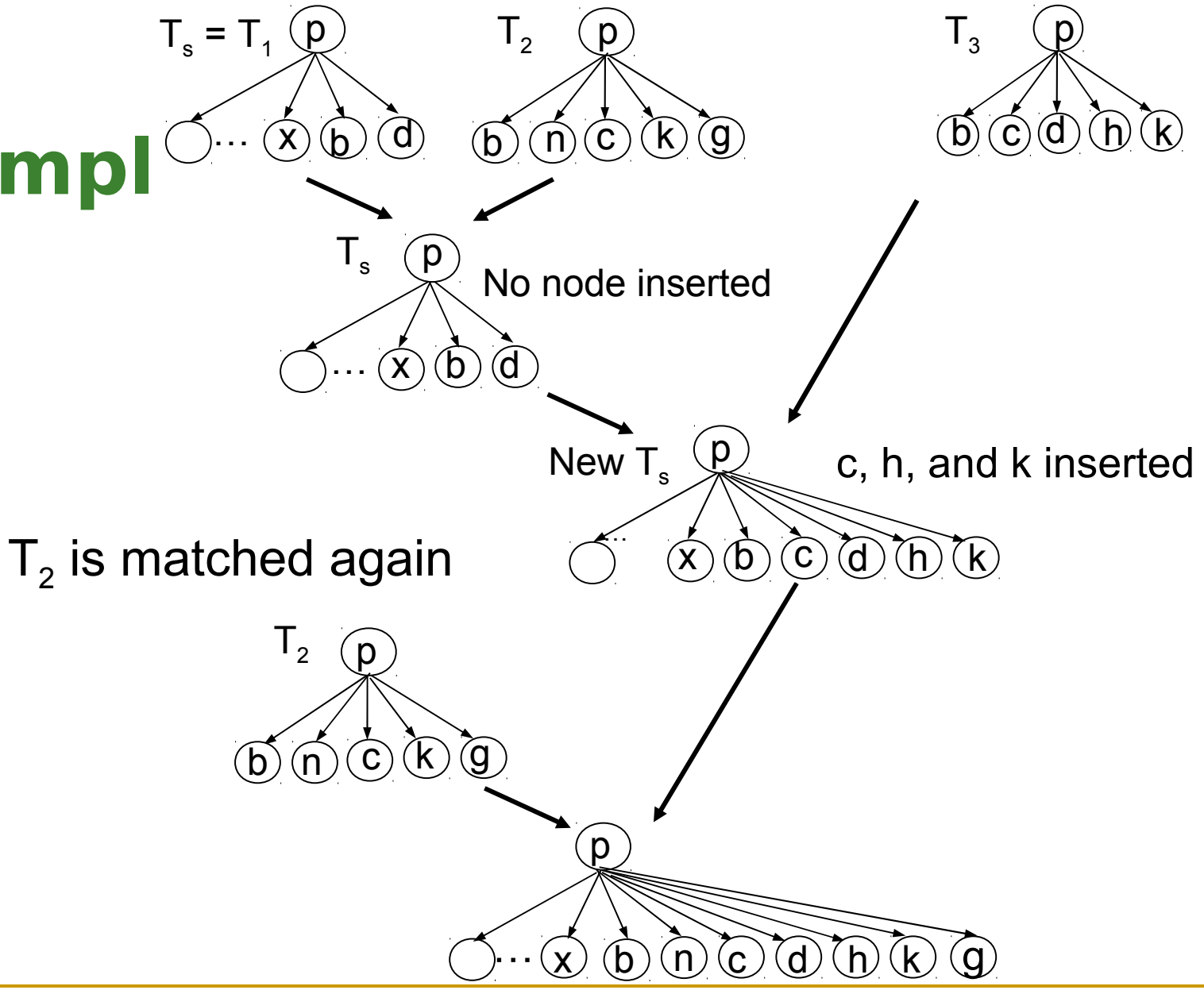
$m(s, \langle\rangle) = m(\langle\rangle, s) = 0$ // s matches any non-empty sub-tree list

$$m(\langle A_1, \ldots, A_k\rangle, \langle B_1, \ldots, B_n\rangle) = \max(m(\langle A_1, \ldots, A_{k-1}\rangle, \langle B_1, \ldots, B_{n-1}\rangle) + W(A_k, B_n),$$
$$m(\langle A_1, \ldots, A_k\rangle, \langle B_1, \ldots, B_{n-1}\rangle),$$
$$m(\langle A_1, \ldots, A_{k-1}\rangle, \langle B_1, \ldots, B_n\rangle)).$$

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


$T_s = T_1$
p
x
b
d
T_2
p
b
n
c
k
g
T_3
p
b
c
d
h
k
T_s
p
No node inserted
x
b
d
New T_s
p
c, h, and k inserted
x
b
c
d
h
k
T_2 is matched again
T_2
p
b
n
c
k
g
p
x
b
n
c
d
h
k
g

Roadmap

- Introduction

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

- Conclusions

Structured data

Semi-structured text

Unstructured text

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

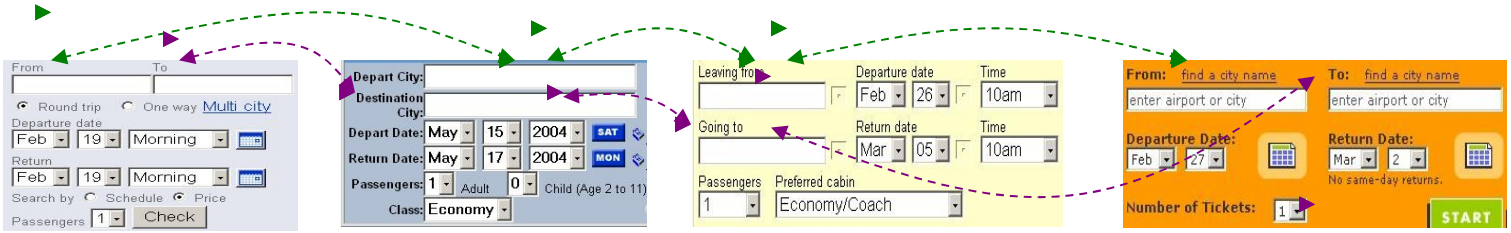
(Wu et al., SIGM[illegible]-04; Dragut et al., VLDB-06)

Global Query Interface

Constructing global query interface (QI)

- 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

- Two-phrased integration
  - ***Interface Matching*** – Identify semantically similar fields



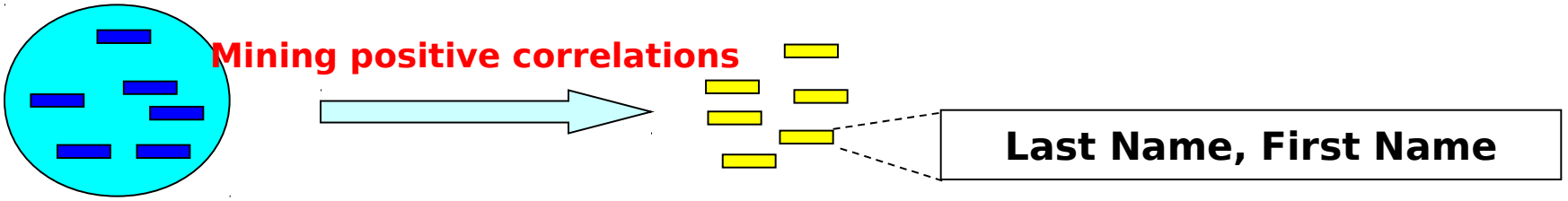
  - Interface Integration – Merge the source query interfaces

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

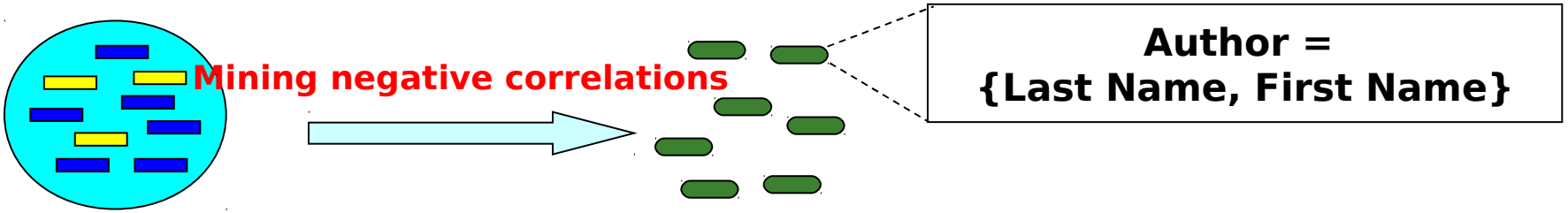
- This technique needs a large number of input query interfaces.

- 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}

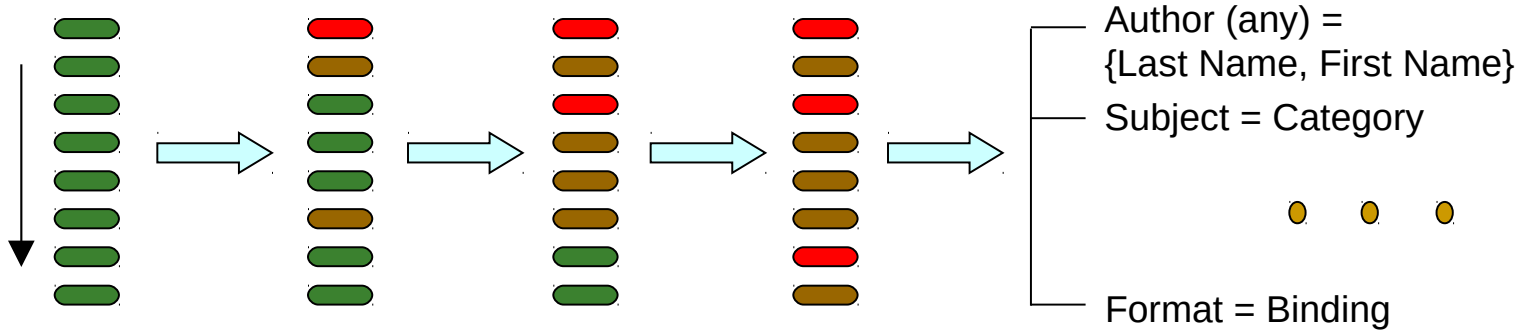
tive correlation mining as potential groups



ve correlation mining as potential matchings

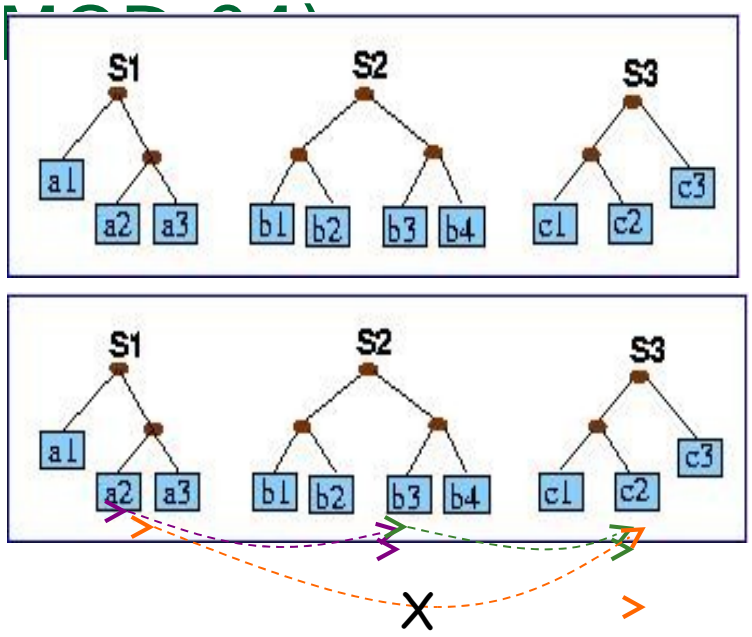


atching selection as model construction



A clustering approach to schema matching (Wu et al. SIGMOD 04)

- 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



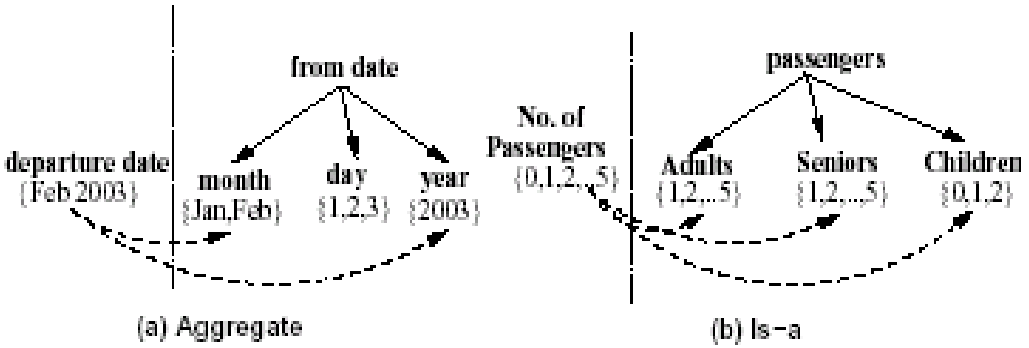
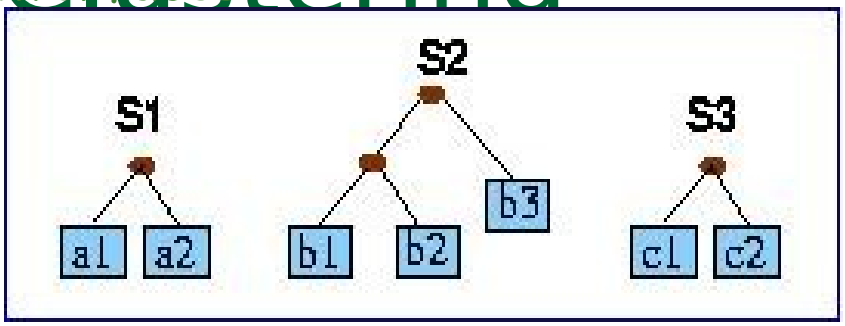


Figure 3: 1:m mappings

Find 1:1 Mappings via Clustering

Interfaces:



Initial similarity matrix:

	a_1	a_2	b_1	b_2	b_3	c_1	c_2
a_1	0	0	.9	0	.85	0	0
a_2	0	0	.15	0	0	0	0
b_1	.9	.15	0	0	0	.8	0
b_2	0	0	0	0	0	.1	.6
b_3	.85	0	0	0	0	0	0
c_1	0	0	0	.1	0	0	0
c_2	0	0	.8	.6	0	0	0

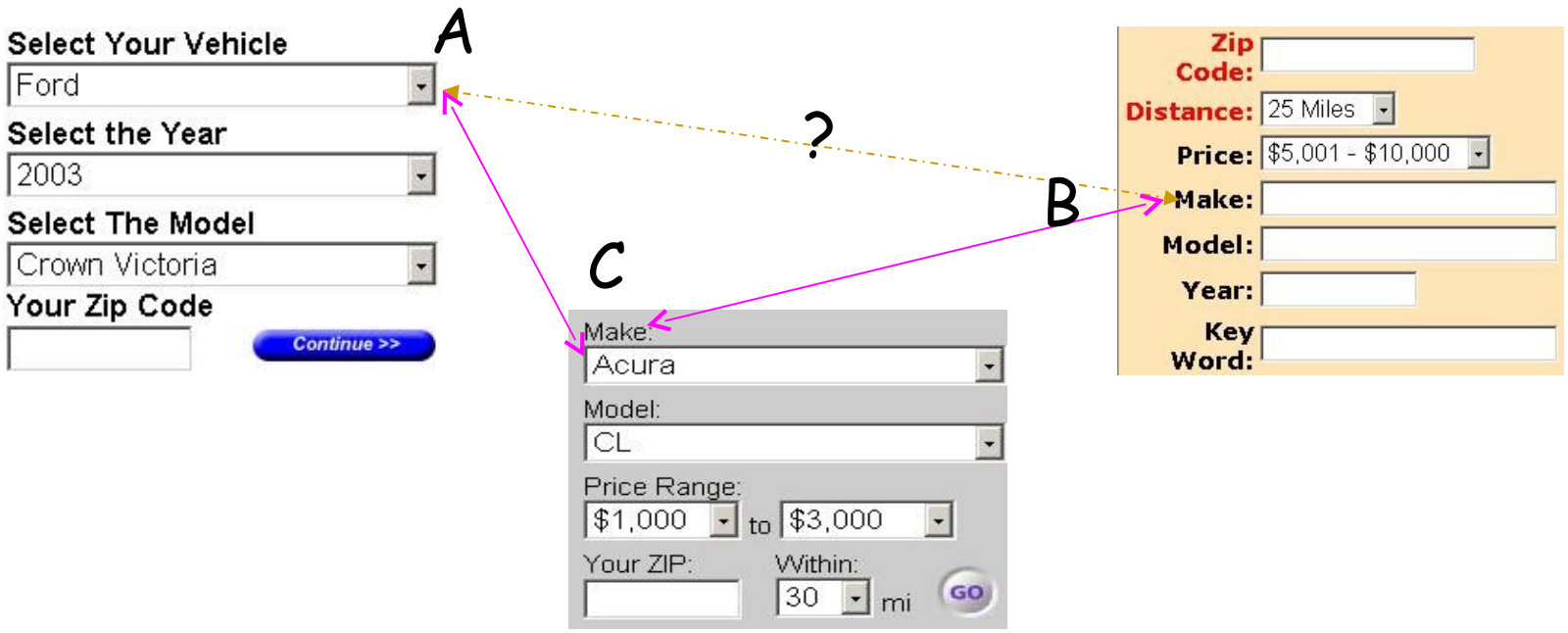
After one merge:

	a_2	b_2	b_3	c_1	c_2	$\{a_1, b_1\}$
a_2	0	0	0	0	0	0
b_2	0	0	0	.1	.6	0
b_3	0	0	0	0	0	0
c_1	0	.1	0	0	0	.8
c_2	0	.6	0	0	0	0
$\{a_1, b_1\}$	0	0	0	.8	0	0

- Similarity functions
  - linguistic similarity
  - domain similarity

…, final clusters: {{a1,b1,c1}, {b2,c2},{a2},{b3}}

"Bridging" Effect

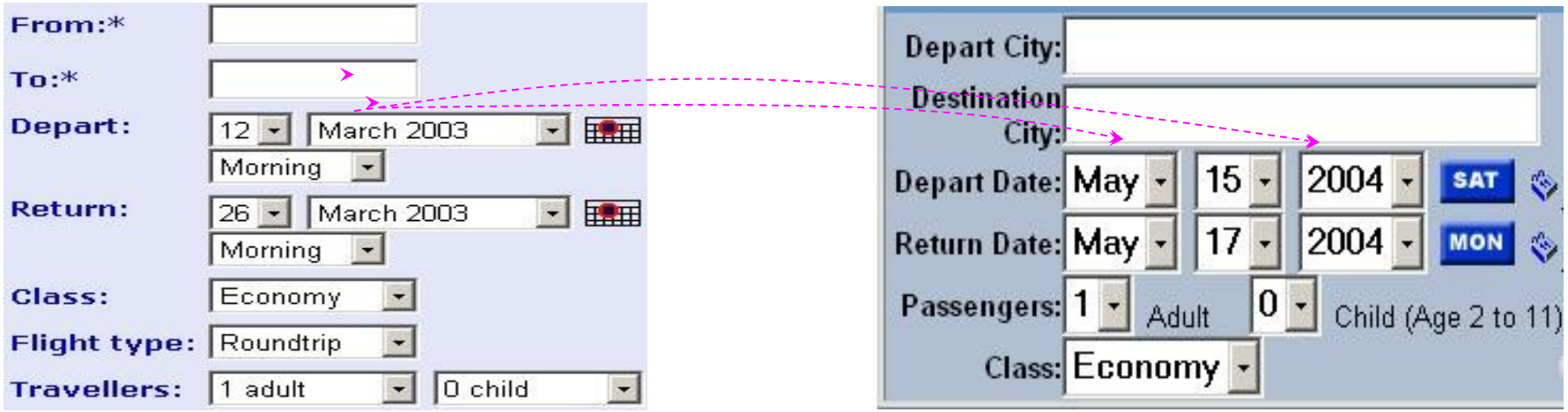


Observations:

- 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

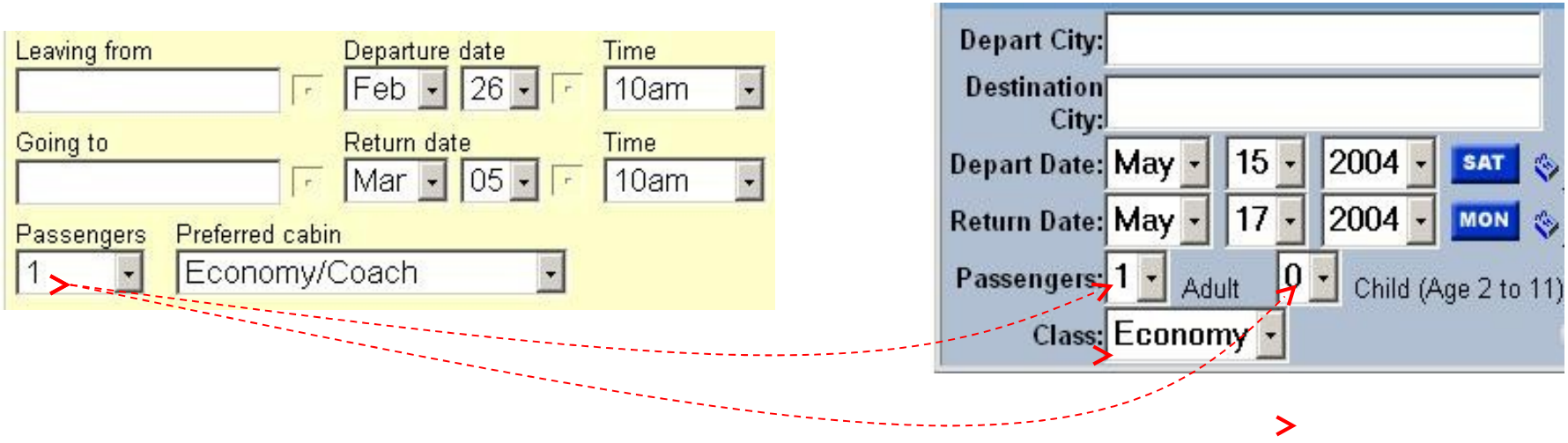
Complex Mappings



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)



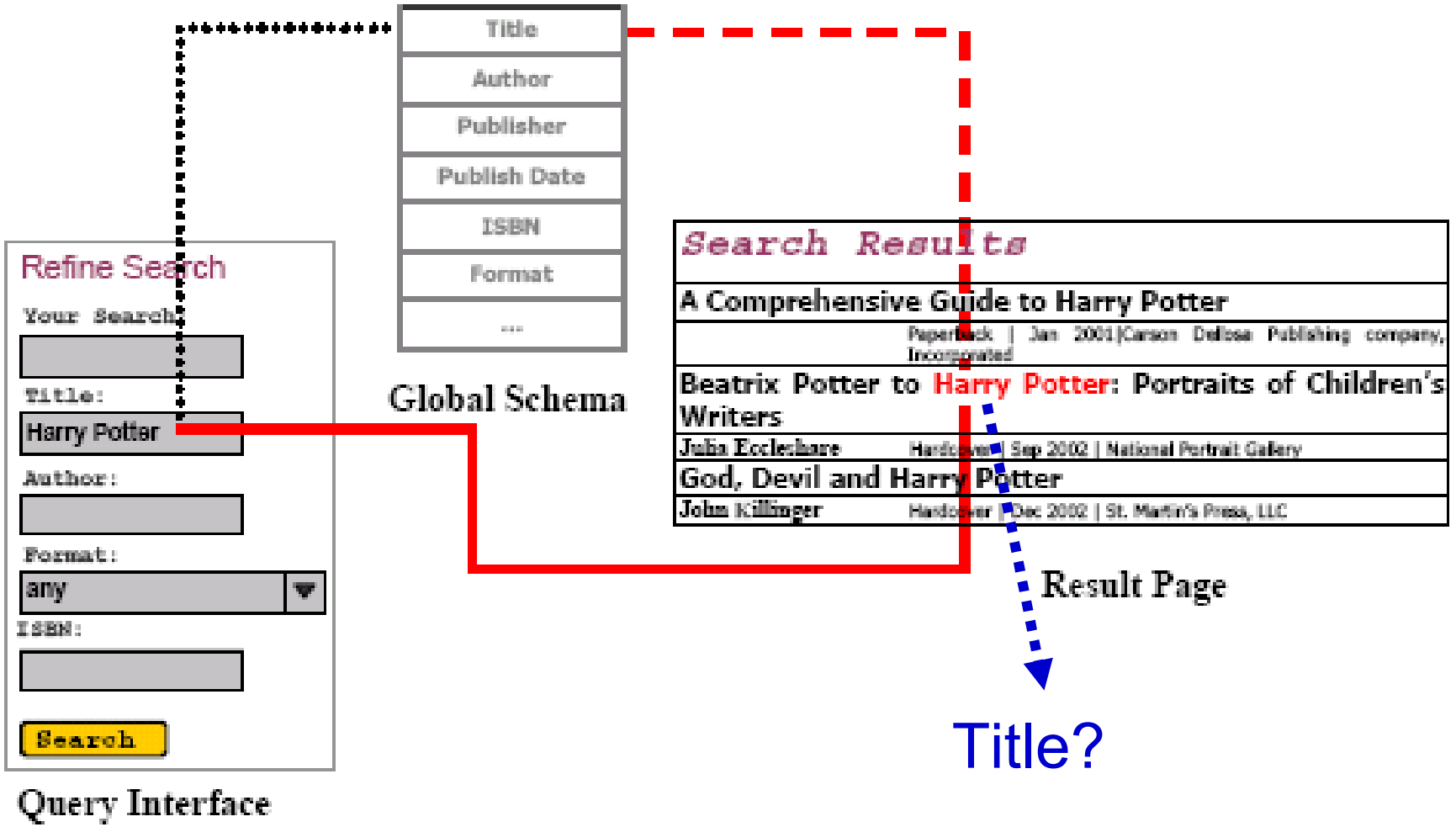
Is-a type – contents of fields on the many side are sum/union of the content of field on the one side

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

Structured data

Semi-structured text

Unstructured text

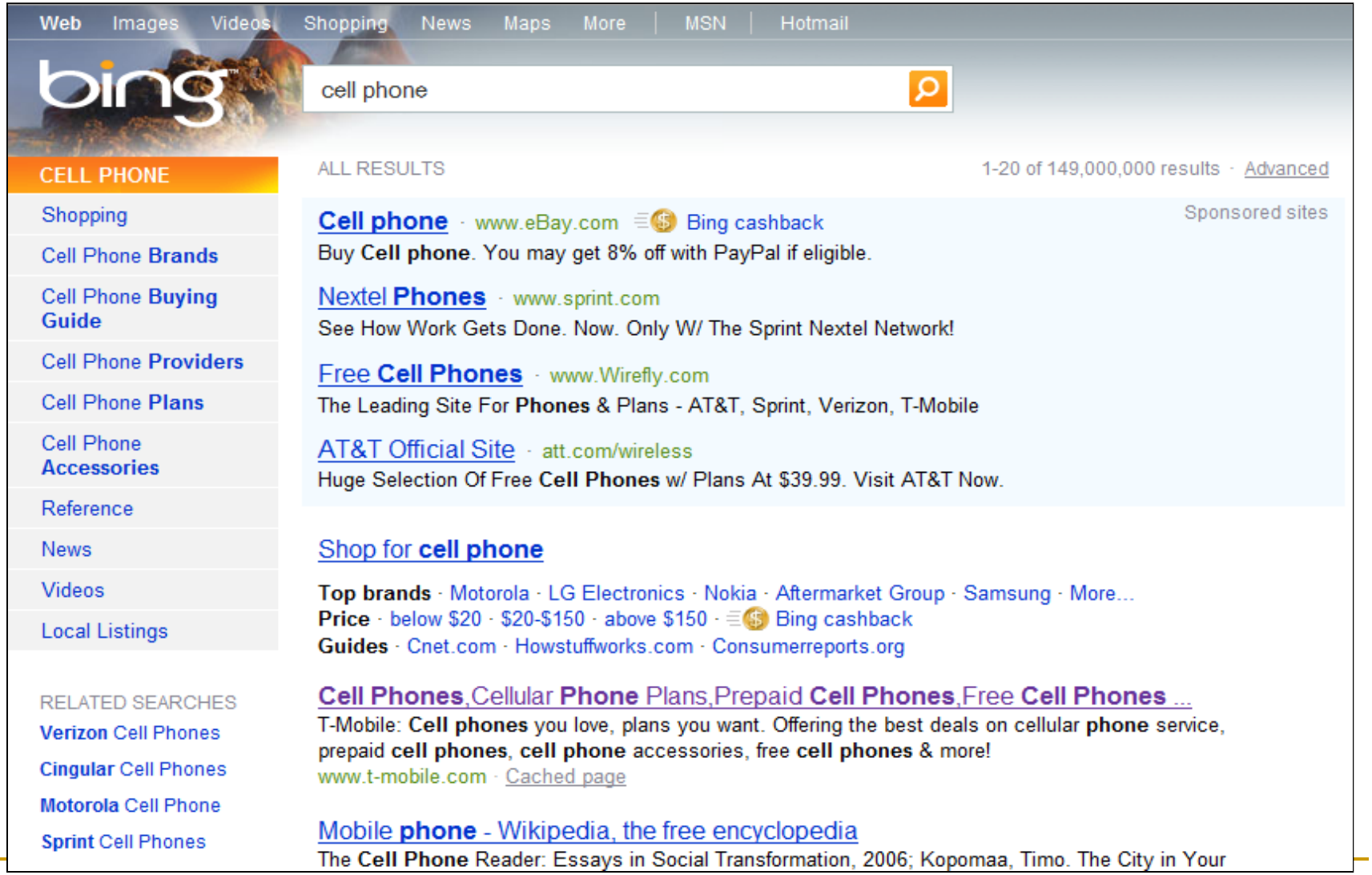
Information/knowledge synthesis

- 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”



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>, , <big>, <i>, , <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} …

Some concepts extraction results

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

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

Structured data

Semi-structured text

Unstructured text

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

- (Turney, ACL-02, Pang et al., EMNLP-02, ...)
  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

- (Hu and Liu, KDD-04)
- 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

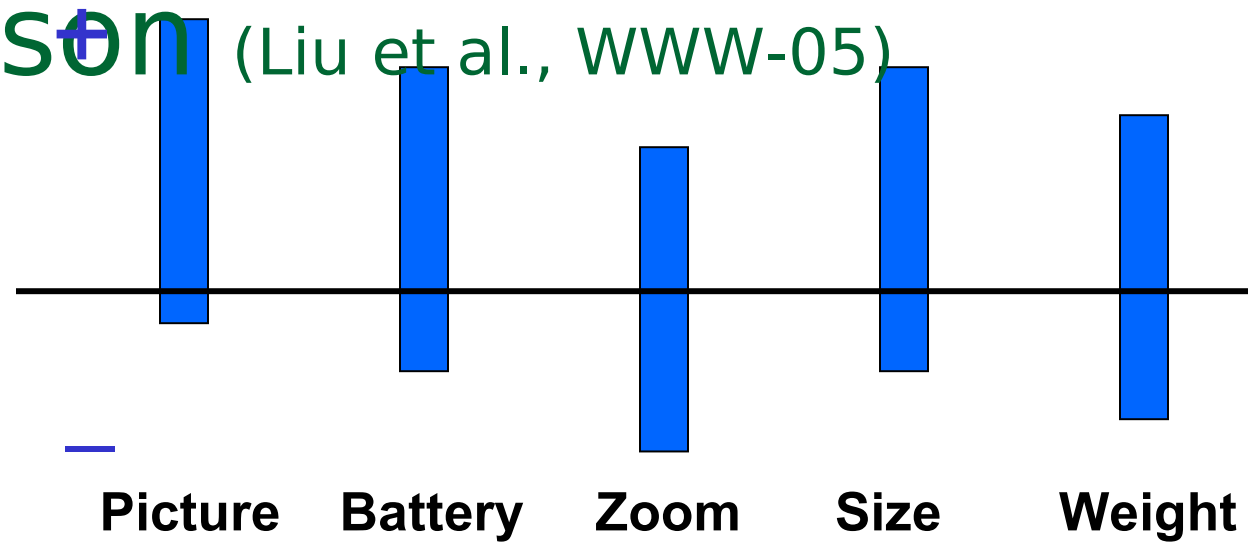
- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Feature2: **battery life**

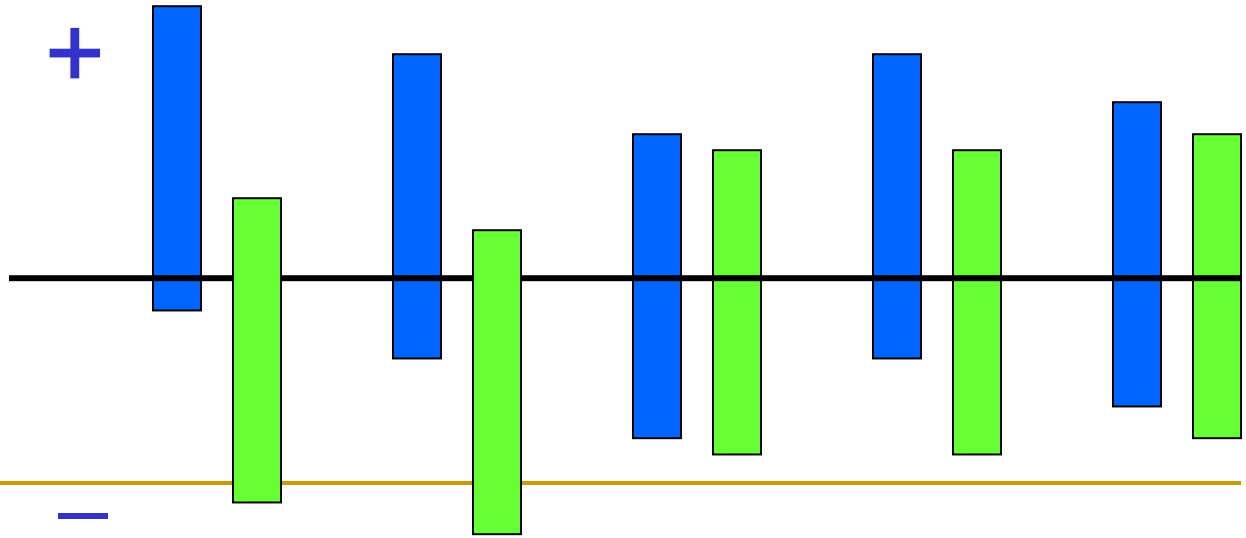
…

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

- Summary of reviews of Digital camera 1



- Comparison of reviews of Digital camera 1 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**

Structured data

Semi-structured text

Unstructured text

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