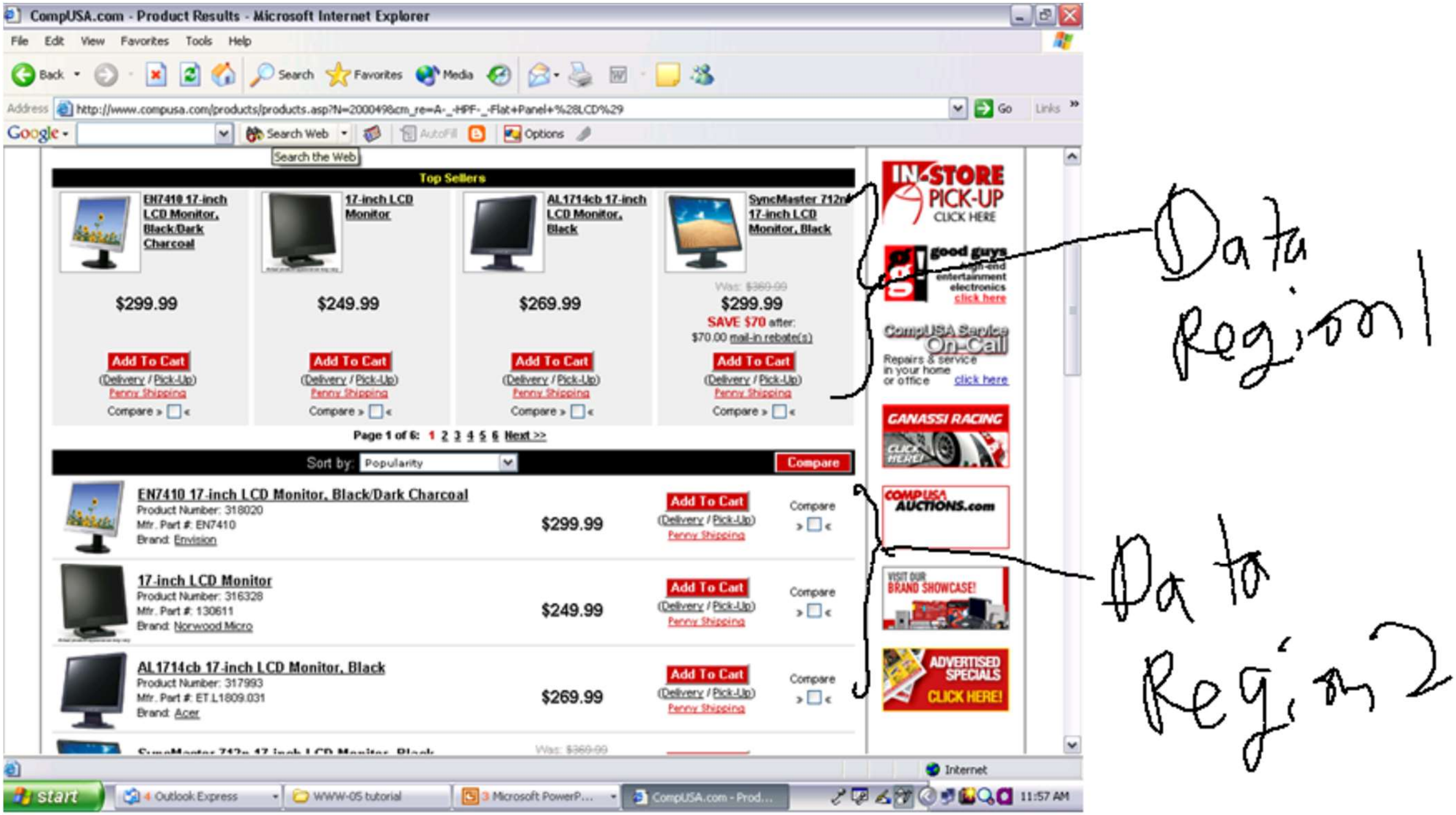
Chapter 10: Information Integration

Combining data regions

Combining Data Regions



CompUSA.com - Product Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.compusa.com/products/products.asp?N=200049&cm_re=A-_-HPF-_-Flat+Panel+%28LCD%29
Go Links
Google Search Web AutoFill Options
Search the Web
IN-STORE PICK-UP CLICK HERE
good guys high-end entertainment electronics click here
CompUSA Service On-Call Repairs & service in your home or office click here
GANASSI RACING CLICK HERE!
COMPUSA AUCTIONS.com
VISIT OUR BRAND SHOWCASE!
ADVERTISED SPECIALS CLICK HERE!
Internet
start Outlook Express WWW-05 tutorial Microsoft PowerP... CompUSA.com - Prod... 11:57 AM
Data Region1
Data Region2

Top Sellers

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

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

Sort by: Popularity Compare

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

Introduction

- At the end of last topic, we identified the problem of integrating extracted data:
  - Column match and instance value match.

- In this part, we introduce
  - Some basic integration techniques, and
  - Web query interface integration

Database integration (Rahm and Berstein 2001)

- Information integration started with database integration, which has been studied in the database community since the early 1980s.
- Fundamental problem: schema matching, which takes two (or more) database schemas to produce a mapping between elements (or attributes) of the two (or more) schemas that correspond semantically to each other.
- Objective: merge the schemas into a single global schema.

Integrating two schemas

- Consider two schemas, *S*1 and *S*2, representing two customer relations, Cust and Customer.

S1	S2
Cust	**Customer**
CNo	CustID
CompName	Company
FirstName	Name
LastName	Phone

Integrating two schemas (contd)

- Represent the mapping with a similarity relation, $\cong$, over the power sets of *S*1 and *S*2, where each pair in $\cong$ represents one element of the mapping. E.g.,

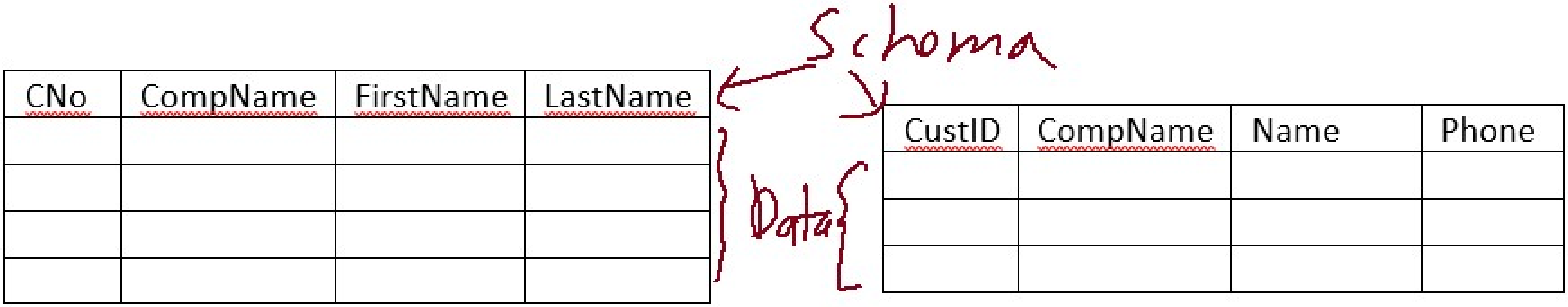
  Cust.CNo $\cong$ Customer.CustID

  Cust.CompName $\cong$ Customer.Company

  {Cust.FirstName, Cust.LastName} $\cong$ Customer.Name

Schema and Domain

- Schema: Header information
- Instance: Data entry
- Domain: A set of possible values of an attribute



CNo	CompName	FirstName	LastName

CustID	CompName	Name	Phone

Different types of matching

- Schema-level only matching: only schema information is considered.
- Domain and instance-level only matching: some instance data (data records) and possibly the domain of each attribute are used. This case is quite common on the Web.
- Integrated matching of schema, domain and instance data: Both schema and instance data (possibly domain information) are available.

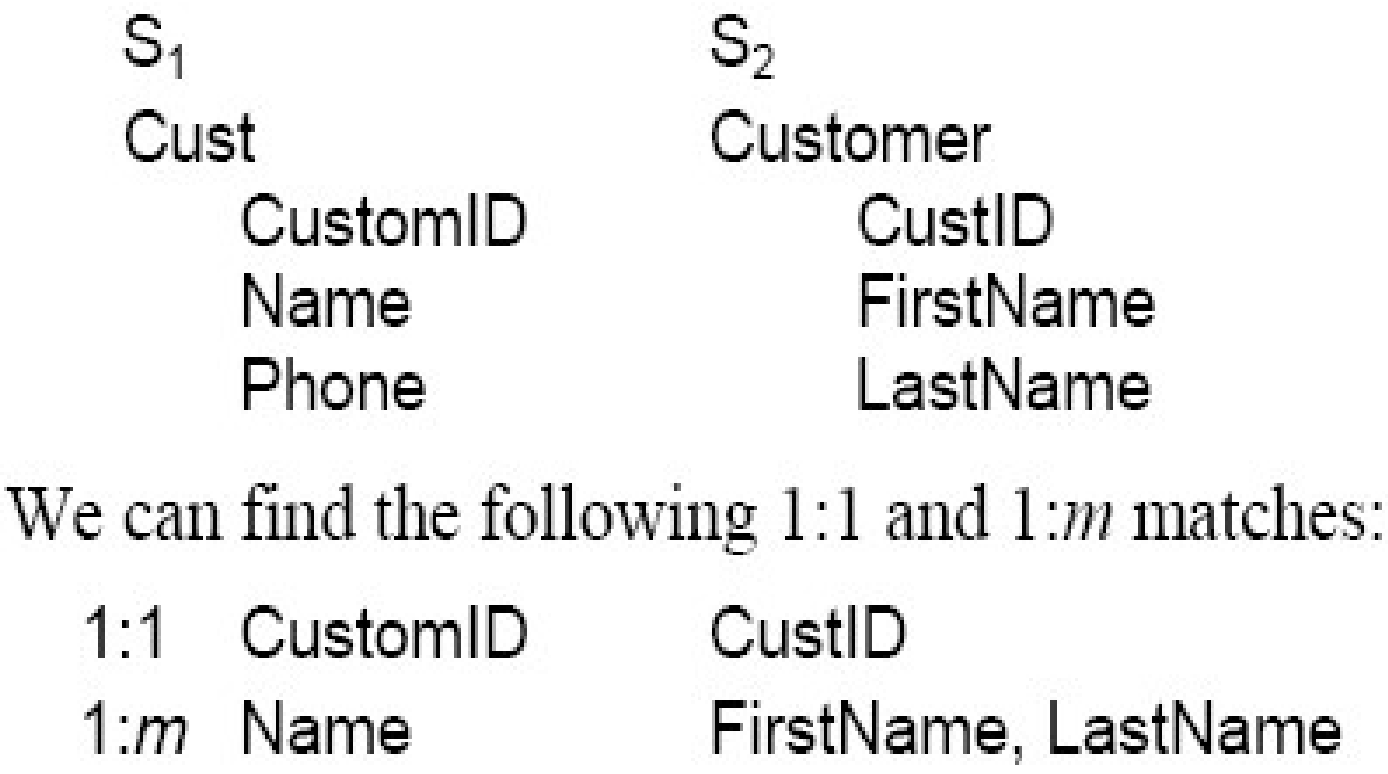
Pre-processing for integration (He and Chang SIGMOG-03, Madhavan et al. VLDB-01, Wu et al. SIGMOD-04

- Tokenization: break an item into atomic words using a dictionary, e.g.,
  - Break “fromCity” into “from” and “city”
  - Break “first-name” into “first” and “name”
- Expansion: expand abbreviations and acronyms to their full words, e.g.,
  - From “dept” to “departure”
- Stopword removal and stemming
- Standardization of words: Irregular words are standardized to a single form, e.g.,
  - From “colour” to “color”

Schema-level matching (Rahm and Berstein 2001)

- Schema level matching relies on information such as name, description, data type, relationship type (e.g., part-of, is-a, etc), constraints, etc.
- Match cardinality:
  - 1:1 match: one element in one schema matches one element of another schema.
  - 1:m match: one element in one schema matches m elements of another schema.
  - m:n match: m elements in one schema matches n elements of another schema.

An example



S_1	S_2
Cust	Customer
CustomID	CustID
Name	FirstName
Phone	LastName

We can find the following 1:1 and 1:*m* matches:

1:1	CustomID	CustID
1:*m*	Name	FirstName, LastName

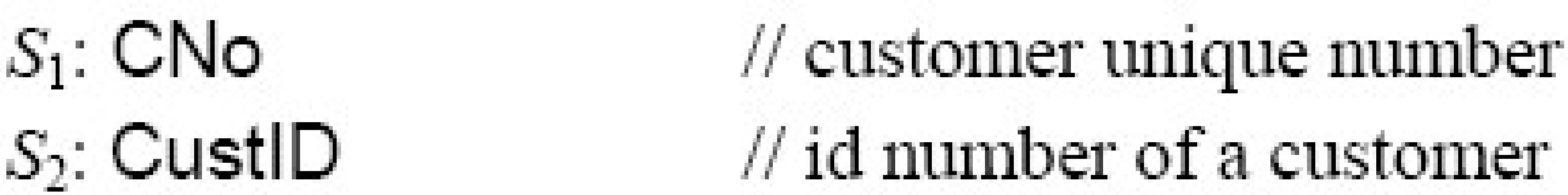
- m:1 match is similar to 1:m match. m:n match is complex, and there is little work on it.

Linguistic approaches (See (Liu, Web Data Mining book 2007) for many references)

- They are used to derive match candidates based on names, comments or descriptions of schema elements:
- Name match:
  - Equality of names
  - Synonyms
  - Equality of hypernyms: A is a hypernym of B is B is a kind-of A. (Example: color is a hypernym of red)
  - Common sub-strings
  - Cosine similarity
  - User-provided name match: usually a domain dependent match dictionary

Linguistic approaches (contd)

- Description match: in many databases, there are comments to schema elements, e.g.,



S_1: CNo // customer unique number

S_2: CustID // id number of a customer

- Cosine similarity from information retrieval (IR) can be used to compare comments after stemming and stopword removal.

Constraint based approaches (See (Liu, Web Data Mining book 2007) for references)

- Constraints such as data types, value ranges, uniqueness, relationship types, etc.
- An equivalent or compatibility table for data types and keys can be provided. E.g.,
  - string ≅ varchar, and (primiary key) ≅ unique
- For structured schemas, hierarchical relationships such as
  - is-a and part-of

  may be utilized to help matching.
- Note: On the Web, the constraint information is often not available, but some can be inferred based on the domain and instance data.

Domain and instance-level matching

(See (Liu, Web Data Mining book 2007) for references)

- In many applications, some data instances or attribute domains may be available.
- Value characteristics are used in matching.
- Two different types of domains
  - Simple domain: each value in the domain has only a single component (the value cannot be decomposed).
  - Composite domain: each value in the domain contains more than one component.

Match of simple domains

- A simple domain can be of any type.
- If the data type information is not available (this is often the case on the Web), the instance values can often be used to infer types, e.g.,
  - Words may be considered as strings
  - Phone numbers can have a regular expression pattern.
- Data type patterns (in regular expressions) can be learnt automatically or defined manually.
  - E.g., used to identify such types as integer, real, string, month, weekday, date, time, zip code, phone numbers, etc.

Match of simple domains (contd)

- Matching methods:
  - Data types are used as constraints.
  - For numeric data, value ranges, averages, variances can be computed and utilized.
  - For categorical data: compare domain values.
  - For textual data: cosine similarity.
  - Schema element names as values: A set of values in a schema match a set of attribute names of another schema. E.g.,
    - In one schema, the attribute color has the domain {yellow, red, blue}, but in another schema, it has the element or attribute names called yellow, red and blue (values are yes and no).

Handling composite domains

- A composite domain is usually indicated by its values containing delimiters, e.g.,
  - punctuation marks (e.g., “-”, “/”, “_”)
  - White spaces
  - Etc.
- To detect a composite domain, these delimiters can be used. They are also used to split a composite value into simple values.
- Match methods for simple domains can then be applied.

Combining similarities

- Similarities from many match indicators can be combined to find the most accurate candidates.
- Given the set of similarity values, $sim_1(u, v)$, $sim_2(u, v)$, …, $sim_n(u, v)$, from comparing two schema elements u (from S_1) and v (from S_2), many combination methods can be used:
  - Max: $CSim(u, v) = \max\{sim_1(u, v), sim_2(u, v), \ldots, sim_n(u, v)\}$
  - Weighted sum: $CSim(u, v) = \lambda_1 * sim_1(u, v) + \lambda_2 sim_2(u, v) + \ldots + \lambda_n * sim_n(u, v)$
  - Weighted average: $CSim(u,v) = \dfrac{\lambda_1 Sim_1(u,v) + \lambda_2 Sim_2(u,v) + \ldots + \lambda_n Sim_n(u,v)}{n}$
  - Machine learning: E.g., each similarity as a feature.
  - Many others.

1:m match: two types

- Part-of type: each relevant schema element on the many side is a part of the element on the one side. E.g.,
  - “Street”, “city”, and “state” in a schema are parts of “address” in another schema.
- Is-a type: each relevant element on the many side is a specialization of the schema element on the one side. E.g.,
  - “Adults” and “Children” in one schema are specializations of “Passengers” in another schema.
- Special methods are needed to identify these types (Wu et al. SIGMOD-04).

Some other issues (Rahm and Berstein 2001)

- Reuse of previous match results: when matching many schemas, earlier results may be used in later matching.
  - Transitive property: if X in schema S1 matches Y in S2, and Y also matches Z in S3, then we conclude X matches Z.
- When matching a large number of schemas, statistical approaches such as data mining can be used, rather than only doing pair-wise match.
- Schema match results can be expressed in various ways: Top N candidates, MaxDelta, Threshold, etc.
- User interaction: to pick and to correct matches.

Web is different from databases

(He and Chang, SIGMOD-03)

- Limited use of acronyms and abbreviations on the Web: but natural language words and phrases, for general public to understand.
  - Databases use acronyms and abbreviations extensively.
- Limited vocabulary: for easy understanding
- A large number of similar databases: a large number of sites offer the same services or selling the same products. Data mining is applicable!
- Additional structures: the information is usually organized in some meaningful way in the interface. E.g.,
  - Related attributes are together.
  - Hierarchical organization.

NLP connection

- Everywhere!
- Current techniques are mainly based on heuristics related to text (linguistic) similarity, structural information and patterns discovered from a large number of interfaces.
- The focus on NLP is at the word and phrase level, although there are also some sentences, e.g., “*where do you want to go*?”
- Key: identify synonyms and hypernyms relationships.

Summary

- Information integration is an active research area.
- Industrial activities are vibrant.
- We only introduced some basic integration methods and Web query interface integration.