Integração de Dados

Processamento Analítico de Dados

Profa. Dra. Cristina Dutra de Aguiar Ciferri

Prof. Dr. Ricardo Rodrigues Ciferri

Integração de Dados

- Problema: dados armazenados nos provedores
  - são heterogêneos
    - seguem diferentes modelos de dados
    - são representados por conceitos diferentes
    - possuem diferentes formatos
    - etc
  - são redundantes, inconsistentes e até mesmo complementares
- Dois níveis: esquema e instância

Integração de Esquemas

- Definição
  - especificação de mapeamentos que descrevem os relacionamento semânticos entre os esquemas dos provedores heterogêneos
- Relativismo semântico
  - conflitos entre duas ou mais representações é relacionado ao fato de que diferentes usuários modelam o mesmo pedaço do mundo real de diferentes formas, de acordo com as suas percepções

Conflito

- Conflito entre duas representações do mesmo conceito pertencentes a esquemas distintos
  - surge quando essas representações não são idênticas
- Representações idênticas
  - usam os mesmos construtores
  - aplicam as mesmas restrições de integridade

Conflito

- Tipos de conflito
  - de nome
  - semântico
  - estrutural

discrepâncias existentes entre os esquemas apresentam mais do que um tipo de conflito

Conflito de Nome

- Definição
  - nomes que representam os diferentes elementos nos esquemas a serem integrados
- Sinônimos
  - diferentes nomes são aplicados ao mesmo elemento
  - ex.: cliente representa, em um esquema, os clientes atendidos por uma loja, enquanto que comprador é usado em outro esquema para representar o mesmo caso

Conflito de Nome

- Homônimos
  - mesmo nome é aplicado a diferentes elementos
  - ex.: nome representa, em um esquema, o nome de um aluno de uma universidade, enquanto que nome representa, em outro esquema, o nome de um produto vendido em uma loja

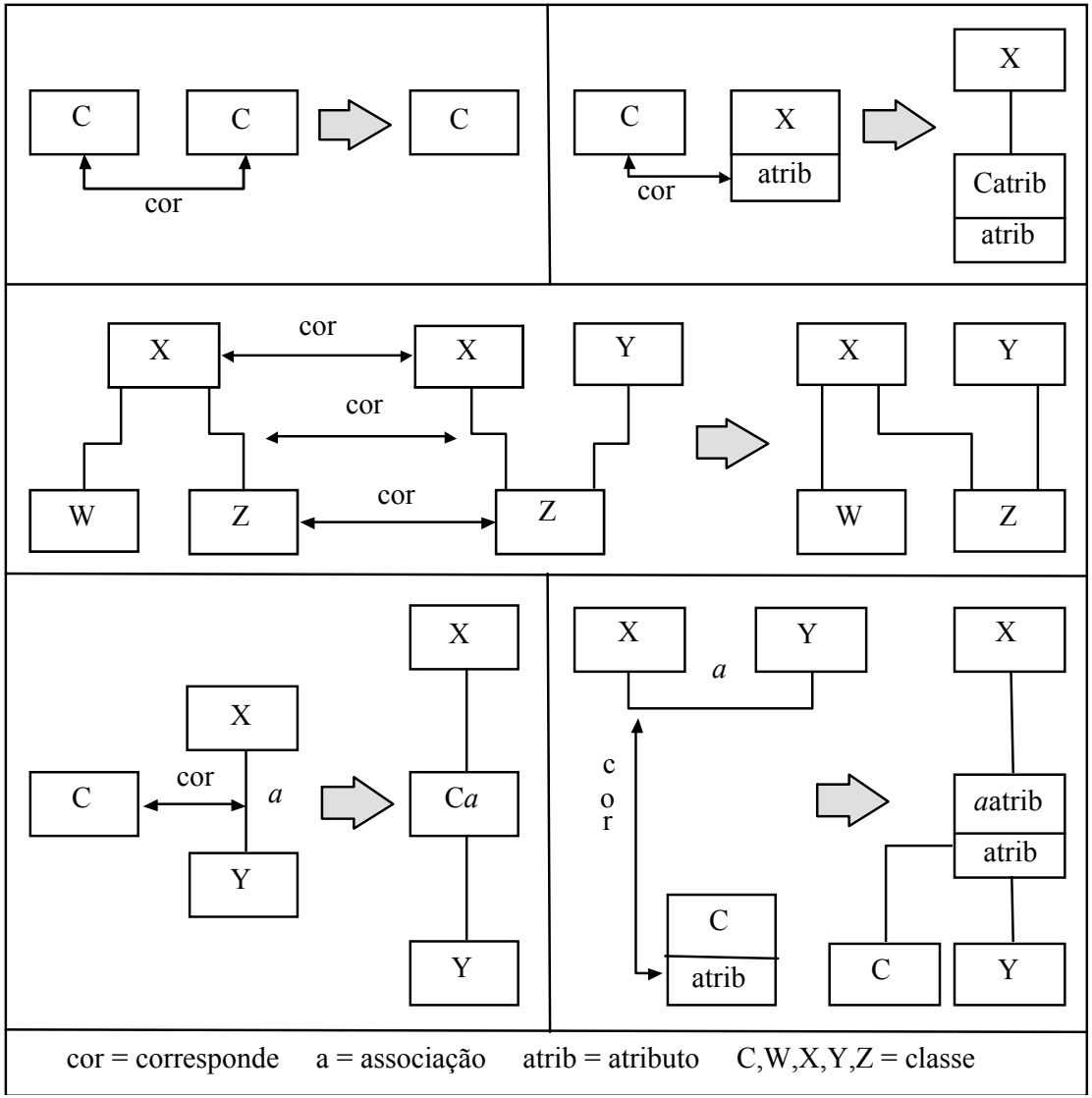
Conflito Semântico

- Definição
  - surge quando o mesmo elemento é modelado em diferentes esquemas, porém representando conjuntos que se sobrepõem
- Exemplo
  - produto representa, em um esquema, todos os produtos de um supermercado, enquanto que produto é usado em outro esquema para representar apenas os produtos da seção de cosméticos

Conflito Estrutural

- Definição
  - surge sempre que diferentes construtores estruturais são utilizados para modelar o mesmo conceito representado em diferentes aplicações
- Exemplo:
  - o mesmo conjunto de objetos do mundo real pode ser representado como um tipo-entidade em um esquema e como um atributo de um tipo-entidade em outro esquema

Exemplo de Mapeamento



SPACCAPIETRA, S., PARENT, C. View Integration: A Step Forward in Solving Structural Conflicts. *IEEE Transactions on Knowledge and Data Engineering*, v.6, n.2, p.258-274, 1994

Integração de Instâncias

- Tipos
  - ambiguidade na identificação de entidades
    - também conhecido como resolução de entidades, reconciliação de referências e deduplicação de dados
  - resolução de conflitos de valores de atributos
    - também conhecido como fusão de dados

Ambiguidade na Identificação de Entidades

- Objetivos
  - identificar quais entidades dos provedores heterogêneos referem-se à mesma entidade do mundo real
  - agrupar essas entidades em agrupamentos de entidades similares

Exemplo

Examples of entities from the class article (a)

a_1 = ({“Distributed query processing in a relational database system”}, {“169-180”}, {p_1; p_2; p_3}; {c_1})

a_2 = ({“Distributed query processing in a relational database system”}, {“169-180”},{p_4; p_5; p_6}; {c_2})

Examples of entities from the class person (p)

p_1 = ({“Robert S. Epstein”}, null, {p_2, p_3}, null)

p_2 = ({“Michael Stonebraker”}, null, {p_1, p_3}, null)

p_3 = ({“Eugene Wong”}, null, {p_1, p_2}, null)

p_4 = ({“Epstein, R.S.”}, null, {p_5, p_6}, null)

p_5 = ({“Stonebraker, M.”}, null, {p_4, p_6}, null)

p_6 = ({“Wong, E.”}, null, {p_4, p_5}, null)

p_7 = ({“Eugene Wong”}, {“eugene@berkeley.edu"}, null, {p_8})

p_8 = (null, {“stonebraker@csail.mit.edu"}, null, {p_7})

p_9 = ({“mike”}, {“stonebraker@csail.mit.edu"}, null, null)

Examples of entities from the class conference (c)

c_1 = ({“ACM Conference on Management of Data”}, {“1978”}, {“Austin, Texas”})

c_2 = ({“ACM SIGMOD”}, {“1978”}, null)

article: title, pages, *authors, *conference
person: name, email, *authors, *emailContact
conference: name, year, local

Exemplo

Grouping from entities of the class article (a)

$grouping_1 = \{a_1, a_2\}$

Grouping from entities of the class person (p)

$grouping_2 = \{p_1, p_4\}$

$grouping_3 = \{p_2, p_5, p_8, p_9\}$

$grouping_4 = \{p_3, p_6, p_7\}$

Grouping from entities of the class conference (c)

$grouping_5 = \{c_1, c_2\}$

DONG, X.; HALEVY, A.; MADHAVAN, J. Reference Reconciliation in Complex Information Spaces. In Proceedings of the ACM International Conference on Management of Data, p.85-96, 2005

Resolução de Conflitos de Valores

- Objetivo
  - resolver inconsistências nos valores dos dados das entidades que referem-se à mesma entidade do mundo real, mas que diferem nos valores dos seus atributos

Exemplo

- Considerando-se p_1 e p_4
  - em p_1, tem-se {"Robert S. Epstein"}
  - em p_4, tem-se {"Epstein, R.S."}
- Perguntas
  - Qual o valor correto?
  - Qual a entidade integrada que representa o agrupamento?

Estratégias

- Objetivo
  - resolver (ou não) os conflitos de valores
- Tipos de estratégia
  - ignorar o conflito
  - evitar o conflito
  - resolver o conflito

BLEIHOLDER, J.; NAUMANN, F. Conflict Handling Strategies in an Integrated Information System. In Proceedings of the International Workshop on Information Integration on the Web, 2006

Ignorar o Conflito

- Objetivo
  - não decidir nada
- Característica
  - não se tem conhecimento de que o conflito exista
- Exemplo
  - PASS IT ON: mostra todos os valores conflitantes ao usuário ou a uma aplicação e deixa o usuário ou a aplicação decidir como resolver os conflitos

Evitar o Conflito

- Objetivo
  - não decidir nada, mas manipular o conflito
- Característica
  - decisão de como manipular o conflito é tomada previamente, sem análise dos dados
- Exemplo
  - TRUST YOUR FRIENDS: permite definir a confiabilidade de cada provedor, e usa as confiabilidades para resolver o conflito

Resolver o Conflito

- Objetivo
  - resolver o conflito
- Característica
  - usa como base os valores de dados
- Exemplos
  - CRY WITH THE WOLVES: escolhe o valor reportado pela maioria dos provedores
  - MEET IN THE MIDLE: escolhe um novo valor, o qual é um valor mediano reportado pelos provedores