Đối sánh lược đồ

Nguyễn Hồng Phương
Email: phuong.nguyenhong@hust.edu.vn
Site: http://is.hut.edu.vn/~phuongnh
Bộ môn Hệ thống thông tin
Viện Công nghệ thông tin và Truyền thông
Đại học Bách Khoa Hà Nội

Nội dung

1. Giới thiệu

- Lược đồ là một cấu trúc siêu dữ liệu, mô tả dữ liệu có thể được lưu trữ, truy cập và thông dịch bởi người dùng và ứng dụng như thế nào.
- Ngoài khía cạnh kĩ thuật liên quan đến quản trị dữ liệu (như định dạng các trường, kiểu dữ liệu), lược đồ cũng thể hiện khía cạnh ngữ nghĩa mở rộng (nội dung và nghĩa của dữ liệu): các giá trị được phép, cardinality, ràng buộc toàn vẹn và tham chiếu.
- Một số ngôn ngữ lược đồ:
  - SQL (Structure Query Language) biểu diễn lược đồ quan hệ
  - DTD (Document Type Definition) và XSD (XML Schema Definition) biểu diễn lược đồ tài liệu XML
  - OWL (Ontology Web Language) biểu diễn ontology

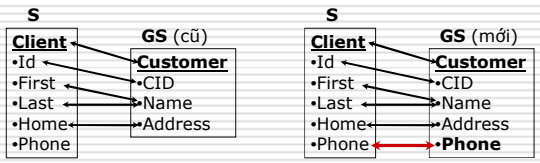
1. Giới thiệu (tiếp)

- Nhiều ứng dụng, như kho dữ liệu, mediating giữa các website, khai phá dữ liệu, quản trị dữ liệu ngang hàng,... cần tích hợp dữ liệu từ nhiều nguồn để hỗ trợ các câu truy vấn và khả năng phân tích.
- Tiến trình này, gọi là tích hợp dữ liệu, nhằm đến việc cung cấp một khung nhìn đồng bộ và nhất quán, gọi là sơ đồ tổng thể (global schema)
- Trên thực tế, việc tích hợp dữ liệu thường được thực hiện tăng trưởng bằng cách bắt đầu với một sơ đồ tổng thể đơn giản rồi thêm các nguồn dữ liệu mới vào khi cần.

1. Giới thiệu (tiếp)

- Việc tích hợp một nguồn dữ liệu mới vào sơ đồ tổng thể hiện tại được thực hiện thông qua hai bước:
  - đối sánh: lược đồ nguồn được so sánh với lược đồ tổng thể để xác định các phần tử tương đồng và các phần tử khác biệt.
  - chuyển đổi dữ liệu: sinh ra các truy vấn để chuyển đổi các dữ liệu thể hiện từ lược đồ nguồn sang lược đồ tổng thể.

Ví dụ: đối sánh lược đồ để tích hợp dữ liệu



A) Đối sánh lược đồ

Ví dụ: đối sánh lược đồ để tích hợp dữ liệu (tiếp)

Id	First	Last	Home	Phone
1	Kristen	Smith	Hurley St. 2	123
…	…	…	…	…

CID	Name	Address	Phone
1	Kristen Smith	Hurley St. 2	123
…	…	…	…

INSERT INTO GS(CID, Name, Address, Phone)
SELECT Id, *Concat*(First, Last), Home, Phone
FROM S

B) Chuyển đổi dữ liệu

7

1. Giới thiệu (tiếp)

- Việc nhận diện các cặp tương quan ngữ nghĩa giữa hai lược đồ được biết đến như là đối sánh lược đồ.

8

2. Khái niệm lược đồ

- Lược đồ có thể hiện hữu trong những định dạng và ngôn ngữ khác nhau: SQL, UML, DTD, XSD, OWL,…
  - SQL cho phép định nghĩa lược đồ cho CSDL quan hệ, truy vấn và thao tác dữ liệu lưu trữ trong lược đồ.
  - XSD mô tả cấu trúc của tài liệu XML. Thành phần chính xủa XSD là các phần tử, thuộc tính và kiểu.

9

2. Khái niệm lược đồ (tiếp)

- OWL thường được sử dụng để đặc tả ontology trên web ngữ nghĩa. Ontology nhắm đến việc khái niệm hóa tri thức miền và hỗ trợ biểu diễn một cách giàu ngữ nghĩa thế giới thực hơn là CSDL hoặc lược đồ tài liệu. OWL cung cấp cấu trúc dựa trên XML để định nghĩa lớp, mối quan hệ giữa chúng, các thuộc tính, miền giá trị của chúng. Miền giá trị của thuộc tính có thể là kiểu dữ liệu nguyên tố hoặc một lớp đã được định nghĩa.Các lớp OWL có thể có các thể hiện, lưu trữ trong cùng tài liệu XML.

10

2. Khái niệm lược đồ (tiếp)

- Một cách tổng quát, lược đồ được định nghĩa đơn giản là một tập các phần tử được nối với nhau bởi một cấu trúc nào đó.
- Ví dụ:
  - Với lược đồ quan hệ, các bảng và cột là các phần tử của lược đồ; mối quan hệ giữa các bảng, các cột và ràng buộc tham chiếu giữa các bảng là cấu trúc lược đồ.

11

2. Khái niệm lược đồ (tiếp)

- Với lược đồ XSD, phần tử lược đồ gồm các phần tử XML và các thuộc tính; cấu trúc lược đồ gồm mối quan hệ giữa phần tử và các phần tử con được xác định bởi kiểu phức hợp
- Với OWL, các lớp và các thuộc tính là các phần tử lược đồ; mối quan hệ giữa các lớp và mối quan hệ giữa các lớp với các thuộc tính hình thành cấu trúc lược đồ.

12

3. Bài toán đối sánh lược đồ

- Vấn đề đối sánh lược đồ được phát biểu như sau:
  - Cho hai lược đồ S_1 và S_2, tìm ra các cặp phần tử tương ứng phù hợp giữa S_1 và S_2, khai thác tất cả thông tin hiện có như lược đồ, dữ liệu thể hiện và nguồn phụ trợ.
- Nếu hai phần tử được cho là tương đồng, thì không nên có sự tương đồng nào giữa một trong hai phần tử này với phần tử thứ 3 khác mà chất lượng đối sánh tốt hơn.

13

Thông tin đầu vào

- Cần khai thác triệt để thông tin hiện có để hiểu được ngữ nghĩa của các phần tử lược đồ, từ đó phát hiện sự tương đồng giữa chúng.
  - Thông tin lược đồ: tên phần tử, mô tả, kiểu dữ liệu, cấu trúc lược đồ, mối quan hệ khác giữa các phần tử.
  - Dữ liệu thể hiện: trong nhiều ứng dụng, dữ liệu thể hiện luôn sẵn có cho các lược đồ.
  - Thông tin bổ trợ: tất cả các thông tin có thể khai thác để phát hiện sự tương đồng giữa các phần tử lược đồ như đồng nghĩa, phân cấp, từ điển,…

14

Thông tin ra

- Cho hai lược đồ S_1 và S_2, thao tác đối sánh trả về ánh xạ giữa chúng, là kết quả của việc đối sánh.
- Ánh xạ là một tập hợp các phần tử ánh xạ, hoặc các tương ứng; mỗi tương ứng xác định chính xác các phần tử của S_1 tương ứng với các phần tử của S_2.
- Mỗi tương ứng có thể có biểu thức ánh xạ, xác định cách mà phần tử của S_1 và S_2 liên quan với nhau.

15

Thông tin ra (tiếp)

- Biểu thức ánh xạ
  - Về ngữ nghĩa, có thể sử dụng các quan hệ đẳng hướng đơn giản, quan hệ thuật ngữ, quan hệ hướng tập, hàm (hàm nối, hàm toán học)
  - Biểu thức ánh xạ có thể có hàm ngược, ví dụ ánh xạ 1:1; hoặc không thể ánh xạ ngược.
- Phần lớn các kỹ thuật đối sánh lược đồ tự động dựa trên heuristic nên rất khó mô hình hóa toán học chính xác.

16

4. Phân loại các phương pháp đối sánh

17

4.1. Một số cách phân loại

- Lược đồ vs. thể hiện: xem xét thông tin
  - mức lược đồ như siêu dữ liệu (tên phần tử, kiểu dữ liệu, thuộc tính,…)
  - dữ liệu thể hiện (nội dung dữ liệu)
- Phần tử vs. cấu trúc:
  - so sánh từng phần tử lược đồ (như là các thuộc tính)
  - kết hợp các phần tử với nhau trong một cấu trúc
- Ngôn ngữ vs. ràng buộc:
  - tiếp cận ngôn ngữ (so sánh tên, mô tả text của phần tử)
  - tiếp cận dựa trên ràng buộc (ràng buộc định nghĩa trên các phần tử như kiểu dữ liệu, tính duy nhất, khóa,…)

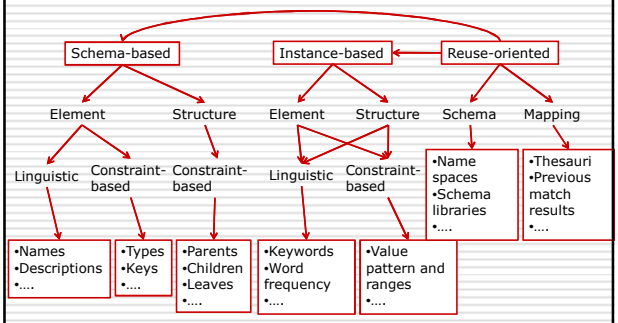
18

4.1. Một số cách phân loại

- Tái sử dụng vs. không tái sử dụng
  - sử dụng các thông tin bổ trợ: từ điển, lược đồ tổng thể, các quyết định đối sánh trước đây, đầu vào người dùng.
- Lai vs. tổng hợp
  - Bộ đối sánh có thể là sự kết hợp của một số tiếp cận riêng

19

Sơ đồ phân loại



20

4.2. Đối sánh dựa trên lược đồ

- Chỉ xem xét thông tin lược đồ
- Dựa trên tính biểu đạt của ngôn ngữ lược đồ, thông tin bao gồm các thuộc tính khác nhau của phần tử lược đồ (tên, mô tả, kiểu dữ liệu, ràng buộc,…) và mối quan hệ giữa chúng (ràng buộc tham chiếu, is-a/part-of)

21

Tiếp cận dựa trên ngôn ngữ

- Khai thác các tính chất dựa trên text của các phần tử: tên, mô tả.
- Sự tương đồng về tên có thể được đánh giá bằng cách so sánh chuỗi tên (ngữ pháp) hoặc ý nghĩa của chúng (ngữ nghĩa)

22

Tiếp cận dựa trên ngôn ngữ (tiếp)

- Đối sánh tên ngữ pháp: thuần túy so sánh hai chuỗi tên
  - đối sánh chuỗi chính xác
  - cùng không gian tên, cùng phản ánh ngữ nghĩa duy nhất
  - đối sánh chuỗi xấp xỉ: tên và tên viết tắt. Vd: Customer và Cust. Một số thuật toán:
    - EditDistance: quy hoạch động; số thao tác soạn thảo để biến một chuỗi thành chuỗi kia
    - N-Gram: Diagram, TriGram
    - SoundEx: tính toán sự tương đồng ngữ âm giữa các tên từ mã SoundEx của chúng

23

Tiếp cận dựa trên ngôn ngữ (tiếp)

- Đối sánh tên ngữ nghĩa: dựa trên mối quan hệ thuật ngữ: đồng nghĩa, phân cấp,…
  - Cần có nguồn thông tin bổ trợ như từ điển, ontology, bảng từ đồng nghĩa, từ điển đa ngôn ngữ WordNet,…
  - Hiện tượng từ đa nghĩa?

24

Tiếp cận dựa trên ngôn ngữ (tiếp)

- Về mô tả của phần tử
  - coi như là đoạn text, tài liệu
  - kỹ thuật xử lý ngôn ngữ tự nhiên, kỹ thuật tìm kiếm thông tin

25

Tiếp cận dựa trên ràng buộc

- Các ràng buộc: khai báo kiểu dữ liệu, các giá trị cho phép, miền giá trị, tính duy nhất, tùy chọn,…
- Nên có bảng so sánh cho các kiểu dữ liệu, ví dụ: string và varchar,…

26

Tiếp cận mức cấu trúc

- Khai thác mối quan hệ giữa các phần tử và đối sánh sự kết hợp của các phần tử xuất hiện cùng nhau trong một cấu trúc.
- Một số kiểu quan hệ dựa trên khả năng mô hình hóa của ngôn ngữ lược đồ
  - is-a/part-of
  - chứa đựng
  - ràng buộc tham chiếu

27

Tiếp cận mức cấu trúc (tiếp)

- Xem xét các phần tử lân cận để ước lượng sự tương đồng: nút cha, nút con, các nút lá,…

28

4.3. Đối sánh dựa trên thể hiện

- Xem xét dữ liệu thể hiện để quyết định các phần tử tương quan
- Kỹ thuật này được sử dụng trong trường hợp
  - Có ít thông tin mức lược đồ
  - Dữ liệu bán cấu trúc
  - Không có thông tin lược đồ =>Trích rút/khôi phục lược đồ
- Kỹ thuật này bổ sung và làm tăng tính chính xác cho kỹ thuật dựa trên lược đồ.

29

4.3. Đối sánh dựa trên thể hiện (tiếp)

- Vấn đề
  - lượng dữ liệu lớn
  - các kỹ thuật khai phá dữ liệu: làm sạch, trích chọn đặc trưng,…

30

Tiếp cận mức phần tử

- Đối với thuộc tính dựa trên text, các kỹ thuật tìm kiếm thông tin:
  - tìm từ khóa, chủ đề dựa trên tần suất tương đối của từ
  - sự kết hợp các từ trong thể hiện thuộc tính.
- Đối với thuộc tính số và chuỗi
  - chiều dài dữ liệu, kiểu dữ liệu, miền giá trị, trung bình, phân bố giá trị, ràng buộc khóa, tần suất các ký tự,…

31

Tiếp cận mức cấu trúc

- Xem xét các thể hiện của nhiều thuộc tính cùng lúc
- Sự kết hợp giữa các thuộc tính: có thể rất lớn

32

4.4. Đối sánh hướng tái sử dụng

- Tái sử dụng các phần lược đồ và các phần tương quan đã được xác định trước đó.
- Sử dụng tất cả các thông tin bổ trợ để cải thiện quá trình đối sánh.

33

Tái sử dụng dựa trên lược đồ

- Các tên đã được sử dụng chung được định nghĩa và duy trì trong từ điển tổng thể hoặc không gian tên.
- Khai thác thêm các đặc điểm khác của lược đồ: kiểu dữ liệu, khóa, ràng buộc.
- Vấn đề: các tổ chức khác nhau khó chấp nhận dùng chung một không gian tên, từ điển chung,…

34

Tái sử dụng dựa trên ánh xạ

- Khai thác mối quan hệ tương đồng đã được quyết định từ trước.

35

4.5. Tiếp cận kết hợp

- Kết hợp nhiều cách tiếp cận
  - Hybrid matcher: tích hợp các cách tiếp cận lại
  - Composite matcher: kết hợp các kết quả của các bộ đối sánh độc lập.

36

4.6. Match cardinality

- Một phần tử của lược đồ S_1 (hoặc S_2) có thể tham gia vào 0, 1 hoặc nhiều tương quan của kết quả đối sánh.
- Một hoặc nhiều phần tử của S_1 có thể đối sánh với 1 hoặc nhiều phần tử của S_2
  - đối sánh mức phần tử: 1:1, 1:n, n:1
  - đối sánh mức cấu trúc: n:m

37

4.6. Match cardinality (tiếp)

- Ví dụ

Cardinality	Phần tử S_1	Phần tử S_2	Biểu thức ánh xạ
1:1	Price	Cost	Price = Cost
n:1	FirstName, LastName	Name	Concat(FirstName, LastName) = Name
1:n	Name	FirstName, LastName	Split(Name) = {FirstName, LastName}
n:m	P.PersName, P.DeptNo D.DeptNo, D.DeptName	A.Person, A.Department	SELECT P.PersName, D.DeptName FROM P, D WHERE P.DeptNo = D.DeptNo = {A.Person, A.Department}

38

5. Vấn đề không thống nhất ngữ nghĩa

- Nguồn thông tin không thống nhất
- Lược đồ và dữ liệu không thống nhất
  - Để hiểu ngữ nghĩa của phần tử: tên phần tử, kiểu dữ liệu, giá trị cho phép, cấu trúc lược đồ, nhóm phần tử
  - Thông tin: không đầy đủ, không đáng tin cậy
  - Lược đồ được phát triển độc lập bởi nhiều người, với nhận thức thế giới thực khác nhau, vì mục đích khác nhau.
  - Một số ví dụ:

39

5. Vấn đề không thống nhất ngữ nghĩa

- Các tên giống nhau chưa chắc đã biểu diễn cùng ngữ nghĩa; các tên khác nhau vẫn có thể chỉ định cùng một khái niệm thế giới thực.
- Tên phần tử có thể được mã hóa hoặc viết tắt
- Ràng buộc toàn vẹn được quy định trong chương trình truy cập dữ liệu, không được khai báo ở mức lược đồ.
- Phần tử có thể được mô hình hóa ở nhiều mức độ khác nhau: thông tin địa chỉ được chia thành phố, mã vùng, thành phố ở trong lược đồ này, nhưng lại chỉ là 1 trường trong lược đồ khác.

40

5. Vấn đề không thống nhất ngữ nghĩa

- Dữ liệu thể hiện cung cấp thêm nội dung và ý nghĩa của phần tử lược đồ; tuy nhiên, thông tin này cũng có thể không nhất quán, khác nhau trên CSDL khác nhau:
  - "F", "Female" chỉ giới tính nữ
  - sử dụng các đơn vị khác nhau (Euro và Dollar,…), định dạng khác nhau,…
  - Có thể chứa lỗi chính tả,…

41

6. Ứng dụng của đối sánh lược đồ

- Tích hợp lược đồ và dữ liệu
- Thương mại điện tử
- Web ngữ nghĩa
- Quản trị mô hình

42

Lời hay ý đẹp

"Không có con đường nào quá dài đối với kẻ bước đi thong thả. Không có thành công nào quá xa vời đối với những ai kiên nhẫn làm việc"

Jean de La Bruyère