MỘT MÔ HÌNH DỮ LIỆU HƯỚNG ĐỐI TƯỢNG THỜI GIAN ĐỐI VỚI TÀI LIỆU CẤU TRÚC

Đỗ Quang Vinh

Quách Tuấn Ngọc

MỞ ĐẦU

Thư viện số đã trở thành một lĩnh vực nghiên cứu tích cực bao gồm lưu trữ khối và các cơ chế truy nhập từ xa, cũng như tổ chức và tìm kiếm thông tin lưu trữ điện tử. Những đề xuất mới đối với thư viên số tiếp cận tới lưu trữ sách, báo, tạp chí định kỳ, bằng sáng chế, hồ sơ y học, sách hướng dẫn v v... Trong nhiều phạm trù, thành công của các đề xuất này thuộc về cách các tài liệu lưu trữ được phân loại và cách thông tin này được sử dung khi tìm kiếm chúng. Trong ngữ cảnh này, thông tin mô tả về một nguồn tin được gọi là siêu dữ liệu. Hầu hết siêu dữ liệu thông thường trợ giúp bởi các hệ thống hiện thời là tác giả, nhan đề, nhà xuất bản, chủ đề, ngày tháng, kiểu, nguồn tin. người đóng góp, vai trò, LSBN v.v... Các đầu mục này được lưu trữ một lần cùng với thông tin tham chiếu tới, chúng được chỉ số hóa và sử dụng trong khi tìm kiếm tài liệu.

Trong công trình này, chúng tôi mở rộng phạm vi của siêu dữ liệu để mô tả hai đặc tính thêm vào của tài liệu: tổ chức lôgíc và chiều thời gian. Cả hai đặc tính được phát triển ở phần tiếp theo. Mục đích cuối cùng của công trình này là cung cấp một mô hình cơ sở dữ liệu thời gian để trợ giúp tất cả siêu dữ liệu và chỉ ra cách sử dụng nó trong khi tìm kiếm tài liệu. Hệ thống đ-ưa ra sẽ cho phép người dùng thư viện số nhận được các lợi ích nâng cao từ các kho tài liệu của nó.

TÀI LIỆU LỊCH SỬ

Chúng tôi coi các tài liệu lịch sử là giữ thông tin liên quan với thời gian chúng được tạo ra và sẽ có ích trong tương lai. Trong nhóm này, có báo, tạp chí định kỳ bằng sáng chế và v v... Để trợ giúp lưu trữ và tìm kiếm chúng, chúng tôi định nghĩa một mô hình cho tài liệu lịch sử được phác thảo dưới đây.

Thứ nhất, ta xem xét cấu trúc phức tạp của tài liệu. Đối với mỗi một lớp tài liệu có định nghĩa cấu trúc chung. Nghĩa là một hệ thống kiểu gốc đối với tài liệu bù hệ thống kiểu hướng đối tượng thông thường ở hai nghĩa: cho phép đối với bộ được sắp thứ tự và cung cấp hợp của các kiểu. Mặt khác, dù rằng các tài liệu không tiến hoá, ba chiều thời gian khác nhau có thể được xác định cho tài liệu lịch sử:

- Tài liệu có thể được nhóm lại bằng kiểu mà các đặc tính của nó có thể thay đổi theo thời gian. Điều này có thể được hiểu như là loại lược đồ hoặc

kiểu tiến hoá nào đó. Ví dụ, báo thường thay đổi cách trình bày và tổ chức của nó.

- Mỗi một tài liệu xuất bản hoặc biên tập tại một thời điểm đã biết. Các đặc tính của chúng phù hợp với định nghĩa kiểu hiện thời tại thời gian đó. Theo ví dụ trước, chúng tôi nhận thấy mỗi một báo được đề ngày tháng và địa phương, cách trình bày của nó tương xứng với định nghĩa hiện thời về kiểu tại ngày tháng đó.

- Chiều thời gian cuối cùng quy về thời gian người dùng tài liệu xem xét nội dung của nó là hợp lệ hoặc cập nhật. Ví dụ, trong một báo, dự báo thời tiết sẽ đúng cho hai ngày tiếp theo trong khi tin về một vụ án sẽ đúng trong toàn bộ quá trình xét xử.

Tất cả thông tin xác định tổ chức và các chiều thời gian của tài liêu lịch sử nằm trong phạm vi của siêu dữ liệu. Trong bài báo này, chúng tôi thảo luân cách biểu diễn dữ liêu. Khi tích hợp siêu dữ liệu và tài liệu vào trong một mô hình dữ liệu đơn giản, một kiến trúc hai mức cần xuất hiện. Ở lớp trên siêu dữ liệu biểu diễn thông tin về tài liệu lưu trữ ở lớp dưới. Điều này tạo thành tính mới lạ chủ yếu của mô hình dữ liệu đã đề nghị.

MÔ HÌNH DỮ LIỆU ĐỐI VỚI TÀI LIỆU LỊCH SỬ

Trong phần này, trình bày một mô hình dữ liệu hình thức đối với tài liệu lịch sử như đã định nghĩa. Bên trong một mô hình dữ liệu hướng đối tượng, chúng tôi dùng siêu lớp để bao hàm các đặc tính thời gian và hành vi mô tả lịch sử của các kiểu tài liệu và đối tượng.

Kiểu hệ thống

Mô hình dữ liệu đã đề nghị dựa vào hai kiểu hệ thống khác nhau, mang tên: MTS và DTS. Kiểu trước được dành cho định nghĩa kiểu siêu dữ liệu, trong khi kiểu sau được dành cho mô tả kiểu tài liệu. Sự phân chia thành hai hệ thống được làm bởi vì các tài liệu và siêu dữ liệu của chúng đưa ra các ký hiệu cú pháp hoàn toàn khác nhau. Tập hợp tất cả định danh duy nhất đối tượng được biểu thị bằng Ql và tập hợp tất cả tên lớp tài liệu bằng Cl.

Kiểu hệ thông MTS

Kiểu hệ thống MTS tuân theo cú pháp sau đây:

τ = ATOMIC {time} [A_1: τ_1,… A_n; τ_n] {τ }

trong đó kiểu nhóm ATOMIC chứa tất cả kiểu nguyên tử như là *integer, real, char, string v* v..., time là kiểu chỉ định miền thời gian, interval là kiểu chỉ định miền khoảng thời gian, A_1: τ_1… A_n: τ_n] là bộ xây dựng và { τ } là tập hợp xây dựng.

Các miền giá trị của kiểu nguyên tử và cấu trúc theo ngữ nghĩa là thông thường cho trước trong mô hình dữ liệu giá trị phức tạp. Về phương

diện này, để nhận được miền của một MTS kiểu τ, hàm *dom(*τ) được định nghĩa. Vì vậy, *dom (time)* là một tập hợp giá trị đẳng cấu với tập số nguyên. Các khoảng thời gian sau đó được dịch như là tập thời điểm liên tục, qua đó các toán tử lý thuyết tập hợp (như $\cap$, $\cup$ và $\subseteq$ được sử dụng để định nghĩa các quan hệ thời gian cơ bản. Miền này được biểu diễn hình thức như sau:

dom (time) $\equiv$ *TIME*= {0,1,2,3,4,5,6,...,now,...}

dom (interval $\equiv$ (INT ={[x,y] 1 x,y $\in$ TIME, x$\leq$ y}

Kiểu hệ thống DTS

Cú pháp của kiểu DTS như sau:

τ := DATA ICLASS { (τ_1{….}(τ_m)}

τ + | τ* | τ? | [A_1 : τ_1…,A_m: τ_m |

Ở đây, mỗi một kiểu DATA là một kiểu multimedia như văn bản, tranh ảnh, đồ thị, .v.v... Mỗi một kiểu CLASS là một tên lớp từ Cl. Cả hai tập hợp kiểu gồm có tập hợp kiểu dữ liệu cơ bản theo quan điểm của các kiểu cấu trúc được định nghĩa sau đây:

1 . Phần tử 1 biểu diễn hợp của các kiểu và cho dãy kiểu dữ liệu τ_1…τ_m

2. Thành phần hậu tố + biểu thị thành phần được kỳ vọng xuất hiện tối thiểu một lần, * chỉ định thành phần có thể xuất hiện 0 hoặc nhiều lần hơn và ? biểu thị thành phần có thể xuất hiện 1 lần hoặc 0 lần.

3. Cấu trúc cây được tạo thành với các *bộ được sắp xếp lồng nhau* dùng các tên thuộc tính Ai như là các nút khái niệm.

Nhận xét rằng kiểu hệ thống DTS tương tự với ngôn ngữ định nghĩa kiểu dữ liệu DTD của SGML. Thật vậy, cả hai chủ yếu nhằm tới biểu diễn cấu trúc chung tài liệu theo quan điểm linh hoạt bằng một cú pháp dựa vào ngữ pháp.

Kiểu hệ thống DTS được có xu hướng trình bày tất cả kiểu tài liệu lịch sử. Điều này nghĩa là mỗi một kiểu tài liệu liên quan tới một khoảng thời gian nó có thể được sử dụng. Một cách nhất quán, các giá trị sinh ra cho mỗi một kiểu lịch sử cũng phải phụ thuộc thời gian. Về phương diện này, {τ } ký hiệu tập giá trị hợp lệ của một DTS kiểu τ tại một thời điểm cho trước t. Hàm này là ấn bản thời gian của hàm *dom* đã định nghĩa cho kiểu hệ thống MTS.

Siêu lớp

Từ một quan điểm khái niệm, một siêu lớp là một lớp của các lớp. Nói cách khác, một siêu lớp trừu tượng hoá các đặc tính thao tác và cấu trúc thông thường của một tập hợp các lớp. Vì vậy, một siêu lớp được định nghĩa như là một bộ 5:

MC = <id, *meta_type, c-meth, min_ type, o_meth*>

trong đó: *id* là định danh của siêu lớp, *meta_type* là một kiểu MTS, *c_meth* là một tập ký hiệu phương thức đã định nghĩa qua kiểu hệ thống MTS, *min_ type là* một kiểu DTS và *o _meth* là một tập ký hiệu phương thức đã định nghĩa qua kiểu DTS.

Trạng thái của một lớp biểu diễn siêu dữ liệu của nó, được định nghĩa trong thành phần của siêu lớp *meta_type.* Mặt khác, hành vi của một siêu lớp được định nghĩa bằng tập phương thức *c_ meth.* Trong số khác, những điều này bao gồm tạo ra các trường hợp và kiểm tra sự nhất quán của chúng. Đồng thời, mỗi một siêu lớp bao gồm một phân cấp lớp chỉ có một gốc liên quan tới kiểu *min_ type* và tập phương thức o_ *meth:* Về hướng này tất cả lớp của một siêu lớp chia sẻ một kiểu và hành vi thông thường.

Do bản chất của các ứng dụng gần kề, hành vi cơ bản của toàn thể mô hình dữ liệu có thể được biểu diễn ở mức siêu lớp. Chú ý rằng hầu hết thao tác chính bao hàm trong thư viện số được rút gọn về chèn tài liệu mới và tìm kiếm tài liệu lưu trữ. Đây là lý do ở phần còn lại của bài báo, phần thao tác của mô hình dữ liệu bị bỏ qua.

Cuối cùng, chúng tôi giả sử rằng các siêu lớp không thể tạo thành hệ thống phân cấp, đó là mật độ siêu lớp hoàn toàn là rời rạc. Điều này đảm bảo rằng mật độ đối tượng không tham gia vào các hệ thống phân cấp khác nhau. Đáng chú ý là tính bất biến này đặc biệt thích hợp với các ứng dụng của chúng ta vì cơ sở tài liệu đối với thư viện số cần quản lý các loại dữ liệu rất khác nhau (như multimedia, tài liệu có cấu trúc, v.v...) phải được sắp xếp trong hệ thống phân cấp rời rạc.

Lớp

Như đã chú ý trước đó, các kiểu tài liệu lịch sử có thể tiến hoá theo thời gian. Nói riêng, các lớp tài liệu đượctrợ giúp để thay đổi như là siêu dữ liệu mới và các thành phần cần được trợ giúp đối với chúng. Đặc tính này dẫn chúng ta tới xác định ngữ nghĩa phụ thuộc thời gian đối với lớp, sự kế thừa và lược đồ cơ sở dữ liệu.

Chúng tôi định nghĩa một ký hiệu lớp như là bộ 5 sau đây:

C = <id, *lifespan, type, history, mc*>

trong đó:

- *id* là định danh lớp.

- *lifespan* là giá trị của INT chỉ định thời gian hợp lệ của lớp,

- *type* là giá trị từ *{historic, static}* chỉ định nếu lớp tiến hoá,

history là một bộ 4 như sau:

(h-type= (τ_1@i_n),c_state = (v_1@j_1…,v_k@j_k)

*i*_ ext= (p_1 @$i_{1...}$pn@i_n), m-ext =(p* 1 @ i_1… p *n @i_n),

trong đó: τ_1…τ_n ký hiệu các kiểu DTS, (1 ...(k ký hiệu các giá trị MTS, p1…'pn , p*$_1$ …p*n là các tập hợp định danh đối tượng từ O1 và i_1 ...in, j1…jk là các khoảng thời gian từ LNT. Ở đây, thành phần *h-type* biểu diễn lịch sử của kiểu C. Thành phần *c-state* biểu diễn lịch sử của trạng thái lớp. Cuối cùng, i-ext chứa mật độ thích hợp và *m*_ ext chứa mật độ mở rộng. Các khoảng của các chuỗi này phải nối tiếp nhau sao cho chúng là rời rạc và thành phần của chúng trùng khớp với lifespan của C. Chú ý rằng các khoảng của h-*type, ext và p*_ ext tạo thành dãy thời gian như nhau.

- *mc* là siêu lớp mà lớp C thuộc vê nó.

Như là một trường hợp siêu lớp, một lớp phải là một trường hợp nhất quán. Vì thế, tất cả trạng thái lịch sử của một lớp phải là các giá trị tương thích với kiểu siêu lớp của nó. Một cách hình thức, với điều kiện M là siêu lớp mà một lớp C thuộc về nó, điều kiện sau đây phải thoả mãn:

đối với mỗi một (v @i) ∈ *C.his- tory. C_state, (v "dom(M.meta_type)*

Đối tượng

Khác lớp, đối tượng của mô hình dữ liệu không tiến hoá theo thời gian. Đặc biệt là, cả sự cập nhật lẫn sự di trú đối tượng là không được phép ở thư viện số tài liệu lịch sử. Tuy nhiên, như đã mô tả ở mở đầu, các đối tượng biểu thị hai chiều thời gian: thời gian biên tập và thời gian hợp lệ. Thời gian trước là một thời điểm phục vụ để gắn mỗi một đối tượng với kiểu của nó trong khi thời gian sau biểu diễn khoảng thời gian đối tượng được coi là cập nhật hoặc hợp lệ. Về các quan niệm trên, một ký hiệu đối tượng được liên kết với bộ 5 sau đây:

O = <*oid, e_ time, vtime, value, c_ id*>

trong đó oid ∈ OI là định danh đối tượng, e_*time* là một thời điểm, *v_ time* là một khoảng thời gian, *value* là một giá trị từ kiểu hệ thống DTS và c_id là tên của lớp đặc trưng nhất mà đối tượng O thuộc về nó tại thời điểm *e_time.*

Định nghĩa tính nhất quán đối tượng: một đối tượng O được coi là nhất quán nếu

O value ∈ {type(O.c_id,t)} t với t = O, e_ *time*

Định nghĩa tập đối tợng nhất quán: một tập OBJ là một tập nhất quán của các đối tượng nếu các điều kiện sau đây thoả mãn:

1 . Đối với tất cả đối tượng o ∈ OBJ, o là một đối tượng nhất quán,

2. Đối với mỗi một cặp o, o’ ∈ OBJ, nếu *o.oid=o'.oid* thì *o.e_time= o'e_time, o.v =o'.v* và *o. vt= o’ vt,*

3. Đối với tất cả đối tượng o ∈ OBJ, mỗi một định danh ở *ref (o)* phải được chứa trong I(OBJ),

4. Đối với mỗi một cặp o,o'∈OBJ, nếu o’.oid∈*ref(o)* thì *o.e_time=o'.e_ time,*

5. Đối với mỗi một cặp o.o' ∈ OBJ, nếu *o'.oid ∈ ref(o)* thì *o’* v-time ⊆ *o.v_time*

Điều kiện 1 tính cho tính nhất quán của các đối tượng cô lập, điều kiện 2 tính cho tính đồng nhất đối tượng và điều kiện 3 tính cho tính toàn vẹn tham chiếu. Các điều kiện 4 và 5 là các ràng buộc về sự kết hợp hệ thống phân cấp của các đối tượng đã mô tả trước đây.

Sự kế thừa

Hệ thống phân cấp lớp cho phép người dùng xác định các quan hệ kế thừa giữa các lớp. Trong mô hình dữ liệu, một lớp có thể chỉ liên quan tới các lớp khác thông qua sự kế thừa nếu quãng đời của nó xuất hiện trong quãng đời của các siêu lớp của nó. Hơn nữa, do sự tiến hoá lớp, quan hệ này không cần giữ gìn trong toàn bộ quãng đời của lớp con.

Ở đây, *một hệ thống phân cấp lớp* được định nghĩa là bộ <CL, <is- A>, trong đó *CL* là một tập tên lớp và < is - A là một quan hệ tam nguyên tạo thành bằng các cặp lớp từ *CL* và khoảng thời gian từ INT. Mỗi một bộ <c, c’, i> từ < is-A được dịch vì c là một lớp con của c' trong khoảng thời gian i. Vì vậy, điều kiện sau đây phải đúng đối với quan hệ này:

i ⊆ lifespan(c) ⊆ (lifespan(c’)

Như quan hệ < is-A, quan hệ kiểu phụ ký hiệu với < t là một quan hệ tam nguyên tạo thành bởi các cặp kiểu từ DTS và thời điểm từ TIME. Cả hai quan hệ< is-A và < t phải liên quan với nhau nhất quán để tạo thành các hệ thống phân cấp lớp nhất quán. Các định nghĩa sau đây định nghĩa các ràng buộc như thế giữa hai quan hệ.

Định nghĩa hợp thành trong: một hệ thống phân cấp lớp <CL, < is- A > là hợp thành trong không chắc chắn đối với mỗi một <c,c' , i >∈ < is-A chứa type (c,t), type(c'.t),t> ∈ < t đối với mọi t ∈ i.

Định nghĩa hợp thành ngoài: một hệ thống phân cấp lớp <CL, <is -A > là hợp thành ngoài không chắc chắn đối với mỗi một <c, c>, i >∈ <is- A chứa π *(c,t) ⊆ (*(c,t) đối với mọi t ∈ i.

Cơ sở dữ liệu tài liệu lịch sử

Phần này liên kết đồng thời tất cả định nghĩa trước đây để đưa vào công thức đối với một cơ sở dữ liệu tài liệu lịch sử. Theo các mô hình cơ sở dữ liệu hướng đối tượng khác, chúng tôi phân biệt giữa lược đồ của cơ sở dữ liệu (tập lớp) và trường hợp của nó (tập đối tượng) . Lược đồ bao gồm các định nghĩa về siêu lớp đem lại các hệ thống phân cấp lớp rời rạc và cấu trúc thông thường đối với các trạng thái lớp của chúng. Như vậy, ký hiệu cho một lược đồ cơ sở dữ liệu được định nghĩa nh sau:

Schema = <MCL, CL- Def, < is- A>

trong đó *MCL* là một tập hợp các định nghĩa siêu lớp. *Cl-Def* là một tập hợp các định nghĩa lớp nhất quán tạo thành một hệ thống phân cấp lớp hợp thành trong nằm dưới quan hệ kế thừa < ls-A. Mặt khác, một cơ sở trường hợp của lược đồ cơ sở dữ liệu ở trên là một tập đối tượng nhất quán OBJ như được định nghĩa ở phần 2.4. Thêm vào, hệ thống phân cấp lớp tạo thành bởi lược đồ phải là hợp thành ngoài đối với OBJ.

KẾT LUẬN

Một thư viện như là một cơ sở dữ liệu tài liệu lịch sử hướng đối tượng thời gian trợ giúp siêu dữ liệu, đóng vai trò nổi bật ở các thư viện số hiện thời. Sự khác nhau với các đề xuất khác là mô hình tích hợp và đồng thời phân biệt giữa siêu dữ liệu và phần còn lại của các thuộc tính và nội dung tài liệu (như khoá, nhan đề v.v...). bỏ qua phần còn lại của thông tin về tài liệu lưu trữ. Mặt khác, các mô hình dữ liệu hướng đối tượng hiện thời không trợ giúp siêu dữ liệu như mô tả ở đây. Cách dùng siêu dữ liệu trong cách tiếp cận cho phép nâng cao các ứng dụng tài liệu trước đây theo một số hướng. Thứ nhất, lưu trữ và chiếu các nội dung toàn phần của tải liệu với tổ chức gốc của chúng. Thứ hai, một chiều thời gian (v-time) đã được thêm vào đề làm cho thuận tiện đặc tả về các quan hệ thời gian giữa các tài liệu được tìm kiếm.

TÀI LIỆU THAM KHẢO

1. C J.DATE, *An Introduction to Database Systems,* 6th Ed., Ad- dision-wesley, 1995.

2. V.S. SUBRAMANIAN, *Principles of Multimedia Database System,* Morgan Kaufmann, 1998.

3. ĐỖ TRUNG TUẤN. *Cơ sở dữ liệu*, Nhà xuất bản Giáo dục, 1998.

4. A.CHILVERS, J.Feather, *The management of digital data: a metadata approach, The Electronic Library.* 16 (1 998) 365-372.