*Tạp chí Tin học và Điều khiển học, T.24, S.3 (2008), 217–228*

# XÂY DỰNG GIẢI PHÁP TÌM KIẾM TRONG CƠ SỞ DỮ LIỆU ẢNH VỚI THÔNG TIN MÔ TẢ MỜ

NGUYỄN THU ANH[1], TRẦN THÁI SƠN[1], LÊ QUỐC THÁI[1], LÊ NGỌC THẮNG[2]

[1] *Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam*
[2] *Trung tâm Công nghệ Hội tụ Đa phương tiện, Viện Khoa học và Công nghệ Việt Nam*

**Abstract.** In this paper, we present a retrieval method in photograph' database with fuzzy information to solve problem of clinical photograph' retrieval. This method is a collective approach, using the results of Hedge Algebra and Similarity Relations of fuzzy data to calculate values of descriptions in the Database, containing Metadata.

**Tóm tắt.** Bài báo trình bày giải pháp tìm kiếm trong cơ sở dữ liệu ảnh với thông tin mô tả mờ để giải quyết bài toán tìm kiếm ảnh bệnh học. Giải pháp là cách tiếp cận tổng hợp, sử dụng các kết quả của Đại số gia tử, quan hệ tương tự của dữ liệu mờ để xử lý các giá trị mô tả trong một cơ sở dữ liệu có chứa các siêu thông tin (Metadata).

## 1. MỞ ĐẦU

Trong nhiều bài toán thực tế, thông tin ta cần quản lý rất phức tạp và đa dạng, có thể là thông tin mờ, không có cấu trúc tốt để chuẩn hoá, dẫn đến việc tìm kiếm không hiệu quả. Trong các mô hình cơ sở dữ liệu đã được xây dựng trên thế giới cũng như ở Việt Nam, như mô hình tập con mờ [3,5], mô hình dựa trên quan hệ tương tự [2,4], mô hình dựa trên lý thuyết khả năng [11]... thông thường là mở rộng mô hình CSDL cổ điển bằng cách lưu thêm hàm thuộc (membership functions) hoặc độ đo khả năng (possibility measure) để tính toán mỗi khi thao tác trên các dữ liệu mờ. Với cách biểu diễn và lưu trữ dữ liệu như vậy, việc tìm kiếm có thể tiến hành được nhưng thường phải thực hiện các tính toán phức tạp và trong nhiều trường hợp gây ra sai số lớn (cho ra quá nhiều kết quả tìm kiếm thiếu chính xác). Mới đây, sử dụng các nghiên cứu về Đại số gia tử (ĐSGT), một số kết quả về xây dựng CSDL mờ đã được công bố [9,10]. CDSL mờ sử dụng theo hướng áp dụng ĐSGT cho ta một cách quản lý và thao tác dữ liệu thuận tiện và cũng hợp lý hơn so với những cách tiếp cận nêu trên. Tuy vậy, trong các nghiên cứu này, dữ liệu mờ đang còn giới hạn trong phạm vi các giá trị của biến ngôn ngữ có cấu trúc, tức là có thể đưa vào trong một ĐSGT tuyến tính.

Trong bài báo này, xuất phát từ nhu cầu thực tế cần quản lý và tìm kiếm thông tin trong cơ sở dữ liệu ảnh bệnh học, bao gồm các thông tin ở dạng giá trị của các biến ngôn ngữ có cấu trúc hoặc không, chúng tôi đưa ra một cách biểu diễn giá trị mờ tổng hợp, có thể sử dụng cho bài toán tìm kiếm thông tin trong CSDL ảnh bệnh học, đồng thời dễ dàng áp dụng

*Bài báo được hoàn thành trong khuôn khổ đề tài nghiên cứu cơ bản cấp nhà nước KC-01

cho các kiểu bài toán có dữ liệu không chuẩn tương tự. Vì cho phép lưu trữ nhiều kiểu dữ liệu khác nhau nên bên cạnh việc quản lý dữ liệu thông thường, trong CSDL này còn lưu các siêu thông tin (metadata - thông tin về thông tin) xác định kiểu dữ liệu, miền trị cũng như các hàm xử lý dữ liệu tương ứng.

Nội dung bài báo gồm 4 phần. Phần đầu giới thiệu về bài toán tìm kiếm ảnh bệnh học. Phần hai chúng tôi tóm tắt một số kiến thức cơ bản sẽ sử dụng trong bài về tập mờ, ĐSGT và quan hệ tương tự mờ cũng như phép kết nhập mờ có trọng số. Sau đó là trọng tâm của bài báo, chúng tôi giới thiệu giải pháp tìm kiếm trong CSDL mờ ảnh bệnh học trên cơ sở áp dụng tổng hợp các kết quả về ĐSGT và quạn hệ tương tự mờ để biểu diễn thông tin mô tả mờ của ảnh. Phần tiếp theo sẽ là các kết quả thực nghiệm tìm kiếm ảnh bệnh học và kết luận.

## 2. BÀI TOÁN TÌM KIẾM CSDL ẢNH BỆNH HỌC

Trong giảng dạy và chẩn đoán các bệnh y tế nhất là các bệnh về da liễu như bệnh phong, các bệnh lây nhiễm qua đường tình dục v.v..,việc sử dụng các ảnh bệnh đặc trưng để giảng cho học viên cũng như để so sánh các hình ảnh trong chẩn đoán là hết sức quan trọng. Số lượng ảnh y tế này là khá lớn và những thông tin mô tả ảnh theo từng bệnh là rất phong phú và đa dạng. Vì vậy để có thể giúp các bác sĩ có thể tìm kiếm các ảnh theo yêu cầu nhằm phục vụ cho giảng dạy cũng như trợ giúp chẩn đoán bệnh cần xây dựng một cơ sở dữ liệu ảnh bệnh và các công cụ tìm kiếm phục vụ giảng dạy và chẩn đoán.

Trong phạm vi bài báo này chúng tôi chỉ xin đề cập đến hai vấn đề cần giải quyết đó là giải pháp biểu diễn các thông tin về ảnh bệnh trong cơ sở dữ liệu và xây dựng thuật toán tìm kiếm với dữ liệu bệnh da liễu. Các thông tin về ảnh bệnh ở đây chỉ bao gồm các thông tin mô tả ảnh được thực hiện bởi các chuyên gia y tế chứ không có các thông tin nội dung ảnh được thực hiện bởi các phương tiện hay các thuật toán xử lý ảnh. Điều này phù hợp với thực tế của ngành y tế Việt Nam thời điểm hiện tại.

Trong ngành da liễu, muốn chẩn đoán được bệnh thì phải căn cứ vào các kết quả thăm khám như: (i) Tổn thương cơ bản; (ii) Triệu chứng lâm sàng kèm theo; (iii) Dấu hiệu kèm theo; và (iv) Xét nghiệm cận lâm sàng. Tuy nhiên, mỗi bệnh lại có những đặc thù riêng, không có mô tả giống nhau dẫn đến cấu trúc dữ liệu rất khác nhau.

- Đối với bệnh phong (leprosy), có 4 thể bệnh khác nhau tương ứng với sự khác biệt về các dạng thương tổn cơ bản. Trong chuyên môn, người ta phân loại bệnh phong thành các thể: bất định (I: Indeterminate), củ (T: Tuberculoid), trung gian (B: Borderline), và u (L: Lepromatous).
- Với nhóm các bệnh lây truyền qua đường tình dục (STD - Sexuall Transmitted Diseases) thì lại có tới 10 bệnh: giang mai (syphilis), lậu (gonorrhea), hạ cam, u hạt bẹn (donovanose), hột xoài (nicolas-favre), éc-pét sinh dục (genital herpes), sùi mào gà (genital wart), rận mu (pediculosis pubis), và ghẻ (scabies).
- Những mô tả từ $(i) \rightarrow (iv)$ cũng khác nhau qua từng loại bệnh.

Ngoài ra, bản thân từng bệnh cũng lại được phân loại theo các cách khác nhau. Ví dụ: bệnh giang mai có 2 cách phân loại, thứ nhất là theo thể bệnh (theo cách này thì có 3 thể

khác nhau), thứ hai là theo giai đoạn (mà theo cách này lại chia thành 2 loại nữa, và đặc điểm của từng loại cũng khác nhau). Trong khi đó bệnh lậu lại phân loại theo đối tượng lây nhiễm (nam, nữ), hoặc cơ quan bị nhiễm bệnh (chẳng hạn ở mắt).
- Các tổn thương cơ bản được thể hiện rõ hơn bởi nhiều đặc điểm. Ví dụ: hình dạng, độ bằng phẳng, mức độ thâm nhiễm, màu sắc, kích thước, số lượng, ranh giới với vùng da lành v.v.. Và số đặc điểm là không giống nhau với mỗi bệnh khác nhau, thậm chí khác nhau trên từng thể bệnh.
- Kiểu dữ liệu đa dạng có khi có cấu trúc (tức là thông tin có thể chuẩn hoá), có trường hợp là ngôn ngữ tự nhiên (không xác định/tuỳ ý), nhưng cũng có lúc vừa nhận dữ liệu kiểu số lại vừa nhận giá trị ngôn ngữ (chẳng hạn, khi mô tả về số lượng tổn thương phát hiện được trên hình ảnh phóng xạ, miền giá trị có thể nói là "1", "2", "3", "4" hay "nhiều", hoặc "tương đối nhiều",... tuỳ vào từng trường hợp cụ thể. Tuy nhiên một đặc thù nổi lên là các điểm đặc trưng của mỗi dạng tổn thương được mô tả bằng ngôn ngữ.

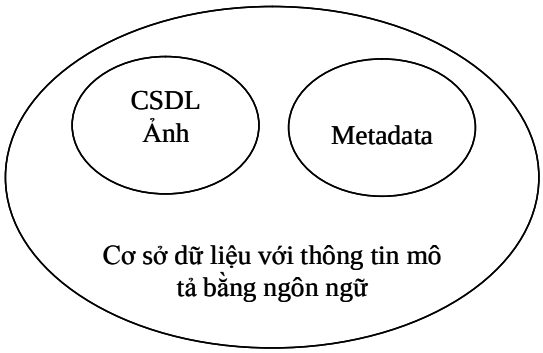
Để thể hiện đầy đủ hơn về từng loại tổn thương, không có cách gì rõ ràng hơn là các ảnh về bệnh học tương ứng thu thập được. Ngược lại, cũng có thể nói rằng mỗi ảnh là một thể hiện trực quan về một tập các đặc trưng bệnh học của một thể bệnh tương ứng. Vì vậy, để minh hoạ cho một bệnh da liễu nào đó, có thể có nhiều ảnh khác nhau được thu thập. Hơn nữa, mỗi ảnh lại được chụp trên một ca bệnh cụ thể nên ngoài thông tin mô tả kèm theo, nó còn mang thêm thông tin về giai đoạn bệnh (khởi phát, toàn phát, giai đoạn cuối, hay đã chuyển sang biến chứng/di chứng).

Tóm lại, dữ liệu về các bệnh da liễu nói chung có cấu trúc tương đối phức tạp. Trong tình hình như vậy, để xây dựng CSDL y tế hỗ trợ trong giảng dạy và chẩn đoán bệnh da liễu, cần phải giải quyết ba vấn đề sau đây trong thiết kế CSDL:
• Một là, phải đảm bảo kiến trúc dữ liệu đơn giản dễ hiểu, nhưng tính linh hoạt cao, khi cần cho phép thay đổi dễ dàng mà không mất nhiều công sức.
• Hai là, phải đề xuất ra các phương thức tính toán, xử lý, đủ linh hoạt để thao tác trên nhiều kiểu dữ liệu khác nhau.
• Ba là, tiến hành chuẩn hoá thông tin sao cho việc thao tác trên các dữ liệu thuận lợi.

Bài toán tìm kiếm ảnh bệnh học được đặt ra là: tìm trong CSDL ảnh bệnh học cho trước những ảnh đáp ứng yêu cầu của người sử dụng . Các yêu cầu này thường ở dạng ngôn ngữ, khá tự do, không chuẩn hoá và do đó các giá trị tìm kiếm do người sử dụng đưa ra có thể không có trong các giá trị lưu trữ trong CSDL. Thí dụ, tìm các ảnh của bệnh nhân bị bệnh phong, vết thương có màu sận đổ (hay tím bầm..) và khá sâu (hay rất sâu, tương đối sâu..). Với những mô hình cơ sở dữ liệu mờ đã nêu trong phần mở đầu, việc tìm kiếm thông tin, như đã nói ở trên sẽ gặp phải những phức tạp về tính toán và tính chính xác của các kết quả đưa ra. Vì khuôn khổ của bài báo, chúng tôi sẽ chỉ ra những nhược điểm của các mô hình CSDL mờ trên thông qua thí dụ CSDL mờ sử dụng hàm thuộc. Ở mô hình này, dữ liệu mờ có thể được biểu diễn dưới dạng bộ $\lfloor A, \mu A(x) \rfloor$, trong đó $A$ là tập mờ (thí dụ "cao"), còn $\mu A(x)$ là độ thuộc của giá trị thuộc tính x tại bộ đó vào tập mờ $A$ (thí dụ người cao khoảng 1m 65 có độ thuộc vào tập "cao" là 0,7). Khi thực hiện câu hỏi tìm kiếm, thí dụ "tìm tất cả những người rất cao" ta phải tính toán các hàm thuộc trong tập giá trị của thuộc tính chiều cao. Với người có độ thuộc vào tập cao là 0,7 thì theo Zadeh sẽ có độ thuộc vào tập "rất

cao" là $0,7^2 = 0,49$. Sau khi tính hết các giá trị độ thuộc vào tập mờ "rất cao" ta có thể cho kết quả là những người mà độ thuộc đạt một ngưỡng nào đó (thí dụ $> 0,5$). Tuy nhiên, các phức tạp nẩy sinh ở đây sẽ là: thứ nhất, việc xác định các hàm thuộc là rất khó khăn, kể cả với các chuyên gia; thứ hai, tính toán trên các hàm thuộc thường là phức tạp (do các hàm này thường là các hàm thực không tuyến tính và không phải lúc nào việc tính cũng đơn giản là lấy bình phương như trong thí dụ trên); thứ ba, sau khi tính toán, việc suy ngược lại tập mờ mà hàm kết quả biểu diễn là không khả thi; và cuối cùng nhưng lại là quan trọng nhất là các kết quả biến đổi hàm thuộc đôi khi cho các kết quả không phù hợp với ngữ nghĩa tập mờ mà nó biểu hiện.Về vấn đề này có thể xem thêm [7, 8, 10].

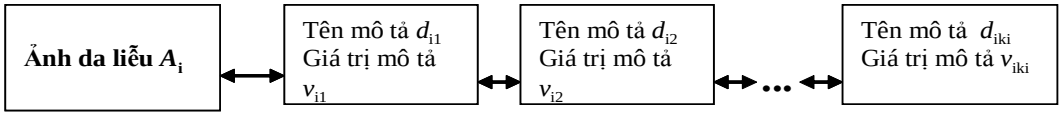
Trong bài báo này chúng tôi lựa chọn một giải pháp là xây dựng CSDL ảnh cùng mô tả của chúng trên một nền cấu trúc dữ liệu và ngữ nghĩa của nó được mô tả bởi các siêu dữ liệu (Metadata). Siêu dữ liệu giúp ta định nghĩa dữ liệu, ngữ nghĩa của chúng, quan hệ giữa các dữ liệu (cấu trúc) và như vậy cho phép thay đổi các thuộc tính, kiểu dữ liệu, cấu trúc một cách linh hoạt khi cần. Siêu dữ liệu sẽ chứa thông tin mô tả về các thuộc tính, dữ liệu về các đại số gia tử và quan hệ tương tự tương ứng với các thuộc tính ngôn ngữ và thuộc tính mờ. Ngoài ra siêu dữ liệu cũng chứa các hàm kết nhập và hệ sẽ cho phép người dùng có thể thay đổi các hàm kết nhập theo ý muốn. Có thể mô hình hoá cơ sở dữ liệu như sau:




*Hình 1.1.* Mô hình cơ sở dữ liệu với thông tin mô tả bằng ngôn ngữ

Về bản chất Metadata cũng chính là một cơ sở dữ liệu quan hệ lưu trữ đầy đủ các tham số về định nghĩa dữ liệu, ngữ nghĩa các thuộc tính mờ và cấu trúc dữ liệu. Cụ thể, Metadata lưu giữ danh sách tất cả các tên mô tả (các thuộc tính dùng để mô tả một ảnh bệnh học), kiểu dữ liệu của tên mô tả ấy (kinh điển, mờ có cấu trúc, mờ không cấu trúc) và miền xác định của từng tên mô tả (với kiểu dữ liệu kinh điển là miền giá trị thực, với kiểu mờ có cấu trúc là ĐSGT tương ứng, với kiểu mờ không cấu trúc - quan hệ mờ tương tự). Ngoài ra Metadata còn lưu giữ các hàm kết nhập để dùng khi tìm kiếm ảnh. Về phần của CSDL ảnh, ở đây sẽ là tập các ảnh da liễu mà mỗi ảnh chứa một tập các mô tả tương ứng. Mỗi mô tả là một bộ bao gồm hai thành phần ⌊ tên mô tả, giá trị mô tả⌋. Sở dĩ ta không dùng lược đồ quan hệ như trên vì khi đó rất nhiều cột (tương ứng với nhiều thuộc tính) sẽ không nhận giá trị (chính xác hơn là nhận giá trị không biết) vì không có giá trị mô tả bệnh ở đó, dẫn đến việc tốn bộ nhớ lưu trữ, đồng thời gây bất tiện cho việc tìm kiếm, xử lý dữ liệu. Ta sẽ chỉ lưu các giá trị của các mô tả được xác định. Như vậy cấu trúc dữ liệu ở đây sẽ là mỗi ảnh da liễu gắn với một dãy các đối tượng, mỗi đối tượng bao gồm ⌊ tên mô tả, giá trị mô tả⌋.

Có thể mô hình hóa một ảnh da liễu như sau.




*Hình 1.2.* Mô hình ảnh da liễu

Ký hiệu $DES$ là tập tất cả các tên mô tả, $DES = \{d_1, d_2, ..., d_n\}$. Mỗi một tên mô tả $d_i$ gắn với miền giá trị $Dom(d_i)$, $i = 1...n$; $Dom(d_i)$, $i = 1...n$ có thể là tập các số thực, tập các giá trị của biến ngôn ngữ có cấu trúc (có thể sắp xếp được) hoặc không. $Dom(d_i)$ cũng có thể là tập hỗn hợp của các kiểu giá trị trên.

Mỗi ảnh $A_i$ tương ứng với tập mô tả $A_i = \{\lfloor d_{i1}, v_{i1}\rfloor, \lfloor d_{i2}, v_{i2}\rfloor, ..., \lfloor d_{iki}, v_{iki}\rfloor\}$ trong đó $d_{ij}$ là tên mô tả, $v_{ij}$ là giá trị mô tả, nằm trong $Dom(d_{ij})$. CSDL ảnh của ta là tập $\{A_i | i = 1...m\}$.

Miền giá trị của các mô tả có thể là số: 1, 2, 3 (số chấn thương), có thể là giá trị biến ngôn ngữ có cấu trúc như tổn thương sâu, rất sâu, khá sâu..., hoặc không có cấu trúc như đỏ, tím bầm. Kỹ thuật xử lý xâu cho phép ta dễ dàng thêm, bớt, sửa đổi và tìm kiếm theo yêu cầu.

Để xử lý các câu hỏi tìm kiếm dạng này trong CSDL ảnh nêu trên, ta sẽ sử dụng các kết quả của lý thuyết tập mờ, cụ thể là ĐSGT và Quan hệ tương tự, sẽ được trình bày ở phần sau.

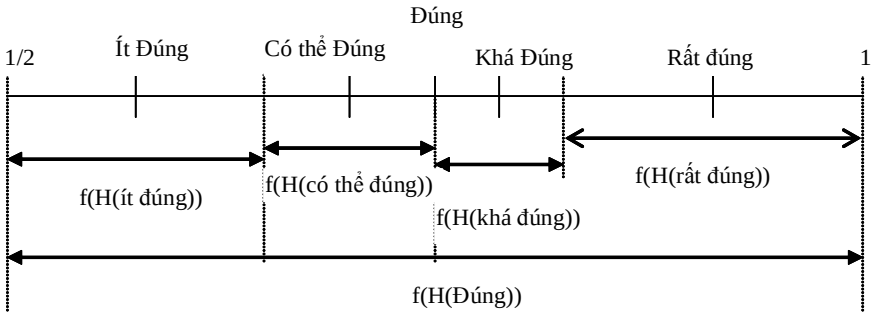
## 2. ĐẠI SỐ GIA TỬ VÀ QUAN HỆ TƯƠNG TỰ

Một số khái niệm về ĐSGT và quan hệ tương tự giữa các dữ liệu mờ.

### 2.1. Đại số gia tử tuyến tính

Một cách không hình thức, ĐSGT là một cấu trúc đại số được đưa vào tập các giá trị của một biến ngôn ngữ (thí dụ biến "chiều cao"), khi ta coi tập các từ nhấn - gia tử, (thí dụ "rất", "tương đối",..) là các toán tử một ngôi, khi tác động lên các phần tử sinh của biến ngôn ngữ (thí dụ "cao", "thấp") cho ta tập các phần tử của ĐSGT (tập rất cao, rất thấp, tương đối thấp, tương đối rất thấp, khá cao...) có thể sắp thứ tự theo ngữ nghĩa của chúng ("rất rất cao" > "rất cao" > "tương đối cao">.. > "tương đối thấp"> "rất thấp"..). Để giải quyết bài toán đặt ra trong bài báo này, chúng ta sẽ chỉ quan tâm đến các ĐSGT tuyến tính hai ngôi, tức là các ĐGGT mà các phần tử của nó được sinh ra từ tập các phần tử sinh gồm hai giá trị đối ngẫu (như "cao" và "thấp", "già và "trẻ", "nông" và "sâu"... và các gia tử như "rất", "tương đối", "khá"...), ký hiệu tương ứng là $c^+$ và $c^-$ đồng thời tất cả các phần tử được sinh ra đều so sánh được với nhau. Như vậy về mặt trực giác, các phần tử của ĐSGT (từ đây trở đi sẽ hiểu là ĐSGT tuyến tính hai ngôi) sẽ được phân bố trên một trục căn cứ vào thứ tự của chúng theo quan hệ $<$ của ĐSGT. Hai phần tử ở hai đầu của ĐSGT trên trục này là rất rất rất...rất $c^-$ và rất rất... rất $c^+$. Sẽ tồn tại các ánh xạ bảo toàn thứ tự từ tập các phần tử của một ĐSGT bất kỳ vào trục số thực ứng một phần tử của ĐSGT (là một giá trị biến ngôn ngữ tức là một từ trong ngôn ngữ thông thường) với một giá trị số

thực mà ta sẽ gọi là giá trị định lượng ngữ nghĩa của phần tử được ánh xạ. Khi có giá trị định lượng ngữ nghĩa của phần tử của ĐSGT, mọi thao tác như so sánh, tính toán,, với các phần tử của ĐSGT sẽ được đưa về các thao tác tương ứng trên trường số thực, một công việc rõ ràng và dễ dàng hơn nhiều. Để xây dựng các ánh xạ định lượng ngữ nghiã, chúng ta sẽ làm quen với khái niệm hàm độ đo tính mờ. Gọi $H(x)$ là tập các phần tử của $X$ (tập cơ sở) sinh ra từ $x \in X$ bởi các gia tử. Nghĩa là $H(x)$ bao gồm các khái niệm mờ mà nó phản ánh ý nghĩa nào đó của khái niệm $x$. Theo lý thuyết của ĐSGT, nếu $x < y$ thì $x' < y'$ với mọi $x' \in H(x)$ và $y' \in H(y)$ và ta viết $H(x) < H(y)$. Ngoài ra, tập các $H(x)$ với tất cả x có cùng độ dài (tức có cùng số các gia tử tác động lên phần tử sinh (thí dụ tương đối rất tốt có độ dài 3) sẽ tạo ra một phân hoạch của $X$ trên trục các phần tử của ĐSGT. Về mặt cảm nhận ta thấy tập $H(x)$ có thể biểu diễn tính mờ của $x$. $H(x)$ càng lớn, tính mờ của càng cao. Nếu $x$ là một khái niệm rõ thì $H(x)$ chỉ chứa một giá trị là $x$ (thí dụ rất 80 tuổi thì cũng là 80 tuổi). Từ đó, ta có thể định nghĩa độ đo tính mờ như là đường kính của tập $f(H(x)) = \{f(u) : u \in H(x)\}$ trong đó $f$ là một hàm độ đo mờ từ $X$ vào $[0, 1]$. Xem Ví dụ minh hoạ Hình 2.1.




*Hình 2.1*

Khi đã có độ đo tính mờ, tức là đã xác định được đường kính của các tập $H(x)$, do tính chất phân hoạch của tập các $H(x)$ có cùng độ dài và tính chất $H(x) < H(y)$ với $x < y$ như đã nói ở trên, ta sẽ xác định được giá trị của một giá trị $x$ bất kỳ thông qua hàm độ đo nếu biết thêm tỷ lệ khoảng cách $\alpha : \beta$ từ $x$ đến hai đầu mút của $f(H(x))$ bằng cách cộng độ dài của tất cả các tập $f(H(y))$, $y < x$ và thêm độ dài từ đầu mút dưới của $f(H(x))$ đến $x$ (lưu ý thêm rằng tỷ lệ a: là giống nhau với mọi x, y theo tiên đề về độ đo tính mờ). Theo các kết quả đã được chứng minh trong lý thuyết về ĐSGT, nếu cho trước f(c+) hoặc f(c-) cùng tỷ số $\alpha : \beta$ hoặc một trong hai giá trị $\alpha, \beta$ (do $\alpha + \beta = 1$) ta sẽ xác định được giá trị định lượng ngữ nghĩa của một giá trị $x$ bất kỳ. Ngược lại, cho trước một giá trị số $0 < r < 1$, có thể xác định được giá trị biến ngôn ngữ thuộc ĐSGT mà giá trị định lượng ngữ nghĩa của nó sai khác $r$ một giá trị e cho trước bất kỳ. Điều này có được do tính trù mật của không gian các phần tử thuộc ĐSGT. Các kết quả này sẽ giúp ta trong giải pháp của bài toán tìm kiếm trong CSDL ảnh được trình bày trong phần sau. Để có các khái niệm về ĐSGT theo cách trình bày chính xác, có thể xem ở các bài báo về ĐSGT đăng trong thời gian gần đây ([6, 7, 8]).

### 2.2. Quan hệ tương tự giữa các giá trị của một biến ngôn ngữ

Cho $U$ là tập vũ trụ, $F$ là một tập mờ chứa đối tượng $u$ thuộc $U$ được xác định bởi hàm thuộc $\mu_F : U \to [0,1]$. Một quan hệ tương tự là một quan hệ mờ hai ngôi đòi hỏi độ tương tự của từng cặp phần tử trong domain $D_i$ thỏa mãn những điều kiện sau

Với mọi $x, y, z$ thuộc $D_i$ thì:

1) Phản xạ: $\mu_S(x,x) = 1$.

2) Đối xứng: $\mu_S(x,y) = \mu_S(y,x)$.

3) Bắc cầu: $\mu_S(x,z) \geq \max\{\min[\mu_S(x,y), \mu_S(y,z)]\}$.

Quan hệ tương tự được cho bởi ma trận biểu diễn quan hệ tương tự

$$Similar(val_i, val_j) = val(i,j)$$

trong đó $val(i,j)$ xác định giá trị của phần tử hàng $i$, cột $j$ của ma trận. Thông thường, các giá trị này được xác định bởi ý kiến các chuyên gia.

### 2.3. Hàm kết nhập

Định nghĩa: Hàm kết nhập là một phép toán $n$ ngôi $\alpha(a_1, a_2, ..., a_n)$ trên đoạn $[0,1]$ thoả mãn các tính chất sau

1) $\alpha(a, a, ..., a) = a$;

2) $\alpha(a_1, a_2, ..., a_n, \alpha(a1, a_2, ..., a_n)) = \alpha(a1, a_2, ..., a_n)$;

3) $\alpha(a_1, a_2, ..., a_n) \leq \alpha(b_1, b_2, ..., b_n)$, với $a_i \leq b_i,\ i = \overline{1n}$.

Tính chất 1) nói rằng kết nhập $n$ ý kiến như nhau phải là chính ý kiến chung đó. Tính chất 2) nêu rõ nếu kết nhập thêm chính ý kiến kết nhập chung thì kết quả kết nhập không thay đổi. Tính chất 3) là tính đơn điệu. Có thể thấy một số hàm thông thường như min, max, trung bình cộng có trọng số đều là các hàm kết nhập.

## 3. GIẢI PHÁP TÌM KIẾM ẢNH DỰA TRÊN THÔNG TIN MÔ TẢ MỜ

### 3.1. Metadata

Về mặt hình thức, Metadata gồm các bảng R1(tên mô tả, kiểu), R2(tên mô tả, ĐSGT), R3(tên mô tả, tên QHTT), R4(tên, hàm kết nhập). Ở đó, thuộc tính kiểu trong R1 nói lên kiểu dữ liệu ứng với tên mô tả, bao gồm kiểu số thực (hoặc nguyên, tự nhiên), kiểu ĐSGT hoặc kiểu mờ không cấu trúc. Nếu là kiểu số thực (hoặc nguyên, tự nhiên) thì trong cột giá trị của kiểu ta lưu luôn miền xác định (thí dụ $[0,1]$). Trong cột giá trị của ĐSGT trong R2 sẽ lưu các giá trị $\alpha, \beta$ tương ứng với tên mô tả của ĐSGT. (lưu ý rằng với việc xác định $\alpha, \beta$, ta sẽ tính được tất cả các giá trị của các phần tử ĐSGT). Các giá trị $\alpha, \beta$ có thể cho mặc định trước hoặc do người sử dụng thay đổi tuỳ ý ($\alpha + \beta = 1$). Tên QHTT trong R3 là tên ma trận quan hệ tương tự ứng với tên mô tả của thuộc tính mờ phi cấu trúc. Ma trận này cũng có thể lưu mặc định sẵn theo ý kiến của chuyên gia hoặc thay đổi theo ý kiến của người sử dụng. Hàm kết nhập trong R4 lưu các hàm kết nhập cho sẵn và có thể thêm vào theo ý kiến của người sử dụng.

### 3.2. Thuật toán tìm kiếm ảnh bệnh học

Bài toán đặt ra là: Cho cơ sở dữ liệu gồm các bản ghi dữ liệu ảnh $\{A_i|i = 1..m\}$ và điều kiện tìm kiếm người dùng đưa vào $q = q(b_1, ..., b_j)$, $q$ là một hàm số, ở đó mỗi $b_k$, $k = 1..j$ có dạng $\lfloor des, val, c \rfloor$ trong đó:

*des*: là một phần tử trong tập các giá trị mô tả ảnh da liễu DES;

*val*: là một phần tử trong miền giá trị của $desDom(des)$;

$c$ : là một giá trị biểu thị trọng số của mô tả *des* trong truy vấn.

Đưa ra các bản ghi dữ liệu ảnh thoả mãn điều kiện tìm kiếm.

Trước hết, điều kiện tìm kiếm $q$ sẽ được biến đổi về dạng chuẩn tuyển (DNF) hoặc chuẩn hội (CNF) của các thành tố $b_k$, $k = 1..j$. Miền giá trị của trọng số được xác định bao gồm miền số hoặc miền ngôn ngữ. Miền số là miền [0, 1], còn miền ngôn ngữ dựa trên biến ngôn ngữ.

Trong khuôn khổ của bài báo này chúng ta chỉ quan tâm đến việc đánh trọng số trên các phần tử đơn của truy vấn. Bằng cách đặt trọng số độ quan trọng vào câu truy vấn người sử dụng chỉ rõ giới hạn chất lượng cần thỏa mãn trong tìm kiếm.

Giải thuật tìm kiếm dữ liệu bao gồm các khối chức năng sau:

1) *Tiền xử lý truy vấn*

Như ta đã biết, mục tiêu của hệ thống tìm kiếm bao giờ cũng bao gồm việc lượng giá dữ liệu tìm kiếm dựa trên mức độ thoả mãn của các điều kiện trong truy vấn đưa ra. Hệ thống con lượng giá với nhiều hơn một điều kiện logic sẽ hoạt động trên cơ sở một quá trình xây dựng bottom-up. Quá trình này được chia làm hai bước:

- Dữ liệu được định giá theo mức độ thoả mãn của chúng đối với từng câu truy vấn con trong câu truy vấn.

- Tiếp theo, dữ liệu được định giá mức độ thoả mãn của chúng bằng việc kết hợp các thành phần truy vấn con và làm việc theo kiểu bottom-up cho đến khi toàn bộ truy vấn được giải quyết.

Để đáp ứng được yêu cầu trên, hệ lượng giá cần phải tiền xử lý truy vấn của người sử dụng để đặt chúng thành dạng chuẩn tuyển hoặc chuẩn hội.

2) *Đối sánh giá trị trong điều kiện truy vấn con với các giá trị tương ứng của từng bản ghi trong CSDL ảnh*

Một truy vấn con có dạng $q_l = \lfloor des, val, c \rfloor$.

Trong quá trình tìm kiếm, từng bản ghi dữ liệu sẽ được lấy ra để đối sánh từng giá trị của thuộc tính trong cơ sở dữ liệu với giá trị tương ứng trong câu truy vấn con. Cụ thể, ta có truy vấn con $\lfloor des, val, c \rfloor$ và CSDL $\{A_i|i = 1..m\}$. Lấy ra ảnh $A_i = \{[d_{i1}, v_{i1}] > [d_{i2}, v_{i2}] > , ..., [d_{iki}, v_{iki}]\}$. Nếu tên mô tả *des* là $d_{ij}$ $(j = 1..k_i)$ thì ta tiến hành so sánh *val* với $v_{ij}$, ở đó $v_{ij}$ là giá trị cụ thể của $d_{ij}$ trong bản ghi dữ liệu ảnh $A_i$. Có thể xảy ra các trường hợp sau:

- Nếu *val* và $v_{ij}$ đều mang giá trị rõ thông thường thì kết quả so sánh có giá trị 1 (khi *val* bằng $v_{ij}$) hoặc 0 (khi *val* không bằng $v_{ij}$) : $r(q_l) = 0$ hoặc 1.

- Nếu *val* và $v_{ij}$ đều mang giá trị ngôn ngữ được biểu diễn bởi đại số gia tử, sử dụng các hàm định lượng để định lượng các giá trị ngôn ngữ . Kết quả đối sánh là hiệu của hai giá trị định lượng của hai giá trị ngôn ngữ trong cơ sở dữ liệu và trong câu truy vấn.

$$r(q_l) = 1 - |f(val) - f(v_{ij}|.$$

- Nếu *val* và $v_{ij}$ đều có giá trị mờ mà quan hệ giữa các giá trị được đưa ra trong ma trận quan hệ tương tự, kết quả đối sánh độ tương tự của hai giá trị trong cơ sở dữ liệu và trong câu truy vấn trong ma trận tương tự.

$$r(q_l) = \mu_{Si}(val, v_{ij}).$$

- Nếu *val* nhận giá trị rõ còn $v_{ij}$ nhận giá trị ngôn ngữ được biểu diễn bởi ĐSGT hoặc ngược lại, ta đưa về trường hợp 2.

- Nếu *val* nhận gia trị rõ hoặc giá trị ngôn ngữ được biểu diễn bởi ĐSGT còn $v_{ij}$ có giá trị được đưa ra trong ma trận quan hệ tương tự hoặc ngược lại, ta đưa về trường hợp 3.

Như vậy, trong mọi trường hợp, phép đối sánh đều cho ra kết quả r(ql) chỉ mức độ tương tự giữa giá trị tìm kiếm trong câu truy vấn con và mô tả ảnh $A_i$.

3) *Kết nhập có trọng số*

Sau khi có các kết quả đối sánh của từng truy vấn con, ta thực hiện quá trình kết nhập thông qua các hàm kết nhập có trọng số tất cả các kết quả nhận được theo trọng số do người sử dụng tự đưa ra hoặc theo ý kiến của chuyên gia cho sẵn . Các trọng số này đánh giá mức độ quan trọng theo ý của người sử dụng hoặc chuyên gia của thuộc tính xuất hiện trong truy vấn con trong câu hỏi tìm kiếm.

Các hàm kết nhập được định nghĩa trong Metadata (CSDL siêu dữ liệu) và được lập thành một danh sách để tuỳ chọn. Người dùng có thể lựa chọn hoặc để hệ tự động kết nhập theo hàm mặc định. Thí dụ nếu dùng hàm kết nhập là hàm trung bình cộng có trọng số, kết quả tìm kiếm sẽ là $r = c_1 r(q_1) + c_2 r(q_2) + ... + c_{ki} r(q_{ki}))$, trong đó các $c_1, c_2, ...$ là các trong số ứng với các mô tả, các $r(q_1), r(q_2), ...$ là các kết quả đánh giá mức độ tương tự giữa giá trị tìm kiếm trong câu truy vấn con và mô tả ảnh $A_i$.

4) *Vấn đề chọn ngưỡng độ thoả của dữ liệu đối với câu hỏi*

Kết thúc quá trình kết nhập, ta đã lượng giá xong bản ghi dữ liệu theo điều kiện tìm kiếm. Vì đây là tìm kiếm theo thông tin mờ nên ta sẽ có kết quả có thể không chỉ là một mà là nhiều ảnh đáp ứng yêu cầu tìm kiếm. Ta sẽ chọn ngưỡng $\lambda$ nào đó và chọn các ảnh kết quả là những ảnh có độ tương tự so với câu truy vấn vượt ngưỡng $\lambda$. Vấn đề còn lại là chọn ngưỡng $\lambda$ để cắt sao cho ra được kết quả tìm kiếm như mong muốn. Tuy nhiên vấn đề chọn ngưỡng cũng có thể gây ra một số khó khăn nếu không có giải pháp. Nếu ngưỡng quá thấp số ảnh tìm ra có thể quá nhiều, ngược lại ngưỡng quá cao có thể sẽ dẫn đến không tìm được ảnh nào và khi đó nếu thay đổi lại ngưỡng giải thuật lại phải tính lại từ đầu sẽ mất nhiều thời gian. Hơn nữa trong thực tế cần để cho người sử dụng không chỉ tự chọn các trọng số mà còn cả tự đặt ngưỡng và ngưỡng có thể thay đổi theo tình huống. Nghiên cứu đã đưa ra "chiến lược" chọn ngưỡng theo thuật toán sau.

Để tránh mất mát, ban đầu cho $\lambda = \lambda_0$ với $\lambda_0$ khá nhỏ, thậm chí bằng 0.

Gọi $N_0$ là một số lượng đối tượng ảnh tối đa thoả mãn các điều kiện của câu hỏi, N là số bản ghi trong cơ sở dữ liệu.

Có thể xảy ra tình huống giải thuật mới duyệt được một phần CSDL nhưng đã có $N_0$ đối tượng thoả mãn câu hỏi đối với ngưỡng $\lambda$. Ý tưởng trực giác ở đây là sẽ tăng ngưỡng vừa đủ để giảm thiểu số lượng bản ghi thoả mãn câu hỏi. Hằng số $k$ sẽ được chọn thông qua thử nghiệm trong một lĩnh vực ứng dụng cụ thể để bảo đảm tăng "vừa đủ". Khi đó, mỗi khi tìm được $N_0$ ảnh trong cơ sở dữ liệu thì $\lambda$ thay đổi như sau

$$\lambda = \min\{\lambda_0 + \lambda^k(1 - (N_0/N)), 1\}, \text{ với } k \geq 1.$$

## 4. THỬ NGHIỆM

Chúng tôi đã tiến hành thử nghiệm phương pháp tìm kiếm trên CSDL ảnh có 356 bản ghi, với 6 yêu cầu tìm kiếm và 4 kiểu hàm kết nhập khác nhau có kết quả thể hiện trong hàng và cột tương ứng của Ví dụ 1 và Ví dụ 2. Chẳng hạn hàng 2 của Ví dụ 1 thể hiện kết quả của yêu cầu tìm các ảnh của bệnh nhân phong, màu sắc chỗ thương tổn là đỏ sậm. 68 là số ảnh tìm được theo phép kết nhập là trung bình có trọng số, cũng là kết quả số ảnh tìm được theo các phép kết nhập khác, Keen Dienes's, Godél và Fordor, trong đó hàm kết nhập trung bình có trọng số là hàm OR có trọng số, còn các hàm kết nhập còn lại là hàm AND. Như vậy, chẳng hạn hàng 3 của Ví dụ 1, ở cột1 69 là kết quả tìm kiếm theo hàm kết nhập trung bình có trọng số các bệnh nhân phong hoặc (OR) vết thương có màu sắc sấn đỏ hoặc mức độ thâm nhiễm là khá sâu còn ở hàng 3 cột 3, 1 là kết quả tìm kiếm theo hàm kết nhập Godél các bệnh nhân phong và (AND) vết thương có màu sắc sẩn đỏ và mức độ thâm nhiễm là khá sâu. Ngưỡng thoả mãn là $\lambda = 0,5$.

**Ví dụ 1.** Kết quả tìm kiếm ảnh bệnh phong

| Thuộc tính \ Hàm kết nhập | Trung bình có trọng số | Keen Dienes's | Godél | Fordor |
|---|---|---|---|---|
| Bệnh : Phong | 76 | 76 | 76 | 76 |
| Bệnh : Phong<br>Màu sắc : sẩn đỏ | 68 | 68 | 68 | 68 |
| Bệnh : Phong<br>Màu sắc : sẩn đỏ<br>Mức độ thâm nhiễm : khá sâu | 69 | 1 | 1 | 1 |

| Thuộc tính \ Hàm kết nhập | Trung bình có trọng số | Keen Dienes's | Godél | Fordor |
|---|---|---|---|---|
| Bệnh : Giang mai | 43 | 43 | 43 | 43 |
| Bệnh : Giang mai<br>Màu sắc : đỏ tươi | 25 | 23 | 23 | 23 |
| Bệnh : Giang mai<br>Màu sắc : đỏ tươi<br>Ranh giới với da lành : rất rõ | 201 | 20 | 20 | 20 |

Trong hai ví dụ trên bệnh là thuộc tính rõ, màu sắc là thuộc tính mờ, mức độc thâm

nhiễm và ranh giới với da lành là thuộc tính ngôn ngữ có cùng một độ quan trọng là rất quan trọng. Khi thay đổi độ quan trọng này, tùy theo từng thuộc tính bị thay đổi và từng hàm kết nhập sử dụng kết quả tìm kiếm sẽ tăng hoặc giảm. Chẳng hạn trong Ví dụ 2, khi tìm kiếm theo 3 thuộc tính với hàm kết nhập là hàm trung bình có trọng số, nếu đặt thuộc tính màu sắc là không quan trọng thì kết quả tìm ra 197 ảnh còn nếu đặt thuộc tính bệnh là không quan trọng thì kết quả tìm ra 207 ảnh. Lưu ý 2 trường hợp ở cột kết quả 1 của cả hai Ví dụ, khi thêm thuộc tính (kết nhập OR) số lượng ảnh lại giảm đi vì giá trị sau kết nhập thấp hơn ngưỡng nên bị cắt đi.

Ngoài ra, để có thể tìm được số ảnh đạt tiêu chuẩn đặt ra với số lượng mong muốn, ta có thể thay đổi ngưỡng kết nhập.

$\lambda = 0,6$ cũng trường hợp vừa nêu trên (hàng 3 cột 1 Ví dụ 2) kết quả còn lại 186 ảnh. $\lambda = 0,8$ cũng trường hợp vừa nêu trên kết quả còn lại 21 ảnh. $\lambda = 0,95$ cũng trường hợp vừa nêu trên kết quả còn lại 12 ảnh.

## 5. KẾT LUẬN

Có thể nói dữ liệu về ảnh bệnh da liễu rất phong phú và đa dạng. Để có thể mô tả các giá trị thuộc tính trong cơ sở dữ liệu, chúng tôi đã sử dụng đại số gia tử và mô hình quan hệ tương tự. Tuy nhiên, các giá trị trong ma trận tương tự (độ tương tự) cũng như một số tham số khi khai báo biến có cấu trúc đại số gia tử đòi hỏi phải có ý kiến của các chuyên gia y tế và qua các số liệu thử nghiệm mới có được giá trị ngày càng chính xác. Thuật toán tìm kiếm dữ liệu ảnh da liễu đã xây dựng qua thử nghiệm đã cho kết quả tìm kiếm tốt. Khi tăng ngưỡng tìm kiếm đã cho kết quả tìm kiếm thoả mãn yêu cầu tìm kiếm cao. Các ảnh tìm ra sẽ giống ảnh so mẫu hơn nữa nếu các giá trị độ tương tự được đưa vào chính xác. Trong thời gian tới chúng tôi sẽ tiếp tục hoàn thiện hệ dựa trên các số liệu thử của bên y tế.

## TÀI LIỆU THAM KHẢO

[1] Adnan Yazici, Murat Koyuncu, Fuzzy object-oriented database modeling coupled with fuzzy logic, *Fuzzy Sets and Systems* **89** (1997) 1–26.

[2] S. K. De, R. Biswas, and A. R. Roy, On extended fuzzy relational database model with proximity relations, *Fuzzy Sets and Systems* **117** (2001) 195–201.

[3] D. A. Chiang, L. R. Chow, and N. C. Hsien, Fuzzy information in extended fuzzy relational database, *Fuzzy Sets and Systems* **92** (1997) 1–20.

[4] Hồ Cẩm Hà, "Một cách tiếp cận mở rộng cơ sở dữ liệu quan hệ với thông tin không đầy đủ", Luận án Tiến sĩ toán học, Đại học Bách khoa Hà Nội (2002).

[5] Le Tien Vuong, Ho Thuan, A relational database extended by application of fuzzy set theory and linguistic variables, *Computer and Artificial Intelligence* **8** (2) 153–168.

[6] Nguyễn Cát Hồ, Quantifying hedge algebras and interpolation methods in approximate reasoning, *Proc. of the 5th Inter. Conf. on Fuzzy Information Processing*, Beijing, March 1-4 (2003) 105–112.

[7] Nguyen Cat Ho, Tran Thai Son, and Le Xuan Viet, Fuzziness measure, quantified semantic mapping and interpolative method of approximate reasoning in medical expert systems, *Tạp chí Tin học và Điều khiển học* **18** (3) (2002) 237–252.

[8] N. C. Hồ, N. V. Long, Cơ sở toán học của độ đo tính mờ của thông tin ngôn ngữ, *Tạp chí Tin học và Điều khiển học* **20** (1) (2004) 64–72.

[9] Nguyễn Cát Hồ, Nguyễn Công Hào, Một cách tiếp cận để đánh giá dữ liệu trong cơ sở dữ liệu mờ, *Tạp chí Tin học và Điều khiển học* **18** (3) (2002) 237–252.

[10] Phương Minh Nam, Trần Thái Sơn, Về một cơ sở dữ liệu mờ và ứng dụng trong quản lý tội phạm, *Tạp chí Tin học và Điều khiển học* **22** (1) (2006) 25–36.

[11] Trần Thiên Thành, "Một số vấn đề về lý thuyết và ứng dụng của cơ sở dữ liệu mờ", Luận án Tiến sĩ toán học, Đại học Khoa học tự nhiên - Đại học Quốc gia Hà Nội (2004).

*Nhận bài ngày 12 - 11 - 2007*
*Nhận lại sau sửa ngày 12 - 5 -2008*