(Hook=Registry) AND
(Host=Service) THEN Cluster2
3. IF (Type=EXE) AND
(Hook=Startup) AND
(Host=App.) THEN Cluster3
4. IF (Type=EXE) AND
(Hook=Startup) AND
(Host=Sevice) THEN Cluster4
5. IF (Type=EXE) AND
(Hook=Registry) AND
(Host=Service) THEN Cluster5
6. IF (Type=EXE) AND
(Hook=WinFile) AND
(Host=App.) THEN Cluster6

3.1.4 Tổ chức tập luật băm theo chỉ mục

Mục đích của phân cụm V-Tree nhằm tách tập mẫu thành các nhóm có cùng đặc tính dữ liệu về mặt *giá trị*. Các cụm kết quả (bố trí ở nút lá của V-Tree) được tách rời và không phủ nhau nên số cụm thu được (m cụm) luôn nhỏ hơn số mẫu dữ liệu (k mẫu). Điều này luôn đúng vì luật có tính khái quát, do đó chỉ cần lập luận trên m luật là có thể bao quát toàn bộ k mẫu (m<k). Như vậy, ngoài chức năng cung cấp tri thức mã độc của chuyên gia, luật nhận dạng còn có thể sử dụng làm chỉ dẫn truy xuất dữ liệu. Chúng tôi bổ sung trường thứ ba vào mỗi nút trên V-Tree nhằm cung cấp giá trị trả về cho hàm băm $H(f(A),k)$ trong phép trích chọn đặc trưng thi hành $f(A)$ của đối tượng.

Đầu tiên, tập luật nhận dạng được tổ chức dạng bảng 2 chiều $R(\psi, £, r)$ có r phần tử luật; mỗi luật ψ liên kết với danh sách £ các thể hiện thỏa luật ψ. Trong ví dụ trên, CSDL luật R có 6 phần tử luật ($r = 6$) được minh họa như sau:

Luật 1: ψ_1 » {1,7,10}
Luật 2: ψ_2 » {4}
Luật 3: ψ_3 » {5}
Luật 4: ψ_4 » {2}
Luật 5: ψ_5 » {3,9}
Luật 6: ψ_6 » {6,8}

Ký hiệu '»' đặc tả phép 'liên kết' (link-to).

Tiếp theo, các bucket luật thành viên ψ_i được biến đổi thành các giá trị tổng kiểm bằng các thuật toán băm thông dụng như CRC (Cyclic Redundancy Check), MD5 (Message - Digest Algorithm 5) hoặc SHA-1 (Secure Hash Algorithm). Sau cùng, sắp xếp tập luật theo chỉ mục checksum (Bảng 2) rồi lưu trữ CSDL luật. Kết thúc giai đoạn luyện trên máy chuyên gia.

Bảng 2: CSDL luật chứa 6 luật thành viên

No.	Rule	Checksum	Link-to
1	ψ_2	58,487,876	4
2	ψ_5	76,455,645	3-9
3	ψ_4	156,496,857	2
4	ψ_1	325,474,326	1-7-10
5	ψ_6	437,665,473	6-8
6	ψ_3	758,355,475	5

3.2 Nhận dạng mã độc dựa vào tri thức

Giai đoạn nhận dạng - xử lý trên máy khách vận dụng các thủ tục suy diễn và lập luận trong động cơ quét (scan engine), căn cứ vào thông tin trong tập chữ ký mẫu và các tri thức mô tả trong CSDL luật để chẩn đoán đối tượng A. Quá trình này được thực hiện qua bốn bước: (1) trích chọn đặc trưng, (2) mã hóa nhân dạng, (3) truy vấn luật và (4), so khớp mẫu.

3.2.1 Trích chọn đặc trưng

Đầu vào của bước này là đối tượng A cần chẩn đoán. Đầu ra là một tập mô tả các đặc trưng thi hành của đối tượng dùng cho bước mã hóa nhân dạng tiếp theo. Để tương thích với khuôn dạng tri thức mô tả mã độc M ở giai đoạn luyện, bước này sử dụng lại hàm trích chọn đặc trưng $f(M)$ của chuyên gia dạng:

$$f(A) = \{\delta_i\} \ \forall \ (i < p)$$

Trong đó, $\delta(q)$ là phép trích chọn tự động tập thuộc tính mã độc ở giai đoạn trước. Ví dụ, áp dụng hàm f trích chọn đặc trưng trên các đối tượng bất kỳ U, V nhận được kết quả như sau:

$f(U)$ = {DLL, Registry, Driver}
$f(V)$ = {EXE, Startup, App.}

Ví dụ này sử dụng $i = 3$ ($i = p' < p$)

3.2.2 Mã hóa nhân dạng

Nhiệm vụ của bước trích chọn đặc trưng là tạo lập nhân dạng (identification) của đối tượng theo khuôn dạng luật trong CSTT. Khái niệm nhân dạng trong tiếp cận học chẩn đoán mã độc dựa vào lập luận: "nếu đối tượng có nhân dạng giống với nhân dạng tội phạm thì có thể đối tượng chính là tội phạm đang truy nã".

Sau khi trích chọn đặc trưng, nhân dạng đối tượng sẽ được mã hóa bằng thuật toán băm h của chuyên gia. Ví dụ, hồ sơ chẩn đoán các đối tượng