**BÁO CÁO KIỂM TRA THUẬT TOÁN TRÊN CÁC BỘ DỮ LIỆU**

(Thực hiện theo hợp đồng số 05/HĐ-NCKH, ngày 06 tháng 5 năm 2014)

Tên đề tài: **HỌC KHÁI NIỆM ĐỐI VỚI CÁC CƠ SỞ TRI THỨC TRONG LOGIC MÔ TẢ DỰA VÀO MÔ PHỎNG HAI CHIỀU,** Mã số: **DHH2013-01-41**.

Người thực hiện hợp đồng: **ThS. Đoàn Thị Hồng Phước**

**NỘI DUNG**

Kiểm tra cụ thể thuật toán trên các tập dữ liệu sau:

**Tập dữ liệu WebKB** chứa thông tin về các trang Web của 4 khoa Khoa học Máy tính (của các Trường Đại học Cornell, Washington, Wisconsin và Texas) bao gồm thông tin 877 trang web (đối tượng) và 1608 liên kết giữa các trang Web của 1 mối quan hệ (cites). Mỗi đối trong tập dữ liệu được mô tả bằng một véc tơ giá trị 1/0 để chỉ ra việc xuất hiện/không xuất hiện của từ tương ứng trong từ điển (gồm 1703 từ). Mỗi đối tượng được gán vào một trong 5 lớp để chỉ thể loại của trang Web: Course, Faculty, Student, Project và Staff . Chúng tôi sử dụng dữ liệu của 2 khoa để huấn luyện (230 đối tượng) và chứng thực (195 đối tượng). Dữ liệu 2 khoa còn lại (452 đối tượng) được sử dụng để kiểm tra.

**Tập dữ liệu Family** chứa thông tin về những người của 5 dòng họ (British Royal, Bush, Roberts, Romanov và Stevens) bao gồm thông tin của 943 người (đối tượng) và 11062 liên kết giữa những người này của 6 mối quan hệ (hasChild, hasSon, hasDaughter, hasWife, hasHusband, hasBrother, hasSister). Mối đối tượng là một thể hiện của khái niệm Male hoặc khái niệm Female. Chúng tôi sử dụng dữ liệu của 2 trong 5 dòng họ để huấn luyện (437 đối tượng) và chứng thực (49 đối tượng). Dữ liệu của 3 dòng họ còn lại (457 đối tượng) được sử dụng để kiểm tra.

**Tập dữ liệu PokerHand** là một tập con được lấy từ UCI Machine Learning Repository [11]. Tập dữ liệu này chứa thông tin của 2542 tay bài, 12710 quân bài, 119 tính chất của quân bài (tổng cộng có 15371 đối tượng) và 65220 liên kết giữa các đối tượng của 6 mối quan hệ (hasCard, hasRank, hasSuit, sameRank, nextRank, sameSuit). Mục đích của chúng ta là dự đoán các tay bài thuộc lớp nào trong 9 lớp “one pair”, “two pairs”, “three of a kind”, “traight”, “flush”, “full house”, “four of a kind”, “straight flush” và “royal flush”. Do số lượng tay bài thuộc các lớp “ four of a kind”, “ straight flush” và “ royal flush” quá ít cho nên chúng tôi loại bỏ những lớp này ra khỏi tập dữ liệu. Tập dữ liệu được chia thành 7 tập con. Dữ liệu của 2 trong 7 tập được sử dụng để huấn luyện (1343 đối tượng) và chứng thực (1343 đối tượng). Dữ liệu của 5 tập còn lại được sử dụng để kiểm tra (12685 đối tượng).

Thuật toán học khái niệm cho các hệ thống thông tin dựa trên logic mô tả được cài đặt bằng ngôn ngữ lập trình Java (JDK 1.6). Thuật toán này thực hiện trong logic ALCIQ sử dụng bộ chọn cơ bản và bộ chọn mở rộng cũng như độ đo gia lượng thông tin.

Qua quá trình kiểm tra, thuật toán đã chạy ổn định và cho kết quả tốt trong các trường hợp cụ thể.

**Người thực hiện**

**ThS. Đoàn Thị Hồng Phước**