ỨNG DỤNG THUẬT TOÁN PHÂN CỤM DỮ LIỆU ĐỂ KHAI THÁC KẾT QUẢ THI NHẰM CHUẨN HÓA CHẤT LƯỢNG ĐỀ THI TRẮC NGHIỆM

ENHANCING THE QUALITY OF MULTIPLE-CHOICE TESTS USING CLUSTERING ALGORITHM TO MINE TEST RESULTS

Đặng Thái Thịnh

Trường Đại học Kinh tế TP.HCM; thinhdt@ueh.edu.vn

Tóm tắt - Công tác ra đề thi hiện nay hầu như phụ thuộc hoàn toàn vào ý chí chủ quan của cá nhân giảng viên hoặc hội đồng ra đề thi. Các phần mềm thi trắc nghiệm có phát sinh đề thi chủ yếu lấy ngẫu nhiên từ các nhóm câu hỏi. Tuy nhiên, kết quả thực tế từ thí sinh có thể phản ánh đúng hoặc không đúng quan điểm và nhận xét trước đó của người ra đề thi. Mục tiêu của nghiên cứu này áp dụng cả ý kiến chuyên gia (phản hồi nhận xét từ giảng viên) và ý kiến của cộng đồng (người dự thi) nhằm đưa ra một cách giải quyết việc trộn đề thi từ cách phân bố ngẫu nhiên chuyển sang phân bố ngẫu nhiên có chủ đích nhằm đạt đến mục tiêu đảm bảo giữa hai đề thi có độ khó tương đương nhau. Thuật toán phân cụm và quá trình phân bố đề thi sau phân cụm được đề xuất để khai thác dữ liệu của kết quả thi. Thực nghiệm được triển khai tại Trường Đại học Kinh tế TP. Hồ Chí Minh phản ánh kết quả của nghiên cứu này.

Từ khóa - khai phá dữ liệu; phân cụm dữ liệu; khai thác kết quả thi; trộn đề đề thi; chất lượng đề thi

Abstract - Currently, working out exam papers depend almost entirely on the subjective opinions of individual faculty members or the exam boards. Multiple test software has given test questions mainly taken randomly from the question groups. However, in some situations, test results from test takers might not reflect the teacher's opinions correctly. This research aims to use rating from teachers and mining from test results in the past to generate new tests with equal level of difficulty . Clustering algorithm combined with proposed test question distribution is used in this study to mine data of test results. The experiment implemented in Ho Chi Minh University of Economics has reflected the result of the research.

Key words - data mining; data clustering; mining test results; mixing test questions; quality of tests

1. Đặt vấn đề

Hiện nay cách thức biểu diễn đề thi chủ yếu phụ thuộc vào phân cấp theo cây [1], tại mỗi node lá chứa nhiều câu hỏi. Mỗi node lá tượng trưng cho một nhóm câu hỏi, khi trộn đề người giảng viên chia tỷ lệ chọn lựa câu hỏi trong mỗi nhóm để có một đề thi. Quá trình này được lặp đi lặp lại để sinh ra nhiều đề thi. Ưu điểm của cách trộn như trên một đề thi luôn giữ được cấu trúc định nghĩa trước về số lượng câu hỏi trong mỗi node lá (phần/chương/mục).

Tuy nhiên với cách truyền thống này, việc chọn câu hỏi trong từng node lá mang tính chất ngẫu nhiên, vì vậy:

- Không thể hiện được độ khó tương đương của các đề thi với nhau;
- Sự trùng lắp nhiều câu hỏi trong các đề thi có thể xảy ra do cách chọn ngẫu nhiên.

Một số cách thức xây dựng ngân hàng câu hỏi có sự phân loại theo mức độ "khó", "dễ", "trung bình"; hoặc sự phân loại theo nhóm câu hỏi thuộc về "phân tích", "kiến thức" hay "kỹ năng" tồn tại trong một số sách của nhà xuất bản Pearson cũng giống tương tự như cách đề cập như trên nghĩa là chia nhỏ số lượng node lá và làm cho người giảng viên vất vả hơn trong quá trình xác định số lượng câu hỏi phân hóa trong đề thi [5].

Nghiên cứu nhằm đưa ra một cách tiếp cận kết hợp giữa cách phân nhóm câu hỏi, đưa ý kiến chuyên gia vào câu hỏi cùng với ý kiến thụ động của đại đa số người dự thi nhằm tự động phân loại và điều chỉnh cách thức chọn câu hỏi nhằm đạt đến mục tiêu giảm thiểu sự trùng lắp câu hỏi giữa các đề thi nhưng đảm bảo độ khó tương đương giữa các đề thi với nhau.

Ứng dụng tại các trường học, phương pháp vừa được đề cập ở trên là cách thức hiện nay đang sử dụng tại trường. Việc khai thác kết quả thi cung cấp cho giảng viên một cái nhìn lại về cách đánh giá của mình qua ngân hàng đề thi. Tính chủ quan có thể đúng hoặc sai, việc nhìn nhận trên dữ liệu thật trên các đối tượng dự thi khác nhau giúp người ra đề có nhiều thông tin để quyết định trong các lần sau, những quyết định có sự hỗ trợ của máy móc để tạo ra những báo cáo cho người ra quyết định [2].

2. Phân tích và đề xuất thuật toán

2.1. Dữ liệu đầu vào

Bước 1: Xây dựng ngân hàng câu hỏi;

Bước 2: Phân nhóm câu hỏi theo các phần/ chương/ mục;

Bước 3: Giảng viên đánh giá mức độ khó/dễ (như ví dụ ở bảng 3) cho từng câu hỏi trong ngân hàng trên thang điểm giá trị thập phân từ 0 đến 1 (tri thức chuyên gia). Trong đó càng khó thì số càng nhỏ (gần 0), càng dễ thì số càng cao (gần 1). Tất nhiên, không nên đánh giá 0 và 1 bởi lẽ, giảng viên đánh giá câu hỏi rất khó không ai trả lời được, hay câu hỏi quá dễ chắc chắn ai cũng trả lời được; thì câu hỏi có vấn đề về nội dung. Mỗi câu hỏi được mang đi thi nhiều lần, thí sinh của một lần thi nào đó có thể xảy ra 2 trường hợp: một là, đánh đúng; hai là, đánh sai.

Tất cả lịch sử này được lưu trữ lại Từ dữ liệu trên ta tính được:

$$Tỷ\ lệ\ trả\ lời\ đúng\ câu\ hỏi\ i\ = \frac{Tổng\ số\ lần\ trả\ lời\ đúng\ câu\ i}{Tổng\ số\ lần\ trả\ lời\ (câu\ i)}$$

Giá trị này được tính từ 0 đến 1.

Quá trình này gọi là quá trình học từ thực tiễn, kết quả ta có dạng như ví dụ ở bảng 1:

Bảng 1. *Ví dụ về tỷ lệ trả lời đúng ở câu hỏi*

Câu hỏi thứ	Tỷ lệ đúng
1	60%
2	30%
…	…
N	25%

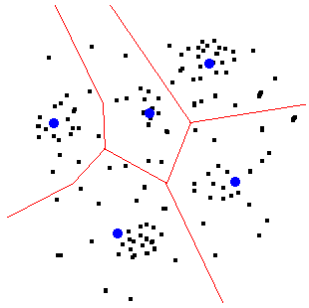
2.2. Biểu diễn phân cụm

Mỗi câu hỏi c_i được biểu diễn thành 1 vector mang 2 tác động (hình 2) là ($c_i(x,y)$), và là 1 điểm trong trục tọa độ Oxy:

Tác động 1: Từ ý kiến chuyên gia

Tác động 2: Từ ý kiến của cộng đồng

Như vậy, n câu hỏi được mô tả thành các điểm giống như trên. Dữ liệu thích hợp cho quá trình phân cụm (clustering). Nghiên cứu này sử dụng thuật toán K-means[6].Hình 1 mô tả cho quá trình phân cụm, tìm những câu hỏi gần tương tự nhau gom thành một nhóm (ở đây là từ 2 tác động ý kiến chuyên gia và ý kiến cộng đồng).



Hình 1. *Biểu diễn phân cụm*

Giải thuật xử lý như sau: trước tiên nó lựa chọn ngẫu nhiên k đối tượng, mỗi đối tượng đại diện cho một trung bình cụm hay tâm cụm. Đối với những đối tượng còn lại, mỗi đối tượng sẽ được ấn định vào một cụm mà nó giống nhất dựa trên khoảng cách giữa đối tượng và trung bình cụm. Sau đó sẽ tính lại trung bình cụm mới cho mỗi cụm. Xử lý này sẽ được lặp lại cho tới khi hàm tiêu chuẩn hội tụ. Bình phương sai số [6] thường dùng làm hàm tiêu chuẩn hội tụ, định nghĩa như sau:

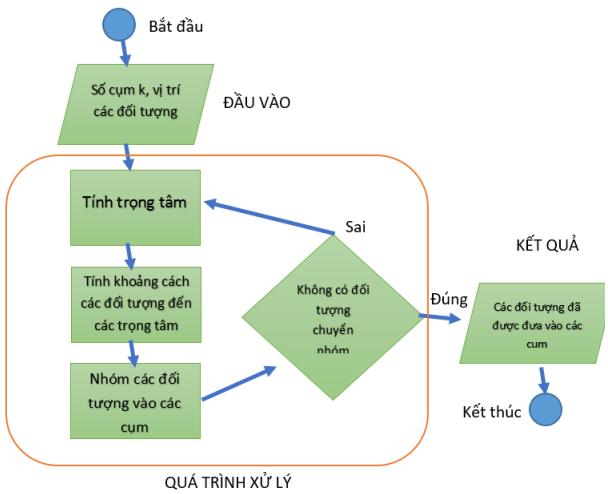
$$E = \sum_{i=1}^{k} \sum_{x \in c_i} |x - m_i|^2 \quad (1)$$

Với k là số cụm, x là điểm trong không gian đại diện cho đối tượng cho trước, m_i là trung bình cụm C_i (cả x và m_i đều là đa chiều). Ta có:

Đầu vào: Số cụm k và hàm E có giá trị theo công thức 1.

Đầu ra: Hàm tiêu chuẩn E đạt giá trị tối thiểu.

Thuật toán được mô tả bằng sơ đồ ở hình 2 như sau:



Hình 2. *Sơ đồ các bước phân cụm*

Thuật toán dừng khi không có đối tượng chuyển nhóm, như vậy ta đã phân các câu hỏi thành k cụm riêng biệt.

2.3. Phân bố câu hỏi sau phân cụm

- Gọi k là số cụm, trước tiên ta tìm tâm của k cụm (chạy thuật toán K-means).
- Tìm tâm chung của C câu hỏi.
- Sắp xếp k cụm thành thứ tự có khoảng cách từ bé nhất đến lớn nhất đến tâm chung của C câu hỏi.
- Với D là tổng số đề thi cần tạo ra, M là số câu hỏi trong 1 đề thi.
- **for** (*d= 1 to D*) **do**//một vòng lặp ở đây ta xây dựng được 1 đề
  - **for** (i=1 to M) **do** //một vòng lặp ở đây ta tìm được 1 câu hỏi cho đề thứ d

    Xét cụm gần thứ i của tâm chung, chọn 1 câu hỏi thỏa các yêu cầu để đưa vào bộ đề thứ d:
    - Chọn ngẫu nhiên;
    - Ưu tiên câu không trùng câu hỏi đã chọn trước, có thể chọn lại câu đó nếu đã chọn hết câu hỏi trong các lần trước);
    - Có tổng khoảng cách đếncác câu hỏi ở *i-1* lần chọn trước bé nhất.

3. Thực nghiệm và đánh giá kết quả

3.1. Một số phương pháp đánh giá

Mỗi đề thi được đánh giá bằng sự tương đồng về độ khó. Giả sử mỗi đề thi có n câu hỏi, mỗi câu hỏi đều có độ khó được biểu diễn bằng 2 vector giá trị của độ khó chuyên gia và độ khó do người dùng định nghĩa. Biểu diễn vector của một đề thi có n câu như sau: ($u_1, u_2, u_3, u_4, \ldots u_n$), ($e_1, e_2, e_3, e_4, \ldots e_n$) với

u_i: độ khó của câu hỏi thứ i do người dự thi quyết định

e_i: độ khó của câu hỏi thứ i do chuyên gia (người ra đề thi) quyết định

Sự tương đồng của 2 đề thi được tính có thể được tính bằng nhiều phương pháp như: Cosine similarity, Pearson correlation [3]. Ví dụ: cosin similarity

$$\text{Cos } Sim\,(x,y) = \frac{\sum_i x_i\, y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x,y\rangle}{||x||||y||} \quad (2)$$

Với đề thi 1 được mô tả: $x_1, x_2, x_3 \ldots x_n$ (x_i)

Với đề thi 2 được mô tả: $y_1, y_2, y_3 \ldots y_n$ (y_i)

Nếu sự tương đồng này cao (giá trị càng tiến về 1) nghĩa là độ khó của đề thi tương đương nhau. Phương pháp này có thể được đánh giá lại kết quả sau khi quá trình trộn đề thi hoàn tất.

Cách đo khoảng cách giữa các vector còn có thể thực hiện qua các phương pháp tính khoảng cách như sau:

Inner product

$$Inner(x,y) = \sum_i x_i\, y_i = \langle x,y\rangle \quad (3)$$

Pearson correlation

$$Corr(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

$$= \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{|| x - \bar{x} || || y - \bar{y} |}$$
$$= CosSim\,(x - \bar{x}, y - \bar{y})\ (4)$$

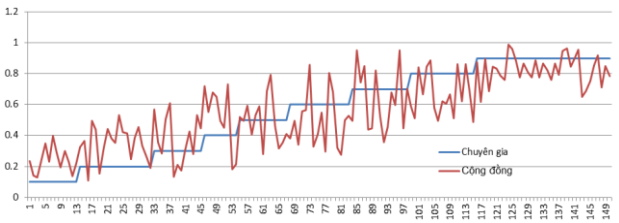
Các công thức đo khoảng cách này đều có thể được thực hiện cho nghiên cứu này. Pearson được sử dụng trong thực nghiệm.

3.2. Thực nghiệm

Thực nghiệm được lấy từ kết quả cuộc thi đánh giá xếp loại đoàn viên của Đoàn thanh niên – Hội sinh viên trường Đại học Kinh tế TP.HCM. Cuộc thi được thực hiện trong học kỳ cuối năm 2014 với ngân hàng 150 câu hỏi và xem như chỉ cần phân loại vào 1 nhóm nội dung thi duy nhất. Nội dung các câu hỏi về chủ đề kiến thức Đoàn Hội. Trung bình mỗi câu hỏi có 203.66 lượt trả lời.

Kết quả chạy thuật toán trên ta có:

Các câu hỏi được sắp xếp theo giá trị chuyên gia tăng dần, ta có phân bố của cộng đồng như sau (hình 3):

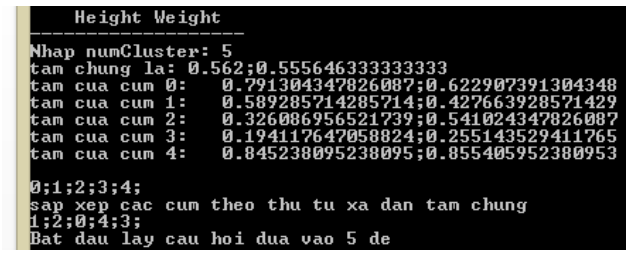


Hình 3. *So sánh độ khó dựa ý kiến chuyên gia và cộng đồng (đã sắp xếp)*

Nhận xét: nhìn chung xu hướng của cộng đồng đi theo xu hướng đánh giá của chuyên gia, như vậy dữ liệu tương đối tốt cho thử nghiệm.

Kết quả sau khi chạy thử nghiệm và chọn đề thi như sau:

Số cụm = 5, số đề = 5, câu hỏi trong 1 đề = 20 (như giao diện ở hình 4)



Hình 4. *Giao diện phần mềm khi làm thực nghiệm*

Đánh giá bằng Pearson độ tương đồng của các đề thi sau khi sinh ra được mô tả ở bảng 2 và bảng 3. Giữa 2 đề thi bất kỳ tồn tại sự tương tự nhau về độ khó dựa trên ý kiến của chuyên gia (người ra đề), hay ý kiến cộng đồng (tỷ lệ người dự thi trả lời đúng). GọiP(x,y) là độ tương quan giữa đề x và đề y có giá trị [-1,1]; P(x,y) càng tiến về 1 thì độ khó của đề x và y tương đương nhau. Nếu P(x,y), P(y,z) càng tiến về 1, thì P(x,z) cũng sẽ tiến về 1. Giả sử P(x,y) gần 1, nhưng P(y,z) lại không gần 1, thì P(x,z) cũng không gần 1. Kết quả được mô tả ở bảng 2 và bảng 3 cho thấy đề thi được phát sinh bằng phương pháp trong bài báo này có giá trị Pearson rất gần 1 (lớn hơn 0.95) , nghĩa là các đề thi được sinh ra từ mô hình của bài nghiên cứu này có độ khó tương đương nhau. Vì tính chất P(x,y) = P(y,z) nên một phần của bảng 2 và bảng 3 được xóa bỏ.

Ý kiến chuyên gia

Bảng 2. *So sánh bằng Pearson ý kiến chuyên gia giữa các đề thi*

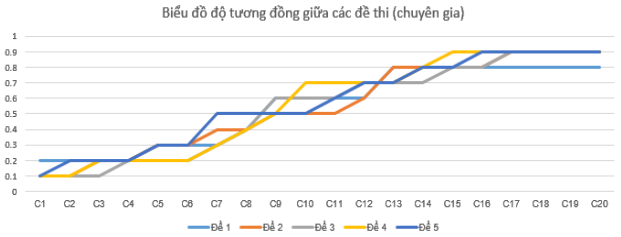
	Đề 1	Đề 2	Đề 3	Đề 4	Đề 5
Đề 1	1	0.972208	0.976262	0.97531	0.953814
Đề 2		1	0.971808	0.961304	0.98156
Đề 3			1	0.984653	0.970552
Đề 4				1	0.965235
Đề 5					1

Ý kiến cộng đồng

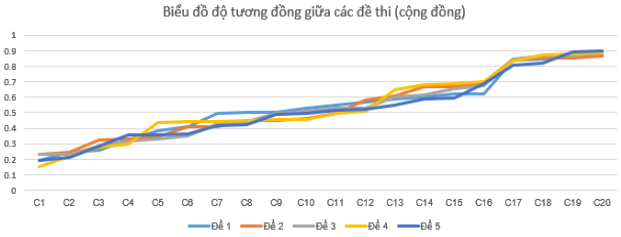
Bảng 3. *So sánh bằng Pearson ý kiến cộng đồng giữa các đề thi*

	Đề 1	Đề 2	Đề 3	Đề 4	Đề 5
Đề 1	1	0.974957	0.986603	0.975203	0.983579
Đề 2		1	0.990853	0.984919	0.98317
Đề 3			1	0.978244	0.992197
Đề 4				1	0.975544
Đề 5					1

Để cụ thể hơn, ta vẽ biểu đồ độ khó (tỷ lệ trả lời đúng) của các đề thi sau chạy thuật toán K-means và cách chọn câu hỏi sau khi phân cụm như sau (hình 5 và 6).



Hình 5.*Biểu đồ độ tương đồng giữa các đề thi (chuyên gia)*



Hình 6. *Biểu đồ độ tương đồng giữa các đề thi (cộng đồng)*

3.3. Đánh giá phương pháp thực hiện

Về thuật toán phân cụm dữ liệu: Nhược điểm của K-means là còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu [6]. Hơn nữa, chất lượng phân cụm dữ liệu của thuật toán K-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k trọng tâm khởi tạo ban đầu. Trong trường hợp các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của K-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất. Đánh giá thuật toán K-means:

Ưu điểm:

- K-means là có độ phức tạp tính toán O(t.k.n) với k là số cụm, n là số lần lặp và t là tổng số lượng phần tử.
- K-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.
- Bảo đảm hội tụ sau một quá trình lặp hữu hạn.

Nhược điểm:

- K-means không khắc phục được nhiễu và giá trị số cụm k phải được cho bởi người dùng.
- Chỉ thích hợp áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu.

Sự trùng lặp câu hỏi trong đề thi:

- Nếu câu hỏi bị trùng nhau nhiều, nghĩa là độ khó sẽ gần nhau nhiều, cách này không phải là mục tiêu chính của nghiên cứu này.
- Giả sử ta tìm được n đề thi, mỗi đề thi có c câu hỏi. Với thuật toán như trên sẽ hạn chế sự trùng nhau trong đề thi, bởi cách chọn được thực hiện trên cơ sở ưu tiên chọn câu hỏi mới.

Điểm mạnh của nghiên cứu:

- Nghiên cứu đề xuất một phương pháp mô tả chi tiết lấy tri thức từ chuyên gia ra đề thi không quá nhiều thông tin phải cung cấp, nhưng đủ cho quá trình đánh giá phân loại đề thi.
- Nghiên cứu cũng đưa ra một mô hình phân loại câu hỏi dựa trên kết quả thi từ cộng đồng và kết hợp tri thức chuyên gia.
- Một phương pháp đánh giá trộn đề thi công bằng giữa các đề thi, các phương pháp trước đây mang nhiều ý kiến chủ quan, hoặc không có sự phân bố dựa trên độ khó mà chỉ dựa trên phân bố ngẫu nhiên.

Điểm yếu của mô hình:

- Bài thi của thí sinh phải khá nhiều trên một câu hỏi, mới có thể đánh giá có ý nghĩa.
- Ý kiến chuyên gia đang được xem xét cùng với ý kiến người dự thi, như vậy chưa chắc đã đúng. Tuy nhiên, ý kiến chuyên gia có thể thay đổi quan điểm sau khi người ra đề xem xét dữ liệu trả về người dự thi.

Nhưng yếu tố khác tác động lên bài thi: như thông tin cá nhân và học thức của người dự thi chưa được xem xét trong mô hình này. Ví dụ: một bài thi tiếng Anh như TOEIC, TOEFL yêu cầu một bài khảo sát nhỏ trước khi thí sinh thực hiện bài thi. Trong đó họ có nghiên cứu các yếu tố ảnh hưởng đến chất lượng bài thi và có thể dùng để phân loại câu hỏi sau này [4].

4. Kết luận

Nghiên cứu này đưa ra một cách tiếp cận dựa trên phương pháp phân cụm dữ liệu, kết quả của quá trình phân cụm được chọn lọc để đưa vào đề thi. Nghiên cứu cũng chỉ ra sự tương đồng giữa các đề thi qua phương pháp đo khoảng cách giữa 2 vector đã trình bày ở trên. Kết quả của phương pháp có thể được áp dụng để cách trộn đề thi vừa phân bố mang yếu tố ngẫu nhiên vừa có độ khó tương đương giữa các đề thi.

Tuy nhiên cách giải quyết trên chưa giải quyết được làm sao đưa ra một tiêu chuẩn chung cho các tiêu chuẩnmà chủ yếu vào kiến thức của người chuyên gia để đưa ra các mức điểm cho các mức độ. Kết quả của nghiên cứu này giúp điểm thi của cộng đồng có xu hướng phân bố rải rác do độ khó được phân bố đồng đều. Điều này hỗ trợ cho quá trình xác định ra quyết định để xác định các mức điển phân loại (khá, giỏi, trung bình, không đạt) dễ dàng hơn.

Đóng góp của nghiên cứu này như thêm một hỗ trợ dựa khai trên khai thác kết quả thi, có sự nhìn nhận đánh giá từ kiến thức của chuyên gia (người ra đề thi) và dữ liệu cộng đồng đánh giá từ kết quả trắc nghiệm khách quan, một cách tiếp cận định lượng.

Thực nghiệm cũng còn thiếu nhiều dữ liệu và các yếu tố khác có thể ảnh hưởng đến kết quả thi. Để có được dữ liệu cộng đồng đủ lớn để giúp quá trình đánh giá có ý nghĩa hơn cũng là điều khó khăn. Ban đầu hệ thống sẽ chạy với dữ liệu chuyên gia hoàn toàn, sau một thời gian dữ liệu cộng đồng có nhiều sẽ kết hợp với dữ liệu chuyên gia để đánh giá. Người ra đề sau khi có kết quả thi sẽ nhìn nhận lại cách đánh giá của mình để xem xét có quá chủ quan khi đưa ra quyết định ban đầu hay không. Từ đó hệ thống được điều chỉnh và học cách làm mới liên tục.

Nghiên cứu có thể được mở rộng bằng cách tăng giảm độ khó của đề thi bằng cách phân bố không đều vào các cụm sau khi phân hoạch. Tuy nhiên cách làm này cần đánh giá lại việc phân loại như thế nào và cần một phương pháp đánh giá khác.

TÀI LIỆU THAM KHẢO

[1] Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna Eds Handbook of test development.

[2] Mahwah: Lawrence Erlbaum Associations. Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks: SAGE Publications.

[3] J.L. Rodgers, W.A. Nicewander, "Thirteen ways to look at the correlation coefficient", Amer. Statist. 42 (1988).

[4] Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. Educational and Psychological Measurement, 63(4), 584-601.

[5] Kane, M. T. (2001). So much remain the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) Setting performance standards. Concepts, methods, and perspectives (pp. 53-88).

[6] Nguyễn Hoàng Tú Anh. (2009). Khai thác dữ liệu & ứng dụng (Data Mining), NXB ĐHQG TP.HCM.

(BBT nhận bài: 18/08/2015, phản biện xong: 29/10/2015)