ユーザに作問を任せるテストの自動評点に向けて

○谷口敦[†]　Ghada Farouk Naiem[†]　井上創造[†]
[†]九州工業大学

Toward Automatic Assessment of User-generated Tests

Atsushi Taniguchi, Ghada Frouk Naiem, and Sozo Inoue
[†] Kyushu Institute of Technology

Abstract: *We focus on the automatic assessment of user-generated tests, where any user can make a question content. We describe a method for assessing an answer of a user for a question, taking into account the difficulty and the validity of the question, and show the result of preliminary experiment performed by a set of people, which implies the positive correlation between the subjective assessment.*

1. はじめに

近年，多くの大学や高等教育機関ではe-Learningシステムが導入されている．その多くの場面で教師のみ，または限られた人間が問題を作成するため，作問の負担が大きい．また，学習者同士がインターネットを介して事前に解答が知られている傾向があり，教師が新しい問題を作成し続ける必要性も直面している．

我々は，学習者自身が問題を作成する方法を構想し，その際においても適切に問題の難易度と妥当性，また評点を与える手法を提案した[1]．提案手法は，各問題の評点と妥当性を再帰的に定義し，収束するまで反復的に計算する方法である．文献[1]では，アルゴリズムが収束することを確認するためにシミュレーションを行い，その結果，難易度と妥当性を評点に反映させるものであることを確認した．

本論文では、まず文献[1]の内容を3節に紹介し，少人数の環境で実際に行った実験を述べる．実験の結果，実際に作問された問題でも主観評価と比べて正しく評点されることが確認できた．一方で，問題の解答数が難易度に影響するという課題も明らかとなった．

2. 背景

我々の手法は，クラウドソーシング[2]と人間計算 [3]に関係がある．クラウドソーシングは物理的な世界とインターネット上の世界両方に活用されている．人間計算ヒューマンコンピュテーションとは"計算機が処理できないタスクの処理に人手を利用する事である．またGoogleで使用されているような画像検索など様々な分野において適用されている．

3. 学習者が作問したテストの自動評点手法

ここでは文献[1]の手法を簡単に説明する．提案手法では，すべてのユーザが問題を作ることができる．また別のユーザが問題に答えることができ，質問は作文したユーザが定義した正解に従い自動的に採点されることが要件である．このような状況で，システムは以下の３つを自動的に算出するようにする．

- 問題の**難易度**：問題の難しさ．
- 問題の**妥当性**：問題が学習目的にそったものどうか．
- 問題へのユーザの解答に対する**評点**：難易度と妥当性を考慮した上で回答者に与えられる評価点．

3.1 評点方法

以下では，Uをユーザ集合，Q を問題集合，qを問題，uを回答者とする．また，問題qに回答者uが回答して自動的に採点された点を$score(q,u)$と表す．ただし簡単のため、採点は0から1の間の値を取るとする．

難易度:

難易度の高い問題は多くの回答者の得点が低く，簡単な問題は点数が高いと考えられるため，難易度はその問題のスコアの平均とする．

$$Difficulty(q) := 1 - \frac{\sum_{u \in U} score(q,u)}{|U|}$$

妥当性:

評点の高い作問者ほど妥当性の高い

問題を作成するものと仮定する．つまり，問題qの妥当性はqの作問者($author(q)$とする)によって解かれた問題の評点の平均と定義する．

$$Validity(q) := \frac{\sum_{q' \in Q} Assessment(q', author(q))}{|Q|}$$

評点:

問題qに対するユーザuの回答に対する評点は，難易度と妥当性によって調整されると考えられる．そのため，$Assessment(q,u) := [score(q,u) + Difficulty(q) + Validity(q)]$と定義する．

ここまでの式で，妥当性と評点の定義は循環している．したがって，ヒューリスティックな手法を用いて収束するまで計算を繰り返す．文献[1]では，各問題の妥当性にばらつきがある場合と難易度にばらつきがある時で計算機シミュレーションを行い，アルゴリズムが収束することと，難易度と妥当性を適切に導くことを確認した．

4. 実験

実際に人間が作文･回答した問題についてシミュレーションと同様の結果が得られるかを評価するために，小規模の実験を行った．

実験での評価項目は以下のとおりである．

- 算出された難易度，妥当性は主観的な難易度，妥当性と一致するか
- 実際の問題でも難易度と妥当性を考慮した評点結果が得られるか

研究室内の3名の学生に，英語の問題を一人あたり15問ずつ作問，45問ずつ解答させた．選択肢の数は制限せず，最小２つ，最大７つの選択肢を持った問題が作問された．下記に問題とその選択肢の例を書き示す．

問題:*I have never met ___ in person before.*
選択肢:*he ／ his ／ him ／ himself*

また，比較対象として問題の主観的な難易度と妥当性を得るためにアンケートを行った．図１の横軸は問題の難易度および妥当性の主観評価，縦軸は実際の難易度および妥当性のグラフである．なお回答数が 2 つ以下の問題を除外した．また直線はその回帰直線である．難易度と妥当性は主観評価と相関していることがわかる．難易度についてはp=0.128となり5%有意水準を満たさないが，妥当性についてはp=0.00257となり5%有意水準を満たすことが分かる．

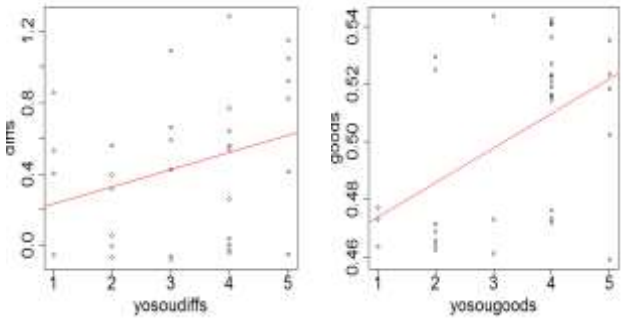

図１　実際の難易度と主観評価(上)，実際の妥当性と主観評価(下)

結果・考察

実際に作成した問題使用した実験において，提案手法で計算された難易度と妥当性は，主観による値に相関することが確認された．また難易度と妥当性を考慮した評点もなされた．実際には解答数のばらつきを考慮しないといけない．また被験者の評点がなされたが差があまりなく，最終的な評点をする際に分散で重みを付けて評点するなど工夫が必要であると考える．これはサンプル数が少なく，また解答数の数のばらつきが大きいために生じた問題である．

5. まとめ

ユーザの問題に対する解答を難易度と妥当性を考慮して評点するための手法を使い，実際に人が作問・解答して実験を行った．その結果，難易度，妥当性は主観的な難易度と妥当性と相関が見られた．また解答数が少なすぎると難易度や妥当性に大きな影響を与えるため実際には考慮しなくてはいけないことがわかった．今後は1問に対しての解答数が多い時，少ない時の比較などを行い，実際の授業での運営を目指す．

参考文献

[1] Ghada Farouk Naiem, 井上創造, "A Method for Assessing User-generated Tests for Online Courses Exploiting Crowdsourcing Concept", IWWISS, September , 2014

[2] Howe, Jeff. "The rise of crowdsourcing." Wired magazine 14.6 (2006): 1-4.

[3] von Ahn, L. Human Computation. Doctoral Thesis. UMI Order Number: AAI3205378, CMU, (2005).