Automated Scoring for Creative Problem Solving Ability with Ideation-Explanation Modeling

Hao-Chuan WANG[1], Chun-Yen CHANG[2], and Tsai-Yen LI[3]

[1]*Institute of Information Science, Academia Sinica, Taiwan*
[2]*Department of Earth Sciences, National Taiwan Normal University, Taiwan*
[3]*Department of Computer Science, National Chengchi University, Taiwan*
haochuan@iis.sinica.edu.tw, changcy@cc.ntnu.edu.tw, li@nccu.edu.tw

Abstract. This paper describes an automated scorer for assessing students' Creative Problem-Solving (CPS) abilities via modeling the intra-structure of students' essays describing their thoughts on solving particular problems. The automated scorer aims to grade students' open-ended responses to an essay-question-type CPS ability test, instead of using typical Likert-type or multiple-choice questions that may be imperfect to assess the *creative* perspective of human problem-solving. The scorer is distinguishable to most generic automated essay scoring systems that a bipartite graph-based representation is explicitly built for the pair-wise relation between a student's ideas and self-explained reasons for a CPS task. This design will enable several analytical approaches for CPS, such as quantitative scoring and qualitative diagnoses. The preliminary empirical evaluation with 20 students' data shows that the scoring results of the scorer is satisfactory and highly correlated with those of human experts (Pearson's r=.67~.82) in terms of quantitative scoring task. The approach provides a promising solution to support large-scaled studies on human creativity and may further enable CPS-aware personalization systems.

Introduction

Human creativity is generally considered as one of the most distinguishable human capacities that *cannot* be easily replaced or imitated by computing machineries. In the era of widespread information explosion, the increasing complexity of decision making for everyday-life problems and scientific challenges has signified the needs of stronger emphases on students' science-process skills and Creative Problem-Solving (CPS) ability, both in educational practice and fundamental research, as proposed by Chang and Weng in [2].

Nevertheless, it is observed that the scientific research of CPS is obscured by certain technical constraints. In finding evidence that supports the existence of hypothesized constructs for CPS, replications across studies with large sample size are demanded [6]. This requirement results in a significant challenge here: *how to* grade open-ended answers for a great quantity of subjects in CPS studies both efficiently and reliably? Since a theoretically valid measuring instrument for CPS abilities is likely to be an open-ended test, such as an essay question, hence the essay scoring task performed *manually* by human graders or coders is inevitably rather time-consuming. The situation may become much severer when the sample size is required to be large. Under this situation, not only the scoring task is itself laborious, but it may also be difficult to maintain a consistent scoring criterion among individual graders when a group of graders are hired to share the labor.

Besides, by only reporting holistic scores, as in the case of quantitative scoring of essays, certain useful features underlying students' answers are not devised to be observable. It appears difficult to investigate some advanced research problems such as "*What are the common misconceptions of these students?*" In other words, with solely holistic scores, *less* diagnostic insight can be derived for amending the real-world instructions. We identify this problem as a *deficiency of holistic scoring*. However, it appears challenging for human graders to locate other informative structures upon a large collection of essays in addition to scoring them holistically.

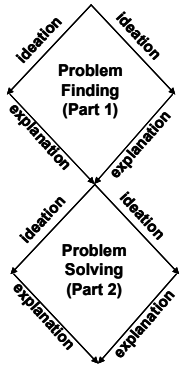


Figure 1. CPS processes with ideation-explanation sub-phases. (a revision of [1])

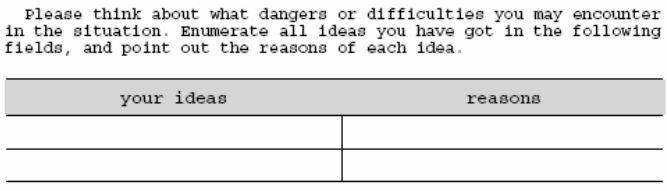
In this work, a novel scheme of automated scoring is developed to automate the scoring task for CPS research. The automated scorer is built based on the bipartite graph-based user modeling framework, called UPSAM (User Problem Solving Ability Modeler) that we have sketched in [9]. The automated scorer is capable of attaining reliable quantitative scoring for an open-ended CPS test, and building human-understandable user profiles for the relation of *ideation-explanation* underlying student's answers using the bipartite graph formalism. The ideation-explanation modeling approach is considered helpful in reaching qualitative diagnostic analyses in the long run, which may heal the aforementioned deficiency of holistic scoring. In this paper, we describe how the proposed automated scoring scheme works, specifically in the perspective of quantitative scoring at the first attempt. A pilot empirical evaluation on the reliability of quantitative scoring is reported to demonstrate the potential of this scheme.

1. Measuring Instrument and User Modeling

1.1 Creative Problem Solving Ability Test

The testing instrument used in this work was previously constructed and analyzed by Wu *et al.* in [10]. The test was developed for the needs of science education research. It specifically emphasizes on students' science-process skills by presenting a context-rich problem-solving scenario in a story-telling manner to test takers, for example, the task of organizing an emergent rescue mission after the happening of a great earthquake.

By referring to models of the Osborn's four-stage problem-solving process: fact-finding, problem-solving, idea-finding, and solution-finding [7], as well as, the ideation-evaluation sub-phases for CPS: divergent thinking and convergent thinking [1], this test is designed to be flexible and versatile for practical uses. Figure 1 shows the underlying model


Please think about what dangers or difficulties you may encounter in the situation. Enumerate all ideas you have got in the following fields, and point out the reasons of each idea.

your ideas	reasons

Figure 2. A snapshot of the answer sheet showing the pair-wise relation between ideas and reasons.

of this test. The Osborn's four-stage model is summarized as a two-stage version here, including stages of problem-finding and problem-solving (see Figure 1), which are believed to be a representative abstraction for most CPS theories. For each stage, its two sub-phases, i.e., ideation and explanation, are also identified. Note that the curtailment of the four-stage model into a two-stage version is mainly for the viability of implementing the test in the classroom, but not a non-negotiable decision. The scheme of user modeling and automated scoring remains flexible to a four-stage CPS process.

Figure 2 depicts the format of the answer form for this test. Testees are required to express their ideas (i.e. the production of divergent thinking) for each CPS stage, and then provide self-explained reasons for the proposed ideas (i.e. the production of convergent thinking). All these responses are expressed in natural language, which is believed to be a valuable source in studying students' conceptions in-depth with better authenticity. Nonetheless, although the test is open-ended, it is not of no structure. The internal structure of pair-wise relation between ideas and reasons, in alignment with CPS theory [1], may help the designer of the automated scorer to make a detour to avoid the challenging open problem of non-structural natural language understanding.

1.2 UPSAM: Bipartite Graph-based CPS Modeling

In a previous work in this series, Wang *et al.* have proposed a bipartite graph-based user modeling framework, called UPSAM (User Problem-Solving Ability Modeler), for capturing the pair-wise relation between divergent and convergent thinking of CPS [9].

The bipartite graph in the graph theory is one whose vertex set can be partitioned into two disjoint subsets. For each edge in the graph, its two ends are from different subsets of the vertex set. For CPS modeling, the user's ideation is represented as a set of ideas $A=\{a_1, a_2, \ldots, a_n\}$, and explanation is represented as a set of reasons $B=\{b_1, b_2, \ldots, b_m\}$. The user model can then be denoted as an undirected bipartite graph $G=(V, E)$ where $V=A \cup B$ and $A \cap B=\phi$. The connections between ideation and explanation entries are represented as $E=\{e_{ij}\}$ that each edge e_{ij} represents a linkage between idea a_i and reason b_j.

With the modeling formalism, we can then implement computer programs capable of building, manipulating, and retrieving CPS models. Note that in UPSAM, all types of users, including experts and testees, are modeled in identical representation formalism, the bipartite graph. Therefore, it makes possible to perform comparative analyses upon users' CPS performance, such as user-user and user-expert comparisons, by inspecting these commensurable user models. Given a collection of user models, there are two possible approaches to perform model comparison. One is on using idioms and algorithms from the graph theory, and casting the problem of user model comparison as a series of operations on graphs. For example, the task of retrieving the common ideas between user U's and the expert D's ideation can be derived by computing the intersection of the two idea sets, that is, the operation of $A_U \cap A_D$. The graph-based approach appeals to the requirements of qualitative fine-structure analyses. Another approach of model comparison is on defining proper metrics to transform graphs into numerical scores. That is, a quantitative and holistic measurement of each user model can be derived for further comparative ordering and analysis. The automated scoring scheme described in the next section is realized using the later approach.

2. Automated Scoring for CPS Ability Test

The two key components in the automated scoring scheme are the *domain model* and the *scoring agent*. Their properties and functions are described in this section.

```
<upsam>
   <part id="part1"
        desc="what are the factors enabling a debris flow to occur?">
      <pairs>
         <pair idea_idref="p1-idea01" reason_idref="p1-reason01"/>
         <pair idea_idref="p1-idea02" reason_idref="p1-reason01"/>
      </pairs>
      <ideas>
         <idea id="p1-idea01" score="4">
            the location is on a steep dip slope
         </idea>
         <idea id="p1-idea02" score="2">
            a rainstorm
         </idea>
      </ideas>
      <reasons>
         <reason id="p1-reason01" score="4">
            the friction force between rocks and the earth surface
            is reduced, such that a debris flow is possible to occur
         </reason>
      </reasons>
   </part>
</upsam>
```

Figure 3. A partial domain model authored in XML

2.1 Domain Model

A domain model is a representation of expert knowledge for a part of the CPS task using the formalism of bipartite graph. In the view of automated scoring, domain modeling exactly refers to organizing the expert's answers (i.e. keys for the CPS test) as a formal model, in which the vertex subsets of ideation and explanation, as well as the edge sets of ideation-explanation linkage are explicitly represented and stored in the system. The task of domain model authoring is performed manually by domain experts using a predefined XML-based data model. Figure 3 shows an XML document for a partial domain model. The `<ideas>` and `<reasons>` elements annotate the expert's ideation and explanation for the CPS task respectively, and the `<pairs>` element annotates the links connecting the two disjoint subsets of ideas and reasons. Note that the incorporation of XML provides *only* syntaxes for the ease of authoring, parsing and processing for implementing the system, but *not* a format of semantics interoperability for domain models. Related issues deserve further exploration but beyond the scope of this paper.

2.2 Scoring Agent

A scoring agent is a software module implementing the functions of automated scoring. The major goal of the scoring agent is to build bipartite graph-based user models based on users' responses to the CPS test.

Given a set of answer pairs $P_U=\{p_1, p_2, \ldots, p_n\}$ from user U, where $p_i=(\alpha_i, \beta_i)$ and α_i and β_i denote the user's natural language entries for ideation and explanation, respectively. Our current approach is to find the most appropriate subsets of the domain model $G_D=(A_D \cup B_D, E_D)$ indicating the overlapping of CPS conceptions between user U and expert D. That is, for each idea entry α, pick $a \in A_D$ such that a's content is most similar to α, and for each reason entry β, use B_D instead of A_D to perform the same procedure.

Inevitably, some assumptions have to be made to control the computational complexity of the procedure. The assumptions include that 1) several user entries are allowed to refer to the same concept node of the domain model, and 2) users' entries of ideation and explanation are not overlapping (i.e., disjoint). In other words, the task of building user models in our scheme is *not* an optimization problem that would signifi-

Procedure: Concept_Identify
Input: I: a natural language entry
π: a subset of vertices, i.e. ideation or explanation subset, in the domain model
T : a collection of synonyms in the domain
Output: χ: a concept node for the user model

1. $tr := 0$ //setting the filtering threshold, tr
2. $I' \leftarrow$ Wording_Processing(I, T) //remove stop words and refine wordings
3. **for** each $c_i \in \pi$ **do**
4. $N \leftarrow$ Desc(c_i) //retrieve the descriptions of ideation/explanation
5. $s_i \leftarrow$ TFIDF(I', N) //compute similarity by using TFIDF method
6. $sim[c_i] \leftarrow s_i$ //associate the simlarity score with c
7. **end for**
8. **if** $\max_{c \in \pi} sim[c] > tr$ **then**
9. $\chi \leftarrow \arg\max_{c \in \pi} sim[c]$ //select c with the highest similairy to I'
10. **else**
11. $\chi \leftarrow$New_Node(I') //instantiate a new node if *no c* meets the threshold
12. **return** χ

Figure 4. The procedure *Concept_Identify* for building the user model

cantly raise the computational cost. Fortunately, this simplification is reasonable for most CPS theories, and its viability for automated scoring is verifiable by empirical corroboration.

Figure 4 shows the procedure for identifying and instantiating concept nodes (i.e., vertices in the bipartite graph) using a natural language entry as the input and the domain model as the foundation of model building. The core idea is to employ token-based similarity metrics used in Information Retrieval (IR) to retrieve the most suitable concept nodes from the domain model for each user entry. In line 2, the user entry is first processed by the function, `Wording_Processing()`, which aims to remove stop words, and refine wording by using a synonyms dictionary T built for the CPS domain. The step of wording refinement aims to make users' and the expert's term usage as consistent as possible. After the pre-processing steps, from line 3 to line 7, an iterative name-matching task of two strings is conducted. Note that the similarity metric used in line 5 is substitutable, and different metrics may affect the correctness of the output. Here we consider two token-based distance functions, *Jaccard* and *TFIDF* [4]. Given the definitions of I' and N as described in line 2 and line 4 respectively, the Jaccard similarity is defined as:

$$Jaccard(I', N) = \left|I' \cap N\right| \Big/ \left|I' \cup N\right| \tag{1}$$

The TFIDF similarity (i.e., a variant of cosine similarity in IR) is defined as:

$$TFIDF(I', N) = \sum_{w \in I' \cap N} V(w, I') \cdot V(w, N) \tag{2}$$

where $V(w, S)$ is the normalized weight of each term w in the string S based on the TFIDF term-weighting scheme; see [4] for more details of the two metrics we used here. Finally, from line 8 to line 12, the task is to select the node with the highest similarity score. If the node does not exceed the minimum *threshold*, as specified in line 1, a new node with the processed text will be instantiated for that entry. Qualitatively speaking, the new nodes may refer to a novel and unique concept, though could be unorthodox or even erroneous, generated by the user beyond the expert's expectation. Clearly, qualitative analysis is

possible by investigating these nodes to make up the deficiency of holistic scoring mentioned at the beginning. The properties of these cases deserve further exploration and study.

The procedure is computationally feasible for practical use. For a single user entry to be identified by using the *Concept_Identify* procedure, its computational cost is linear to $|\pi|$. Given the maximum possible number of entries in a single sub-phase (i.e., a ideation/explanation phase) h, and the number of total sub-phases k (e.g. k=4 for the model shown in Figure 1), the time complexity for the model building task is bounded by $O(kh|\pi|)$. For h and $|\pi|$, the values are mostly of the order of tens, and for k, the value would not exceed 10 based on most CPS theories. So the total cost at this level is acceptable. Note that the major unit cost of computation refers to the overhead of similarity score evaluation by using different distance functions, e.g. Jaccard or TFIDF, so the actual running time would also depend on the implementation of similarity metrics. However, since these metrics have been employed as standard methods in large scaled information retrieval systems, efficient and reliable implementations are available. The SecondString toolkit developed by Cohen et al. [4][11] is utilized currently.

2.3 Automated Scoring

Given a pre-authored domain model $G_D=(A_D \cup B_D, E_D)$, different ideas, reasons, and combinations of ideation-explanation linkages can be given difference scores indicating the quality of conceptions. Therefore, the *scoring functions* are assigned to A_D, B_D, and E_D , respectively to transform each subset of the graph into scores. That is,

$$Sc = \{good\ answer,\ regular,\ no\ credit\}$$
$$f_A : A_D \rightarrow Sc,\ f_B : B_D \rightarrow Sc, \text{and } f_E : E_D \rightarrow Sc$$

where S_C denotes the range of these scoring functions, and each ordinal value (e.g. "*regular*") is associated with numeric scores. Subsequently, the score of a user model $G_U=(A_U \cup B_U, E_U)$ can be derived via:

$$Score(G_U) = \sum_{a \in A_U \cap A_D} f_A(a) + \sum_{b \in B_U \cap B_D} f_B(b) + \sum_{e \in E_U \cap E_D} f_E(e) \quad (3)$$

We have implemented an automated scorer based on the scoring scheme described above. Next, we will describe the comparative evaluation of the automated scorer by utilizing human graders' scoring results as the criteria.

3. Evaluation

3.1 Method

The evaluation aims to corroborate the viability of the automated scoring scheme empirically. The substantive question is to inquire whether the scores reported by the automated scorer are reliable in comparison with the results of CPS essay grading performed by human graders.

A group of 20 Taiwan tenth grade students was employed as participants for the pilot evaluation. The topic of the CPS test is about debris flow in the area of Earth sciences. The test is partitioned into two parts according to the aforementioned CPS theoretical model shown in Figure 1. In the first part of the test, the phase of problem finding, students were required to describe their thoughts (i.e. ideation and explanation) on *what* are the possible

Table 1. Pearson product-moment correlations, *r* among two human graders and different versions of automated scorers. Grayed areas highlight the part of human-computer correlation, and *tr* denotes the threshold parameterized in the *Concept_Identify* procedure.

	Human1	Human2	TFIDF, *tr*=0	TFIDF, *tr*=.1	Jaccard, *tr*=0
Human1	1	.89**	.74**	.82**	.69**
Human2		1	.71**	.78**	.67**
TFIDF, *tr*=0			1	.95**	.88**
TFIDF, *tr*=.1				1	.80**
Jaccard, *tr*=0					1

**$p<.01$

factors to make the disaster happen. Then in the second part, the phase of problem solving, students were asked *how* to prevent the hazard from harming people again. When scoring a student's responses to the test, quantitative scores were given to each part of the answers, and a total score was derived by combining them.

These students' answers to the CPS test were graded by the automated scorer, and two domain experts respectively. More precisely, the same set of data (i.e. responses from the 20 students) was repeatedly graded by several graders, including domain experts and the computer program, *without* informing them the scoring results from other graders *a priori*. The scores reported by each grader are then employed in the correlation analysis.

3.2 Results and Discussion

We utilize the statistics of Pearson product-moment correlation as an estimation of the inter-rater reliability. Table 1 shows the results of the correlation analysis. Note that since the adjustment of the parameter *tr* and the choice of different similarity metrics may affect the performance of automated scoring, three different versions of automated scorer were evaluated experimentally to shed some lights upon such design decisions. The three versions are, TFIDF metric with threshold *tr*=0, TFIDF with *tr*=.1, and Jaccard metric with *tr*=0.

The major findings revealed in Table 1 include: 1) the correlation of human-to-human comparison is positive, high, and statistically significant ($r=.89$, $p<.01$), and the associated *effect size*, $r^2=.79$, is large according to Cohen(1988) [3]. The result shows that the scoring results reported by the two human graders are reliable to be used as criteria for evaluating automated scorers. 2) the correlation statistics *r* for human-to-computers comparison range from .67 to .82, which are positive, high, and statistically significant ($p<.01$), as shown in the grayed area of Table 1. The effect size coefficients $r^2=.45\sim.67$ are also large ones.

Overall, the preliminary empirical evaluation conducted with a small sample size has manifested the promise of the proposed automated scoring scheme for CPS. The best outcome is the version using TFIDF with *tr*=.1, which results in performance of $r=.78\sim.82$. Inter-rater reliability at this level has been a satisfactory one meeting the needs of most educational studies. Notably, the performance of the proposed automated scorer is also comparable with other *generic* automated essay grading systems in the literature. Most systems were developed for identifying the structures of writing or the correctness of semantics [5]. Their human-to-computer correlations seem to vary, from as good as $r \cong .9$ [5], to as poor as $r \cong .4$ on short essays scoring [8]. Comparatively speaking, the performance of our scorer seems among the top. However, our approach is distinguishable to those generic scorers both in methods and purposes. Unlike other works of automated scoring, our scheme takes the structure of CPS theoretical models into consideration, and reciprocally, dedicates

to CPS studies. Nevertheless, it merits further exploration in incorporating methods developed by generic automated scoring to improve our scheme.

4. Conclusion

In this paper, we have described our approach of developing an automated scorer for CPS. The scheme illustrates an integral method by considering the intra-structure of the CPS test and adopting the bipartite graph-based formalism to model the relation of ideation-explanation characterized by CPS theories. A pilot empirical corroboration was conducted upon the reliability of this automated scorer, and the result shows that the scheme is promising in automating the laborious task of CPS test scoring.

One point mentioned but not elaborated in-depth in this paper is the function of graph-based qualitative analyses. Related works are now underway, and will be reported soon. Other future works include conducting replicate evaluations with a large sample size, refining the performance of the scheme, and using the CPS model as a basis for personalizing the pedagogy of science education.

Acknowledgement

We thank Yi-Chun Chen at the Department of Earth Sciences, National Taiwan Normal University for her help on data collection and pre-processing. We also thank the Chinese Knowledge and Information Processing (CKIP) group at Academia Sinica for providing the toolkit of Chinese word segmentation.

References

[1] Basadur, M. (1995). Optimal Ideation-Evaluation Ratios. *Creativity Research Journal*, 8(1), pp.63-75.

[2] Chang, C-Y., Weng, Y-H. (2002). An Exploratory Study on Students' Problem-Solving Ability in Earth Sciences. *International Journal of Science Education*, 24(5), pp. 441-451.

[3] Cohen, J. (1988) *Statistical power analysis for the behavioral sciences, 2 ed.* Lawrence Erlbaum.

[4] Cohen, W. W., Ravikumar, P., Fienberg, S. E. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *Proceedings of 18th International Joint Conference on Artificial Intelligence, Workshop on Information Integration on the Web*, Mexico, pp.73-78.

[5] Hearst, H. A. (2000) The Debate on Automated Essay Grading. *IEEE Intelligent Systems*, 15(5), pp. 22-37.

[6] MumFord, M. D. (2000-2001) Something Old, Something New: Revisiting Guilford's Conception of Creative Problem Solving. *Creativity Research Journal*. 13(3&4), pp. 267-276.

[7] Osborn, A. (1963) *Applied Imagination: Principles and Procedures of Creative Problem Solving*. New York: Charles Scribner's Sons.

[8] Ventura, M.J., Franchescettie. D.R., Pennumatsa, P., Graesser, A.C., Jackson, G.T. (2004) Combining Computational Models of Short Essay Grading for Conceptual Physics Problems. *Proceedings of Intelligent Tutoring Systems*, *Lecture Notes of Computer Science*, 3220, pp.423-431, Springer-Verlag.

[9] Wang, H-C., Li, T-Y., Chang, C-Y. (2005). A User Modeling Framework for Exploring Creative Problem-Solving Ability. *Proceedings of AIED Conference*, Amsterdam, The Netherlands.

[10] Wu, C-L., Chang, C-Y. (2002). Exploring the Interrelationship Between Tenth-Graders' Problem-Solving Abilities and Their Prior Knowledge and Reasoning Skills in Earth Science. *Chinese Journal of Science Education,* 10(2), pp. 135-156.

[11] SecondString project. Available at: http://secondstring.sourceforge.net/ , last access: May 17, 2005.