

Некоторые аспекты качественной оценки ответов тестируемых в системах контроля знаний

И.П. Карпова

Московский государственный институт электроники и математики,
доцент кафедры "Вычислительные системы и сети"

karpova_ip@mail.ru

Адрес публикации:

<http://masters.donntu.org/2013/fknt/sedakov/library/karpova.htm>

В статье проанализированы некоторые методы оценки ответов, используемые при автоматизированном контроле знаний. Введена метрика для определения степени сходства ответов обучаемого и эталонных ответов. Показаны преимущества использования предложенных методов.

1. Введение

В исследованиях, касающихся автоматизированного контроля знаний, большое внимание уделяется методам оценки знаний (см., например, обзор [2]). Все существующие методы оценки знаний базируются на оценках ответов на контрольные вопросы. В то же время методам оценки самих ответов уделяется очень мало внимания. Данная статья призвана хотя бы частично восполнить этот пробел.

Контроль знаний в современных системах контроля знаний (СКЗ) осуществляется, в основном, с помощью тестов. Тест должен включать вопросы и эталонные ответы на них (далее – эталоны).

Качество теста не в последнюю очередь определяется методами оценки ответов тестируемых. К этим методам предъявляются обычно следующие требования:

1. Адекватность (можно трактовать как близость оценки, полученной с помощью данного метода, к оценке, выставленной за данный ответ преподавателем).
2. Универсальность (независимость от предметной области).
3. Простота в использовании (метод не должен требовать от составителя заданий каких-то специфических знаний, не имеющих отношения к его предметной области).

Основное требование – адекватность. Два других требования являются не обязательными, но желательными.

2. Оценка ответов тестируемых

2.1. Типы ответов

Метод оценки ответа зависит от его типа. Один из основоположников научного подхода к проблеме тестирования В.С. Аванесов показал, что все ответы могут быть сведены к следующим формам [1]:

1. Выбор одного или нескольких вариантов ответов в произвольном порядке. Такие ответы называют выборочными.
2. Определение порядка вариантов ответов.
3. Определение соответствия элементов двух множеств.
4. Открытые: свободно-конструируемые ответы (числовые и текстовые).

Наиболее универсальной является, естественно, свободно-конструируемая форма ответа, но она и требует самых сложных методов анализа. Отчасти поэтому в современных системах контроля знаний (СКЗ) в качестве типов свободно-конструируемых ответов используются числовые ответы, ответы на специализированных языках (например, химические формулы) или тексты на ограниченном естественном языке. И если с оценкой числовых ответов дело обстоит просто, то возможности анализа специализированных и естественных языков могут быть предоставлены только с помощью достаточно сложного программного обеспечения.

Типы ответов (2,3) с определением порядка вариантов и соответствия элементов двух множеств проще в использовании. Вопрос сопровождается перечнем возможных ответов и нужно упорядочить их или сформировать пары элементов из двух множеств. Для таких вопросов существует вероятность угадывания правильного ответа, гораздо *большая*, чем при использовании свободно-конструируемых ответов. Для снижения этой вероятности в перечень возможных вариантов ответа включают "лишние" элементы, которые в правильный ответ не входят.

Выборочная форма ответа всё ещё продолжает активно использоваться в СКЗ как самая простая в реализации. К вопросу "прилагается" список ответов и нужно выбрать один или несколько правильных. Вероятность угадывания (т.е. возможность неадекватной оценки) здесь ещё больше.

Теперь рассмотрим методы оценки ответов, которые наиболее часто используются в существующих СКЗ, не касаясь методов анализа ответов на специализированных и естественных языках.

Самый примитивный метод оценки ответов – недифференцированный: если тестируемый дал ответ, совпадающий с эталоном, то он получает максимальную оценку, в противном случае – минимальную. При этом

вариантов неполного или частично правильного ответа не предусматривается. Даже в тех системах, которые позволяют указывать несколько эталонов для одного вопроса, оценка не дифференцирована. Попытки уйти от подобного примитивизма предпринимаются регулярно.

В некоторых СКЗ есть возможность не только перечислить возможные ответы на вопрос, но и указать оценку за каждый из них в зависимости от правильности и полноты ответа. Очевидным недостатком такого подхода является необходимость перечислять все возможные ответы и придумывать правдоподобные, но неправильные ответы. В других системах обучаемому предоставляется N попыток ответа: если он сразу дал правильный ответ, он получает максимальный балл, если с i -й попытки ($i \leq N$), то оценка уменьшается пропорционально значению i . Последний вариант – это просто "угадай-ка": если не угадал с первого раза, то оценка будет ниже.

Но можно пойти другим путём: вопрос с готовым набором вариантов ответов заменить свободно-конструируемым, который включает одно или несколько слов (терминов, понятий), а при тестировании использовать формальный подход к оценке таких ответов, чтобы не усложнять работу составителя тестов. Такой подход должен имитировать общую методику оценивания ответа тестируемого, которую используют преподаватели.

2.2. Методика оценивания ответа

При оценивании ответа преподаватель выступает в качестве эксперта. В своё время было установлено, что баллы, присваиваемые экспертами при оценке, как правило, измерены в порядковой шкале [6].

Ситуация с оцениванием ответа на вопрос несколько отличается от стандартных задач по оценке объектов (процессов, явлений), стоящих перед экспертами. Она характеризуется тем, что эксперт знает правильный ответ на этот вопрос. Наличие эталона (правильного ответа) позволяет не просто ранжировать ответы, т.е. определить, какой из полученных ответов хуже, а какой лучше. Он позволяет определить степень правильности (полноты, точности) ответа по сравнению с эталоном. Если ответ тестируемого совпадает с эталоном, то преподаватель выставляет максимальную оценку; если ответ не имеет ничего общего с эталоном, то оценка минимальна; и чем больше сходство ответа и эталона, тем выше оценка.

В статьях [3, 4] был предложен формальный метод дифференцированной оценки ответов тестируемых в автоматизированных СКЗ (дельта- или Д-метод). В Д-методе ответ на вопрос рассматривается как множество или список элементов. В первом случае правильность ответа определяется количеством элементов ответа, совпадающих с элементами

эталонного ответа. Во втором случае правильность ответа зависит также от порядка элементов. Формальный подход здесь заключается в ранжировании ответов на основе понятия *расстояния* между правильным (эталонным) ответом и тем ответом, который дал тестируемый.

Расстояние должно удовлетворять следующим условиям:

1. $|aa| = 0, \forall a$;
2. $|ab| = |ba|, \forall a, b$;
3. $|ab| + |bc| \leq |ac|, \forall a, b, c$;

где a, b, c – произвольные однотипные объекты (точки, множества или др.).

В том случае, если объектами являются множества элементов, примем $[0,1]$ за интервал изменения расстояния r таким образом, что:

- $r = 0$, если $M_a \equiv M_e$;
- $r = 1$, если $M_a \cap M_e = \emptyset$;
- $0 < r < 1$, если $M_a \cap M_e \neq \emptyset$.

где M_a, M_e – множества элементов.

Для списков элементов примем $[0,1]$ за интервал изменения расстояния r таким образом, что:

- $r = 0$, если списки совпадают;
- $r = 1$, если множества элементов списков не пересекаются или списки упорядочены в противоположном порядке;
- $0 < r < 1$, если списки частично совпадают.

В соответствии с таким определением расстояния введём понятие степени сходства множеств (списков). **Степенью сходства множеств (списков)** назовем величину, обратную к расстоянию:

$$\delta = 1 - r \quad (1)$$

Теперь вкратце изложим основные положения Д-метода, который позволяет ввести метрику для определения степени сходства ответов обучаемого и эталонных ответов.

2.3. Суть Д-метода

Предлагается рассматривать ответы тестируемых как множество или список элементов. Рассмотрим вопрос:

ВОПРОС: Какие существуют модели представления знаний?

ЭТАЛОННЫЙ ОТВЕТ: логическая, продукционная, фреймовая, семантическая

В общем случае количество правильных элементов в ответе тестируемого может варьироваться от 0 до N , где N – количество элементов в эталоне ответа. Если нас не интересует порядок, в котором перечислены элементы ответа, то мы получаем представление ответа в виде *множества*

элементов. Степень сходства δ_{mn} ответа тестируемого (А) и эталонного ответа (Е) можно определить как

$$\delta_{mn} = \frac{K_A}{L_E + K'} \quad (2)$$

где L_E – мощность эталонного множества, K_A – количество элементов из ответа тестируемого, входящих в эталон, K' – количество элементов из ответа тестируемого, не входящих в эталон. Таким образом, оценка ответа снижается как при наличии "лишних" элементов в ответе, так и при отсутствии требуемых, а степень сходства эталона и ответа изменяется в пределах $[0,1]$.

Если правильность ответа зависит от порядка входящих в него элементов, то мы имеем дело со *списком*. Для сравнения ответа и эталонного списка можно воспользоваться процедурой сортировки [5], которая заключается в приведении одного списка к другому путем попарных перестановок его элементов. Предполагается, что нумерация элементов эталонного списка такова, что он образует полностью упорядоченную подстановку.

Максимальное число перестановок для списка длиной n равно

$$K_n = \frac{n(n-1)}{2}$$

и тогда степень сходства δ_{cn} списков можно определить как

$$\delta_{cn} = 1 - \frac{K_i}{K_n} \quad (3)$$

где K_i – количество перестановок (инверсий) в исходном списке (ответе).

При использовании списка элементов также надо учитывать возможность присутствия в ответе "лишних" элементов, которых нет в эталоне. Для таких элементов не определён правильный порядок следования. В этом случае процедуру сравнения списков можно разбить на два этапа:

1. Сначала списки сравниваются как множества, при этом подсчитывается степень сходства δ_{mn} (2).
2. Затем из ответа исключаются все "лишние" элементы (не входящие в эталонный список), а остальные упорядочиваются путем попарных перестановок. Подсчитывается степень сходства δ_{cn} (3).

Степень сходства списков δ является функцией от δ_{mn} и δ_{cn} . Эта функция также должна быть определена на интервале $[0,1]$. Вид этой функции зависит от "вклада" значений δ_{mn} и δ_{cn} , который в общем случае может быть разным. Если же степени сходства δ_{mn} и δ_{cn} равноценны, то в качестве такой функции можно взять, например, среднее арифметическое или произведение δ_{mn} и δ_{cn} . Нетрудно убедиться в том, что при $\delta_{mn} \in [0,1]$, $\delta_{cn} \in [0,1]$ обе эти функции остаются в прежних пределах $[0,1]$.

Итак, оценка ответа определяется степенью сходства ответа и эталона: чем выше степень сходства, тем выше оценка. Степень сходства изменяется в границах $[0,1]$, что дает нам возможность нормировать оценку и не привязываться к конкретной шкале: максимальная оценка ставится при степени сходства 1, минимальная – при степени сходства 0.

Примечание: для реализации типа ответа на соответствие элементов двух множеств можно ввести двухуровневую схему "множество списков", используя для оценки Д-метод. Подробнее это изложено в статье [3].

2.4. Сравнение результатов Д-метода с оценкой преподавателя

Основным критерием качества метода оценки ответов является близость полученной оценки и той оценки, которую выставил бы преподаватель за этот ответ. За какой ответ следует выставить более высокую оценку: за ответ, не содержащий одного из необходимых элементов или за ответ, в котором один элемент лишний?

Попробуем ответить на этот вопрос с формальной точки зрения. Будем считать все элементы ответа тестируемого равноценными. Начнём с ответа типа МНОЖЕСТВО.

Какими формальными правилами руководствуется преподаватель при выставлении оценки? Во-первых, абсолютно правильный ответ оценивается максимальной оценкой, а абсолютно неправильный – минимальной. Во-вторых, чем ближе ответ к эталону, тем выше его оценка. В-третьих, если два ответа обнаруживают одинаковую степень неправильности, то они должны оцениваться одинаково. Запишем это формально для эталона $E = (a_1, a_2, \dots, a_n)$ и различных ответов A :

$$1. A = (a_1, a_2, \dots, a_n): E \equiv A \Rightarrow \delta(A) = 1$$

$$A = (a_{n+1}, a_{n+2}, \dots, a_{n+m}): E \cap A = \emptyset \Rightarrow \delta(A) = 0$$

$$2. A \rightarrow E \Rightarrow \delta(A) \rightarrow 1$$

$$3. A^1 \approx A^2 \Rightarrow \delta(A^1) \approx \delta(A^2)$$

Очевидно, что первым двум правилам предложенный метод удовлетворяет. Сложнее проверить соответствие третьему правилу. Неправильный ответ отличается от правильного отсутствием каких-либо элементов, наличием "лишних" элементов или заменой одних элементов другими.

Рассмотрим варианты ответов, ближайших к эталону, т.е. отстоящих от него "на один шаг" (на один элемент). Возможны три способа получения таких ответов: убрать один элемент из эталона, добавить один элемент или заменить один элемент другим.

Если эталонный ответ содержит три элемента (a, b, c), то ответы (a,b) и (a,c), в каждом из которых отсутствует один из требуемых элементов, должны быть оценены одинаково. Другой вариант ответа – (a, b, c, d) – содержит все элементы эталонного ответа, но один элемент в нём лишний. Очевидно, что такой ответ должен быть оценен не максимальной оценкой, но выше, чем вариант с одним отсутствующим элементом. И ещё один вариант – (a, b, d): вместо одного правильного элемента один лишний. Это комбинация первых двух вариантов, т.е. обе ошибки (одного не хватает и один лишний) присутствуют одновременно. Понятно, что такой ответ должен быть оценен ниже, чем два предыдущих.

Опишем эти варианты формально и посчитаем их оценки:

$$E = (a_1, a_2, \dots, a_n), \delta(E) = 1$$

$$A^1 = (a_1, a_2, \dots, a_{n-1}), \delta(A^1) = \frac{n-1}{n}$$

$$A^2 = (a_1, a_2, \dots, a_n, a_{n+1}), \delta(A^2) = \frac{n}{n+1}$$

$$A^3 = (a_1, a_2, \dots, a_{n-1}, a_{n+1}), \delta(A^3) = \frac{n-1}{n+1}$$

$$\delta(A^3) < \delta(A^1) < \delta(A^2)$$

Эти рассуждения можно распространить на ситуацию с двумя и более отсутствующими или лишними элементами.

Перейдём к ответу типа СПИСОК. Для таких ответов правильность зависит от совпадения элементов эталона и ответа и от совпадения порядка этих элементов. Пусть ответ совпадает с эталоном по элементам, но есть ошибки в упорядочении. Для эталона [1,2,3,4] какой ответ должен быть оценен выше: [1,4,2,3], или [2,1,4,3], или [2,3,1,4]? Здесь сложно высказать какие-либо мотивированные соображения в пользу того или иного ответа, если не учитывать семантику вопроса. В одном случае важно, с чего начинать список, и тогда первый вариант должен получить более высокую оценку. В другом случае сдвиг одного элемента на две позиции хуже, чем сдвиг двух элементов на одну позицию каждый.

Но с формальной точки зрения все эти ответа содержат по две ошибки:

- [1,4,2,3] – 4-й элемент сдвинут на две позиции,
- [2,1,4,3] – первый и третий элемент сдвинуты на одну позицию каждый,
- [2,3,1,4] – первый элемент сдвинут на две позиции.

В соответствии с Д-методом количество ошибок при упорядочении определяется количеством операций перестановки элементов при изменении порядка элементов ответа на правильный. В результате все вышеперечисленные ответы будут оценены одинаково.

Более интересной представляется ситуация с наличием лишних или отсутствием требуемых элементов в списке. Примем за единичную ошибку отсутствие одного элемента в списке, наличие одного лишнего элемента или нарушение порядка следования двух соседних элементов.

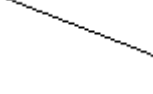
ЭТАЛОН $e = [1, 2, 3, 4]$,

ответы: $\mathbf{a} = [1, 2, 3]$ – отсутствие одного элемента;

$\mathbf{b} = [1, 2, 3, 4, 5]$ – один лишний элемент;

$c = [1, 2, 4, 3]$ – изменение порядка следования двух элементов.

Таблица 1. Оценки ответов типа СПИСОК

 Ответ Оценка	a	b	c
δ_{MH}	3/4	4/5	1
$\delta_{СП}$	1	1	5/6
δ	3/4	4/5	5/6

Интерполяция этих расчётов на более серьёзные расхождения эталона и ответов в случае со списками не так очевидна, как в случае с множествами. Вариантов с расхождением в двух элементах существенно больше:

- [1, 2] – двух элементов не хватает,
 [1, 2, 3, 4, 5, 6] – два элемента лишние,
 [1, 4, 2, 3] – смещение на две позиции,
 [2, 1, 3] – одного элемента не хватает и одна перестановка,
 [2, 1, 3, 4, 5] – один элемент лишний и одна перестановка,
 [1, 2, 3, 5] – один элемент лишний и одного не хватает.

Не будем приводить числовые оценки (при желании это легко можно сделать самостоятельно). Вместо этого сравним "вклад" отдельных видов расхождений в окончательный результат.

Результаты расчётов для $n=10$ (список из 10 элементов) приведены на рис. 1. Моделирование проводилось для случая изменения значения одного параметра при фиксированных значениях двух других параметров. Как видно из графиков на рис. 1, сильнее всего влияет отсутствие требуемых элементов, а неправильная упорядоченность списка влияет в меньшей степени.

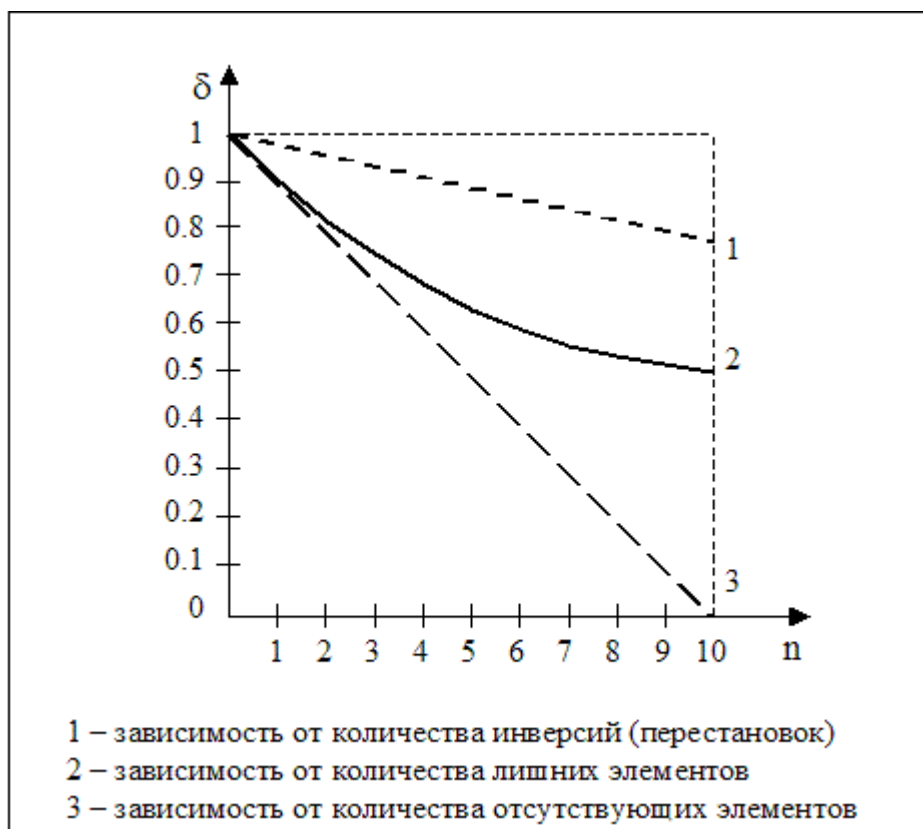


Рис. 1. Зависимость степени схождения ответа и эталона от различных параметров

Очевидно, что такое положение вещей не очень подходит для тех ситуаций, когда последовательность элементов ответа чрезвычайно важна, и перестановка хотя бы двух элементов должна приводить к выставлению минимальной оценки. Но в таких случаях вполне достаточно использовать сравнение на полное совпадение двух списков.

Примечание: в реальной системе контроля знаний TS (<http://rema.44.ru/razrabotki/> , [3]), использующей данный метод оценки ответов, предлагаются дополнительные средства для реализации правильного поведения системы в подобных ситуациях.

3. Заключение

Рассмотренный Д-метод имеет несомненные достоинства, позволяющие снизить трудозатраты составителя и при этом повысить качество тестирования, заставляя тестируемого думать над ответом, а не просто пытаться выбрать наиболее правдоподобный из готовых вариантов.

Результаты анализа предложенного Д-метода показывают, что он демонстрирует близость к оценкам преподавателей. Формально эти выводы можно сформулировать следующим образом:

- на снижение оценки ответов при использовании Д-метода сильнее всего влияет отсутствие требуемых элементов ответа;
- наличие лишних элементов влияет в меньшей степени;
- неправильная упорядоченность элементов списка влияет ещё меньше.

Данные результаты подтверждаются личным преподавательским опытом автора.

Литература

1. Аванесов В.С. Научные проблемы тестового контроля знаний / Монография. – М.: Исследовательский центр проблем качества подготовки специалистов, 1994. – 135 с.
2. Зайцева Л.В., Прокофьева Н.О. Модели и методы адаптивного контроля знаний. – Educational Technology & Society 7(4) 2004. – pp. 265-277.
http://ifets.ieee.org/russian/depository/v7_i4/html/1.html
3. Карпов В.Э., Карпова И.П. Язык описания системы контроля знаний. – Компьютеры в учебном процессе, 2000, №4. – с.147-155.
4. Карпова И.П. Анализ ответов обучаемого в автоматизированных обучающих системах. – Информационные технологии, 2001, № 11. – с.49-55.
5. Кнут Д. Искусство программирования для ЭВМ / т. 3. "Сортировка и поиск" / Пер. с англ. Под ред. Баяковского и Штаркмана. – М.: "Мир", 1978. – 848 с.
6. Патругин Ю.А. Экономика и математические методы. – 1970. Т. VI. № 6. – с. 887-893.