УДК 005.935.3 + 025.2
ББК 65.050.2-73

О РЕЙТИНГЕ ЖУРНАЛОВ И АГРЕГИРОВАНИИ НЕПОЛНЫХ БАЛЛЬНЫХ ОЦЕНОК

Чеботарев П. Ю.[1]
(*Учреждение Российской академии наук
Институт проблем управления РАН, Москва*)

Обсуждается метод обработки данных, предложенный в статье О. В. Федорца «Коллективная экспертиза научных журналов: методика агрегирования экспертных оценок и построения рейтинга», опубликованной в этом же выпуске Сборника. Предложен подход к агрегированию неполных балльных экспертных оценок. Он использует взвешенное среднее оценок, включающих нейтральную априорную оценку.

Ключевые слова: балльные экспертные оценки, неполные данные, априорная оценка, рейтинг журналов

1. Важность темы

С интересом прочел статью О. В. Федорца «Коллективная экспертиза научных журналов: методика агрегирования экспертных оценок и построения рейтинга». Тема ее должна привлечь внимание многих. Действительно, практически каждый научный работник строит для себя неформальный рейтинг журналов в своей области. Этот рейтинг помогает ему отвечать на важные вопросы: «Статьи в каких журналах читать в первую очередь?», «Куда послать свою новую статью?», «Будет ли эта публикация замечена коллегами и достойным образом оценена?» и т. д. Кроме того, интересно знать, чем руководствуются

[1] *Павел Юрьевич Чеботарев, доктор физико-математических наук, ведущий научный сотрудник* (*pchv@rambler.ru*).

библиографы при комплектовании фондов библиотек, а также составители баз данных ВИНИТИ и международных реферативных журналов, баз данных цитирования *Web of Science*, *Science Citation Index* и др. Для ответа на эти вопросы статья дает много информации и полезных ссылок.

Весьма интересна она и специалистам по экспертному оцениванию. Дело в том, что в работе сравниваются объекты, получившие разное количество оценок, т. е. экспертная информация неполна, а обработка неполной информации – одна из интереснейших задач анализа экспертиз.

2. *О методике агрегирования экспертных оценок, предложенной в статье О. В. Федорца*

На мой взгляд, с методологической точки зрения работа заслуживает довольно серьезной критики; этой критике и посвящен данный раздел.

Эвристическая методика, предложенная автором для построения рейтингов научных журналов, не производит впечатления тщательно, в деталях, продуманной. Полагаю, что ее использование при определенных исходных данных может приводить к ошибочным результатам. Далее я это поясню.

Первоначально автор ранжирует журналы в порядке убывания суммы экспертных оценок. Он замечает, что результат оказывается неадекватным из-за того, что количество оценок для разных журналов различно, и журналы, получившие только высокие оценки, но в небольшом количестве, «оказались значительно ниже журналов, получивших множество средних оценок». Вместо того чтобы скорректировать этот критерий за счет учета количеств оценок, автор добавляет второй критерий: максимум из всех оценок журнала, и этот критерий назначает главным, лексикографически подчиняя ему рассмотренный ранее критерий суммы оценок. Использование такой лексикографии вызывает ряд возражений.

Во-первых, данный подход дает слишком большие полномочия каждому отдельному эксперту. Если один эксперт поставил журналу высшую оценку, то этот журнал автоматически оказывается в лидирующей группе. Но, как мы понимаем, необъективность эксперта, в частности, его заинтересованность в «продвижении» определенного журнала, не исключена. Автор отмечает, что «в лидирующем кластере оказалось 45,5% журналов, получивших хотя бы одну высокую оценку», и делает из этого вывод, что «влияние одного эксперта невелико». С этим трудно согласиться: влияние эксперта, поставившего одну из нескольких высших оценок, полученных журналом, действительно, умеренно, но эксперт, поставивший журналу единственную его высшую оценку, радикально повышает место журнала в общем рейтинге. И получается, что качественные журналы, которым все оценившие их эксперты поставили высокую (но не высшую) оценку, оказываются в лучшем случае в середине рейтинга. А могут (если получили не очень много оценок) оказаться ниже середины.

В конце статьи О. В. Федорца помещен «TOP–20 научных сериальных изданий в 2008 г. по результатам обработки экспертных оценок». И в этом рейтинге журнал «Известия высших учебных заведений (вузов). Северо-Кавказский регион. Технические науки» обгоняет все отечественные журналы за исключением «Докладов РАН» и пропускает вперед лишь 7 международных изданий. Места 3-4 среди российских журналов занимают «Известия Томского политехнического университета» и «Горный информационно-аналитический бюллетень». Лишь за ними следует «Известия РАН. Серия физическая». Хотелось бы понять, чем объясняется такой высокий рейтинг упомянутых журналов, занявших 2-4 места среди российских: их высокой популярностью среди экспертов, специфичностью предмета экспертизы («желательность его [издания] обработки в ВИНИТИ РАН и отражения в РЖ») или же не вполне удачной методикой агрегирования оценок.

Двух критериев оказывается недостаточно: «образовались значительные кластеры журналов с повторяющимися значениями» этих критериев. Действительно, как отмечено выше, главный критерий почти половину журналов помещает в лидирующий кластер; для этих журналов нередко повторяются и суммы оценок (второй критерий). Поэтому автор вводит третий критерий, лексикографически подчиненный первым двум. Он назван «высшим рангом» и, как и главный критерий, позволяет отдельному эксперту продвинуть вперед свой «любимый» журнал. Строится критерий «высшего ранга» так: возвращаемся к балльным оценкам, которые *j*-ый эксперт приписал журналам, и ранжируем журналы по убыванию этих баллов, в случае равных баллов ранжируя журналы по убыванию суммы экспертных оценок (т. е. по второму критерию). Поскольку разные эксперты оценили разное число журналов, вычисляются *нормированные* ранги, т. е. каждый ранг делится на максимальный для данного эксперта ранг. Третий критерий есть критерий минимизации минимального (по экспертам) нормированного ранга.

Для «продвижения» своего любимого журнала критерий «высшего ранга» может быть использован так: эксперту нужно назначить этому журналу, единственному, наивысшую оценку, а всего оценить как можно больше журналов. Если эксперт обгоняет по количеству оцененных журналов остальных экспертов, то его «любимый» журнал обгонит по критерию «высшего ранга» все остальные журналы.

В связи с обсуждением здесь такого стратегического поведения экспертов можно, конечно, вспомнить, что само понятие «эксперт» несовместимо с личной заинтересованностью и лоббированием. Тем не менее, во-первых, гарантировать идеальность экспертов нельзя, а во-вторых, случаются и «добросовестные» грубые ошибки у вполне квалифицированных специалистов. Поэтому сама процедура агрегирования должна быть хотя бы в некоторой степени защищена «от дурака» и от желающего ею манипулировать (полной защиты, разумеется, не существует). Общим принципом построения агрегирующих процедур

является их устойчивость, т. е. разумное ограничение влияния отдельного эксперта на результат групповой экспертизы.

Следует отметить, что главная особенность предложенной лексикографической 3-критериальной процедуры, а именно, ее «оптимистическое» доверие к максимальной оценке, полученной каждым журналом, получает в доработанной версии статьи некоторое обоснование. Автор пишет: «Конечно, существует риск, что эксперт может поставить высокую оценку журналу случайно. Присвоена хотя бы одна высокая оценка – и журнал автоматически оказывается в лидирующем кластере. В этом случае журнал может быть ошибочно включён в подписку, запущен в технологию обработки, и ошибка будет исправлена не раньше чем через год. Однако многоотраслевой реферативной службе, стремящейся отражать наиболее передовые научные издания в каждом тематическом РЖ, гораздо важнее не упустить ценный монотематический журнал». Вместе с тем бюджет всегда ограничен и ошибочное включение журнала в подписку исключит из нее более важный журнал. Поэтому лучше не облегчать до предела манипулирование результатами экспертизы.

Что касается поддержки монотематических журналов, получивших мало оценок, предложенная методика, вообще говоря, не решает эту задачу. Действительно, представим себе ведущий монотематический журнал, которому поставлены высшие оценки, но число этих оценок весьма умеренно, и мультидисциплинарный журнал с большим количеством средних и низких оценок. Тогда, если у второго журнала есть всего один горячий сторонник среди экспертов, поставивший ему высшую оценку, то данный журнал попадет в лидирующий кластер и по критерию суммы оценок оставит далеко позади первый журнал. Не по этой ли причине второе и третье места среди всех российских изданий заняли мультидисциплинарные региональные журналы?

Второй основной недостаток предложенной методики состоит в том, что сумма оценок при разном их количестве у разных журналов – вообще малоосмысленный критерий: в

соответствии с ним много низких оценок лучше, чем одна высокая. Представим себе, что опрос дополнен несколькими самыми низкими (равными 0,1)[1] оценками журнала X, полученными от экспертов, которые ранее этот журнал не оценивали. Здравый смысл требует, чтобы при большом числе таких добавочных оценок положение X в рейтинге существенно ухудшилось. Что же дает методика? Максимальные оценки журналов (главный критерий) не меняются. А сумма оценок журнала X (второй критерий) увеличится, и в нарушение здравого смысла он может занять более высокое место!

Таким образом, в целом предложенную эвристическую методику построения рейтинга можно назвать причудливой и не свободной от явных недостатков. Перечислим их еще раз: 1) процедурой очень легко манипулировать, так как главным критерием выбрана максимальная оценка журнала; 2) сумма оценок (второй критерий) при неполных данных – показатель, искажающий качество объектов; в частности, добавление к множеству оценок журнала самых низких оценок повышает, а не понижает его оценку; 3) «наилучший ранг» журнала (третий критерий, являющийся аналогом 1-го критерия, учитывающим также 2-й критерий и неоднородность экспертов по количеству оценок) дополнительно облегчает манипулирование результатами опроса.

Для освобождения от этих недостатков и поддержки сильных монотематических журналов есть довольно простое средство: критерий взвешенного среднего всех оценок, включая априорную. Его использованию посвящен следующий раздел.

---

[1] *В шкале есть и нулевая оценка, но она закреплена за сравнительно нейтральной семантической градацией* («*журнал выдан эксперту случайно и не соответствует его тематике*», *либо* «*журнал соответствует тематике эксперта, но статьи по его тематике в этом журнале носят научно-популярный характер*»).

3. Агрегирование неполных балльных оценок: взвешенное среднее и презумпция нейтральности

Имеет смысл в первую очередь так изменить критерий суммы оценок, чтобы он адекватно работал с неоднородными (по количеству оценок) данными. Замена суммарной оценки на среднюю не решает проблемы, так как при этом совокупность оценок (5, 5, 5) (5 – наивысшая оценка[1]) оказывается не лучше одиночной оценки 5.

Предложим иной подход: использование взвешенного среднего с априорной оценкой. Рассмотрим функцию s_t от экспертных оценок $c_1, \ldots, c_m$ следующего вида:

$$(1) \quad s_t(c_1,\ldots,c_m) = \frac{1}{m+t}\sum_{i=1}^{m} c_i \ ,$$

где m – число экспертов, оценивших объект (отказ от оценивания не считается оценкой), t – положительная константа либо функция от m. Например, выбор $t = 1/3$ приводит к следующему упорядочению некоторых наборов оценок: $s_t(5, 5, 5) > s_t(5, 5) > s_t(5, 5, 4) > s_t(5, 5, 3) > s_t(5, 4) > s_t(5) > s_t(5, 3) > s_t(4)$, что приемлемо. При использовании в качестве критерия функции s_t низшая оценка должна «оцифровываться» не нулем, а отрицательным числом. Это связано с тем, что добавление новых наихудших оценок должно ухудшать положение объекта в рейтинге, а «оценить» набор (0, 0, 0) ниже, чем (0) с помощью критерия вида (1) невозможно. В то же время, чтобы «оценить» набор (–5, –5, –5) ниже, чем (–5), функция s_t подходит. При данной оцифровке нулевое числовое значение должно соответствовать «средней» оценке, равноотстоящей от минимальной и максимальной.

---

[1] *Здесь для наглядности в качестве оценок взяты номера́ уровней, а не сопоставленные им* «*численные значения*» *из таблицы* 3, *но тот же подход применим и для численных значений* (*кроме нулевого, которое в методике О. В. Федорца семантически выбивается из ряда остальных и скорее должно трактоваться как отказ от оценивания*).

Легко убедиться, что использование симметричной относительно нуля шкалы и критерия максимизации функции s_t вида (1) – процедура, свободная от недостатков, перечисленных в конце предыдущего раздела. Проблемы с мультидисциплинарными «журналами-середняками» при этом также не возникает: они не получают неадекватно высоких оценок.

Теперь следует объяснить, почему функции s_t относятся к классу взвешенных средних с априорной оценкой. Взвешенными средними арифметическими называются функции вида

$$(2)\quad s_{\mathbf{w}}(c_1,\ldots,c_m)=\frac{\sum_{i=1}^{m} w_i c_i}{\sum_{i=1}^{m} w_i},$$

где $\mathbf{w} = (w_1, \ldots, w_m)$, $w_i \geq 0$, $i = 1, \ldots, m$, – ненулевой вектор весов наблюдений.

Предположим, что исследователь придерживается «презумпции нейтральности», заключающейся в том, что все объекты исходно равноценны, и, ничего о них не зная, он относится к ним как к *средним*, т. е. (при симметричной относительно нуля шкале оценок) приписывает им априорную оценку 0. Усредняя полученные позже экспертные оценки объекта, исследователь добавляет к ним и свою априорную нулевую оценку, но с весом, возможно, отличным от весов «реальных» наблюдений. Такой подход выражает определенную осторожность (консерватизм) исследователя: например, если объект получил единственную максимально возможную оценку, исследователь оценивает его не этим максимальным значением, а взвешенным средним между ним и априорной нулевой оценкой.

Нетрудно убедиться, что если исследователь приписывает всем «реальным» оценкам единичные веса, а своей априорной нулевой оценке вес t, то формула (2) приобретает вид (1), и значит, функции s_t принадлежат классу взвешенных средних с априорной оценкой.

Усреднение с априорной информацией отвечает духу байе-

совского статистического оценивания. Кстати, в обсуждаемой статье, к сожалению, вообще не упоминается большой класс статистических методов обработки экспертных оценок.

В случае, когда значения функции s_t (для которой можно специально поработать над выбором константы или функции t) у нескольких журналов совпадают (что будет реже, чем для суммы оценок), можно ранжировать их по убыванию максимальной оценки, если исследователь признает этот показатель важным. Еще лучше для устойчивости и снижения шансов совпадения значений использовать сумму двух (трех) максимальных оценок. А можно в качестве единственного критерия использовать взвешенную сумму усредненной оценки вида (1) и максимальной оценки журнала, первой присваивая значительно больший вес.

4. К организации экспертного опроса

Мне кажется весьма возможным, что качество работы экспертов повысилось бы в результате следующей модификации процедуры опроса.

Вспомним, что характеристика, которую оценивают эксперты, – это «желательность его [издания] обработки в ВИНИТИ РАН и отражения в РЖ». Но для большого массива журналов вопрос так, вероятно, не стоит: всем ясно и без опроса, что эти издания необходимо обрабатывать и отражать. Не сто́ит ли тогда ограничить опрос зоной неопределенности, включив в него только те издания, относительно которых существуют сомнения? Это заставило бы экспертов отнестись к новым и недостаточно известным им изданиям с повышенным вниманием. Возможно, при этом оказалась бы полезной более дифференцированная (в области средних значений) вербальная шкала оценок. Образно говоря, зону неопределенности полезно было бы «рассмотреть под увеличительным стеклом», сосредоточившись на сравнительных достоинствах попадающих в нее журналов.

В качестве резюме следует отметить, что рассматриваемая

статья О. В. Федорца посвящена важной прикладной задаче, представляющей практический интерес для многих научных работников. Важно знать, как она решается – тогда можно обсуждать и альтернативы. Ничего нет хуже практиковавшегося одно время ВАК РФ появления непонятно как составленных списков международных журналов, в которых, в отличие от других (нередко, ведущих), «могут быть опубликованы основные научные результаты» докторских и кандидатских диссертаций.

JOURNALS EVALUATION AND AGGREGATION OF INCOMPLETE POINT JUDGMENTS

Pavel Chebotarev, Institute of Control Sciences of RAS, Moscow, D.Sc., leading researcher (pchv@rambler.ru).

Abstract: *We discuss the method of data analysis proposed in the article "Collective expert examination of scientific journals: Procedure of expert judgments aggregation and rating construction" by O. V. Fedorets published in the same issue. An alternative approach to the aggregation of expert point judgments with non-uniform number of estimates is proposed. This approach involves the weighted mean of the judgments including a neutral prior judgment.*

Keywords: expert judgments, incomplete expert data, point score, prior judgment, rating of scientific journals.