УДК 004.052.42

В.И. Кузовлев, А.О. Орлов

ВЕРОЯТНОСТНЫЙ ПОДХОД К ОЦЕНКЕ ПОКАЗАТЕЛЯ ДОСТОВЕРНОСТИ ЭЛЕМЕНТОВ РЕЗУЛЬТАТОВ ПРОФИЛИРОВАНИЯ

Рассмотрены показатели достоверности информационных элементов как результатов процесса профилирования данных. Описаны методы оценки показателей достоверности и их недостатки. Предложены оператор Tr, упорядочивающий элементы внутри отдельно взятого класса профилирования, а также формула для расчета значения оператора.

E-mail: forewar@gmail.com

Ключевые слова: *профилирование, классы профилирования, показатели достоверности, стратегии повышения качества данных.*

Введение. Для анализа и контроля достоверности данных разработаны средства и методы, позволяющие оценить базовые показатели достоверности на основе принятых в этих методах моделей. Средствами контроля качества данных являются репозитории метаданных, средства профилирования информации, системы управления базами данных и др. Также разработаны методы, основанные как на анализе самих данных, так и процессов их формирования и преобразования в процессе функционирования систем.

В работе [1] для оценки показателей достоверности используется набор графов: граф ошибок $\varepsilon(\pi)$; индикаторный граф $J(\pi)$; информационный граф $I(\pi)$. Информационный граф $I(\pi)$ определяет общую технологию обработки данных. На его основе строится индикаторный граф $J(\pi)$, который отображает события возникновения ошибок в обрабатываемых информационных элементах. Граф ошибок $\varepsilon(\pi)$ формируется на базе индикаторного графа, его вершинами также являются индикаторы переменных логических функций, а дуги отображают причинно-следственные связи между индикаторами событий ошибки.

Схема технологии возникновения и распространения искажений, основными элементами которой являются процессы формирования информационных элементов и связывающие их потоки искажений, предложена в работе [2]. Схема делится на два уровня: уровень обобщения потоков и уровень детализации процессов. Первый уровень содержит информацию о процессах формирования информационных элементов и потоках искажений, связывающих эти процессы. Выделяются первичные и целевые процессы, учитывается степень детализации проведения исследования, определяются общие особенности технологии возникновения и распространения искажений в ав-

томатизированной системе обработки информации. На втором уровне находятся показатели достоверности обработки информации для всех элементов схемы потоков искажений. Степень детализации проведения исследования варьируется в зависимости от решаемых прикладных задач.

Средства профилирования информации оперируют метаданными, но, в отличие от репозиториев, не только хранят метаданные, но обрабатывают и изменяют их. По сравнению со средствами управления базами данных у средств профилирования имеется более широкий спектр возможностей анализа данных. На рис. 1 приведена схема преобразования данных в процессе профилирования.

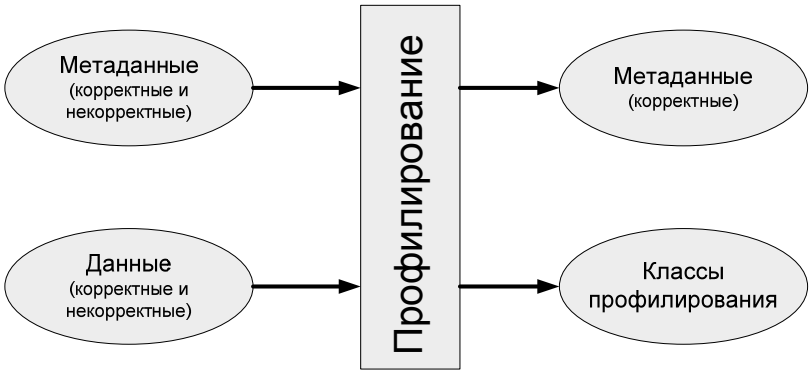

Рис. 1. Схема преобразования данных в процессе профилирования

На вход процесса профилирования подаются данные и метаданные [3]. Они могут быть корректными и некорректными. В результате профилирования формируются корректные метаданные и классы профилирования. Профилирование состоит из нескольких шагов, в которых данные анализируются различными методами (анализ столбцов, структурный анализ, анализ правил, анализ значений). После профилирования остается некоторое количество некорректных данных, не классифицированных процессом (рис. 2).

На всех шагах профилирования происходит анализ документированных (отраженных в метаданных) свойств объектов и выявление их недокументированных свойств. Далее путем проверки свойств определяются некорректные данные. Для столбцов анализируются свойства значений атрибута (домен, текстовые правила, шаблоны, интерпретация спецсимволов и т. п.) и свойства хранилища (длина атрибута, тип данных и т. п.).

В процессе анализа выявляются дублирующиеся данные. Для поиска дубликатов текстовых данных используются различные алгоритмы, наиболее распространенный из которых — алгоритм шинглов. Суть алгоритма заключается в следующем: текстовая строка разбивается на подстроки одинаковой длины с определенным шагом,

меньшим длины строки. По набору полученных подстрок строится сигнатура документа. Документы считаются дубликатами в том случае, если их сигнатуры совпадают. Развитием алгоритма шинглов является метод ключевых слов [4]. По определенным параметрам из текста выбираются ключевые слова, к которым потом применяется алгоритм шинглов. Такой подход сокращает объемы анализируемых данных, сохраняя при этом высокое качество построения сигнатуры.

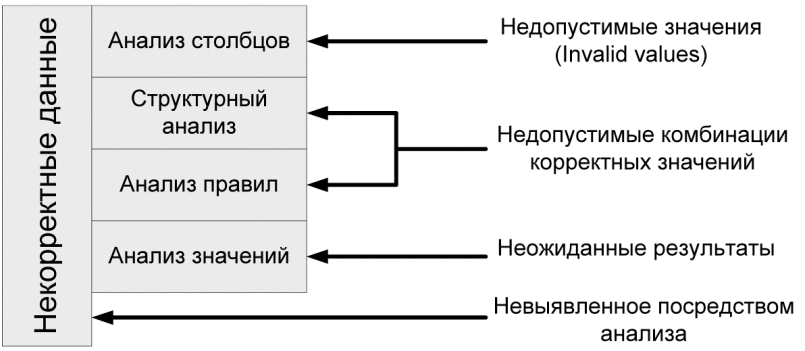

Рис. 2. Типы значений, определяемые в процессе профилирования

В процессе профилирования фильтры f рассматриваются как отдельные бизнес-задачи, несущие в себе смысловую нагрузку. Тогда с точки зрения качества информации производительность системы можно определять по количеству решаемых бизнес-задач, т. е. по количеству успешно обрабатываемых фильтров.

При анализе фильтра данные подразделяются на несколько классов профилирования. Под данными понимаются объекты r_i отношения R, а под классами — множество $S_1, \ldots, S_i$ где $i \geq 2$ (рис. 3). Часть полученных классов является проблемными, в них попадают объекты r, не удовлетворяющие набору условий фильтра. Классы с объектами, соответствующими условиям фильтра, называются успешными классами.

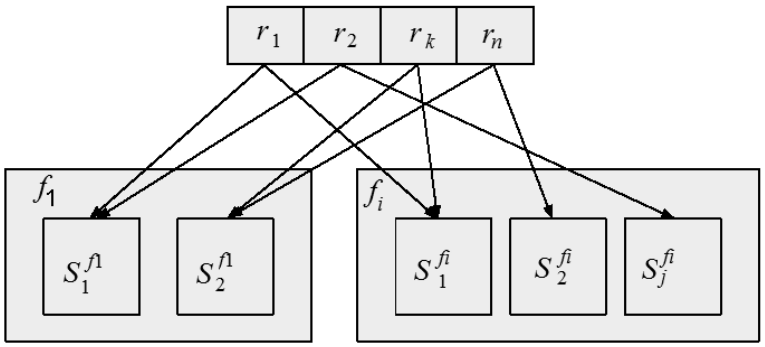

Рис. 3. Схема миграции объектов при профилировании

В процессе повышения качества данных объекты из проблемных классов должны переместиться в успешные классы. Таким образом, все данные в результате анализа фильтра f_i попадут в один или несколько успешных и только успешных классов. Тогда фильтр считается успешно обработанным, а система — корректно работающей в рамках соответствующей бизнес-задачи.

Пусть $CF(r_j)$ — количество проблемных классов, в которые попал объект r_j по результатам обработки всех фильтров; $CT(r_j)$ — количество успешных классов, содержащих объект r_j. Следовательно, общее количество классов по всем фильтрам $\sum_j \left(CF(r_j) + CT(r_j) \right) = CS$. Очевидно, что объект r_j содержится во всех классах n раз, где n — общее количество фильтров.

Степень искаженности объекта r_j

$$\frac{CF(r_j)}{CS}.$$

Объект является корректным, если $CF(r_j)/CS = 0$.

Если $S_k^{f_i}$ – k-й класс по i-му фильтру, то объекты r_j внутри каждого класса $S_k^{f_i}$ можно сортировать по критерию степени искаженности объекта r_j. Оператор, упорядочивающий объекты r внутри отдельного класса S:

$$Tr(r_j) = 1 - \frac{\left| CF(r_j) \right|}{CS}.$$

В первую очередь, будут обрабатываться те объекты r класса S, которые наименее искажены.

Мощность отношения R, т. е. количество экземпляров в этом отношении

$$\left| S_F^{f_i} \right| + \left| S_T^{f_i} \right| = |R|,$$

где $\left| S_F^{f_i} \right|$ — мощность множества проблемных классов по результатам обработки фильтра f_i; $\left| S_T^{f_i} \right|$ — мощность множества успешных классов для фильтра f_i.

В результате профилирования будут выводиться те проблемные классы, которые имеют минимальное количество объектов. Объекты в свою очередь распределяются внутри класса так, что сначала идут объекты, попавшие в наименьшее количество классов. После выравнивания всех объектов в текущем проблемном классе S_j фильтра f_i возникает ситуация, когда по фильтру f_i не осталось проблемных классов, т. е. фильтр обработан успешно.

В то же время событие попадания объекта r_j в тот или иной класс носит вероятностный характер. В случае попадания объекта r_j в какой-либо проблемный класс CF можно утверждать о недостоверности информации в объекте r_j в условиях данного фильтра f_i. Следовательно, необходимо разработать показатель достоверности классов в фильтрах. В работе [5] в качестве основного показателя достоверности предлагается использовать вероятность искажения единицы информации:

$$P_{\text{ед}} = \frac{n_{\text{ед}}}{N_{\text{ед}}},$$

где $n_{\text{ед}}$ — число искаженных единиц информации; $N_{\text{ед}}$ — общее число обрабатываемых единиц информации. Таким образом, $P_{\text{ед}}$ — вероятность события, когда рассматриваемый фильтр содержит хотя бы одну искаженную единицу информации. Данный показатель не дает информации об уровне искажений рассматриваемого элемента, то есть о характере возникающих в нем ошибок.

В работе [6] рассматривается показатель, отражающий вероятность возникновения ошибки h-го класса при обработке i-го информационного элемента. Если $L_i = \{l_{i,z}\}$ — множество всех ошибок, возникающих при обработке i-го информационного элемента, то $v_{i,h} \in L_i$; $q(v_{i,h}) = 1 - \Pi(1 - q_{i,z})$, где $q_{i,z}$ — вероятность возникновения z-й ошибки при обработке i-го элемента, $l_{i,z} \in v_{i,h}$.

В показателе, предложенном в работе [6], не учитывается возможность возникновения нескольких разнородных ошибок в единственном рассматриваемом элементе, а сами элементы не соотносятся друг с другом.

Оператор Tr, упорядочивающий элементы внутри одного класса профилирования, предложен в работе [7]. Значение оператора Tr можно вычислить как вероятность события, при котором в j-м обрабатываемом элементе присутствует минимальное количество ошибок по всем фильтрам. Другими словами, Tr — вероятность такого события, когда рассматриваемый элемент принадлежит классу CF в условиях фильтра f_i и не принадлежит никаким другим классам CF в условиях множества других фильтров:

$$Tr = q\left(v_{i,h}\right)\left(1 - q\left(L_i / v_{i,h}\right)\right) =$$

$$= \left(1 - \prod\nolimits_{l_i,\, z \,\in v_{i,h}} \left(1 - q_{i,z}\right)\right)\left(1 - \left(1 - \prod\nolimits_{l_i, z \in L_i / v_{i,h}} \left(1 - q_{i,z}\right)\right)\right) =$$

$$= \prod\nolimits_{l_i,\, z \,\in L_i / v_{i,h}} \left(1 - q_{i,z}\right) - \prod\nolimits_{l_i, z \in L_i} \left(1 - q_{i,z}\right).$$

Значение оператора *Tr* равно вероятности события, при котором в обрабатываемом информационном элементе произойдет хотя бы одна ошибка из класса h и только из класса h, т. е. элемент попадет в класс *CF* в условиях фильтра f_h и только его. Таким образом, если элемент содержит ошибки по всем фильтрам, оператор *Tr* для него будет равен нулю, т. е. вес элемента будет считаться минимальным, а элемент — наименее полезным с точки зрения смысловой нагрузки ввиду максимальной искаженности. И наоборот, если элемент включает в себя ошибку только по данному фильтру, тогда оператор *Tr* сводится к показателю $q(v_{i,h})$ из работы [6].

Специфика функционирования автоматизированных систем обработки информации такова, что искажение отдельных элементов данных не останавливает работу системы в целом. Оценка достоверности данных проводится внутри процессов функционирования системы. При этом важно оценивать данные, полученные в процессе анализа, по уровню критичности искажений, а также по степени распространения однотипных искажений между различными информационными элементами.

В процессе профилирования проводится анализ данных, после которого в результате экспертной оценки итогов профилирования выбирается стратегия повышения качества данных. Введем следующее предположение: смысловая ценность полностью достоверного информационного элемента много выше ценности частично достоверного информационного элемента. Под полностью достоверным информационным элементом понимается такой элемент, который не содержит ошибок ни одного класса из всего множества классов v_i данного элемента. Полностью достоверный элемент не попадет ни в один класс *CF*. Напротив, за частично достоверный элемент принимается такой элемент, который содержит хотя бы одну ошибку из какого-либо класса v_i. Из данного предположения следует, что лучшей стратегией *W* повышения качества данных является такая стратегия, при которой в первую очередь обрабатываются информационные элементы с наименьшим количеством ошибок минимальной стоимости. Под стоимостью ошибки понимается размер убытков, выраженный в абсолютных или относительных единицах, которые повлечет за собой обработка информационного элемента с такой ошибкой.

Рассмотрим следующие варианты распределения ошибок в элементах результатов профилирования. В целях наглядности допустим, что имеется всего два класса искажений.

На рис. 4 представлено множество всех элементов профилирования *X*. Подмножество X_1, проецируется на множество *X* и выделяет элементы, содержащие ошибку из первого класса искажений. Аналогично, подмножество X_2 показывает элементы с ошибкой из второго класса искажений.

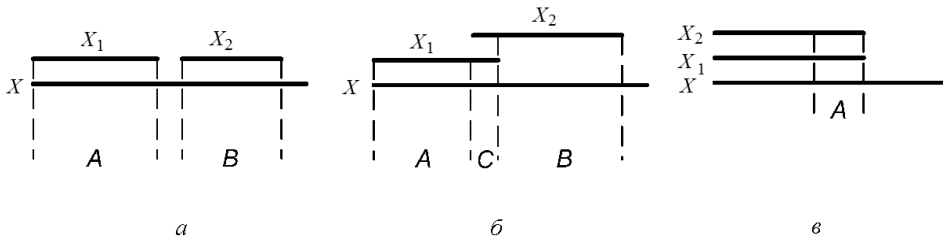

Рис. 4. Варианты распределения ошибок в элементах профилирования

Вариант, приведенный на рис. 4, *а*, соответствует такой ситуации, когда не существует информационных элементов, одновременно содержащих ошибки обоих классов искажений. В таком случае стоимость устранения ошибок для показателя $q(v_{i,h})$ составит:

$$C_q = C_1 X_1 q(v_{i,h1}) + C_2 X_2 q(v_{i,h2}),$$

где C_1, C_2 — средние стоимости ошибки первого и второго классов искажений; X_1 и X_2 — количество информационных элементов с ошибками первого и второго классов искажений. Очевидно, что для вариантов, представленных на рис. 4, *б* и *в*, стоимость устранения ошибок при показателе $q(v_{i,h})$ будет аналогичной, так как не учитывается взаимное расположение множеств X_1 и X_2. В таком случае стратегия повышения качества данных для всех трех вариантов будет одинакова при использовании показателя $q(v_{i,h})$.

Стоимость устранения ошибок при использовании оператора Tr для варианта, приведенного на рис. 4, *а*:

$$C_{Tr} = C_1 X_1 Tr(v_{i,h1}) + C_2 X_2 Tr(v_{i,h2}).$$

Поскольку элементы из подмножества X_1 содержат ошибки лишь по данному классу искажений h_1, а элементы из подмножества X_2 — только по классу искажений h_2, тогда $Tr(v_{i,h1}) = q(v_{i,h1})$, а $Tr(v_{i,h2}) = q(v_{i,h2})$. Откуда $C_q = C_{Tr}$, т. е. стоимости устранения ошибок в информационных элементах для варианта, приведенного на рис. 4, *а*, равны при использовании любого из показателей.

Однако для вариантов, приведенных на рис. 4, *б* и *в*, картина меняется. Для варианта на рис. 4, *б*, стоимость устранения ошибок составит

$$C_{Tr} = C_1 ATr(v_{i,h1}) + C_2 BTr(v_{i,h2}) + C_1 CTr(v_{i,h1}) + C_2 CTr(v_{i,h2}),$$

где A, B, C — подмножества множества X, содержащие ошибки только по классам искажений h_1, h_2 и (h_1, h_2) соответственно.

Исходя из предположения, выбирается такая стратегия, которая позволит получить максимальное количество полностью достоверных информационных элементов при минимальных затратах:

$$C_{Tr}(Q) = C_1 ATr(v_{i,h1}) + C_2 BTr(v_{i,h2}),\ C_{Tr}(Q) < C_{Tr},$$

следовательно, $C_{Tr}(Q) < C_q$.

Для варианта на рис. 4, *в*, когда все информационные элементы содержат одновременно ошибки из разных классов искажений, оператор *Tr* будет равен нулю. В этом случае целесообразно выбирать такую стратегию, при которой информационные элементы будут обрабатываться последовательно в целях получения максимального количества полностью достоверных элементов при постоянных затратах.

Эксперименты показали, что порядок выбора стратегии повышения качества данных сохраняется при использовании любого из показателей $q(v_{i,h})$ или *Tr*, однако конечный выбор альтернативы зависит от порядка весов ошибок.

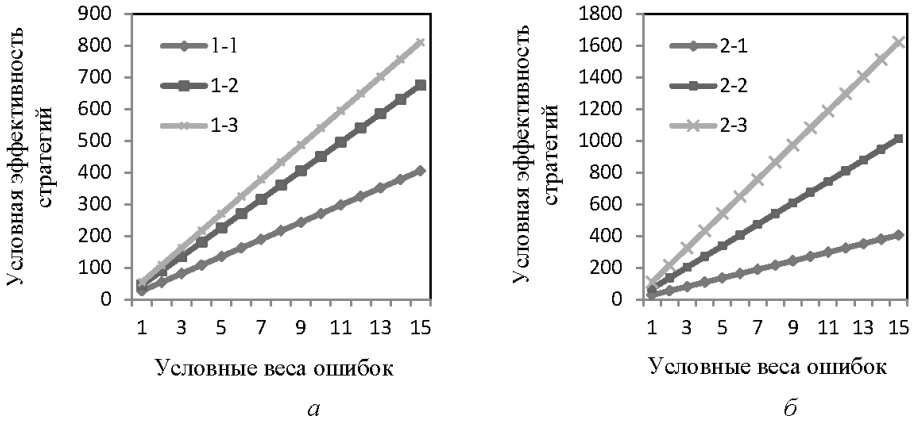

Рис. 5. Распределение альтернативных стратегий по затратам для показателя $q(v_{i,h})$ (*а*) и оператора *Tr* (*б*)

На рис. 5 приведены распределение альтернативных стратегий для показателя $q(v_{i,h})$ и для оператора *Tr* при одинаковых вероятностях возникновения ошибок в информационных элементах. Для решения проблемы выбора альтернативы необходимо оценить прогнозируемую эффективность стратегий при различных условных весах ошибок. Так, для показателя $q(v_{i,h})$ эффективность стратегии 1—1 при значении условного веса ошибки 7 равна эффективности стратегии 1—2 при значении условного веса ошибки 4. Такая кучность стратегий влечет за собой относительно низкую устойчивость выбора к изменению веса ошибки. Другими словами, вес ошибки превалирует над вероятностью ее появления. Напротив, при использовании оператора *Tr* стратегии обладают большим разбросом, что обеспечивает преимущество вероятности возникновения ошибки над ее относительным весом. Согласно результатам экспериментов, использование оператора *Tr* при оценке выбора стратегии повышения качества

данных позволяет добиться выигрыша в итоговом качестве информационных элементов до 30 % по сравнению с выбором стратегий, основанном на оценке показателя $q(v_{i,h})$.

Заключение. Предложенный оператор *Tr* позволяет оценивать распределение ошибок в информационных элементах и выбирать наилучшую стратегию в целях получения максимального количества полностью достоверных информационных элементов при минимальных затратах.

СПИСОК ЛИТЕРАТУРЫ

1. Мамиконов А.Г., Кульба В.В., Шелков А.Б. Достоверность, защита и резервирование информации в АСУ. — М.: Энергоатомиздат, 1986. — 304 с.
2. Кузовлев В.И., Липкин Д.И. Формализованное описание процессов возникновения и распространения искажений в АСОИУ с помощью схемы потоков искажений. — М., 2001. Деп. в ВИНИТИ. № 1093-В2001. — 22 с.
3. Olson J.E. Data Quality — The accuracy dimension. — San Francisco, CA: Morgan Kaufmann Publishers, 2003.
4. Кузовлев В.И., Орлов А.О. Методы нечеткого поиска дубликатов данных в электронных хранилищах. — М., 2011.
5. Николаев Ф.А., Фомин В.И., Хохлов Л.М. Проблемы повышения достоверности в информационных системах. — Л.: Энергоиздат, 1982. — 142 с.
6. Кузовлев В.И., Липкин Д.И. Определение базовых показателей достоверности обработки информации проектных решений АСОИУ. — М., 2001. Деп. в ВИНИТИ. № 1094-В2001. — 12 с.
7. Кузовлев В.И., Орлов А.О. Учет взаимосвязей результатов профилирования. — М., 2012.

Статья поступила в редакцию 4.07.2012