электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
издатель ФГБОУ ВПО «Московский государственный технический университет им. Н.Э. Баумана»

Учет взаимосвязей между объектами результатов профилирования

77-48211/482766

08, август 2012

Кузовлев В. И., Орлов А. О.

УДК 004.052.42

Россия, МГТУ им. Н.Э. Баумана
kuzovlevvi@iu5.bmstu.ru

Процесс, ставящий своей целью задачу повышения качества информации, включает в себя несколько последовательных подпроцессов. Первый из них – это процесс анализа данных, который называется также процессом профилирования. Объектом профилирования является набор данных. Это могут быть как сущности в реляционной базе данных, так и данные в других формах. Профилирование включает в себя набор аналитических процедур, применяемых к объекту анализа. Основными процедурами профилирования являются:

1. Процедуры проверки полноты данных. Они проверяют наличие заданных значений в атрибутах сущности реляционного отношения.
2. Процедуры проверки точности данных. Эти процедуры устанавливают корректность значения того или иного атрибута на основе таких методов, как сверка значения атрибута с эталоном, поиск значения среди элементов проверочного множества, проверка соответствия типа атрибута некоторому предполагаемому типу и других.

Параллельно с основными процедурами в процессе профилирования выполняется целый ряд сопутствующих подпроцессов, формирующих в том числе следующие статистические данные (применительно к реляционным базам данных):

а) разброс длин строк атрибутов сущностей;

б) процент пустых значений атрибутов;

в) максимальные, минимальные и средние значения числовых значений атрибутов;

г) зависимости значений одних атрибутов от значений других атрибутов;

д) наборы шаблонов и процент записей, удовлетворяющих этим шаблонам.

Собранные в процессе профилирования данные позволяют вести дальнейший анализ качества информации.

Существует достаточно большое количество программ профилирования информации в базах данных. Все программы в процессе анализа делят содержимое базы данных на несколько классов S_n, где $n \geq 2$. Объектами анализа являются экземпляры отношений, то есть строки таблиц. Экземпляр r_i сущности R состоит из кортежа атрибутов A_1, A_2, … A_j. Объекты причисляются к тому или иному классу на основании некоторого условия, называемого фильтром f.

Результатом работы программы профилирования данных является множество:

$$Q = \{ S^{fi}_1 \ldots S^{fi}_n;\ n_i \geq 2; \},$$ где S^{fi}_k – k-тый класс по i-му фильтру, k=1…n.

$S^{fi}_i \cap S^{fi}_j = \varnothing$, то есть объект может принадлежать только одному классу при данном фильтре f_i.

Объектом класса S^{fi}_k является экземпляр r_j отношения R такой, что:

$$f_i(r) = \begin{cases} 1, \text{если } r \text{ удовлетворяет условиям фильтра } f \\ 0, \quad \text{в противном случае} \end{cases}$$

Фильтр задается набором условий, например: $f = \begin{cases} a1 & \neq 0 \\ a2 & \subset N, \\ a3 & \subset Z \end{cases}$

где a1, a2, a3 – это атрибуты сущности, а Z может быть любым заранее заданным множеством.

Если хотя бы одно из условий не выполняется для r_i, тогда $f(r_i) = 0$. На основании значения $f(r_i)$ объект r_i причисляется к классу S_1 или S_2.

Фильтры можно классифицировать по двум категориям:

1. Фильтр-валидатор;

2. Фильтр-шаблонизатор.

Фильтр-валидатор сортирует объекты по двум классам по условию истина/ложь.

Фильтр-шаблонизатор выделяет конечное число шаблонов, далее объекты анализируются по степени соответствия тому или иному шаблону. Количество классов соответствует количеству шаблонов, то есть для фильтров-шаблонизаторов число классов профилирования будет не менее двух.

Для фильтров-шаблонизаторов набор условий может быть задан следующим образом:

$$U_f = \begin{cases} A1 \wedge \neg A2 \vee A3 \rightarrow S1 \\ \neg A1 \rightarrow S2 \end{cases},$$

где A1= $(a1 \neq 0)$, A2= $(a2 \subset N)$, A3= $(a3 \subset Z)$.

Подобная запись ставит в соответствие классам наборы логических цепочек. Объекты в процессе анализа проверяются на удовлетворение заданным условиям и распределяются в соответствующий класс. Поскольку $S^{fi}_i \cap S^{fi}_j = \emptyset$, значит, каждый объект должен соответствовать одному и только одному классу для заданного фильтра. Отсюда следует недопустимость противоречий между цепочками правил в заданном наборе условий.

В конце процесса профилирования для каждого фильтра определяются корректные и проблемные классы. Объекты, попавшие в корректные классы, считаются объектами надлежащего качества на данном этапе работы с учетом заданных критериев фильтрации. Объекты проблемных классов анализируются и корректируются на последующих шагах процесса повышения качества данных.

На рынке продуктов программного обеспечения, занимающегося профилированием данных, представлен широкий спектр решений от компаний различного масштаба. Тем не менее, общие функциональные характеристики схожи у многих программ. Сравнение программных продуктов выявляет существенные сходства в результатах профилирования, несмотря на разную политику компаний-производителей.

Название	Datiris Profiler	DataCleaner
Разработчик	Datiris.com	Human Inference
Тип	Платная	Бесплатная (с открытым исходным кодом)

Описание	Профилирование данных. Позволяет проводить анализ данных, хранящихся в источнике данных (datasource). Источником данных является база данных, реализованная в одной из популярных СУБД или в виде листов Excel.	
Объект анализа	Сущность и ее экземпляры. Для реляционной базы данных это таблица и строки в ней.	
Возможность анализа отдельных атрибутов сущности	Да	Да

Принцип анализа в программе Datiris Profiler основан на работе с доменами. Домен – это некая область, включающая только те бизнес-объекты, которые удовлетворяют набору бизнес-правил. Под бизнес-объектом понимаются значения отдельных параметров одной записи и/или совокупность значений всех параметров записи (вся запись). В данном случае объектом считается атрибут сущности. Например: фамилия клиента. Datiris Profiler поддерживает версионность доменов. На момент запуска задачи профилирования учитывается текущая версия настроек каждого домена. При изменении домена производится сохранение его параметров под номером очередной версии. Для домена задаются бизнес-правила.

Для каждого значения определяется, удовлетворяет оно правилам домена или нет, а также анализируется удовлетворение каждому отдельному правилу домена. Правила задают ограничения для значений атрибута напрямую или в виде шаблона.

Программа DataCleaner работает следующим образом. Из источников собираются метаданные, которые будут анализироваться. Помимо основных метаданных существуют дополнительные данные: словари, каталоги синонимов, шаблоны строк.

Перед началом анализа создаются объекты условий, такими объектами являются фильтры. DataCleaner предлагает обширный набор фильтров: сравнение со словарем, проверка на идентичность, проверка на пустые значения, проверка на попадание в диапазон, проверка на длину строки и другие. Результатом работы фильтра для каждого экземпляра сущности (для каждой строки таблицы) является одно из двух значений: истина или ложь.

Помимо фильтров существует еще два типа объектов, обрабатывающих данные в процессе анализа: трансформаторы и анализаторы. Трансформаторы преобразуют исходные метаданные в соответствии с заданными условиями. Примеры трансформаторов: объединение строк, дат, чисел; преобразование форматов;

стандартизация имен, адресов; удаление пробелов и иных символов; вывод дополнительных данных. Анализаторы исследуют метаданные с целью формирования статистических дынных. Примеры анализаторов: анализ на совпадение значений со словарем; поиск шаблонов; анализ разброса значений и другие. Анализаторы и трансформаторы могут использовать созданные ранее фильтры.

После проведения операции профилирования оба продукта предоставляют множество статистических данных о характеристиках информации в источнике данных. Основные статистические характеристики отдельного атрибута:

1. Доли значений, удовлетворяющих и неудовлетворяющих правилам домена.
2. Процент уникальных значений.
3. Процент записей, значения которых не заданы.
4. Наличие записей, имеющих значения, выходящих за пределы значений атрибута (пределы задаются правилами домена, которому принадлежит атрибут).
5. Минимальные, максимальные и средние длины строк значений атрибута.
6. Количество шаблонов, обнаруженных среди значений атрибута.

На основании полученных результатов профилирования проводятся шаги, направленные на устранение выявленных проблем и повышение качества данных. Проблемой, наблюдающейся, в рассмотренных методах профилирования является отсутствие корреляции как между классами разных фильтров, так и между объектами внутри одного отдельно взятого класса.

Множество Q классов профилирования не является упорядоченным. Классы в свою очередь формируются в зависимости от результатов проверок условий, заданных в том или ином фильтре. Фильтры выбираются экспертом. Существующие программы профилирования генерируют большой набор фильтров, предлагающих различные условия профилирования. В результате множество Q можно рассматривать как набор множеств $Q_1 \ldots Q_i$, не связанных между собой. Можно показать, что процесс профилирования информации дает не один, а множество невзаимосвязанных ответов, каждый из которых требует последующего экспертного анализа.

Для решения проблемы логического связывания классов вводится оператора T_s такой, что $T_s(S^{fi}_n)>T_s(S^{fi}_{n+1})$. Данный оператор справедлив для фильтров-шаблонизаторов, когда количество классов больше двух. Смысл T_s в том, чтобы задавать относительный вес классов S и позиционировать классы соответствующим образом для дальнейшей работы. Весом класса S можно считать «цену» выполнения условий A_i. Для этого необходимо каждому условию A_i поставить в соответствие некую оценочную характеристику, определяющую степень важности данного правила. В результате A_i будет состоять из самого условия L и веса правила V_a.

$$A_i = \{L, V_a\}.$$

Аналогично оператору T_s вводится оператор T_r – оператор, упорядочивающий экземпляры r внутри отдельного класса S. $T_r(r_n)>T_r(r_{n+1})$. Это позволяет в рамках одного класса S получить упорядоченное по критерию приоритетности множество экземпляров.

Объект r_n может принадлежать только одному классу для заданного фильтра f_i, однако этот объект r_n в то же время может принадлежать классу для другого фильтра f_j. Поскольку каждый фильтр f заключает в себе оценку отдельной значимой характеристики всего множества объектов, можно утверждать, что факт попадания некоторого объекта r в проблемные классы нескольких фильтров одновременно должен фиксироваться системой, и оценка значимости данного объекта должна корректироваться с учетом этого события.

$V(r_i)$ – это значимость объекта r_i по результатам процесса профилирования. При расчете $V(r_i)$ необходимо учитывать сумму весов объекта r_i среди классов всех фильтров.

$$V(r_i) = \sum_f T_r(r_i)$$

Таким образом, результатом процесса профилирования будет:

$$Q = \{ S^{fi}_1 \dots S^{fi}_n, V_r, T_s, T_r\}, \text{ где } V_r = \{ V(r_1) \dots V(r_n) \}.$$

Процесс повышения качества информации может включать одно или несколько действий:

1. Получение или обновление информации из внешнего источника, считающегося достоверным;
2. Сбор новой информации путем проведения статистических исследований;
3. Экспертная оценка имеющихся данных с целью устранения ошибок.

Сбор новых данных или экспертный анализ объектов могут потребовать значительных временных и административных ресурсов. При этом работа с данными может продолжаться параллельно. Очевидно, что эффективность этапа профилирования прямым образом влияет на качественные и временные результаты этапа сбора и обновления информации.

Предложенный подход позволяет обеспечить такую работу, когда в первую очередь будут обрабатываться те классы и те объекты, значимость которых определены как наивысшая. Таким образом, обеспечивается интенсивность повышения качества данных в условиях параллельной эксплуатации системы.

Список литературы

1. Battini C, Scannapieca M. Data Quality. Concepts, Methodologies and Techniques.
2. Olson Jack E. Data Quality – The Accuracy dimension. 2003.
3. DataCleaner. Reference documentation [Электронный ресурс]. URL: http://datacleaner.eobjects.org/resources/docs/2.3/pdf/datacleaner-reference.pdf (дата обращения 22.08.2011)
4. Datiris Profiler Tutorial [Электронный ресурс]. URL: http://www.datiris.com/campaign/C1-1.htm (дата обращения 20.08.2011).
5. Data profiling. The foundation for data management [Электронный ресурс]. URL: http://infoimpact.com/articles/Data%20Profiling%20White%20Paper1003-final.pdf (дата обращения 11.09.2011)