УДК 004.522

Д.А. Суранова

Использование естественного языка для формирования запросов в биллинговых системах

Предложена структура программного комплекса, использующего для формирования запросов естественный язык. Определены основные структурные элементы представления данных для биллинговой системы. Приводится описание реализации.

Ключевые слова: биллинг, естественный язык, автоматизированная система.

В современных автоматизированных системах расчета главными инструментами аналитики являются пользовательские запросы и отчеты, позволяющие получить интересующую клиента информацию. В биллинговых системах, используемых для ежемесячного расчета платы за жилищно-коммунальные ресурсы, такой информацией чаще всего являются данные о начислениях, потреблениях, оплатах, счислениях и перерасчетах, полученные в виде суммы в определенном разрезе.

Как правило, для построения запросов и отчетов в биллинговых системах предусмотрен специальный графический инструментарий в виде формы задания входных параметров с последующим запуском формирования результата. В качестве примеров таких систем можно рассмотреть UserGate Billing для Windows XP, расширяющая функции биллинга UserGate – программы, контролирующей доступ пользователей в Интернет. Утилита позволяет выполнить ограниченный набор действий с балансом абонента. Другим примером является биллинговая система АСР Ideco3, используемая для учета интернет-трафика. Программа содержит набор статических отчетов, которого пользователям системы не всегда хватает[*]. Если учитывать основные пожелания предполагаемых пользователей при проектировании системы, проблема не исчезнет, так как требования могут меняться и уточняться со временем. Как следствие, возникает необходимость в получении новых инструментов для аналитики. К тому же графический интерфейс не всегда интуитивно понятен конечному пользователю.

Решением проблемы могла бы стать возможность запрашивать данные у системы в привычной для человека форме. Например, в виде фразы, построенной с использованием ключевых слов и определением интересующих параметров. Проанализировав фразу и разобрав ее на составляющие части, система сформирует запрос и выдаст результат в виде таблицы.

Построение системы предполагает решение следующих задач:

– создание единой исчерпывающей структуры для хранения данных, необходимых для выполнения аналитических запросов;

– взаимодействие с системой посредством естественного языка.

Таким образом, описаны алгоритмы и структуры данных для анализа фраз, сформированных пользователем посредством естественного языка, их последующей обработки и формирования выходных форм по результатам полученных запросов.

Описание технологии

Запросы, сформированные на основе фразы пользователя, должны иметь максимально простую структуру и высокую производительность. Для этого можно воспользоваться технологией OLAP [1] и хранить нужные данные в таблице в виде числа и набора атрибутов, определенных для такого числа. Пример представлен в табл. 1.

Таблица 1

Пример хранения данных с использованием OLAP-технологии

Тип суммы	Лицевой счет	Услуга	Сумма
Начисление	1234	Отопление	200
Начисление	2345	Вода	300
Оплата	1234	Отопление	200
Оплата	2345	Отопление	700

[*]Информация получена на основе отзывов пользователей о программе на официальном сайте http://ideco-software.ru/products/billing/index_casestudies.html

Здесь «Тип суммы», «Лицевой счет», «Услуга» являются «измерениями», или атрибутами, а «Сумма» – «фактом», или значением. Так, посмотрев на таблицу, можно заметить, что ее легко можно перевести в три измерения: по одной оси отметить «Тип суммы», по другой – «Лицевой счет», по третьей – «Услугу». «Фактами» в таком трехмерном массиве будут соответствующие «измерениям» суммы, представленные на рис. 1.

В данном примере заполнены не все значения на плоскостях, т.к., например, нет значений для начисления по отоплению для лицевого счета 2345. Таким образом, имея агрегированную проиндексированную таблицу, можно получить интересующую информацию, задав значения атрибутов. Размерность массива может быть числом от 1 до *N*, где *N* – целое положительное число.

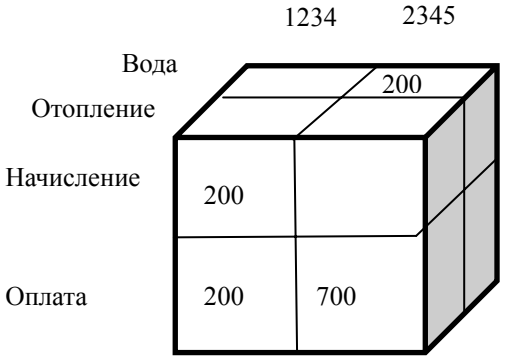

Рис. 1. Пример представления данных в виде куба

Описание реализации

Рассмотрим способ получения необходимой пользователю информации на основе работы с объемами данных по услугам лицевых счетов биллинговой системы. Потребление, начисление, оплаты, счисления, перерасчеты – все это числовые значения, определяемые для набора атрибутов. Даже если в системе они хранятся по-разному, для более эффективного построения пользовательских запросов их можно объединить в одну агрегированную таблицу с использованием технологии OLAP.

В биллинговых системах для расчета платы за жилищно-коммунальные ресурсы минимальным отрезком времени для получения отчетности является календарный месяц. Следовательно, необходимые для отчета данные можно записывать в виде суммы для каждого уникального набора атрибутов в разбивке по периодам времени, равным месяцу. Все атрибуты такой таблицы, назовем ее «BillingData», являются внешними ключами к справочникам системы. Для удобства работы справочники имеют единообразную структуру, содержащую поля «уникальный ключ», «название» и некоторый набор дополнительных полей. Так, например, таблица «Service» («Услуги») будет содержать поля «Id_Service» и «Name».

Пользователь сможет получить интересующие его данные по услугам лицевого счета, сформировав соответствующую фразу [2]. Фраза должна отвечать на вопросы о том, какие данные выбрать и каким условиям они должны удовлетворять. Если определить правила для построения такой фразы, то ее можно преобразовать в SQL-запрос к таблице «BillingData». Запрос должен содержать: ключевые слова, справочные слова и значения справочных слов. Ключевые слова должны делить фразу на смысловые части, справочные слова и их значения должны определять конкретные значения выборки и ее условия.

SQL-запрос к таблице «BillingData» будет сформирован из следующих частей:

1. Select …
2. From BillingData
3. Where …
4. Group by …

В первой части будут заданы поля, которые необходимо выбрать. Вторая часть содержит источник данных – таблицу «BillingData». Третья часть определяет условия, которым должна удовлетворять выборка. Четвертая часть задает разрез данных.

Например, фраза может быть построена так:

отчет выбрать начисление адрес город Барнаул улица Попова дом 194 месяц январь разрез лицевой и услуга.

Данная фраза содержит ключевые слова «отчет», «разрез», «выбрать». Справочными словами являются: «адрес», «город», «улица», «дом», «месяц», «лицевой и услуга». Значения справочных слов – это «Барнаул», «Попова», «194», «январь», «начисление». Проанализировав слова в той последовательности, в которой они записаны, и сверив данные с соответствующими справочниками, получаем готовую структуру, представленную в табл. 2.

Таблица 2

Структура данных после предварительного разбора

Ключевое/справочное слово	Значение справочного слова	Значение	Признак корректности
Отчет		0	1
Выбрать	Начисление	1	1
Адрес		0	1
Город	Барнаул	1	1
Улица	Попова	1	1
Дом	194	1	1
Период	Январь	201201	1
Разрез	Лицевой и услуга	4	1

Для указанного разреза данных получаем из справочника «Разрезы данных» список полей для группировки и отображения:

Таблица 3

Параметры для отображения и группировки полей

Разрез	Отображаемые поля
Id_Acct, Id_Service	(select Name from Acct A where A.Id_Acct = R. Id_Acct) as «Лицевой», (select Name from Service A where A.Id_Service = R.Id_Service) as «Услуга», SUM R.Summa) as «Сумма»

В структуру в процессе сверки со справочниками заносится идентификатор объекта в базе данных и ставится признак корректности, если значение в справочнике найдено. По ненайденным словам происходит уточнение через диалог с пользователем и заносится правильное значение.

На основании данной структуры и данных справочников «ключевые слова» и «разрез данных» динамически формируется SQL-запрос :

```
select (select Name from Acct A where A.Id_Acct = R.Id_Acct) as "Лицевой",
            (select Name from Service A where A.Id_Service = R.Id_Service)
                  as "Услуга", SUM(R.Summa) as "Сумма"
  from BillingData R
where Id_House = 1
and Id_Period = 201201
and Id_Summatype = 1
 group by Id_Acct, Id_Service
```

Далее на основе запроса происходит формирование отчетной страницы в формате PDF. Для получения возможности генерации файла средствами PHP был использован открытый и свободно распространяемый класс tFPDF [3], поддерживающий кодировку UTF-8 и кириллицу. Исходный код класса и документацию можно скачать по адресу: http://www.fpdf.org/.

Для корректного отображения таблицы используются php-функции определения количества столбцов и их параметров, таких как ширина столбца, его название и тип.

Результаты запроса выводятся на странице в виде сформированной таблицы:

Таблица 4

Результат выполненного запроса

Лицевой	Услуга	Сумма
4567	Холодная вода	200
1234	Электроэнергия	150
3456	Электроэнергия	1180

Заключение

В результате работы был создан и описан алгоритм, осуществляющий семантический разбор и последующий анализ фразы, введенной пользователем для получения информации о системе. На основании данного алгоритма разработаны программный модуль и интерфейс пользователя, позволяющий уточнять информацию и выводить результат в виде отчета. С применением технологии OLAP разработана и описана структура данных для получения информации, интересующей абонента. Использование данной технологии позволяет уменьшить время выполнения запроса и сформировать сам запрос без написания сложного кода.

Алгоритм и технологию хранения данных, разработанные и описанные в рамках статьи, можно использовать для формирования запроса через голосовой ввод данных. В дальнейшем планируется расширить возможности написания запросов, а также опробовать ввод фраз посредством голоса [4].

Литература

1. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб: БХВ-Петербург, 2007. – 384 с.

2. Батура Т.В. К вопросу об анализе текстов на естественном языке// Новые информационные технологии в науке и образовании / Т.В. Батура, О.Н. Еркаева, Ф.А. Мурзин. – Новосибирск, 2003. – С. 7–58.

3. Олищук А. Введение в FPDF [Электронный ресурс]. – Режим доступа: http://www.php.su/articles/?cat=others&page=004, свободный (дата обращения: 24.05.2010).

4. Мещеряков Р.В. Структура систем синтеза и распознавания речи // Известия ТПУ. – 2009. – Т. 315, № 5. – С. 121–126.

Суранова Дарья Александровна
Аспирант каф. теоретической кибернетики и прикладной математики
Алтайского государственного университета
Тел.: (385-2) 367-18
Эл. почта: daria@suranova.ru

Suranova D.A.

Using natural language to query the billing system

The software structure is used to generate natural language queries. The main structural elements of the submission of data to the billing system. The description of the implementation.

Keywords: billing, natural language, the automated system.
