Исследование возможностей методов Data Mining для прогнозирования персонального трафика с использованием системы WizWhy.

Кораблин М.А., Пальмов С.В.(inerrren@rambler.ru)

Поволжская Государственная Академия Телекоммуникаций и Информатики

В настоящее время телекоммуникационные биллинговые системы (CBOSS, ИБС) содержат множество функций для работы с абонентами [2, 3], но практически не используют индивидуальную информацию о клиенте. Вместе с тем, такая информация может быть использована для прогнозирования спроса на определённые услуги, разработки персональной ценовой тактики, корректирования рекламной политики и т.д. Такой подход давно используется в сфере торговли, в первую очередь электронной.

Основу современной технологии выявления взаимосвязей между персонифицирующей информацией и целевым направлением бизнеса составляют методы Data Mining [1]. Эти методы используют концепцию шаблонов, отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в наглядной форме.

Важное положение методов data mining – нетривиальность разыскиваемых шаблонов. Это означает, что такие шаблоны должны отражать неочевидные, неожиданные регулярности в данных, составляющие так называемые скрытые знания. Можно сказать, что Data Mining (интеллектуальный анализ данных) – это процесс обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний (закономерностей), необходимых для принятия решений в различных сферах человеческой деятельности.

Методы Data Mining в нашей стране не получили большого распространения, в то время как опыт многих зарубежных предприятий показывает, что отдача от использования Data Mining может достигать 1000 %. Например, известны сообщения: об экономическом эффекте, в 10 – 70 раз превысившем первоначальные затраты; о проекте в $20 млн., который окупился всего за 4 месяца; о годовой экономии $700 тыс. за счёт внедрения Data Mining в сети универсамов в Великобритании [1]. Таким образом, применение методов Data Mining может обеспечить большие экономические выгоды, в том числе и при использовании этой технологии для анализа биллинга в телекоммуникационных системах.

Чтобы проиллюстрировать возможности методов Data Mining в этой области, был разработан и проведён компьютерный эксперимент, связанный с установлением оценочных зависимостей между индивидуальными характеристиками клиента и величиной (и соответственно оплатой) его персонального трафика. В настоящее время обычной практикой телекоммуникационных компаний является формирование файлов персонального трафика клиентов. При этом клиент сам выбирает тот или иной тариф на основе собственных предпочтений, компания при этом фактически не делает индивидуального предложения клиенту по выбору тарифного плана. Целью описываемого ниже эксперимента является иллюстрация возможностей таких предложений на основе изучения и анализа личностных характеристик и предпочтений клиента. В конечном

счёте, такая стратегия способна сформировать новый баланс между спросом и предложением на рынке телекоммуникационных услуг.

В качестве личностных характеристик были выбраны: пол, образование, семейное положение (СП), количество детей (Дети), работа, доход в месяц, возраст. Эти характеристики, на наш взгляд, достаточно полно определяют «портрет» клиента в аспекте его персонального трафика.

Имитация файла индивидуальных характеристик клиента связана с агрегированием двух семантических компонент: личностных характеристик и персонального трафика.

В качестве характеристик персонального трафика были выбраны:

исходящие вызовы на мобильный телефон (день/ночь),
исходящие местные вызовы (день/ночь),
входящие вызовы с мобильного телефона,
входящие остальные (день/ночь),
общее время (трафик за месяц),
количество вызовов (количество вызовов за месяц),
средняя продолжительность разговора,
SMS (исходящее/входящее),
передача данных (день/ночь).

Такие характеристики персонального трафика являются типичными для компаний-операторов сотовой связи.

Все перечисленные характеристики по выбранным условиям эксперимента могут принимать следующие значения (или находиться в следующих диапазонах):

Личностные характеристики:

«Пол»: «1» - мужской, «0» - женский.
«Образование»: «1» - среднее, «2» - высшее, «3» - два высших, «4» - учёная степень.
«Семейное положение»: «1» - не состоит в браке, «2» - состоит в браке.
«Дети»: «0» - детей нет, «1» - один, «2» - двое, «3» - трое, «4» - более трёх.
«Работа»: «1» - не работает и не учится, «2» - учится, «3» - работает и учится, «4» - работает.
«Доход в месяц»: «1» - низкий, «2» - средний, «3» - высокий.
«Возраст»: «1» - от 20 до 30 лет, «2» - от 30 до 40 лет, «3» - от 40 до 50 лет, «4» - от 50 до 60 лет, «5» - больше 60 лет.

Персональный трафик:

а) Исходящие вызовы:

«Исходящие вызовы на мобильный телефон (день)»: от 1 до 350 мин.
«Исходящие вызовы на мобильный телефон (ночь)»: от 1 до 350 мин.
«Исходящие местные вызовы (день)»: от 1 до 350 мин.
«Исходящие местные вызовы (ночь)»: от 1 до 350 мин.

б) Входящие вызовы:

«Входящие вызовы с мобильного телефона»: от 1 до 360 мин.

«Входящие остальные (день)»: от 1 до 720 мин.
«Входящие остальные (ночь)»: от 1 до 720 мин.

в) SMS:

«Исходящее SMS»: от 1 до 140 сообщений.
«Входящее SMS»: от 1 до 140 сообщений.

г) Передача данных:

«Передача данных (9600 бит/с.) (день)»: от 1 до 100 мин.
«Передача данных (9600 бит/с.) (ночь)»: от 1 до 150 мин.

д) Прочее:

«Общее время»: сумма значений атрибутов «Исходящие вызовы на мобильный телефон (день)», «Исходящие вызовы на мобильный телефон (ночь)», «Исходящие местные вызовы (день)», «Исходящие местные вызовы (ночь)», «Входящие вызовы с мобильного телефона», «Входящие остальные (день)», «Входящие остальные (ночь)».
«Количество вызовов»: от 100 до 500 вызовов в месяц.
«Средняя продолжительность разговора»: результат деления значения атрибута «Общее время» на «Количество вызовов».

С учётом персонального трафика (кроме атрибутов группы «Прочее») и характеристик тарифов рассчитываются атрибуты «Оплата по тарифу «Дебют», «Оплата по тарифу «Фаворит 100», «Оплата по тарифу «Приём».

Между личностными характеристиками и персональным трафиком существуют скрытые зависимости, которые имеют нечёткий характер и, поэтому, трудно формализуемы. Поэтому, для имитации таких зависимостей мы использовали субъективные причинно-следственные связи, которые представляются вполне правдоподобными. Механизм имитации строился на основе использования набора эвристических правил, связывающих личностные характеристики клиента с атрибутами персонального трафика. Такого рода правила широко используются в эвристическом программировании. Ниже приведены три примера правил установления связей (всего же их в модели было введено 25):

- ***Если*** («Образование» = «4») **и** («Работа» = «4») **и** («Доход в месяц» = «3») ***то*** (диапазон изменения значений атрибута «Общее время» лежит в интервале 1500… 2100 мин).
- ***Если*** («Пол» = «0») **и** («Семейное положение» = «1») **и** («Дети» = «1») ***то*** (диапазон изменения значений всех атрибутов в группе «Исходящие вызовы» лежит в интервале 50… 100 мин).
- ***Если*** («Пол» = «1») **и** («Семейное положение» = «2») **и** («Дети» = «0») ***то*** (диапазон изменения значений всех атрибутов в группе «Входящие вызовы» лежит в интервале 50… 80 мин) и т.д.

Массивы личностных характеристик клиентов и персонального трафика формируются программой (реализована в среде Delphi 5), содержащей генераторы целых случайных чисел. Каждая запись файла индивидуальных характеристик клиента состоит из трёх компонент: личностных характеристик, персонального трафика и тарифов.

В качестве тарифов были выбраны следующие:

«Дебют»:

Ежемесячная абонентская плата – 83,58 руб.
Ежемесячная минимальная оплата за трафик – 0 руб.
Стоимость одной минуты:
Исходящие вызовы на мобильный телефон: 4,42 руб. (день);
2,45 руб. (ночь).
Исходящие местные вызовы: 4,42 руб. (день); 2,45 руб. (ночь).
Входящие вызовы с мобильного телефона: бесплатно.
Входящие остальные: 4,42 руб. (день); 2,45 руб. (ночь).
Исходящее SMS: 0,98 руб. (за сообщение)
Входящее SMS: бесплатно.
Передача данных (9600 бит/с.) (день): 1,47 руб.
Передача данных (9600 бит/с.) (ночь): 0,49 руб.

«Фаворит 100»:

Ежемесячная абонентская плата (городской номер) – 98,33 руб.
Ежемесячная минимальная оплата за трафик (включает 100 минут местных вызовов) – 442,50 руб.
Стоимость одной минуты:
Исходящие вызовы на мобильный телефон: 3,93 руб. (день);
2,45 руб. (ночь).
Исходящие местные вызовы: 3,93 руб. (день); 2,45 руб. (ночь).
Входящие вызовы с мобильного телефона: бесплатно.
Входящие остальные: 3,93 руб. (день); 2,45 руб. (ночь).
Исходящее SMS: 0,98 руб. (за сообщение)
Входящее SMS: бесплатно.
Передача данных (9600 бит/с.) (день): 1,47 руб.
Передача данных (9600 бит/с.) (ночь): 0,49 руб.

«Приём»:

Ежемесячная абонентская плата (городской номер) – 0 руб.
Ежемесячная минимальная оплата за трафик (включает 1000 минут местных вызовов) – 800 руб.
Стоимость одной минуты:
Исходящие вызовы на мобильный телефон: 4 руб. (день);
4 руб. (ночь).
Исходящие местные вызовы: 4 руб. (день); 4 руб. (ночь).
Входящие вызовы с мобильного телефона: бесплатно.
Входящие остальные: 1 руб. (день); 1 руб. (ночь).
Исходящее SMS: 0,98 руб. (за сообщение)
Входящее SMS: бесплатно.
Передача данных (9600 бит/с.) (день): 1,47 руб.

Передача данных (9600 бит/с.) (ночь): 0,49 руб.

Значения тарифной части записи о клиенте рассчитываются на основе смоделированных атрибутов персонального трафика. Таким образом, общая структура записи файла иллюстрируется таблицей 1.

Таблица 1

Значение атрибута	Название атрибута	
0	Пол	Личностные характеристики клиента
1	Образование	
2	СП	
1	Дети	
2	Работа	
3	Доход в месяц	
2	Возраст	
80	Исходящие вызовы на мобильный телефон (день), мин.	Персональный трафик
8	Исходящие вызовы на мобильный телефон (ночь), мин.	
55	Исходящие местные вызовы (день), мин.	
69	Исходящие местные вызовы (ночь), мин.	
104	Входящие вызовы с мобильного телефона, мин.	
154	Входящие остальные (день), мин.	
278	Входящие остальные (ночь), мин.	
748	Общее время, мин	
164	Количество вызовов, шт.	
5	Средняя продолжительность разговора, мин.	
53	Исходящее SMS	
1	Входящее SMS	
38	Передача данных (9600 бит/с.) (день), мин.	
64	Передача данных (9600 бит/с.) (ночь), мин.	
2369	Оплата по тарифу «Дебют», руб.	Оплата по тарифам
2358	Оплата по тарифу «Фаворит 100», руб.	
1787	Оплата по тарифу «Приём», руб.	

Второй этап описываемого эксперимента связан с анализом данных методами Data Mining. Для проведения этого этапа использовалась система WizWhy 3.08 компании WizSoft.

Программный продукт WizWhy является наиболее ярким современным представителем систем, которые используют алгоритмы ограниченного перебора (хотя авторы не раскрывают специфику алгоритма; вывод о наличие ограниченного перебора был сделан по результатам тщательного тестирования системы). Они вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: $X < a$, $X = b$, $X > c$ и т.д., где X – какой-либо параметр, a, b и c - константы. На основе анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации прогнозирования и т.п. WizWhy является на сегодняшний день одним из лидеров на рынке продуктов Data Mining [1].

Целью эксперимента являлась оценка эффективности обнаружения скрытых закономерностей в файле личностных характеристик.

Каждая из введенных нами закономерностей (25 правил) обнаруживается системой с определением характеристик качества такого обнаружения.

Любое правило, обнаруживаемое системой WizWhy, представляется в виде условного суждения ЕСЛИ (А) ТО (В) и имеет две основные характеристики – точность и полноту. Точность правила – это доля случаев, когда правило подтверждается, среди всех случаев его применения (доля случаев В среди случаев А). Полнота правила – это доля случаев, когда правило подтверждается, среди всех случаев, когда имеет место объясняемый исход В (доля случаев А среди случаев В) [1]. Значение 100% показывает, что системе задаются требования обнаруживать правила, которые не будут давать ошибок.

В итоге, система обнаружила 187 правил. Они выводятся в отчёте о правилах (rule report).

Искомое правило находится в нём под номером «6»:

6) *If* **Образование** *is* **4**
and **Работа** *is* **4**
and **Доход в месяц** *is* **3**
Then
Общее время *is* ***between*** **1 500,00** ***and*** **2 000,00**
Rule's probability: **1,000**
The rule exists in **17** *records.*
Significance Level: Error probability is almost 0
Positive Examples (records' serial numbers):
41, 50, 80, 108, 144, 151, 177, 259, 376, 419

Параметр *Rule's probability* означает точность правила; *the rule exists in…records* – число записей, на которые распространяется правило; *Significance Level: Error probability…* - статистическая оценка уровня значимости правила (в данном случае для всех правил доверие практически равно 100%); *Positive Examples (records' serial numbers)* – «положительный примеры», которые затем представлены как номера записей в наборе данных [1].

При поиске остальных 24 известных закономерностей система WizWhy работала аналогичным образом и нашла их все.

Таким образом, можно сказать, что система WizWhy полностью справилась с поставленной задачей (нашла 25 из 25 известных правил). Обнаруженные правила обладают высокой точностью (*Rule's probability:* **1,000)** и большим уровнем значимости (*Significance Level: Error probability is almost 0*). Исходя из этого, можно сделать вывод, что система WizWhy пригодна для использования в сфере телекоммуникационных услуг. Она может быть использована для решения задач изучения клиентов и предоставления им индивидуальных видов обслуживания, например:

- Построение профиля абонента компании, выявления характерных черт для выработки целенаправленноё маркетинговой политики;
- Прогноз «доходности» и характеристик нового абонента сети при заключении договора на основе нахождения аналогичных случаев в прошлом;
- Повышение эффективности рекламных компаний путём выделения целевой аудитории;
- Персонализация предоставления цифровых информационных услуг абонентам оператора (sms информационные сообщения, mobile Internet контент);
- On-line рекомендации по приобретению совместно покупаемых услуг на терминале оператора Интернет-сайте индивидуально каждому клиенту;
- Анализ совместно приобретаемых услуг и др. [4].

Список литературы:

1. В. Дюк, А. Самойленко. Data Mining: учебный курс. – С.- Пб, Питер, 2001.
2. Г. Большова. Биллинг в телекоммуникациях // Сети. 1998. №5.
3. И. Елисеев. Биллинг в бизнесе телекоммуникаций // Computerworld. 2000. № 41.
www.megaputer.ru. Data Mining в телекоммуникациях. Москва, 2001.