

(12) МЕЖДУНАРОДНАЯ ЗАЯВКА, ОПУБЛИКОВАННАЯ В СООТВЕТСТВИИ С
ДОГОВОРом О ПАТЕНТНОЙ КООПЕРАЦИИ (РСТ)

(19) Всемирная Организация
Интеллектуальной Собственности
Международное бюро



WIPO | РСТ



(10) Номер международной публикации
WO 2013/055257 A1

(43) Дата международной публикации
18 апреля 2013 (18.04.2013)

(51) Международная патентная классификация:
G06Q 10/04 (2012.01)

(21) Номер международной заявки: РСТ/RU2012/000787

(22) Дата международной подачи:
27 сентября 2012 (27.09.2012)

(25) Язык подачи: Русский

(26) Язык публикации: Русский

(30) Данные о приоритете:
2011141635 14 октября 2011 (14.10.2011) RU

(71) Заявитель (для всех указанных государств, кроме US):
**ОБЩЕСТВО С ОГРАНИЧЕННОЙ
ОТВЕТСТВЕННОСТЬЮ "ЛАЙФСТАЙЛ
МАРКЕТИНГ"** (OBSHESTVO S OGRANICHEN-
NOY OTVETSTVENNOSTYU "LAIFSTAIL MAR-
KETING") [RU/RU]; ул. Смольная, 20 А, Москва,
125445, Moscow (RU).

(72) Изобретатель; и

(71) Заявитель (только для US): **ДРОБЫШЕВ, Максим
Андреевич** (DROBYSHEV, Maksim Andreevich)
[RU/RU]; Угловой переулок, 12 Московская область,
Клин, 141600, Klin (RU).

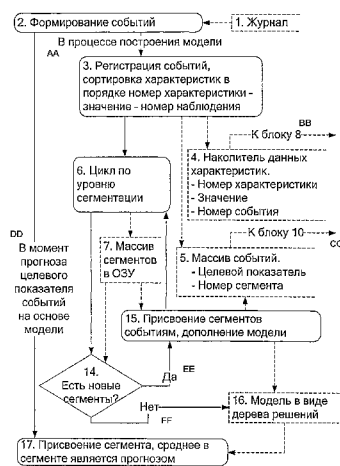
(74) Агент: **ОБЩЕСТВО С ОГРАНИЧЕННОЙ
ОТВЕТСТВЕННОСТЬЮ ПАТЕНТНО-ПРАВОВАЯ
ФИРМА "ЮС"** (PATENT & LAW FIRM "YUS",
LIMITED LIABILITY COMPANY); Проспект Мира,
д. 6, Москва, 129090, Moscow (RU).

(81) Указанные государства (если не указано иначе, для
каждого вида национальной охраны): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP,
KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD,
ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,

[продолжение на следующей странице]

(54) Title: METHOD FOR PREDICTING A TARGET FOR EVENTS ON THE BASIS OF AN UNLIMITED NUMBER OF CHARACTERISTICS

(54) Название изобретения : СПОСОБ ПРОГНОЗА ЦЕЛЕВОГО ПОКАЗАТЕЛЯ СОБЫТИЙ ПО НЕОГРАНИЧЕННОМУ КОЛИЧЕСТВУ ХАРАКТЕРИСТИК



Фиг. 1

- 1 Log
- 2 Events generated
- 3 Events recorded, characteristics sorted by order of number of a characteristic, value and observation number
- 4 Storage medium for data about characteristics.
 - Characteristic number
 - Value
 - Event number
- 5 Set of events.
- 6 Cycle according to segmentation level
- 7 Set of segments in the working memory
- 14 New segments?
- 15 Segments assigned to events, model updated
- 16 Model in the form of a decision tree
- 17 Segment assigned, segment average represents prediction
- AA During construction of a model
- BB To block 8
- CC To block 10
- DD At time of prediction of an event target on the basis of the model
- EE Yes
- FF No

(57) Abstract: The invention relates to a method for predicting a target for events. In the claimed method, event data is obtained using a computer and showing a target for each event and an arbitrary set of numerical characteristics of the event, segmentation is performed, a predictive model is created, and the target of a subsequent similar event is predicted, wherein the events are numbered by the computer, the continuous numbering of the characteristics of the events is carried out, the values of the characteristics are sorted for events in the order of the number of an event characteristic and the value of said event characteristic and are recorded in this order in a data storage medium of the computer with the numbers of the events being indicated, after which, in the segmentation cycle, sets of events are divided into subsets, i.e. next level segments, corresponding to the nodes of a decision tree, during which process the characteristics, the values thereof and the numbers of the events are consecutively read from the data storage medium. The value of an optimum characteristic is calculated for each segment, and the segment to which a subsequent similar event relates is determined, the average value of the target of that segment serving as a prediction for the subsequent similar event.

(57) Реферат:

[продолжение на следующей странице]



RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(84) Указанные государства (если не указано иначе, для каждого вида региональной охраны): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), евразийский (AM, AZ, BY, KG, KZ, RU, TJ, TM), европейский патент (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,

Опубликована:

- с отчётом о международном поиске (статья 21.3)
- до истечения срока для изменения формулы изобретения и с повторной публикацией в случае получения изменений (правило 48.2(h))

Изобретение относится к способу прогноза целевого показателя событий. В заявленном способе получают компьютером данные о событиях с указанием для каждого события целевого показателя и произвольного набора его числовых характеристик, производят сегментацию, создают прогнозную модель, и прогнозируют целевой показатель последующего аналогичного события, при этом посредством компьютера нумеруют события, проводят сквозную нумерацию характеристик событий, сортируют значения характеристик для событий в порядке номер характеристики события - значение характеристики события, запоминают их в этом порядке в накопителе данных компьютера с указанием номеров событий, после чего в цикле сегментации выполняют последовательное деление множества событий на подмножества - сегменты следующего уровня, соответствующие узлам дерева решений, в ходе которого выполняют последовательное чтение характеристик, их значений и номеров событий из накопителя данных. Для каждого сегмента вычисляют значение наилучшей характеристики и определяют сегмент, к которому относится последующее аналогичное событие, среднее значение целевого показателя такого сегмента служит прогнозом для последующего аналогичного события.

СПОСОБ ПРОГНОЗА ЦЕЛЕВОГО ПОКАЗАТЕЛЯ СОБЫТИЙ ПО НЕОГРАНИЧЕННОМУ КОЛИЧЕСТВУ ХАРАКТЕРИСТИК

Область техники

Изобретение относится к вычислительной технике и может быть использовано для автоматизированного проведения адресных кампаний в отношении покупателей, прогноза снижения и увеличения количества потребителей продукции, прогноза результатов реализации событий для различных объектов и направлений хозяйственной деятельности.

Предшествующий уровень техники

Известны прогностические методы на основе нейросетей, карт Кохонена, деревьев решений, общим для которых является получение на входе процесса моделирования «вектора признаков» фиксированной размерности после этапа «подготовки данных» (этап Data Preparation, см. Cross Industry Standard Process for Data Mining – CRISP-DM).

Однако реальные данные, накапливаемые компаниями, содержат историю событий исследуемых объектов с переменным и часто большим количеством событий по каждому объекту, и только с большими потерями информации могут быть преобразованы в вектор. При этом каждое событие может содержать целый ряд характеристик, потенциально полезных для прогноза.

Например, информация о покупке товара по дисконтной карте одновременно содержит сумму покупки, точную дату, название и категорию товара, сумму скидки, название и адрес магазина, название отдела – всего до 100 и более полей. При этом значимыми для прогноза поведения покупателей в отношении конкретных акций могут оказаться совершенно различные характеристики, например:

- Сумма покупок со словом 'ШАМПАНСКОЕ' в 'НАЗВАНИЕ ТОВАРА' УТРОМ
- Сумма покупок со словом 'МАРТИНИ' ПО ПЯТНИЦАМ,
- Количество покупок со словом 'ПАМПЕРС' ДНЕМ,
- Количество слов 'ВЫСШЕЕ' В АНКЕТЕ,
- Количество посещений web-страницы 'ШУРУПЫ И САМОРЕЗЫ', и т.д.

Таким образом, как значения текстовых полей, так и слова в них, в том числе – в комбинации с числовыми полями – могут быть использованы для формирования отдельных потенциально полезных характеристик. При ассортименте торговых сетей

в десятки и сотни тысяч наименований, количество потенциально полезных характеристик покупателей часто превышает миллион.

Объем данных всех возможных характеристик событий миллионов покупателей превышает размер доступной оперативной памяти современных компьютеров, что вынуждает разработчиков информационных систем «уменьшать размерность» анализа до нескольких десятков тысяч переменных, то есть экспертным методом отбрасывать потенциально ценную информацию.

Известен также алгоритм ассоциативных правил, использующийся для поиска часто встречающихся комбинаций товаров в «корзинах» покупок (Association Rule and Quantitative Association Rule Mining among Infrequent Items MDM'07, Ling Zhou, Stephen Yau, 2007, San Jose, California, USA). Но этот алгоритм не позволяет использовать наряду с основным идентификатором (например, название товара) другую дополнительную информацию (дата и время покупки, адрес магазина, группа, подгруппа, категория и подкатегория товара).

Известен способ (US, 6839682) предсказания финансового поведения потребителей (например, покупателей), включающий получение ряда входных транзакций для множества потребителей относительно множества торговцев; определение, по крайней мере, одного торгового сегмента, в котором каждый торговец объединяется по крайней мере с одним из определенных торговых сегментов; и по крайней мере для одного потребителя, применяя входные транзакции потребителя компьютером к каждой по крайней мере одной торговой прогнозирующей модели сегмента, каждому торговому определению прогнозирующей модели сегмента для торгового сегмента прогноз функционирует между входными транзакциями в прошлом временном интервале и финансовым поведением в последующем временном интервале, чтобы произвести для каждого потребителя предсказанное поведение в каждом, по крайней мере, подмножестве торговых сегментов.

Этот известный способ реализует автоматизированный прогноз целевого показателя (ЦП) событий, включающий получение компьютером данных о событии с указанием по каждому событию целевого показателя и набора характеристик этого

события, сегментацию, построение прогнозной модели, и прогноз целевого показателя для последующих аналогичных событий.

В данном способе событием является покупка, целевым показателем – сумма продаж в торговом сегменте, характеристикой – вектор сумм продаж покупателя в
5 других торговых сегментах, при этом все транзакционные данные о покупателе предварительно преобразуются в вектор сумм выручки по «сегментам торговли», после чего проводится компьютерный анализ только по этим упрощенным данным.

Таким образом, в этом известном способе не учитывается и не используется для прогноза информация о времени покупок, адресах магазинов, торговых сетях
10 внутри «сегмента», анкетных данных покупателей и каких-либо других доступных данных.

Ограничением известного способа являются недостаточно высокая точность и достоверность прогноза целевых показателей (ЦП) событий, а также длительное время и сложная процедура при обработке больших объемов данных о событиях.

15 Раскрытие изобретения

Решаемая изобретением задача – улучшение технико-эксплуатационных характеристик и расширение функциональных возможностей прогнозирования при упрощении его процедуры.

Техническим результатом является повышение точности и достоверности
20 прогноза целевых показателей событий, а также сокращение времени и упрощение процедуры при больших объемах данных наблюдений о различных событиях за счет оптимизации распределения данных о событиях по типам памяти компьютера и алгоритмов их обработки, что дает возможность полного анализа всех необходимых характеристик событий.

25 Современные процессоры компьютеров позволяют выполнять миллиарды операций в секунду, но их производительность ограничивается быстродействием устройств памяти и накопителей данных.

По этой причине в основу изобретения положен полный анализ всех возможных характеристик о событиях, что становится возможным за счет
30 оптимизации распределения данных о событиях по типам памяти компьютера и алгоритмов их обработки.

Построение модели проводится путем последовательного построения уровней дерева решений путем разделения подмножества событий, отнесенных к узлу дерева (сегмента) на каждом уровне на сегменты следующего уровня.

Такой подход позволяет при промышленном применении использовать только
5 основные начальные уровни дерева решений из построенной модели, дающее максимальный экономический эффект с учетом расходов на их хранение и использование для сотен и тысяч промо-акций, в том числе – на точках продаж в устройствах с ограниченным объемом памяти.

10 Для решения поставленной задачи с достижением указанного технического результата в способе получают компьютером данные о событиях с указанием для каждого события целевого показателя и произвольного набора его числовых характеристик, посредством компьютера нумеруют события, проводят сквозную нумерацию характеристик событий, сортируют значения характеристик для событий
15 в порядке номер характеристики события – значение характеристики события, запоминают их в этом порядке в накопителе данных компьютера с указанием номеров событий, после чего в цикле сегментации выполняют последовательное деление множества событий на подмножества – сегменты следующего уровня, соответствующие узлам дерева решений, в ходе которого выполняют
20 последовательное чтение характеристик, их значений и номеров событий из накопителя данных, и для каждого сегмента вычисляют агрегированные данные на основе целевых показателей и количества из массива событий из оперативной памяти компьютера, и при отличии следующей характеристики события от предыдущей или ее значения от предыдущего рассчитывают экономический эффект деления на
25 сегменты между предыдущим и следующим значениями предыдущей характеристики, запоминают наилучшую характеристику события и значение, имеющие максимальный экономический эффект деления на сегменты, после чтения последнего значения характеристики из накопителя данных проводят деление сегментов, в которых экономический эффект деления на сегменты следующего
30 уровня положителен, на сегменты следующего уровня по наилучшей характеристике и ее значению, если обнаружены новые сегменты, то повторяют цикл сегментации, если новые сегменты не обнаружены, то для прогноза целевого показателя последующего аналогичного события последовательно сравнивают значения

характеристик аналогичного события со значениями наилучших характеристик, по которым было проведено деление сегмента на сегменты следующего уровня, и по значению наилучшей характеристики определяют сегмент, к которому относится последующее аналогичное событие, среднее значение целевого показателя такого сегмента служит прогнозом для последующего аналогичного события.

Возможен дополнительный вариант способа, согласно которому при вычислении экономического эффекта деления на сегменты, экономический эффект уменьшают на величину доверительного интервала экономического эффекта, для чего вычисляют дополнительные агрегированные данные, в частности, сумму квадратов целевых показателей событий. В качестве критерия наилучшего деления на сегменты в этом случае используется максимальный пессимистический экономический эффект. Экономическим эффектом может быть, например, экономия в связи с отказом от проведения промо-акции в нецелевом (убыточном по среднему целевому показателю) сегменте. Пессимистический экономический эффект корректируется в меньшую сторону на величину оценки статистической погрешности экономического эффекта путем расчета доверительного интервала для заданной пользователем требуемой доверительной вероятности.

Возможны также варианты способа, когда деление сегментов выполняют не более заданного количества уровней и/или не менее заданного количества событий в каждом сегменте следующего уровня. Ограничение на количество событий в сегменте может устанавливаться исходя из компромисса между требуемой достоверностью сегментов и имеющимся объемом данных, которых может оказаться недостаточно для удовлетворения этих требований.

Возможен также вариант способа, когда деление сегмента производится не более чем на два сегмента следующего уровня.

Указанные преимущества, а также особенности настоящего изобретения поясняются лучшим вариантом его выполнения со ссылками на прилагаемые фигуры.

Перечень чертежей

Фиг. 1 изображает диаграмму формирования и обработки данных для построения модели и прогноза целевого показателя заявленным способом;

Фиг. 2 – цикл по уровню сегментации на фиг. 1;

Фиг. 3 – экран формирования событий (наблюдений) на основе истории покупок покупателя для заявленного способа;

Фиг. 4 – экран результата построения модели в виде дерева решений для проведения промо-акции;

5 Фиг. 5 – то же, что фиг. 4, для персональных рекомендаций фильма «Pretty Woman» на данных компании NetFlix.

Лучший вариант реализации изобретения

Заявленный способ может быть реализован посредством следующего алгоритма (фиг. 1 и фиг. 2) работы вычислительного устройства, в процессе которого строится бинарное дерево решений, то есть каждый сегмент делится не более чем на два сегмента следующего уровня. На фиг. 1 и 2 штрихом выделены блоки данных и путь их передачи, сплошными линиями показаны блоки обработки данных, блоки управления и путь. Блоки, выделенные на схеме жирными линиями (фиг. 2), характеризуют основной цикл, обеспечивающий максимальную производительность.

15 На основе журнала 1 событий, полученных компьютером, (фиг. 1) блок 2 формирует данные о событиях, состоящие для каждого события из числового целевого показателя (ЦП) и произвольного набора числовых характеристик с текстовыми названиями. События формируются в соответствии с необходимой для конкретной задачи бизнес-логикой, например, экономикой промо-акции, почтовой или СМС рассылки, раздачи купонов или персональных предложений. Здесь и далее под понятием «число» понимаются вещественные числа, которые как частный случай могли быть получены в результате присвоения сравнимым нечисловым характеристикам следующих друг за другом числовых значений, например, датам, уровням образования или владения иностранным языком из анкеты покупателя и
25 прочее.

В процессе построения модели при поступлении в обработку каждого события оно нумеруется. Каждая вновь обнаруженная характеристика (нумерация сквозная по всем событиям) также последовательно нумеруется блоком 3, в результате чего событие может выглядеть, например, так:

Событие № 928712	Целевой показатель = -1023.3
Характеристика № 1000200	= 56.4

7

Характеристика № 56223222	= 20.1
...	
Характеристика № 23001234	= -23112.8

После этого значение каждой характеристики для каждого события сортируется и запоминается в накопителе 4 данных в порядке «Номер характеристики – значение характеристики – номер события», например:

№ Характеристики	Значение	Номер события
1	112334.1	899772
1	111222.2	100223
1	110445.0	223344
1	99223.9	887778
1	12899.0	22333
2	56.3	822109
2	-0.1	112223
2	-12.9	122
...		
56223222	23.6	766553
56223222	22.6	122233
56223222	22.5	666554
56223222	20.1	928712
56223223	190.6	956712

- 5 Жирным шрифтом выделено значение характеристики из примера события для блока 3 (сортировка по значению проведена в порядке убывания).

После этого в оперативном запоминающем устройстве (ОЗУ) компьютера создают блок 5 массива событий, индексированного по номеру события с указанием целевого показателя и номера сегмента.

- 10 Далее в цикле сегментации блока 6 подсчитывают суммарные значения агрегированных данных сегментов блока 7 (фиг. 2). Агрегированными данными может быть общее количество событий, сумма целевых показателей, сумм квадратов целевых показателей и другие, выбранные разработчиком.

Пример массива сегментов и агрегированных данных о них (блок 7):

Поле \ Сегмент №	1	2	...
ID Сегмента	R0110	R00	
К-во событий, всего	65422	875092	
К-во событий, текущее	56	78921	
Сумма ЦП всего	34982.67	187430.0	
Сумма ЦП текущая	-12932.1	45638.0	
Сумма квадратов ЦП всего (двойная точность)	425166667	566033498.0	
Сумма квадратов ЦП текущая (двойная точность)	5368123.01	45984023.8	
Лучший пессим. Результат	2941.45	98.88	
Лучшая характеристика, №	5	87469	
Лучшее значение характеристики	-56.12	Отсутствует	
Обновлен	Да	Нет	

Для этого в блоке 8 (фиг. 2) в цикле до конца данных из накопителя блока 4 производится буферизованное чтение следующей записи значений характеристик. Каждая запись значения характеристики проверяется на совпадение номера характеристики и значения характеристики от предыдущей записи в блоке 9 принятия решения. Если они не отличаются, то целевой показатель события, полученный из массива блока 5, сразу используется в блоке 10 для накопления агрегированных данных сегмента блока 7, к которому принадлежит событие. Например, к текущему значению количества событий добавляется единица, к сумме целевых показателей (ЦП) сегмента добавляется значение целевого показателя события, к сумме квадратов ЦП добавляется квадрат текущего ЦП и т.д. При этом сегмент помечается как обновленный установлением флага Обновлен = «Да» в блоке 7.

Блоки 8, 9 и 10 в процессе построения модели используются чаще всех других при выполнении операций, поскольку данные отсортированы по номерам и значениям характеристик, и высока вероятность, что соседние записи в накопителе блока 4 имеют одинаковые значения характеристик (например, тысяча покупателей купили по одной единице товара и соответствующая характеристика у всех таких событий будут одинакова и равна единице). Для этих операций требуется только сравнение данных в ОЗУ компьютера и выполнение арифметических операций его процессором для обновления агрегированных данных, что позволяет увеличить

производительность и скорость построения модели, несмотря на то что объем данных характеристик для реальных объемов событий, как правило, не помещается в оперативной памяти.

В качестве критерия наилучшего деления на сегменты используется

5 Пессимистический экономический эффект.

Экономическим эффектом может быть, например, экономия в результате отказа от проведения промо-акции в нецелевом (убыточном по среднему целевому показателю) сегменте, в этом случае экономический эффект будет равен сумме целевых показателей в убыточном сегменте со знаком минус. Другой пример
10 экономического эффекта – сумма превышения ЦП в дочернем сегменте по сравнению со средним по родительскому сегменту.

Могут быть использованы и другие показатели оптимального деления сегмента, не являющиеся строго экономическими, но в любом случае для промышленной применимости они так или иначе должны в итоге приводить к расчету
15 экономического эффекта, поэтому мы их рассматриваем как частный случай экономических критериев.

Для получения Пессимистического экономического эффекта, Экономический эффект корректируется в меньшую сторону на величину оценки статистической погрешности Экономического эффекта путем расчета доверительного интервала для
20 заданной пользователем доверительной вероятности.

Таким образом, [Результат деления на сегменты] =
= [Пессимистический экономический эффект] =
= [Экономический эффект] – [Доверительный интервал Экономического эффекта]

[Доверительный интервал Экономического эффекта] может вычисляться,
25 исходя из стандартного отклонения целевого показателя и количества событий в сегменте с учетом коэффициентов Стьюдента, или другими статистическими методами с использованием необходимых для них и накапливаемых в блоке 7 агрегированных данных, в зависимости от задачи, поставленной разработчику алгоритма.

30 Деление на сегменты производится, если пессимистический экономический эффект от него больше нуля.

Если номер или значение характеристики отличается (а также после считывания последней записи из накопителя 4), то, прежде чем обновить данные сегмента, в цикле блока 11 для всех обновленных сегментов в блоке 12 проверяется, улучшен ли Результат возможного деления данного сегмента на сегменты
5 следующего уровня по значению между отличающимися значениями одной характеристики, сравнение которых проводилось в блоке 9, или, если в блоке 9 считана запись с отличающейся характеристикой, то по значению «Присутствует-Отсутствует» предыдущей характеристики.

Например, если предыдущая и следующая запись в накопителе 4 относятся к
10 одной характеристике:

Номер характеристики	Значение	Номер наблюдения
94512	8,456	305612
94512	7,232	2309145

- то деление должно проводиться по значению характеристики № 94512, между 8,456 и 7,232.

Поскольку между предыдущим и следующим значением характеристики при используемой сортировке значений быть не может, то более точная граница между
15 сегментами следующего уровня не может быть определена, и можно ее устанавливать в этом интервале. Например, точная граница может определяться как среднее с округлением до минимального количества значащих цифр, которые сохраняют ее в границах интервала, в данном случае, 8,0.

Для варианта расчета экономического результата с отбросом сегментов с
20 отрицательным средним значением Целевого показателя, Экономический эффект от деления сегмента на два сегмента следующего уровня по значению характеристики 94512, равному 8,0, будет равен 12932.1 при следующих агрегированных значениях в массиве сегментов блока 7:

К-во событий, всего	65422
К-во событий, текущее	56
Сумма ЦП всего	34982.67
Сумма ЦП текущая	-12932.1

Если результат возможного деления сегмента (для простоты в данном примере рассчитывается без доверительного интервала) при этом улучшен, то в данных этого сегмента блока 7 могут быть установлены следующие значения:

Поле	Значение
Лучший пессим. результат	12932.1
Лучшая характеристика, №	94512
Лучшее значение характеристики	8,0
Обновлен	Да

5 Если же предыдущая и следующая запись в накопителе 4 относятся к разным характеристикам:

Номер характеристики	Значение	Номер наблюдения
94512	8,456	305612
94513	96205467,223	2309145

- то деление должно проводиться по значению «Присутствует» - «Отсутствует» характеристики № 94512, и, если результат возможного деления сегмента при этом также улучшен до 12932.1, то в данных этого сегмента блока 7 должны быть установлены следующие значения:

Поле	Значение
Лучший пессим. результат	12932,1
Лучшая характеристика, №	94512
Лучшее значение характеристики	Присутствует
Обновлен	Да

10

В случае улучшения Экономического результата возможного деления сегмента в блоке 13 для сегмента устанавливаются найденные новые значение результата, лучшей характеристики и ее лучшего значения.

15 Вычисления блока 12 требуют большего количества операций процессора компьютера, чем цикл блоков 8, 9 и 10, но выполняются реже, что позволяет не обращаться в процессе вычислений к накопителю данных и обеспечивает скорость вычислений, соизмеримую со скоростью последовательного буферизированного чтения данных с жесткого диска, то есть близкую к максимально возможной.

После каждого цикла по уровню сегментации блока 6 (фиг. 1) в блоке 14 принятия решений проверяется, выявлены ли на этом уровне новые сегменты, если они обнаружены, то новые сегменты в блоке 15 добавляются к блоку 16 прогнозной модели в виде бинарного дерева решений.

5 После этого проводится деление сегментов, в которых выявлены сегменты следующего уровня, то есть присвоение событиям этих сегментов в блоке 5 новых сегментов, для чего считываются данные из накопителя 4 по характеристикам, по которым производится деление (вошедшим в модель блока 16), сравниваются со значением, по которому производится деление, и в зависимости от результата
10 сравнения событие относится к соответствующему сегменту. Одновременно вычисляются средние значения целевого показателя в сегментах, которые указываются для каждого сегмента в модели блока 16, после чего она сохраняется для дальнейшего использования.

Если новых сегментов не выявлено, построение модели заканчивается.

15 В момент прогноза целевого показателя (возможно в реальном времени), данные о событии генерируются блоком 2 по таким же данным и правилам, что и для построения модели, что гарантирует применимость построенной модели блока 16 для прогноза. Для прогноза целевого показателя такого события, в блоке 17 соответствующие значения его характеристик последовательно сравниваются со
20 значениями этих характеристик в узлах дерева блока 16, пока событие не будет отнесено к сегменту, не имеющему сегментов. Среднее значение целевых показателей событий в этом сегменте, использованных для построения модели, является прогнозом целевого показателя такого события.

На фиг. 3 отображен пример потенциального события промо-акции по
25 продвижению товара «БАТОН СОЛНЕЧНЫЙ 400 ГР.» (поле 30) потенциальному клиенту, похожему на существующего клиента № 5 (поле 31), и порядок вычисления целевого показателя (поле 32) и характеристик (таблица 33).

В данном случае бизнес-модель акции состоит в бесплатном предоставлении покупателям продвигаемого (целевого) товара («батон солнечный») на пробу.
30 Целевым показателем в поле 32 является оценка прибыли от промо-акции в отношении покупателя, похожего на существующего, но который еще не покупал

данный товар, вычисляемого в полях 34 и 35. Все вошедшие на экран характеристики события таблицы 33 сформированы путем автоматического вычисления сумм покупок за различные временные периоды с указанием названия товара, содержащегося в первой записи 36 в журнале покупок таблицы 37: «ТОРТ
5 «ТВОРОЖНИК» ШОКОЛАДНЫЙ 340 ГР», или слов в названии этого товара. Для остальных товаров в таблице покупок также формируются соответствующие характеристики, они не вошли на экран.

Для построения модели в соответствии с заявленным способом из каждого события указывается целевой показатель в поле 38, и характеристики события в виде
10 таблицы 33 вещественных значений с текстовыми названиями. Пользователем должны быть введены только параметры бизнес-модели в разделах «Доходы» 39 и «Расходы» 40, а также в одном из вариантов способа может быть указана доверительная вероятность.

В результате компьютером построена прогнозная модель (фиг. 4) в виде дерева
15 решений. В поле 50 «Дерево решений - информация» отображено наличие 200 событий в сегменте из 687 в вышестоящем узле дерева (родительском сегменте), при этом средний экономический эффект в поле 51 в виде ожидаемой прибыли на 1 контакт в сегменте составляет 0.51 рубля, а в вышестоящем узле дерева – отрицателен и равен -1.39 рубля. Деление на сегменты проведено по характеристике «Сумма
20 покупок со словом 'Ф/П' в названии», по значению 'больше 68.71'. Указание 'Ф/П' в названии товаров – сокращение от «Фин. Пакет», то есть полиэтиленовая упаковка для бюджетных молочных продуктов, таким образом, «БАТОН СОЛНЕЧНЫЙ» покупают те же клиенты, которые покупают и бюджетные молочные продукты.

В таблице 52 «Лучшие делящие параметры» показаны агрегированные данные
25 потенциального деления на сегменты по характеристикам «Сумма покупок Продуктов со словом в названии 'ГР' по понедельникам». Сокращение 'ГР' используется для указания веса в граммах фасованных товаров в упаковке. При этом количество наблюдений в возможных сегментах следующего уровня равно, соответственно, 137 и 63, а средняя прибыль в них равна +1,39 и -1,41. Однако, с
30 учетом стандартного отклонения и коэффициента Стьюдента, Пессимистический экономический эффект от возможного деления на сегменты (ПДП) 53 оказался

отрицательным, в результате чего дальнейшее деление на сегменты не было произведено.

Для клиентов, ранее не покупавших целевой товар, но у которых характеристика «Сумма покупок со словом 'Ф/П' в названии» присутствует и больше 5 68.71, среднее значение в этом сегменте, то есть 0.51 рубля на контакт, является прогнозом целевого показателя – экономического эффекта.

Деление на сегменты также может регулироваться параметром 54 «максимальное количество уровней сегментации».

Аналогичный пример построенной модели – дерево решений для 10 персональных рекомендаций фильма «PRETTY WOMAN» на данных компании NetFlix (фиг. 5). Отличие этого примера в том, что в качестве базы для числовых характеристик вместо суммы покупки взята оценка фильма зрителем по 5-бальной шкале. В этом случае минимальный рейтинг, при котором нужно давать рекомендацию, принят равным 4 и указан в поле 60. Наилучший критерий 15 (характеристика и значение) в узле дерева решений 61 для рекомендаций фильма «PRETTY WOMAN» на первом уровне сегментации - рейтинг фильма «DIRTY DANCING» больше 3,5. При этом средний прогнозный рейтинг в поле 62 в этом сегменте превышает требуемый, равный 4, на 0,37, при том, что в вышестоящем узле дерева (родительском сегменте) он ниже требуемого на 0,09 балла.

20 Пессимистический эффект деления этого сегмента на сегменты следующего уровня в поле 63 также положителен, и наилучшим критерием для деления стала та же характеристика фильма «DIRTY DANCING», но уже по значению 4,5.

В обоих примерах никакого ручного ввода или подготовки данных не производилось, сегментация проводилась по всем автоматически сформированным 25 характеристикам, общее число которых достигало сотен тысяч и более, что явствует из значений в полях 54 и 64 «Дерево решений – количество уникальных параметров+значений».

Таким образом, в заявленном способе для построения следующего уровня прогнозной модели в виде бинарного дерева решений используется 30 пессимистический экономический эффект деления каждого сегмента на сегменты следующего уровня с учетом доверительного интервала.

Значения характеристик событий сохраняются в сортированном виде в накопителе данных, что позволяет высвободить ОЗУ компьютера для целевых показателей и сегментов, а также за счет последовательного чтения данных из накопителя максимально быстро провести полный перебор всех значений множества характеристик для всех сегментов каждого уровня бинарного дерева решений.

Построенная модель может использоваться для моментального прогнозирования целевого показателя событий, в том числе, не участвовавших в ее построении. В качестве прогноза используется среднее значение целевого показателя в сегменте, к которому событие будет отнесено, путем последовательного сравнения характеристик события со значениями в узлах дерева решений.

Статистические методики вычисления доверительного интервала, в частности, на базе коэффициентов Стьюдента, как показали исследования, на большом объеме данных позволяют гарантировать выявление зависимостей в данных по миллионам показателей, автоматизируя процесс «уменьшения размерности» путем анализа всех вычисляемых характеристик и всех их значений, одновременно практически исключая при прогнозировании событий и их оценки человеческий фактор.

Промышленная применимость

Наиболее успешно изобретение применимо для формирования адресных предложений целевым сегментам покупателей на промышленных объемах данных в различных отраслях торговли.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ прогноза целевого показателя событий, заключающийся в том, что получают компьютером данные о событиях с указанием для каждого события целевого показателя и произвольного набора его числовых характеристик, производят сегментацию, создают прогнозную модель, и прогнозируют целевой показатель последующего аналогичного события, отличающийся тем, что посредством компьютера нумеруют события, проводят сквозную нумерацию характеристик событий, сортируют значения характеристик для событий в порядке номер характеристики события – значение характеристики события, запоминают их в этом порядке в накопителе данных компьютера с указанием номеров событий, после чего в цикле сегментации выполняют последовательное деление множества событий на подмножества – сегменты следующего уровня, соответствующие узлам дерева решений, в ходе которого выполняют последовательное чтение характеристик, их значений и номеров событий из накопителя данных, и для каждого сегмента вычисляют агрегированные данные на основе целевых показателей и количества из массива событий из оперативной памяти компьютера, и при отличии следующей характеристики события от предыдущей или ее значения от предыдущего рассчитывают экономический эффект деления на сегменты между предыдущим и следующим значениями предыдущей характеристики, запоминают наилучшую характеристику события и значение, имеющие максимальный экономический эффект деления на сегменты, после чтения последнего значения характеристики из накопителя данных проводят деление сегментов, в которых экономический эффект деления на сегменты следующего уровня положителен, на сегменты следующего уровня по наилучшей характеристике и ее значению, если обнаружены новые сегменты, то повторяют цикл сегментации, если новые сегменты не обнаружены, то для прогноза целевого показателя последующего аналогичного события последовательно сравнивают значения характеристик аналогичного события со значениями наилучших характеристик, по которым было проведено деление сегмента на сегменты следующего уровня, и по значению наилучшей характеристики определяют сегмент, к которому относится последующее аналогичное событие,

среднее значение целевого показателя такого сегмента служит прогнозом для последующего аналогичного события.

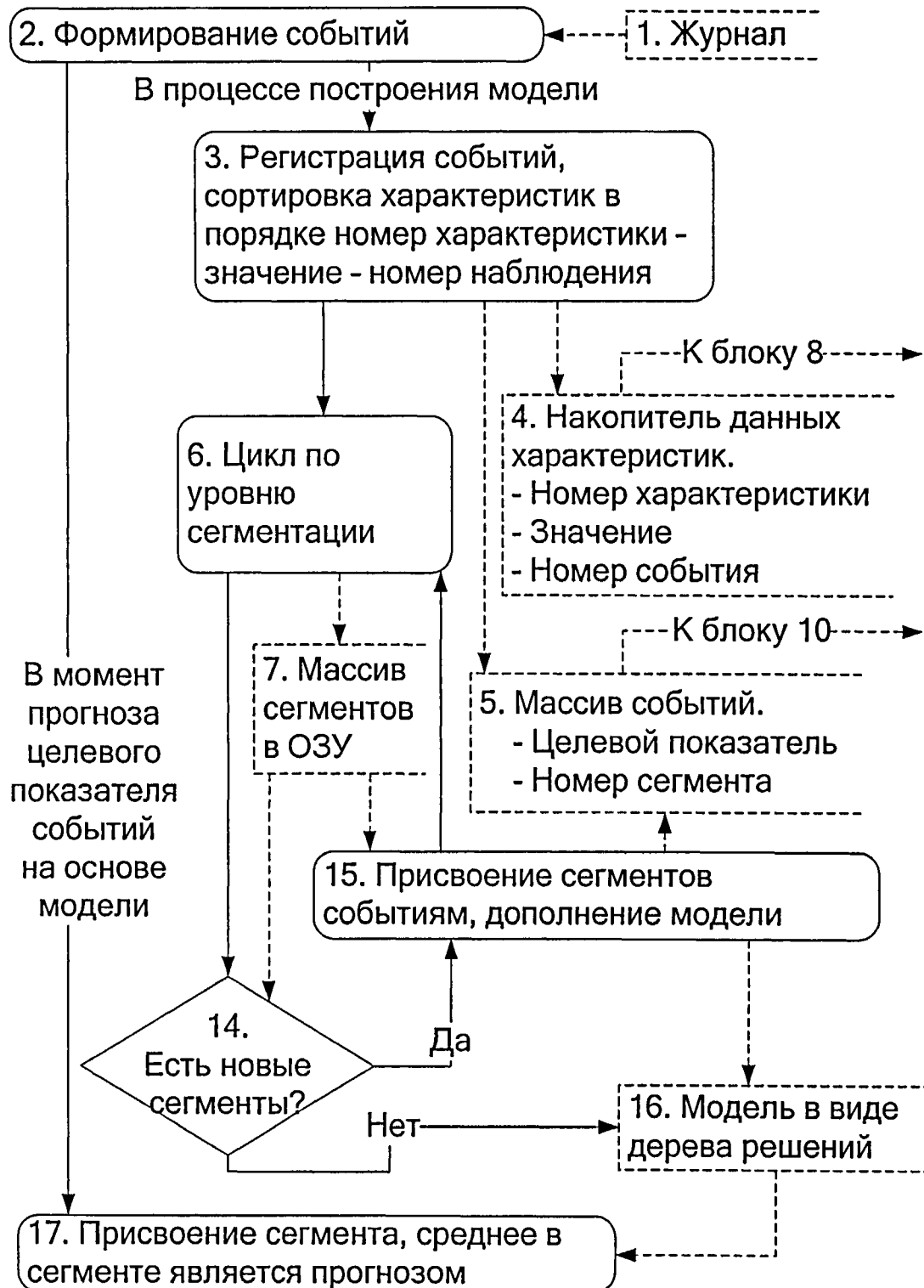
2. Способ по п. 1, отличающийся тем, что при расчете экономического эффекта деления на сегменты экономический эффект уменьшают на величину доверительного интервала экономического эффекта.

3. Способ по п. 1, отличающийся тем, что деление сегментов выполняют не более заданного количества уровней.

4. Способ по п. 1, отличающийся тем, что деление сегментов выполняют при условии, что каждый сегмент следующего уровня содержит не менее заданного количества событий.

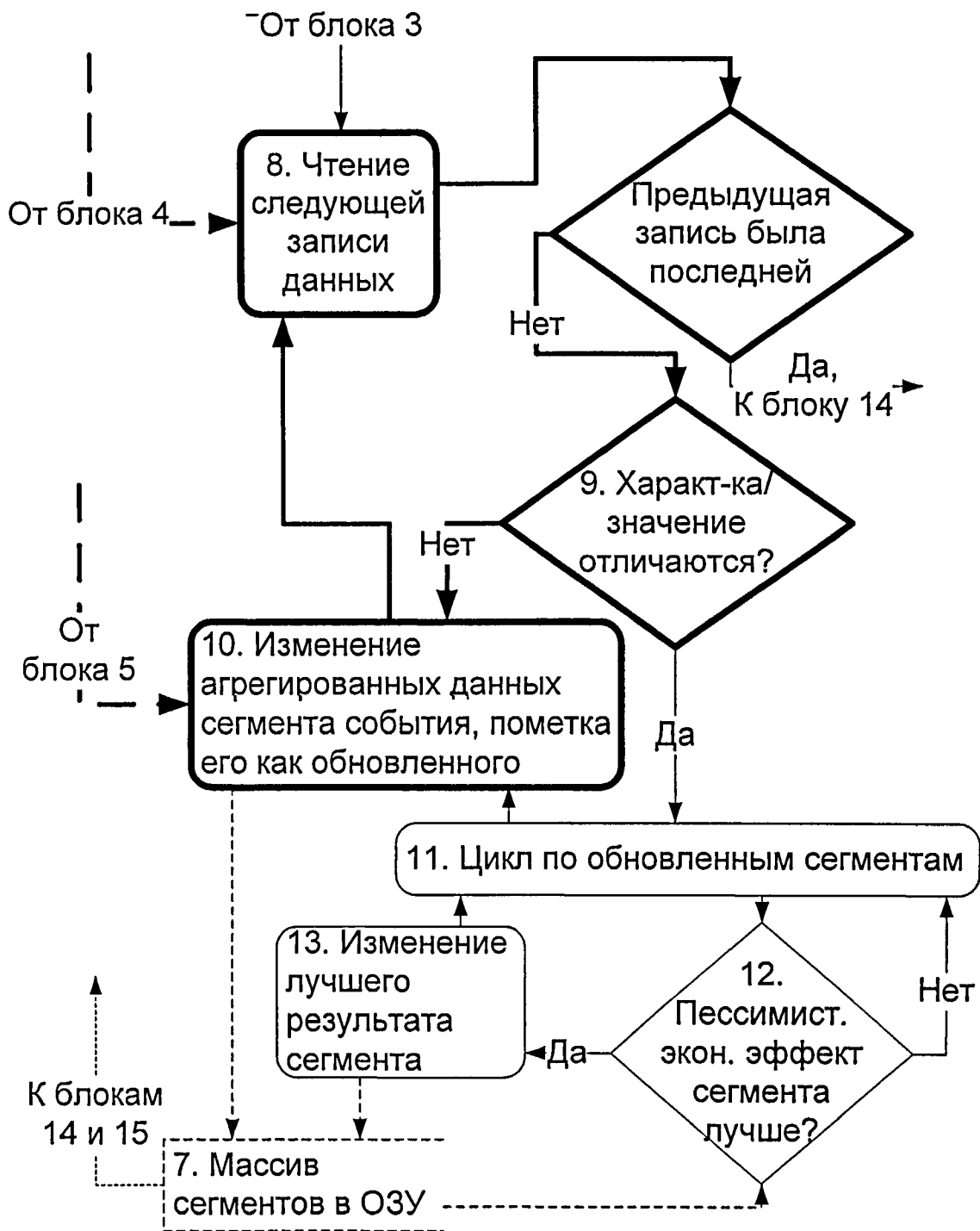
5. Способ по п. 1, отличающийся тем, что последовательное деление множества событий на подмножества – сегменты следующего уровня производят путем разделения сегмента предшествующего уровня на два сегмента следующего уровня.

1/5



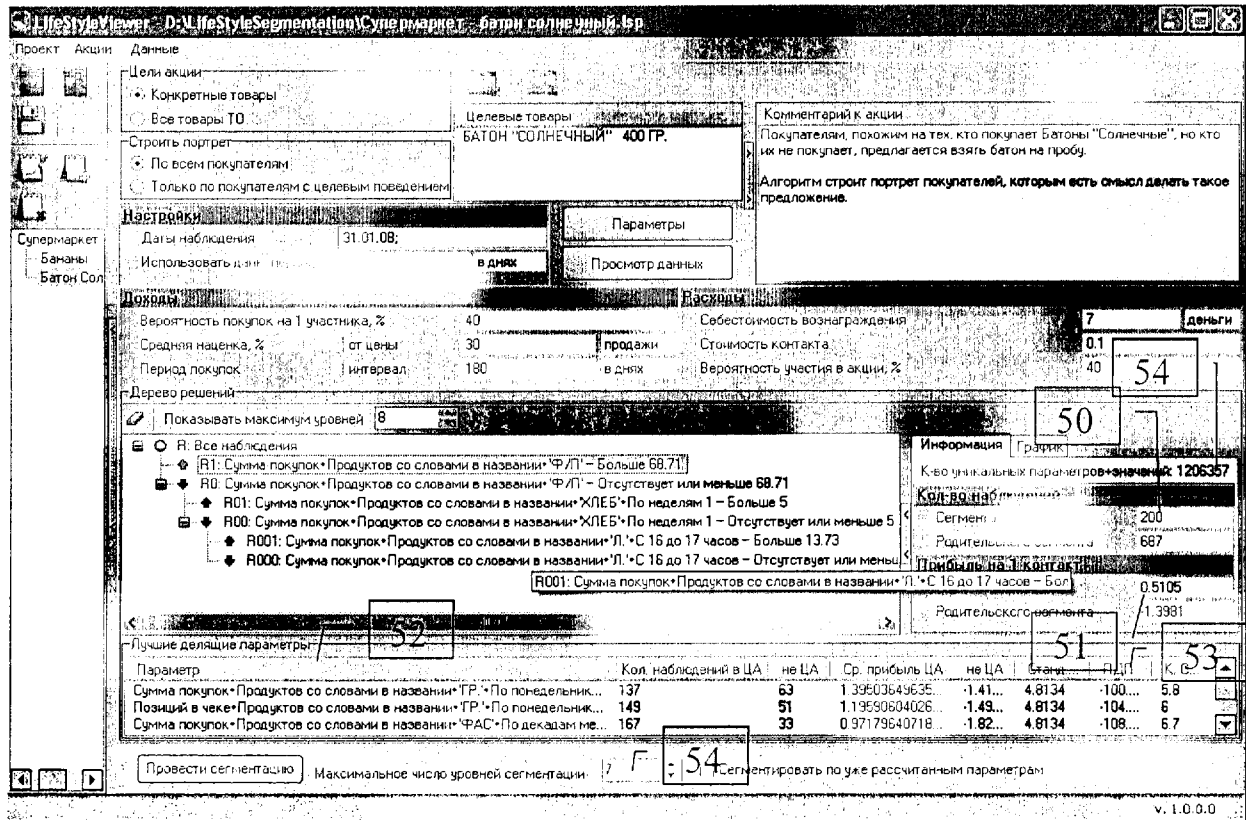
Фиг. 1

2/5



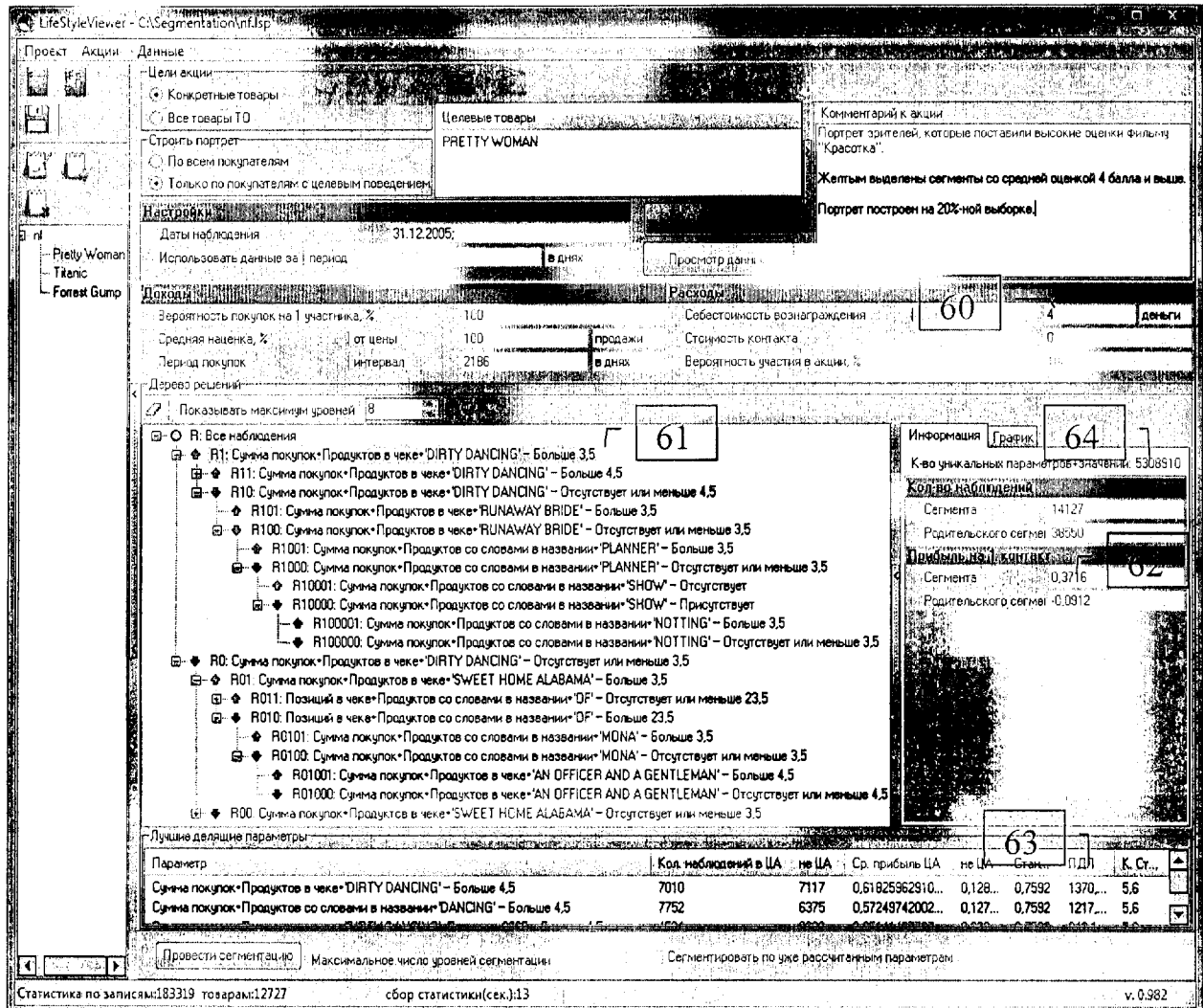
Фиг. 2

Фиг. 3



ФИГ. 4

5/5



Фиг. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/RU 2012/000787

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 10/04 (2012.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06Q 10/00, 10/04, 10/06, G06F 17/00, 17/10, 17/30, 17/40

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatSearch (RUPTO internal), USPTO, PAJ, Esp@cenet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 6839682 B1 (FAIR ISAAC CORPORATION) 04.01.2005, abstract, col.3, line 20- col.4, line 24	1-5
Y	US 2008/0103808 A1 (KEVIN E. BERRY et al) 01.05.2008, [0043]	1-5
A	US 7542881 B1 (JEAN-MARIE BILLIOTTE et al) 02.07.2009	1-5
A	US 2011/0191140 A1 (ALAN B.NEWMAN et al) 04.08.2011	1-5



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 January 2013 (18.01.2013)

Date of mailing of the international search report

07 February 2013 (07.02.2013)

Name and mailing address of the ISA/

Authorized officer

Facsimile No.

Telephone No.

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ G06Q 10/04 (2012.01) Согласно Международной патентной классификации МПК		
В. ОБЛАСТЬ ПОИСКА Проверенный минимум документации (система классификации с индексами классификации) G06Q 10/00, 10/04, 10/06, G06F 17/00, 17/10, 17/30, 17/40 Другая проверенная документация в той мере, в какой она включена в поисковые подборки Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины) PatSearch (RUPTO internal), USPTO, PAJ, Esp@cenet		
С. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:		
Категория*	Цитируемые документы с указанием, где это возможно, релевантных частей	Относится к пункту №
Y	US 6839682 B1 (FAIR ISAAC CORPORATION) 04.01.2005, abstract, col.3, line 20- col. 4, line 24	1-5
Y	US 2008/0103808 A1 (KEVIN E. BERRY et al) 01.05.2008, [0043]	1-5
A	US 7542881 B1 (JEAN-MARIE BILLIOTTE et al) 02.07.2009	1-5
A	US 2011/0191140 A1 (ALAN B. NEWMAN et al) 04.08.2011	1-5
☐ последующие документы указаны в продолжении графы С. ☐ данные о патентах-аналогах указаны в приложении		
* Особые категории ссылочных документов:	“Т” более поздний документ, опубликованный после даты международной подачи или приоритета, но приведенный для понимания принципа или теории, на которых основывается изобретение “Х” документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает новизной или изобретательским уровнем, в сравнении с документом, взятым в отдельности “У” документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает изобретательским уровнем, когда документ взят в сочетании с одним или несколькими документами той же категории, такая комбинация документов очевидна для специалиста “&” документ, являющийся патентом-аналогом	
“А” документ, определяющий общий уровень техники и не считающийся особо релевантным		
“Е” более ранняя заявка или патент, но опубликованная на дату международной подачи или после нее		
“L” документ, подвергающий сомнению притязание(я) на приоритет, или который приводится с целью установления даты публикации другого ссылочного документа, а также в других целях (как указано)		
“О” документ, относящийся к устному раскрытию, использованию, экспонированию и т.д.		
“Р” документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета		
Дата действительного завершения международного поиска	Дата отправки настоящего отчета о международном поиске	
18 января 2013 (18.01.2013)	07 февраля 2013 (07.02.2013)	
Наименование и адрес ISA/RU: ФИПС, РФ, 123995, Москва, Г-59, ГСП-5, Бережковская наб., 30-1 Факс: (499) 243-33-37	Уполномоченное лицо: Володин В. Телефон № 499-240-25-91	