РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 469 389** (13) **C1**

(51) МПК
G06F 17/30 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21)(22) Заявка: **2011145077/08, 08.11.2011**

(24) Дата начала отсчета срока действия патента:
08.11.2011

Приоритет(ы):
(22) Дата подачи заявки: **08.11.2011**

(45) Опубликовано: **10.12.2012** Бюл. № **34**

(56) Список документов, цитированных в отчете о поиске: **US 2008/0133716 A1, 05.06.2008. US 7970912 B2, 28.06.2011. US 7680770 B1, 16.03.2010. RU 2344474 C2, 20.01.2009.**

Адрес для переписки:
119331, Москва, а/я 88, В.Н. Рослову, рег.№ 18

(72) Автор(ы):
Бартунов Сергей Олегович (RU),
Коршунов Антон Викторович (RU),
Турдаков Денис Юрьевич (RU),
Кузюрин Николай Николаевич (RU),
ПАРК Сеунг-Таек (KR),
РЫУ Вонхо (KR),
ЛИ Хыунгдонг (KR)

(73) Патентообладатель(и):
Учреждение Российской академии наук Институт системного программирования РАН (RU),
Корпорация "САМСУНГ ЭЛЕКТРОНИКС Ко., Лтд." (KR)

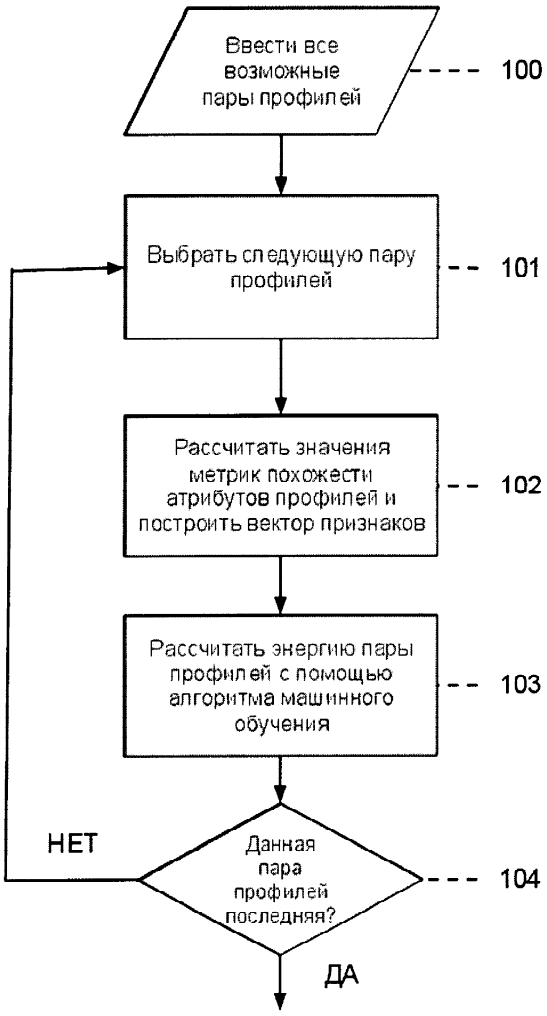
(54) СПОСОБ ИНТЕГРАЦИИ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ ОНЛАЙНОВЫХ СОЦИАЛЬНЫХ СЕТЕЙ

(57) Реферат:

Изобретение относится к области обработки пользовательских данных, полученных из графов онлайновых социальных сетей, с целью интеграции данных различных профилей, принадлежащих одному пользователю. Техническим результатом является повышение эффективности интеграции профилей пользователей онлайновых социальных сетей. Способ содержит этапы, на которых вводят все возможные пары профилей, строят модель Условных Случайных Полей из всех профилей и связей между ними, для каждой пары профилей рассчитывают значения похожести их атрибутов с помощью метрик строковой и графовой похожести, из полученных значений метрик похожести строят вектор признаков, который передается алгоритму машинного обучения, который осуществляет расчет унарной энергии, либо бинарной энергии, для каждой пары профилей, в которой профили принадлежат различным социальным графам, рассчитывается похожесть профилей, проверяется, превышает ли полученное значение похожести профилей заданное пороговое значение, в случае положительного ответа пара профилей заносится в список кандидатов, из полученного списка кандидатов выбираются априорно верные проекции, модель Условных Случайных Полей разбивается на независимые компоненты, для каждой компоненты модели производится поиск оптимальной конфигурации проекций, производится объединение списков найденных проекций для всех компонент модели. 3 ил., 2 табл.

RU 2469389 C1

RU 2469389 C1



Вид 1.1

Фиг. 1

RU 2469389 C1

RU 2469389 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 469 389** (13) **C1**

(51) Int. Cl.
G06F 17/30 (2006.01)

(12) ABSTRACT OF INVENTION

(21)(22) Application: **2011145077/08, 08.11.2011**

(24) Effective date for property rights:
08.11.2011

Priority:
(22) Date of filing: **08.11.2011**

(45) Date of publication: **10.12.2012 Bull. 34**

Mail address:
119331, Moskva, a/ja 88, V.N. Roslovu, reg.№ 18

(72) Inventor(s):
Bartunov Sergej Olegovich (RU),
Korshunov Anton Viktorovich (RU),
Turdakov Denis Jur'evich (RU),
Kuzjurin Nikolaj Nikolaevich (RU),
PARK Seung-Taek (KR),
RYU Vonkho (KR),
LI Khyungdong (KR)

(73) Proprietor(s):
Uchrezhdenie Rossijskoj akademii nauk Institut sistemnogo programmirovanija RAN (RU),
Korporatsija "SAMSUNG EhLEKTRONIKS Ko., Ltd." (KR)

(54) METHOD OF INTEGRATING USER PROFILES OF ONLINE SOCIAL NETWORKS

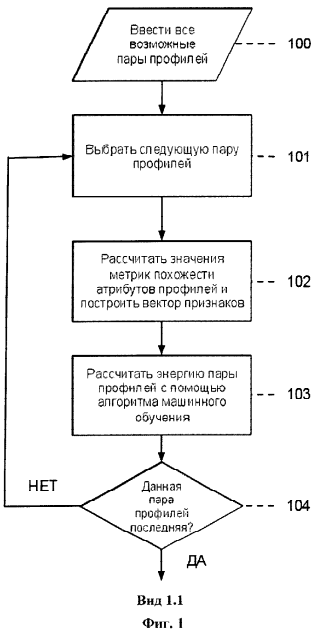
(57) Abstract:

FIELD: information technology.

SUBSTANCE: method comprises steps of entering all possible pairs of profiles; constructing a Conditional Random Field model for all profiles and connections between them; for each pair of profiles, calculating the similarity value of their attributes using a string or graph similarity metric; constructing a feature vector from the obtained similarity metric values, which is sent to a machine learning algorithm which calculates unary energy or binary energy for each pair of profiles, wherein the profiles belong to different social graphs; calculating profile similarity; checking whether the obtained profile similarity value exceeds a given threshold value; if so, the pair of profiles is entered into a list of candidates; a priori true projections are selected from the obtained list of candidates; the Conditional Random Field model is broken into independent components; for each component of the model, the optimum configuration of projections is sought; the lists of the found projections are merged for all components of the model.

EFFECT: high efficiency of integrating user profiles of online social networks.

6 dwg



Вид 1.1
Фиг. 1

RU 2469389 C1

RU 2469389 C1

Изобретение относится к области обработки пользовательских данных, полученных из графов онлайновых социальных сетей, с целью интеграции данных различных профилей, принадлежащих одному пользователю. Может быть использовано для построения баз данных пользовательской информации, полученной из различных источников, в частности, для построения расширенного социального графа, содержащего данные о пользователе, полученные из нескольких различных социальных графов. Подобный расширенный социальный граф может быть использован для улучшения качества результатов в ряде задач, таких как поиск информации в Интернете, онлайн-реклама товаров и услуг, построение рекомендаций товаров и услуг пользователям и др.

Рассмотрим основные понятия, необходимые для понимания представленного изобретения:

1. Интеграция данных включает объединение данных, находящихся в различных источниках и предоставление данных пользователям в унифицированном виде. Изобретение обеспечивает интеграцию данных на логическом уровне с целью обеспечения возможности доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных свойств.

2. Социальный граф является цифровым представлением взаимоотношений пользователей онлайновых социальных сетей, которое явно задается различными типами отношений связи между пользователями (например, отношение дружбы, отношение следования и т.д.). Данные пользователя в социальном графе представлены в виде профиля, который представляет собой находящуюся на материальном носителе либо в памяти вычислительной машины совокупность атрибутов (в основном, строковых) в виде пар "имя - значение", которые содержат различную информацию о пользователе (например, имя, пол, адрес, номер телефона и т.д.). Изобретение предназначено для объединения различных социальных графов путем сравнения атрибутов профилей пользователей и интенсивного использования информации, скрытой в связях между профилями.

3. Условные Случайные Поля - это графическая вероятностная модель, в которой в виде ненаправленного графа представлены зависимости между случайными величинами. Узлы графа делятся на два непересекающихся множества - наблюдаемые переменные, которые задаются в качестве входных данных, и скрытые переменные. Ребра графа соответствуют вероятностным взаимосвязям между узлами. Вершинам и ребрам могут быть назначены численные значения, называемые энергиями. Вычисление значений скрытых переменных называется выводом из модели.

4. Машинное обучение - раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. В представленном изобретении используется разновидность машинного обучения, именуемая обучением с учителем: алгоритм генерирует функцию, которая связывает входные данные с выходными определенным образом (задача классификации). В качестве обучающих данных используются примеры связи входных данных с выходными. В алгоритмах машинного обучения широко используется понятие признака. Признаки - индивидуальные измеримые свойства наблюдаемого феномена, которые используются для создания его численного представления (например, значения метрик строковой похожести для пар атрибутов профилей).

5. Метрики строковой похожести возвращают численное значение похожести пары строк, основываясь на порядке расположения составляющих их символов (например,

расстояние Джаро-Винклера [17]).

6. Метрики графовой похожести возвращают численное значение похожести пары узлов графа, основываясь на структуре связей между ними (например, коэффициент Дайса [18]).

Крупнейшим исследованием, посвященным интеграции профилей пользователей социальных сетей, является диссертация Veldman [1]. В ней предлагается множество эвристик, использующих как данные профилей пользователей, так и существующие связи между ними. Результаты подобных исследований также представлены в работах Motoyama et al [2], Gae-won et al [3], Raad et al [4] и Vozecky et al [5]. В работе [2] авторы сравнивают профили пользователей MySpace и Facebook. В работе [3] авторы делают то же самое с профилями из Twitter и EntityCube. Авторы [4] генерируют синтетические профили пользователей и применяют к ним различные сложные эвристики, стараясь использовать любой потенциально полезный источник данных в социальной сети. В работе [5] профили пользователей Facebook и StudiVZ представлены в виде n-мерных векторов, которые впоследствии сравниваются с помощью различных техник, включая нечеткое сравнение. Авторы также исследуют влияние различных атрибутов профиля на точность результатов сравнения.

Интерес также представляют проекты Foaf-o-matic [6] и Okkam [7], целью которых является интеграция социальных профилей с помощью формальной семантики FOAF (Friend-of-a-friend). Проект Stanford Entity Resolution Framework [8] также предназначен для решения задач, подобных данной. Помимо исходных кодов фреймворка, доступно множество работ, посвященных теоретическим аспектам интеграции данных, таким как масштабируемость, оценка качества и др.

Несмотря на успехи, достигнутые авторами вышеперечисленных работ, в них используется слишком простая модель сравнения профилей, основанная, в основном на попарном сравнении с помощью строковой похожести отдельных атрибутов. Кроме того, существующие связи между профилями учитываются недостаточно либо вообще не берутся во внимание, то же касается особенностей сравниваемых социальных графов.

Наиболее близким к представленному изобретению является способ, предложенный Singla et al [9] для выявления дубликатов в сети цитирования научных работ с помощью модели Условных Случайных Полей. Авторы формулируют задачу в терминах марковской случайной логики и строят модель из фактов и утверждений о сравниваемых объектах, после чего рассчитывают их вероятности. Для оптимизации скорости работы алгоритма авторы производят разбиение модели на пересекающиеся компоненты. Вместе с тем, представленный подход обладает следующими недостатками, препятствующими его использованию для решения рассматриваемой задачи:

- в модели предусмотрено наличие только одного источника данных (граф сети цитирования научных работ);
- узлами модели являются факты и утверждения о сравниваемых объектах, а не сами объекты. В частности, определено два типа узлов: узлы-записи и узлы-атрибуты. Первый тип узлов предназначен для хранения вопроса "Идентичен ли данный объект другому объекту?", тогда как второй тип хранит информацию о похожести атрибутов объектов;
- для сравнения объектов используются только метрики строковой похожести их атрибутов, тогда как метрики графовой похожести не используются.

Настоящее изобретение обладает следующими преимуществами по сравнению с

ранее предложенными подходами:

- позволяет упростить представление социального графа в памяти вычислительной машины, поскольку модель Условных Случайных Полей строится на основе одного из сравниваемых графов, при этом узлами модели являются непосредственно профили, а ребрами - связи между ними. Затем рассчитываются энергии связей модели исходя из данных о строковой и графовой похожести профилей, после чего производится поиск оптимального решения в виде конфигурации проекций профилей одной социальной сети на другую;

- с помощью модели Условных Случайных Полей учитывается вся доступная информация о связях между профилями, что позволяет использовать латентную информацию, скрытую в этих связях, для уточнения результатов. Таким образом, способ позволяет производить интеграцию профилей, атрибуты которых содержат лишь незначительное количество полезной информации, что существенно усложняет применение общепринятого подхода, основанного на строковой близости атрибутов;

- адекватность и эффективность различных метрик строковой и графовой близости, а также относительная значимость атрибутов оцениваются с помощью методик машинного обучения на предварительно составленном наборе реальных экспериментальных данных, что позволяет учесть особенности выбранных социальных графов и оптимизировать параметры алгоритма для достижения лучших результатов.

Технический результат использования предлагаемого изобретения состоит в том, что изобретение позволяет ранее неизвестным способом получать список пар профилей пользователей онлайновых социальных сетей, в котором профили в каждой паре содержат информацию об одном и том же пользователе, основываясь только на информации, содержащейся в атрибутах профилей и связях между ними. Также предложен универсальный подход, позволяющий учитывать особенности любой пары социальных графов и оптимизировать параметры метода для получения лучших результатов.

Для лучшего понимания заявленного изобретения далее приводится его подробное описание с соответствующими чертежами.

Фиг.1 представляет собой блок-схему алгоритма работы изобретения

Фиг.2 представляет собой схему расчета значений похожести атрибутов профилей и построение вектора признаков

Фиг.3 представляет собой Модель Условных Случайных Полей

Рассмотрим два социальных графа A и B. Для вершины $v \in A$ соответствующий профиль из графа B будем называть проекцией $pr(v)$ вершины $v \in A$ на граф B. Если для вершины $v \in A$ не определена проекция из B, то ей назначается нейтральная проекция: $pr(v)=N$.

Задача интеграции профилей пользователей заключается в определении максимально возможного количества верных проекций (v, u), $v \in A$, $u \in A$. в терминах модели Условных Случайных Полей проекции $pr(v)$ для каждого $v \in A$ являются скрытыми переменными, значение которых нужно установить.

Представленное изобретение основано на следующих основных утверждениях:

- задача определения верной проекции для узла из графа A связана с задачами определения верных проекций для всех смежных узлов из графа A;

- если два узла в графе A связаны, то их проекции в графе B должны иметь как можно более высокое значение графовой похожести.

Предлагаемый способ интеграции профилей пользователей онлайновых

социальных сетей содержит алгоритм, включающий в себя ввод всех возможных пар профилей, 19 основных шагов и вывод результата в виде списка пар профилей, в котором каждая пара содержит информацию об одном и том же пользователе, при этом составляющие проекцию профили относятся к разным социальным графам.

Рассмотрим более подробно шаги алгоритма (фиг.1).

Шаг 100. Ввод всех возможных пар профилей.

В сравнении участвуют все возможные пары профилей, поскольку используемый в изобретении алгоритм вывода из модели Условных Случайных Полей дает лучшие результаты при наличии информации об энергиях всех возможных парных комбинаций узлов. Набор профилей из графов A и B со всеми известными связями между ними образуют модель Условных Случайных Полей.

Шаг 101. Выбор следующей пары профилей.

Шаг 102. Расчет значений похожести атрибутов профилей и построение вектора признаков.

Сравнение атрибутов профилей из графов A и B производится с помощью схемы соответствия, которая задает порядок сравнения атрибутов и применяемые метрики похожести (примеры схем соответствия для расчета похожести атрибутов между узлами из различных графов и между двумя узлами графа ν даны в Табл. 1 и Табл. 2 соответственно).

На фиг.2 изображен пример сравнения профилей из Twitter и Facebook. T_i и F_i соответствуют двум сравниваемым атрибутам из профилей T и F, где i - порядковый номер содержащей данные атрибуты записи в схеме соответствия. К каждой паре атрибутов применяется метрика похожести sim_i. Значения метрик похожести для всех пар атрибутов составляют вектор признаков, который передается на следующий шаг.

Шаг 103. Расчет энергии пары профилей с помощью алгоритма машинного обучения.

Определим понятия энергий, предварительно определив необходимые переменные.

Граф A используется для построения модели Условных Случайных Полей. Пусть ν и u являются пользователями из графа A, а pr(ν) и pr(u) - их проекциями из графа B.

Тогда энергия узла (унарная энергия) определяется как

$\Phi(pr(\nu)|\nu)=\alpha(\nu)\cdot(1\text{-}profile_{similarity(\nu,pr(\nu))})$,

энергия связи (бинарная энергия) определяется как

$\psi(pr(\nu),pr(u)|\nu,u)=\beta(pr(\nu),pr(u))\cdot(1\text{-}network_{similarity(pr(\nu),pr(u))})$,

а полная энергия определяется как

$$E=\sum_{\nu\in A}\phi(pr(\nu)\,|\,\nu)+\sum_{(\nu,u)\in A}\psi(pr(\nu),pr(u)\,|\,\nu,u).$$

$\alpha(\nu)$ и $\beta(pr(\nu),pr(u))$ являются коэффициентами, регулирующими влияние каждого типа энергий на итоговую модель и результаты работы. Функции $profile_{similarity}$ и $network_{similarity}$ являются функциями похожести профилей, которые нормированы и увеличиваются с увеличением похожести сравниваемых профилей.

Для расчета унарных и бинарных энергий с целью максимизации точности результатов путем анализа реальных данных используется методика машинного обучения, которая получает на вход набор значений метрик похожести для данной пары профилей в виде вектора признаков и возвращает значение энергии связи данной пары профилей. В качестве методики машинного обучения предусмотрено использование одного из способов:

- Взвешенная линейная комбинация признаков, где веса подбираются при помощи линейной регрессии [15], исходя из того, что унарная энергия для правильных

проекций должна быть равна 0, а для неправильных - 1;

- Алгоритм машинного обучения MultiBoostAB [13] над решающими деревьями C 4.5 [14];

- Алгоритм машинного обучения LogitBoost над решающими деревьями M5P [16].

Перечисленные алгоритмы перед использованием проходят процедуру обучения, т.е. получают на вход множество векторов признаков, полученных при расчете метрик похожести для атрибутов профилей, входящих в состав набора данных, в котором верные проекции заранее заданы вручную. Вектор признаков, используемый для обучения, включает в себя все значения метрик похожести для сравниваемой пары профилей и дополнительное измерение, содержащее значение булевского типа и указывающее, содержат ли данные профили информацию об одном и том же пользователе. Такой набор данных может быть составлен для любой пары социальных графов.

Шаг 104. Является ли данная пара профилей последней? Если НЕТ - переход к шагу 101, если ДА - переход к шагу 105.

Шаг 105. Выбор следующей пары профилей.

Производится последовательный перебор всех возможных пар профилей, в которых один из профилей принадлежит графу A а второй - графу B.

Шаг 106. Расчет похожести пары профилей.

Производится сравнение атрибута выбранной пары профилей, который с наибольшей степенью вероятности однозначно идентифицирует профиль, с помощью метрики строковой похожести.

Шаг 107. Превышает ли значение похожести пары профилей заданное пороговое значение? Если ДА - переход к шагу 108, если НЕТ - переход к шагу 109.

Шаг 108. Добавить пары профилей в список кандидатов.

Шаг 109. Является ли данная пара профилей последней? Если НЕТ - переход к шагу 105, если ДА - переход к шагу 110.

Шаг 110. Выбор наилучших пар профилей из списка кандидатов и составление списка априорно верных проекций.

Для выбора наилучших пар профилей к списку кандидатов применяется алгоритм Куна-Манкреса [19], который производит последовательный перебор всех пар из списка кандидатов и выдает в качестве результата часть из них, выбранных таким образом, чтобы каждый профиль встречался только в одной паре, а профили, составляющие пару, взаимно максимизировали похожесть друг на друга. Результатом является набор априорно верных проекций (пар профилей), которые заносятся в модель с целью улучшения качества результатов.

Шаг 111. Разбиение модели на независимые компоненты путем удаления априорно верных проекций.

С целью уменьшения вычислительной сложности алгоритма исходная задача разбивается на подзадачи путем разбиения исходной модели. Это достигается путем удаления найденных априорно верных проекций.

Шаг 112. Выбор следующей компоненты модели.

Шаг 113. Поиск оптимальной конфигурации проекций для выбранной компоненты.

Для решения данной задачи полная энергия модели должна быть минимизирована. Для этого производится вывод из построенной модели Условных Случайных Полей. К задаче вывода сначала применяется квадратичная релаксация [10], затем задача аппроксимируется с применением методов Power Iteration [11] и Singular Value Decomposition [12], после чего решается как задача квадратичного программирования.

Результатом данного шага является конфигурация модели (набор проекций), которая минимизирует полную энергию модели и содержит максимальное количество верных проекций профилей.

На фиг.3 схематично изображена модель с набором различных проекций. Узлы графа A изображены непрерывной линией, тогда как узлы графа B - штрихпунктирной линией. Узлы в форме квадратов соответствуют априорно верным проекциям, узлы в форме треугольников являются компонентами проекций, найденных в результате вывода, в то время как для всех остальных узлов подходящих проекций не найдено. Пара узлов, составляющих проекцию, соединена с помощью пунктирной линии.

Шаг 114. Является ли данная компонента модели последней? Если НЕТ - переход к шагу 112, если ДА - переход к шагу 115.

Шаг 115. Объединение результатов для всех компонент модели.

Объединяются списки найденных проекций для всех компонент модели.

Шаг 116. Выбор следующей проекции.

Шаг 117. Построение вектора признаков для выбранной проекции и перенаправление его классификатору в качестве входных данных.

Для выбранной проекции строится вектор, состоящий из следующих признаков:

- унарная энергия вершины;
- средняя бинарная энергия связи с априорно верными проекциями;
- доля априорно верных проекций в списке вершин, связанных с данной;
- качество набора априорно верных проекций.

Качество набора априорно верных проекций вычисляется как сумма наибольших N весов, назначаемых априорно верным проекциям, связанным с рассматриваемой вершиной, где вес вершины вычисляется как средняя бинарная энергия связи между данной вершиной и другими взвешиваемыми вершинами.

Шаг 118. Является ли выбранная проекция верной?

Для уточнения результатов применяется классификаторный алгоритм машинного обучения (бустинг MultiBoostAB [13] над решающими деревьями C 4.5 [14]). Алгоритм классификации принимает построенный вектор признаков в качестве входных данных и возвращает решение о том, является ли данная вершина корректно спроецированной. В случае, если проекция неверная (решение классификатора не совпадает с решением алгоритма), то переход к шагу 119, в противном случае - переход к шагу 120.

Шаг 119. Удаление выбранной проекции из результатов.

Шаг 120. Является ли данная проекция последней? Если НЕТ - переход к шагу 116, если ДА - переход к шагу 121.

Шаг 121. Вывод списка проекций, в котором каждая проекция содержит информацию об одном и том же пользователе, при этом составляющие проекцию профили относятся к различным социальным графам.

Пример работы

На фиг.3 изображены модели двух социальных графов, где профили представлены узлами, а связи между ними - ребрами. Будем считать, что узлы 1-8 принадлежат графу A (Twitter), а узлы 9-14 принадлежат графу B (Facebook). Все связи между узлами внутри каждого из графов заданы изначально, тогда как связи между узлами разных графов (объединение узлов в пары) являются результатом работы. Граф Twitter с полным набором узлов и ребер используется для построения модели Условных Случайных Полей, а узлы из графа Facebook считаются скрытыми переменными

модели. При этом связью в графе Twitter считается отношение взаимного следования (каждый профиль пары следует за другим профилем, т.е. получает уведомления о появлении новой информации в профиле; такое отношение устанавливается путем односторонней активации уведомлений владельцем того профиля, который следует за другим профилем), а в графе Facebook - отношение дружбы (каждый профиль пары дружит с другим профилем; такое отношение устанавливается путем отправки владельцем одного из профилей запроса на установление отношения и получения явного подтверждения запроса).

На шаге 102 рассчитываются значения похожести атрибутов профилей, а также строится вектор признаков. Сравнение атрибутов профилей из графов - A и B производится с помощью схемы соответствия, которая задает порядок сравнения атрибутов и применяемые метрики похожести (примеры схем соответствия для расчета похожести атрибутов между профилями Twitter и Facebook и между профилями Facebook даны в Табл. 1 и Табл. 2 соответственно).

Таблица 1

Схема соответствия для расчета похожести атрибутов профилей между узлами 2 и 11

Атрибут профиля Twitter (узел 2)	Значение атрибута профиля Twitter	Атрибут профиля Facebook (узел 11)	Значение атрибута профиля Facebook	Метрика похожести	Значение метрики
Name	John Smith	Name	J Smith	VMN	0,66
User place	New York, US	Current city	New York	Jaro	0,45
URL	www.my.site	Website	www.no.site	URL measure	0

Таблица 2.

Схема соответствия для расчета похожести атрибутов профилей узлами 11 и 14

Атрибут профиля Facebook	Значение атрибута первого профиля Facebook (узел 11)	Значение атрибута второго профиля Facebook (узел 14)	Метрика похожести	Значение метрики
Список контактов	9, 10, 12, 13, 14	11, 12,13	Bidirectional Contact Score	1
Список контактов	9, 10, 12, 13, 14	11,12,13	Weighted Dice	0,3

На шаге 103 рассчитываются значения унарных и бинарных энергий.

Для расчета унарных и бинарных энергий используется методика машинного обучения, которая получает на вход вектор признаков и возвращает значение энергии связи данной пары профилей. К примеру, для профилей из Табл.1 вектор признаков выглядит следующим образом:

[0,66; 0,45; 0].

Алгоритм машинного обучения возвращает значение унарной энергии для данной пары профилей, равное 0,52.

На шаге 106 производится попарное сравнение атрибута "Name" всех пар профилей Twitter - Facebook с помощью метрики строкой похожести VMN. Результатом является набор троек вида "номер профиля - номер профиля - значение похожести". Значение порога похожести примем равным 0,51.

На шаге 107 производится сравнение значения метрики строковой похожести с заданным порогом.

На шаге 108 все пары, для которых значение метрики строковой похожести превысило заданный порог, заносятся в список кандидатов.

К примеру, начальный список пар имеет вид:

1	9	0
1	10	0,65
1	11	0,55
2	9	0,5
2	10	0,6
2	11	0,66
3	9	0,35
3	10	0,7
3	11	0

Тогда список пар-кандидатов будет следующим:

1	10	0,65
1	11	0,55
2	10	0,6
2	11	0,66
3	10	0,7

На шаге 110 к списку кандидатов применяется алгоритм Куна-Манкреса для выбора наилучших проекций профилей, взаимно максимизирующих похожесть друг на друга. Результатом является набор априорно верных проекций, которые заносятся в модель:

2	11
3	10

На фиг.3 априорно верные проекции соответствуют парам узлов в форме квадратов, соединенных пунктирной линией (пары 2-11,3-10 и 5-13). Каждая такая проекция вносит дополнительную полезную информацию в модель и, таким образом, не только помогает уменьшить количество необходимых для получения оптимальной конфигурации вычислений, но и уменьшает вероятность неверных ответов в результатах работы.

На шаге 111 исходная модель разбивается на независимые компоненты путем удаления априорно верных проекций. На рассматриваемом примере после удаления узлов 2, 3 и 5 остаются 3 независимых подграфа (1, 7), (8) и (4). Соответствующие узлы 10, 11 и 13 из графа Facebook также в дальнейшем не рассматриваются и не участвуют в подборе проекций для оставшихся в модели узлов.

На шаге 113 для каждой из образовавшихся независимых компонент модели производится вывод путем применения итеративного алгоритма, перебирающего возможные конфигурации проекций модели и минимизирующего ее полную энергию.

На шаге 115 результаты для всех компонент модели объединяются.

Пример результатов работы данного шага - набор проекций 1-9, 1-12 и 6-14.

На следующих шагах производится уточнение результатов.

На шаге 117 для каждой найденной на шаге 115 проекции составляется вектор признаков, который подается на вход классификатора.

На шаге 118 классификатор принимает решение о верности каждой из проекций. Пусть вектор для проекции 1-9 будет [0,85; 0,5; 0,4; 0,5], а для проекции 1-12 [0,7; 0,6; 0,6; 0,35]. Тогда для проекции 1-9 классификатор дает ответ "истина", а для проекции 1-12 - ответ "ложь".

На шаге 119 проекция 1-12 исключается из результатов работы, поскольку ответ классификатора для данной проекции не совпадает с решением алгоритма.

На шаге 121 выводится список проекций, в котором каждая проекция содержит информацию об одном и том же пользователе, при этом составляющие проекцию профили относятся к разным социальным графам. Окончательный результат работы изобретения на рассматриваемом примере изображен на фиг.3 и содержит 2 проекции: 1-9 и 6-14. Это означает, что вершины 1 и 9 принадлежат различным социальным графам, но при этом содержат информацию об одном и том же пользователе. То же касается вершин 6 и 14.

Список источников

[1] I. Veldman. *Matching Profiles from Social Network Sites.* Master's thesis, University of Twente, 2009.

[2] Motoyama, M., Varghese, G. *I Seek You - Searching and Matching Individuals In Social Networks.* WIDM '09: Proceeding of the eleventh international workshop on Web information and data management.

[3] Gae-won You, Seung-won Hwang, Zaiqing Nie, Ji-Rong Wen. *SocialSearch: Enhancing Entity Search with Social Network Matching.* EDBT 2011.

[4] Raad, E., Chbeir, R., Dipanda, A. *User Profile Matching in Social Networks.* 13th International Conference on Network-Based Information Systems (NBiS), 2010.

[5] Vosecky, J., Dan Hong, Shen, V.Y. *User identification across multiple social networks.* In Proc. of First International Conference on Networked Digital Technologies, 2009.

[6] http://www.foaf-o-matic.org/

[7] http://www.okkam.org/

[8] http://infolab.stanford.edu/serf/

[9] Parag Singla, Pedro Domingos. *Multi-relational Record Linkage.* KDD Workshop on Multi-Relational Data Mining (pp. 31-48), 2004. Seattle, WA.

[10] Pradeep Ravikumar, John Lafferty. *Quadratic Programming Relaxations for Metric Labeling and Markov Random Field MAP Estimation.* Proceedings of the 23rd International Conference on Machine Learning (Pittsburgh, Pennsylvania, June 25 - 29, 2006). ICML '06, vol. 148. ACM, New York, NY, 737-744.

[11] *Solution directe de l'équation séculaire et de quelques problèmes analogues transcendants*, C. R. Acad. Sci. Paris, 156 (1913), 43-

46.

[12] Golub, G. H. and Van Loan, C. F. *The Singular Value Decomposition*. §2.5.3 in Matrix Computations, 3rd ed. Baltimore, MD: Johns Hopkins University Press, pp. 70-71, 1996.

[13] Geoffrey I. Webb (2000). *MultiBoosting: A Technique for Combining Boosting and Wagging*. Machine Learning, 40(2): 159-196, Kluwer Academic Publishers, Boston.

[14] Ross Quinlan (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

[15] Friedman, J., T. Hastie and R. Tibshirani (1998). *Additive Logistic Regression: a Statistical View of Boosting*.

[16] Ross J. Quinlan: *Learning with Continuous Classes*. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348, 1992.

[17] Winkler, W. E. (1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods (American Statistical Association): 354–359.

[18] Dice, Lee R. (1945). *Measures of the Amount of Ecologic Association Between Species*. Ecology 26 (3): 297–302.

[19] J. Munkres, *Algorithms for the Assignment and Transportation Problems*, Journal of the Society for Industrial and Applied Mathematics, 5(1):32—38, 1957 March.

Формула изобретения

Способ интеграции профилей пользователей онлайновых социальных сетей, отличающийся тем, что вводят все возможные пары профилей, строят модель Условных Случайных Полей из всех профилей и связей между ними, после чего для каждой пары профилей рассчитывают значения похожести их атрибутов с помощью метрик строковой и графовой похожести, основываясь на схемах соответствия для профилей из различных социальных графов, после чего из полученных значений метрик похожести строят вектор признаков, который передается алгоритму машинного обучения, который осуществляет расчет унарной энергии по формуле

$$\Phi(pr(v)|v)=\alpha(v)\cdot(1-profile_{similarity(v,pr(v))})$$

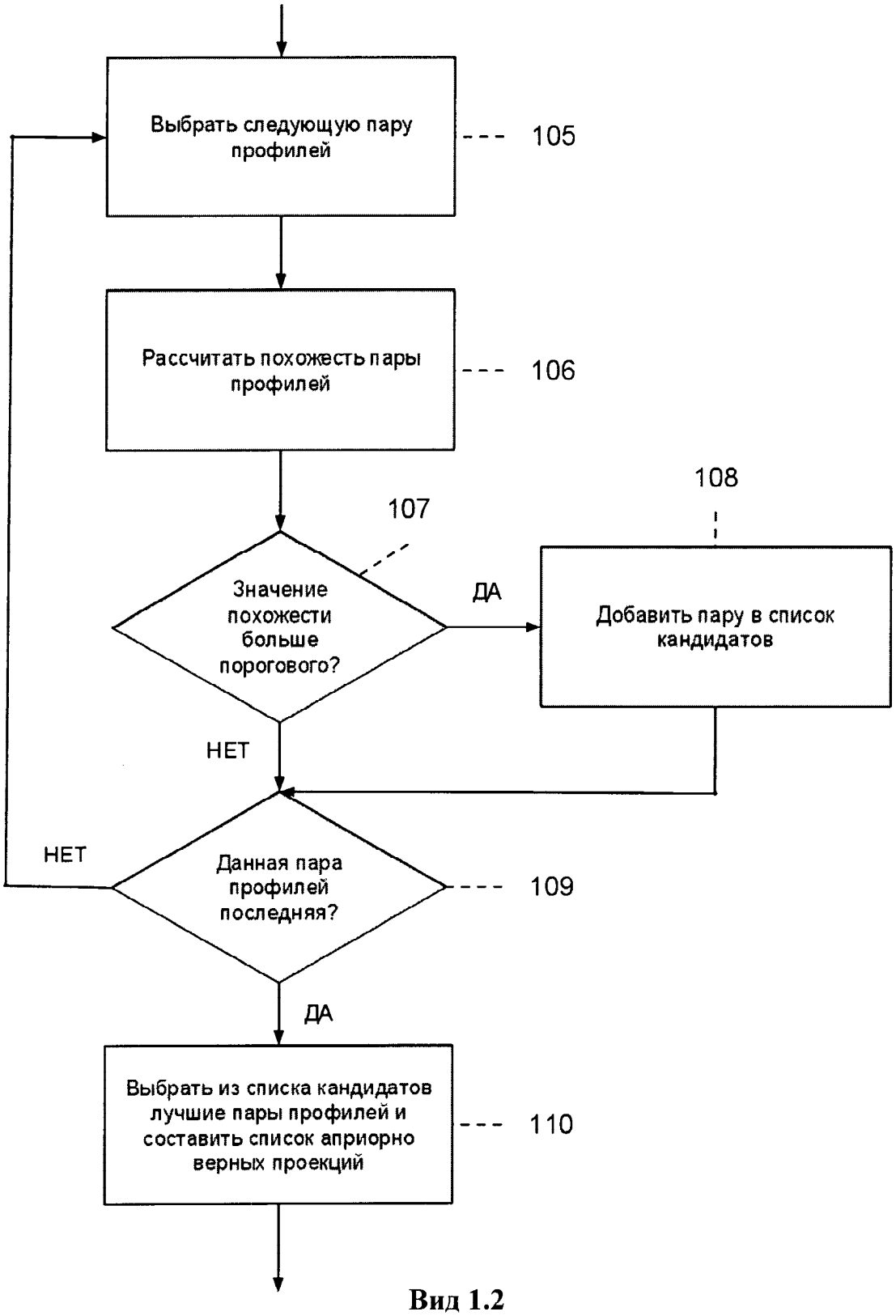
для пары профилей v и pr(v), принадлежащих различным социальным графам, либо бинарной энергии по формуле

$\Psi(pr(v),pr(u)|v,u)=\beta(pr(v),pr(u))\cdot(1-network_{similarity(pr(v),pr(u))})$

для пары профилей pr(v) и pr(u), принадлежащих одному и тому же социальному графу, после чего для каждой пары профилей, в которой профили принадлежат различным социальным графам, рассчитывается похожесть профилей путем применения метрики строковой похожести к атрибуту, который с наибольшей степенью вероятности однозначно идентифицирует профиль, затем проверяется, превышает ли полученное значение похожести профилей заданное пороговое значение, в случае положительного ответа пара профилей заносится в список кандидатов, затем из полученного списка кандидатов выбираются априорно верные проекции путем применения к списку кандидатов алгоритма, который производит последовательный перебор всех пар из списка кандидатов и выдает в качестве результата часть из них, выбранных таким образом, чтобы каждый профиль встречался только в одной паре, а профили, составляющие пару, взаимно максимизировали похожесть друг на друга, затем модель Условных Случайных Полей разбивается на независимые компоненты путем удаления априорно верных проекций, после чего для каждой компоненты модели производится поиск оптимальной конфигурации проекций путем применения итеративного алгоритма, минимизирующего полную энергию модели

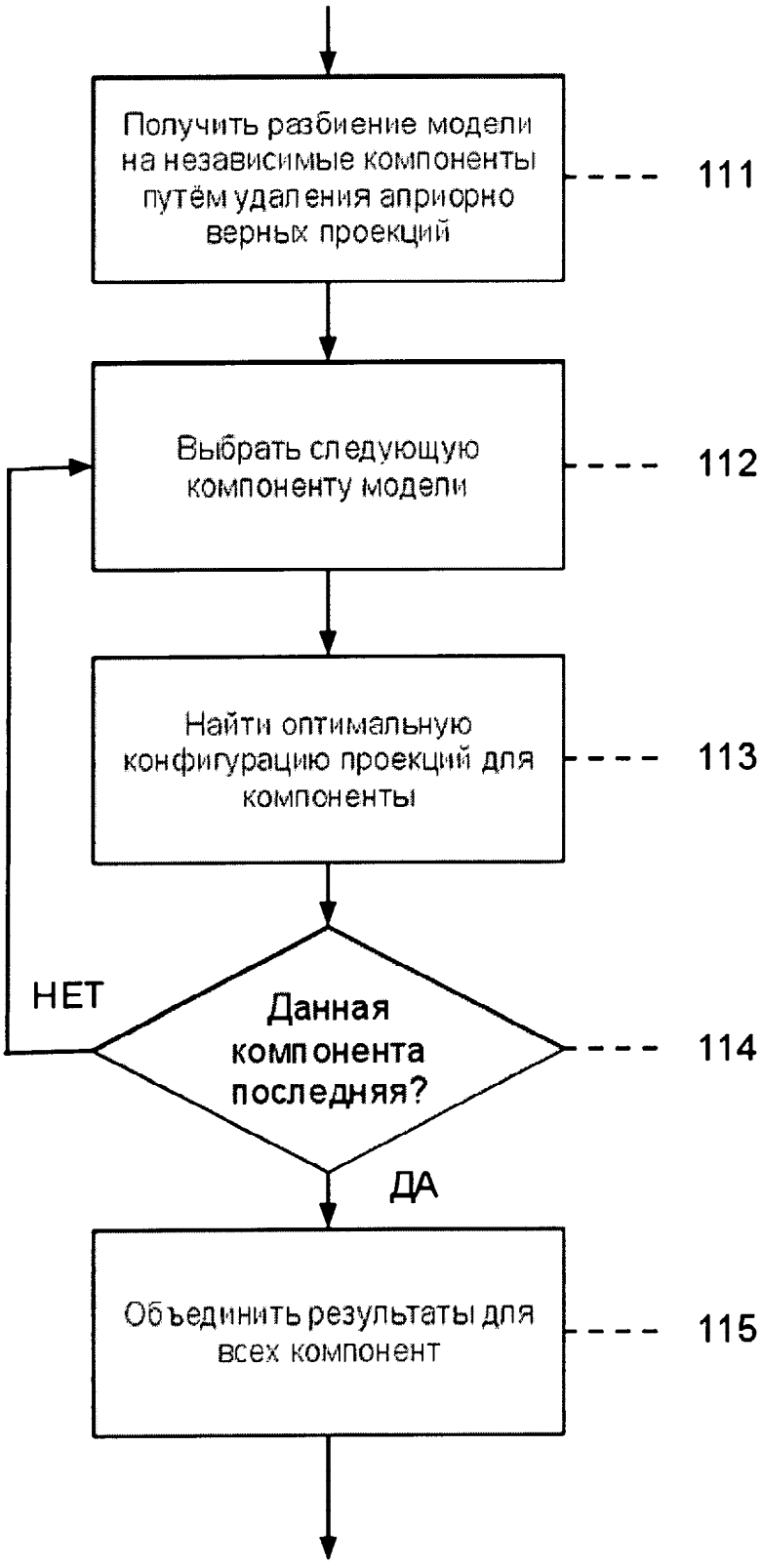
$$E = \sum_{v \in A} \Phi(pr(v) \mid v) + \sum_{(v,u) \in A} \psi(pr(v), pr(u) \mid v, u),$$

с целью получения максимального количества верных проекций профилей, после чего производится объединение списков найденных проекций для всех компонент модели, затем для каждой найденной проекции строится вектор признаков, который передается классификатору, после чего классификатор определяет, является ли данная проекция верной, в случае отрицательного ответа проекция исключается из результатов, после чего выводят список проекций (пар профилей), каждая из которых содержит информацию об одном и том же пользователе, при этом составляющие проекцию профили относятся к различным социальным графам.


Выбрать следующую пару профилей
105
Рассчитать похожесть пары профилей
106
107
108
Значение похожести больше порогового?
ДА
Добавить пару в список кандидатов
НЕТ
НЕТ
Данная пара профилей последняя?
109
ДА
Выбрать из списка кандидатов лучшие пары профилей и составить список априорно верных проекций
110

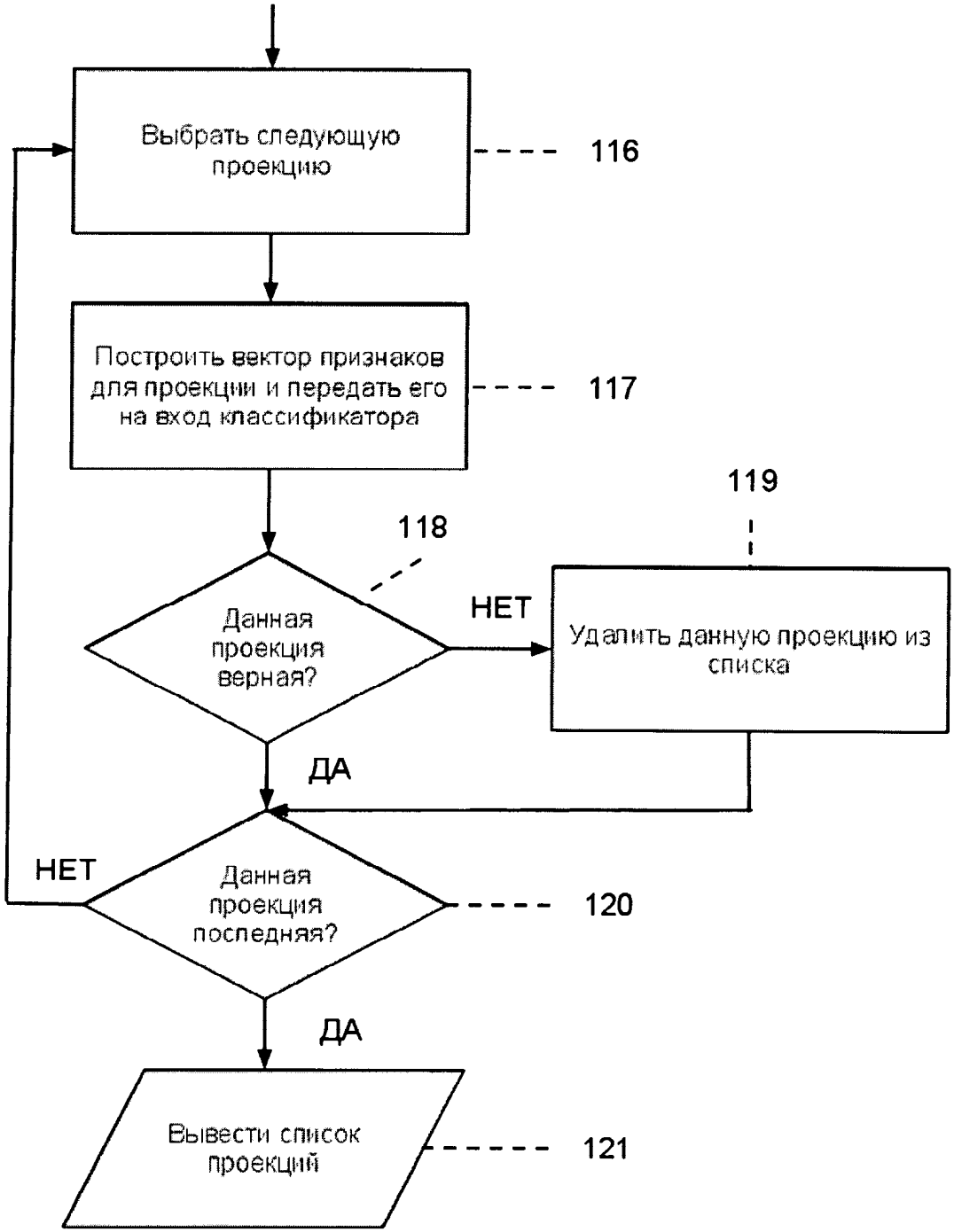
Вид 1.2

Фиг. 1


Получить разбиение модели на независимые компоненты путём удаления априорно верных проекций
111
Выбрать следующую компоненту модели
112
Найти оптимальную конфигурацию проекций для компоненты
113
НЕТ
Данная компонента последняя?
114
ДА
Объединить результаты для всех компонент
115

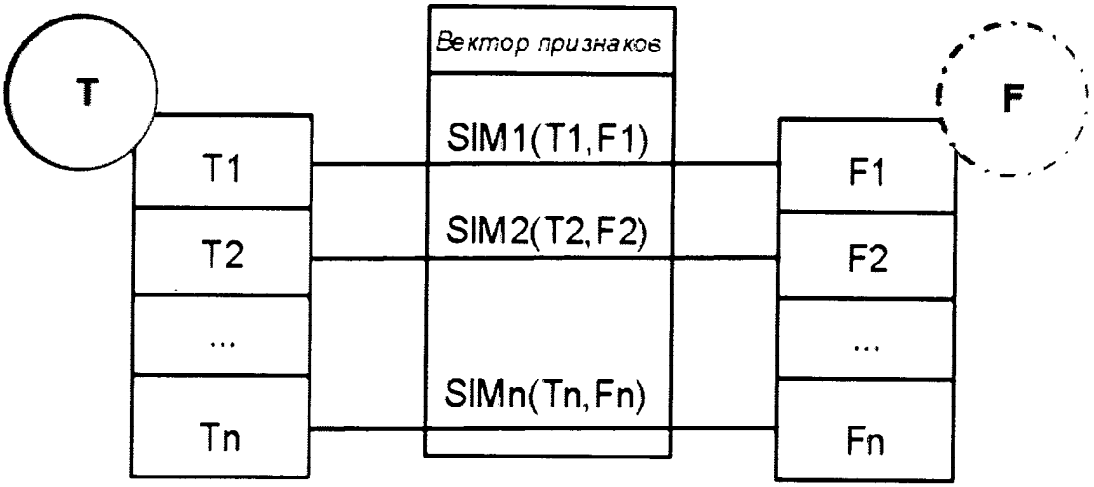
Вид 1.3

Фиг. 1

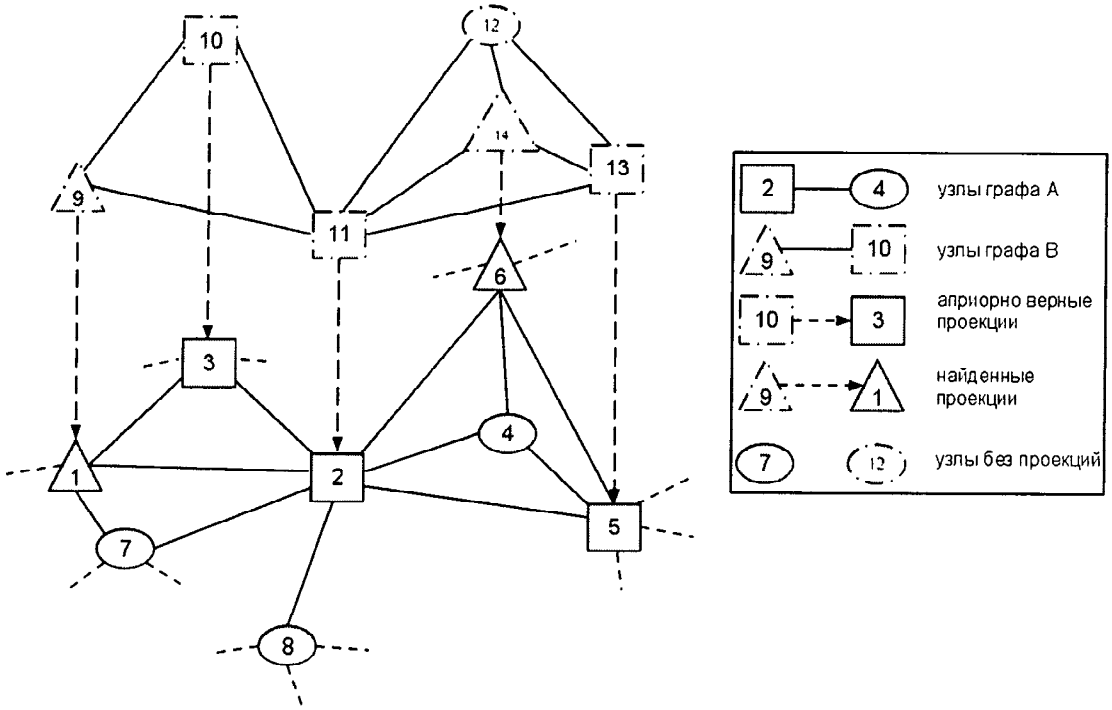

Выбрать следующую проекцию
116
Построить вектор признаков для проекции и передать его на вход классификатора
117
119
118
Данная проекция верная?
НЕТ
Удалить данную проекцию из списка
ДА
НЕТ
Данная проекция последняя?
120
ДА
Вывести список проекций
121

Вид 1.4

Фиг. 1


T
Вектор признаков
F
T1
T2
...
Tn
SIM1(T1,F1)
SIM2(T2,F2)
SIMn(Tn,Fn)
F1
F2
...
Fn

Фиг. 2


узлы графа А
узлы графа В
априорно верные проекции
найденные проекции
узлы без проекций

Фиг. 3