



ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21), (22) Заявка: 2005114658/09, 13.05.2005

(24) Дата начала отсчета срока действия патента:
13.05.2005

(30) Конвенционный приоритет:
14.05.2004 US 10/846,396

(43) Дата публикации заявки: 20.11.2006

(45) Опубликовано: 20.04.2010 Бюл. № 11

(56) Список документов, цитированных в отчете о поиске: Berlin J. et al., Database Schema Matching Using Machine Learning with Feature Selection. Proceedings of Advanced Information Systems Engineering, 14 th International Conference, Caise 2002. RU 2001110088 A, 10.01.2003. Kang J. et al., On Schema Matching with Opaque Column Names and Data Values, SIGMOD 2003, Jun. 9-12, 2003, San Diego, CA. Wang J. (см. прод.)

Адрес для переписки:

129090, Москва, ул. Б.Спасская, 25, стр.3,
ООО "Юридическая фирма Городисский и
Партнеры", пат.пов. Ю.Д.Кузнецову,
рег.№ 595

(72) Автор(ы):

ВЭНЬ Цзи-Жун (US),
МА Вэй-Ин (US)

(73) Патентообладатель(и):

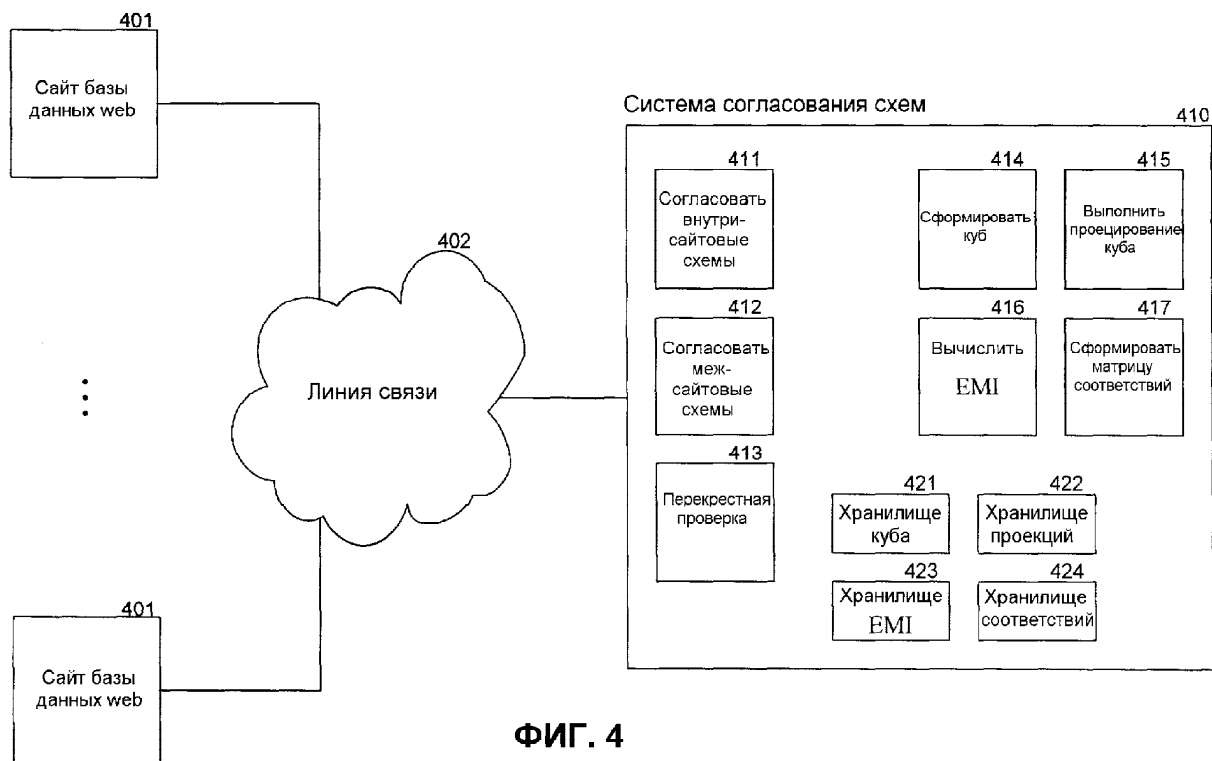
МАЙКРОСОФТ КОРПОРЕЙШН (US)

(54) СПОСОБ И СИСТЕМА СОГЛАСОВАНИЯ СХЕМ БАЗ ДАННЫХ WEB

(57) Реферат:

Изобретение относится к идентификации атрибутов схем баз данных web. Изобретение позволяет идентифицировать атрибуты различных схем, которые представляют семантически один и тот же контент. Подают запрос к базе данных с интерфейсным атрибутом, установленным в значение глобального атрибута. Подсчитывают количество вхождений значения глобального атрибута в качестве значения результирующего атрибута результата запроса для каждой

комбинации глобального атрибута глобальной схемы домена, интерфейсного атрибута интерфейсной схемы и результирующего атрибута результирующей схемы базы данных. Оценивают взаимную информацию между парами схем на основе предоставленных значений подсчета. На основе оценки взаимной информации идентифицируют, какие атрибуты соответствуют друг другу. Сохраняют указание соответствующих атрибутов. 2 н. и 13 з.п. ф-лы, 13 ил., 5 табл.



ФИГ. 4

(56) (продолжение):

et al., Data Extraction and Label Assignment for Web Databases, WWW2003, May 20-24,2003, Budapest, Hungary. Arasu A. et al., Extracting Structred Data from Web-pages, SIGMOD 2003.



FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(19) **RU** (11) **2 386 997** (13) **C2**

(51) Int. Cl.

G06F 7/00 (2006.01)

G06F 17/30 (2006.01)

(12) ABSTRACT OF INVENTION

(21), (22) Application: **2005114658/09, 13.05.2005**

(24) Effective date for property rights:
13.05.2005

(30) Priority:
14.05.2004 US 10/846,396

(43) Application published: **20.11.2006**

(45) Date of publication: **20.04.2010 Bull. 11**

Mail address:

**129090, Moskva, ul. B.Spasskaja, 25, str.3, OOO
"Juridicheskaja firma Gorodisskij i Partnery",
pat.pov. Ju.D.Kuznetsovu, reg.№ 595**

(72) Inventor(s):

**VEhN' Tszi-Zhun (US),
MA Vehj-In (US)**

(73) Proprietor(s):

MAJKROSOFT KORPOREJShN (US)

(54) METHOD AND SYSTEM FOR COORDINATING WEB DATABASE SCHEMES

(57) Abstract:

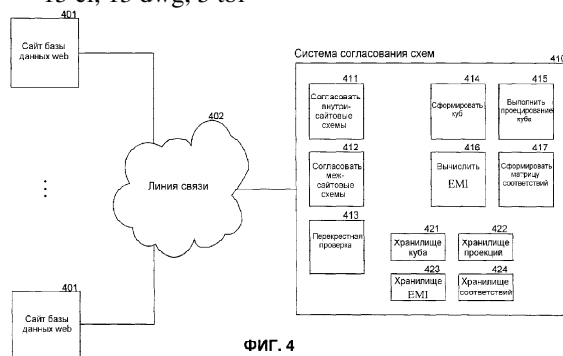
FIELD: information technology.

SUBSTANCE: query is sent to a database with an interface attribute set to the value of the global attribute. The number of times the global attribute value occurs as the value of the resultant attribute of the query result for each combination of the global attribute of the global domain scheme, the interface attribute of the interface scheme and the resultant attribute of the resultant database scheme is counted. Mutual information between pairs of schemes based on the counting values is assessed. Based on the assessment of the mutual information, attributes which correspond to each other are identified. Indicators of corresponding attributes

are stored.

EFFECT: invention enables identification of attributes of different schemes, which semantically offer the same content.

15 cl, 13 dwg, 5 tbl



ФИГ. 4

ОБЛАСТЬ ТЕХНИКИ, К КОТОРОЙ ОТНОСИТСЯ ИЗОБРЕТЕНИЕ

Описываемая технология в целом относится к определению схемы базы данных web.
ПРЕДШЕСТВУЮЩИЙ УРОВЕНЬ ТЕХНИКИ

Всемирная паутина ("web") предоставляет значительные объемы информации,
5 которая доступна через web-страницы. Web-страницы могут содержать либо
статический контент (информационно значимое содержимое), либо динамический
контент. Статический контент обычно относится к информации, которая может
оставаться одинаковой по многим доступам к web-страницам. Динамический контент
10 обычно относится к информации, которую хранят в базе данных web и добавляют
к web-странице в ответ на поисковый запрос. Динамический контент представляет то,
что упоминается как глубокая web или скрытая web.

Многие услуги поисковой машины дают возможность пользователям осуществлять
поиск статического контента web. После того как пользователь подает поисковый
15 запрос или запрос, который включает в себя поисковые термины, услуги поисковой
машины идентифицируют web-страницы, которые могут быть связанными с этими
поисковыми терминами. Эти web-страницы являются результатом поиска. Чтобы
быстро идентифицировать связанные web-страницы, услуги поисковой машины могут
20 поддерживать соответствие ключевых слов web-страницам. Это соответствие может
быть сформировано посредством "ползания" (crawling, процесс автоматического
подбора web-страниц в целях поиска) по web с целью идентификации ключевых слов
каждой web-страницы. Для осуществления "ползания" по сети услуги поисковой
машины могут использовать перечень корневых web-страниц для идентификации
25 всех web-страниц, которые являются доступными через эти корневые web-страницы.
Ключевые слова для любой конкретной web-страницы могут быть идентифицированы
с использованием различных широко известных методик информационного поиска,
например идентификация слов из заголовка, слов, поставляемых в метаданных web-
30 страницы, выделенных слов и так далее.

Эти услуги поисковой машины, однако, обычно не предусматривают поиск
динамического контента, который также считается контентом, в отношении которого
отсутствует возможность осуществления "ползания". Одна проблема, связанная с
поиском динамического контента, состоит в том, что является трудным или
35 невозможным непосредственно получать схемы соответствующих баз данных web без
взаимодействия с web-сайтом, который обеспечивает базу данных web. Схема задает
информацию или атрибуты, которые хранятся в базе данных. Например, база
данных web для продавца книг может содержать схему для каталога ее книг (то есть,
40 базы данных web), которая включает в себя атрибут «заглавие» (title) и атрибут
«автор» (author) для каждой книги. Без знания схемы услугам поисковой машины
будет очень трудно осуществить "ползание" в отношении контента базы данных web,
чтобы определить, какая информация является доступной для поиска. Даже если бы
схема базы данных web была известна, услугам поисковой машины все равно
45 потребовалось бы определять, как осуществить "ползание" в отношении базы
данных web, чтобы извлечь ее контент. Предположив, что поисковая машина может
извлекать контент из баз данных web, услугам поисковой машины все еще потребуется
идентифицировать, когда атрибуты различных схем представляют семантически
50 эквивалентные атрибуты. Например, web-сайты продавца книг могут содержать
каталоги, которые описывают, является ли книга книгой в мягкой обложке, книгой в
твердом переплете или компакт-диском. Web-сайт одного продавца книг может
именовать этот атрибут "типом", а web-сайт другого продавца книг может именовать

этот же самый атрибут "форматом". Чтобы предоставить возможность эффективного поиска динамического контента по многим web-сайтам, услуги поисковой машины должны понимать смысл или семантику атрибутов баз данных web.

Желательно иметь методику, которая позволила бы автоматически идентифицировать схемы, ассоциированные с базами данных web, и идентифицировать атрибуты различных схем, которые представляют семантически один и тот же контент.

КРАТКОЕ ОПИСАНИЕ СУЩНОСТИ ИЗОБРЕТЕНИЯ

Обеспечены способ и система идентификации схем баз данных web. Система согласования схем формирует отображение между интерфейсной схемой и результирующей схемой базы данных web, которую используют для представления основной схемы базы данных. Система согласования схем также формирует отображение, интерфейсных атрибутов и результирующих атрибутов базы данных web на глобальные атрибуты глобальной схемы, семантика которых известна.

Используя эти отображения, услуги поисковой машины могут формулировать запросы с использованием глобальных атрибутов, отображать эти запросы на соответствующие интерфейсные атрибуты, подавать запрос и извлекать значения из результирующих атрибутов, которые соответствуют требуемым глобальным атрибутам.

ПЕРЕЧЕНЬ ЧЕРТЕЖЕЙ

Фиг.1 - схема, иллюстрирующая различные схемы базы данных web для продавца книг.

Фиг.2 - иллюстрация внутрисайтового и межсайтового согласования в одном варианте осуществления.

Фиг.3 - иллюстрация одного прохода разбиения системы согласования схем в одном варианте осуществления.

Фиг.4 - блок-схема, иллюстрирующая компоненты системы согласования схем в одном варианте осуществления.

Фиг.5 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента внутрисайтового согласования в одном варианте осуществления.

Фиг.6 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента формирования куба в одном варианте осуществления.

Фиг.7 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента обновления куба в одном варианте осуществления.

Фиг.8 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента проецирования куба в одном варианте осуществления.

Фиг.9 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента вычисления ЕМІ в одном варианте осуществления.

Фиг.10 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента формирования матрицы соответствия в одном варианте осуществления.

Фиг.11 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента межсайтового согласования в одном варианте осуществления.

Фиг.12 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента вычисления оценки подобия векторов в одном варианте осуществления.

Фиг.13 - блок-схема последовательности операций, иллюстрирующая обработку данных для компонента перекрестной проверки в одном варианте осуществления.

ПОДРОБНОЕ ОПИСАНИЕ СУЩНОСТИ ИЗОБРЕТЕНИЯ

Обеспечены способ и система идентификации схем баз данных web. В одном варианте осуществления система согласования схем формирует отображение между интерфейсной схемой и результирующей схемой базы данных web, которую используют для представления основной схемы базы данных. Интерфейсная схема базы данных web представляет атрибуты базы данных, которые могут использоваться для поиска. Результирующая схема базы данных web представляет атрибуты базы данных, которые выводятся на экран дисплея в качестве части результата поиска.

Отображение указывает, какие интерфейсные атрибуты имеют тот же смысл (также называют «соответствует» или «совпадает»), что и результирующий атрибут. Система согласования схем также формирует отображение интерфейсных атрибутов и результирующих атрибутов базы данных web на глобальные атрибуты глобальной схемы, семантика которых известна. Используя эти отображения, услуги поисковой машины могут формулировать запросы с использованием глобальных атрибутов, отображать эти запросы на соответствующие интерфейсные атрибуты, подавать запрос и извлекать значения из результирующих атрибутов, которые соответствуют требуемым глобальным атрибутам. Таким образом, система согласования схем идентифицирует схемы базы данных web, которые могут использоваться для поиска по базе данных web.

На Фиг.1 показана блок-схема, на которой проиллюстрированы различные схемы базы данных web для продавца книг. База данных web включает в себя схему 101 базы данных, интерфейсную схему 102 и результирующую схему 103. Схема базы данных представляет основную схему базы данных web, которая в этом примере включает в себя атрибуты «заглавие», «автор», «издательство» (Publisher), «стандартный международный номер книги» (ISBN), «формат», и «дата публикации». Web-сайт обеспечивает web-страницу поиска так, чтобы пользователь мог посещать ее для поиска книг. Интерфейсная схема для этой базы данных web включает в себя атрибуты «заглавие», «автор», «формат» и «ISBN». Пользователь может задавать поисковые строки для любой комбинации интерфейсных атрибутов для поиска по базе данных книг. Поле "Ваш поиск" web-страницы дает возможность пользователю осуществлять поиск в пределах всех атрибутов базы данных web. Результат поиска выводят на экран на web-странице результата. Результирующая схема для этой базы данных web включает в себя «заглавие», «автор», «издательство», «формат» и «дата публикации». Результат поиска обычно будет предоставлять множество элементов для каждого элемента базы данных, соответствующего поисковому запросу. Каждый элемент результата обычно содержит значение для каждого из результирующих атрибутов. В этом примере интерфейсная схема содержит атрибут (то есть, ISBN), который не включен в результирующую схему, и результирующая схема содержит атрибут (то есть, «дата публикации»), который не включен в интерфейсную схему.

В дополнение к использованию для базы данных web интерфейсной схемы и результирующей схемы система согласования схем использует также глобальную схему, специфическую для конкретного домена. Глобальная схема для домена представляет набор атрибутов, которые являются общеиспользуемыми базами данных web внутри домена. Например, базы данных web в домене для книг обычно имеют атрибуты, которые включают в себя «заглавие», «автор» и «издательство», а базы данных web в домене для автомобилей обычно имеют атрибуты, которые включают в себя «наименование изготовителя», «модель» и «год». Глобальная схема также может иметь ассоциированные с ней типовые значения глобальных атрибутов.

Например, атрибут «издательство» из домена книг может иметь значения глобального атрибута, которые включают в себя наименования издательств "Random House" и "MIT Press".

5 Чтобы сформировать отображения, система согласования схем первоначально идентифицирует глобальную схему для домена базы данных web и интерфейсную
схему и результирующую схему базы данных web. (Способы идентификации этих схем
описаны ниже). Система согласования схем формирует запросы на основе значений
10 глобальных атрибутов (например, на основе типового набора значений) для
глобальных атрибутов и подает эти запросы через интерфейсную web-страницу на
базу данных web (например, посылая запрос по протоколу передачи
гипертекста (HTTP), что соответствует подаче запроса через web-страницу поиска).
Система согласования схем анализирует результат, представленный web-страницей
15 результата, чтобы определить, какие интерфейсные атрибуты каким результирующим
атрибутам соответствуют ("соответствие интерфейсный-результирующий"), какие
глобальные атрибуты каким интерфейсным атрибутам соответствуют ("соответствие
глобальный-интерфейсный") и какие глобальные атрибуты каким результирующим
атрибутам соответствуют ("соответствие глобальный-результирующий"). Эти
20 соответствия называют "внутрисайтовым" согласованием, поскольку интерфейсная и
результирующая схемы соответствуют схемам одного web-сайта. Система
согласования схем идентифицирует, что интерфейсный атрибут может соответствовать
результирующему атрибуту, на основании того, что значение результирующего
атрибута совпадает со значением интерфейсного атрибута, используемого при поиске.
25 Например, если интерфейсному атрибуту «заглавие» задано значение "Harry Potter"
(Гарри Поттер), то многие элементы результата будут вероятно иметь значение "Harry
Potter" в результирующем атрибуте «заглавие». Напротив, если интерфейсному
атрибуту «автор» задают для поиска значение "Harry Potter", только небольшое
30 количество элементов результата будут, вероятно, иметь значение "Harry Potter" в
интерфейсном атрибуте «заглавие». Как таковой, интерфейсный атрибут «заглавие»,
вероятно, соответствует результирующему атрибуту «заглавие», но интерфейсный
атрибут «автор», вероятно, не соответствует результирующему атрибуту «заглавие».

В одном варианте осуществления система согласования схем может также
35 формировать соответствия между интерфейсными схемами и результирующими
схемами различных web-сайтов. Система согласования схем анализирует результаты
запросов, поданных как описано выше, и идентифицирует, какие интерфейсные
атрибуты схемы одного web-сайта каким интерфейсным атрибутам схемы другого web-
40 сайта соответствуют ("соответствие интерфейсный-интерфейсный") и какие
результирующие атрибуты схемы одного web-сайта каким результирующим
атрибутам схемы другого web-сайта соответствуют ("соответствие результирующий-
результирующий"). Например, система согласования схем может идентифицировать,
что интерфейсный атрибут «тип» одного web-сайта может соответствовать
45 интерфейсному атрибуту «формат» другого web-сайта. Эти соответствия называют
"межсайтовым" согласованием, поскольку схемы согласовываются между
различными web-сайтами. Информация о межсайтовом согласовании может
использоваться при поиске нескольких баз данных web внутри домена. Информация о
50 межсайтовом согласовании также может использоваться, чтобы оказывать помощь в
проверке того, является ли внутрисайтовое согласование корректным.

На Фиг.2 проиллюстрировано внутрисайтовое и межсайтовое согласование в
одном варианте осуществления. В овале 202 представлены схемы, относящиеся к

базам данных web в домене для книг. Каждый из web-сайтов 1...N имеет интерфейсную схему ("IS") и результирующую схему ("RS"), и домен имеет глобальную схему ("GS"). Линии между представлениями схем представляют межсайтовое и внутрисайтовое согласование. Например, линия между IS web-сайта 1 и GS представляет внутрисайтовое соответствие глобальный-интерфейсный, линия между IS web-сайта 1 и RS web-сайта 1 представляет внутрисайтовое соответствие интерфейсный-результирующий, и линия между IS web-сайта 1 и IS web-сайта 2 представляет межсайтовое соответствие интерфейсный-интерфейсный между web-сайтом 1 и web-сайтом 2.

В одном варианте осуществления система согласования схем формирует куб вхождений, который для каждой комбинации глобального атрибута, интерфейсного атрибута и результирующего атрибута базы данных web идентифицирует количество раз, которое глобальное значение атрибута для этого глобального атрибута встречается в этом результирующем атрибуте в случае, когда значение глобального атрибута используется при поиске в качестве значения этого интерфейсного атрибута. Для каждого интерфейсного атрибута система согласования схем подает несколько запросов. Каждый запрос содержит значение этого интерфейсного атрибута, установленное в отличающееся значение глобального атрибута. Например, если глобальные атрибуты включают в себя атрибут «формат», имеющий значения «книга в мягкой обложке», «книга в твердом переплете» и «компакт-диск», и атрибут «автор», имеющий значение «Rowling», то система согласования схем подает запрос, в котором атрибут «заглавие» установлен в «книга в мягкой обложке», запрос, в котором атрибут «заглавие» установлен в «книга в твердом переплете», запрос, в котором атрибут «заглавие» установлен в «компакт-диск», и запрос, в котором атрибут «заглавие» установлен в «Rowling». Для каждого отличающегося интерфейсного атрибута система согласования схем подает запросы для значений глобального атрибута «книга в мягкой обложке», «книга в твердом переплете», «компакт-диск», и «Rowling». Для каждого результата запроса система согласования схем вычисляет количество раз, которое значение глобального атрибута из запроса встречается в качестве значения в каждом результирующем атрибуте. Например, в случае, когда подают запрос, в котором интерфейсный атрибут «заглавие» установлен в «книга в мягкой обложке», вероятно, что будут найдены очень немногие или никакие совпадения, что означает, что интерфейсный атрибут «заглавие» вероятно не соответствует глобальному атрибуту «формат». Напротив, когда подают запрос, в котором интерфейсный атрибут «формат» установлен в «книга в мягкой обложке», вероятно, что многие совпадения будут найдены, а поисковый термин «книга в мягкой обложке» будет найден во многих элементах результата в результирующем атрибуте «формат», что означает, что глобальный атрибут «формат», интерфейсный атрибут «формат» и результирующий атрибут «формат», вероятно соответствуют друг другу. Большое значение счета для конкретной комбинации глобального атрибута, интерфейсного атрибута и результирующего атрибута, особенно относительно других комбинаций, может означать, что эти атрибуты, вероятно, соответствуют, то есть, они представляют один и тот же семантический контент.

После формирования куба вхождений система согласования схем создает матрицы вхождений для соответствия глобальный-интерфейсный, соответствия глобальный-результирующий и соответствия интерфейсный-результирующий. В одном варианте осуществления система согласования схем создает матрицу вхождений посредством проецирования на плоскость некоторого измерения куба вхождений. Чтобы

сформировать матрицу вхождений для соответствия глобальный-интерфейсный, система согласования схем суммирует значение счета вхождений для всех результирующих атрибутов для каждой комбинации глобального атрибута и интерфейсного атрибута. Система согласования схем формирует матрицы вхождений для соответствия глобальный-результатирующий и соответствия интерфейсный-результатирующий таким же образом. В Таблице 1 показан пример матрицы вхождений для соответствия глобальный-интерфейсный.

Таблица 1				
	Title _{GS}	Author _{GS}	Publisher _{GS}	ISBN _{GS}
Author _{IS}	93	[498]	534	0
Title _{IS}	[451]	345	501	0
Publisher _{IS}	62	184	[468]	2
Keyword _{IS} (ключевое слово)	120	248	143	[275]
ISBN _{IS}	0	0	0	258

Несмотря на то, что величина счета является показателем соответствия между парами атрибутов, относительная величина является более показательной для соответствия, чем абсолютная величина. В частности, большое значение счета вхождений может не представлять соответствующие атрибуты. Например, элемент матрицы для Author_{IS} и Publisher_{GS} (534) является наивысшим значением в матрице, но Author_{IS} и Publisher_{GS} семантически не соответствуют друг другу. В целом, для заданного конкретного элемента m_{ij} матрицы его относительная величина между всеми элементами для его интерфейсного атрибута i и глобального атрибута j более важна, чем его абсолютная величина. Например, Keyword_{IS}, который может включать в себя поле "ваш поиск" и который не является действительным атрибутом для домена книг, имеет подобную характеристику для всех глобальных атрибутов, что означает, что не может быть хорошего согласования для любого из глобальных атрибутов. Элемент Publisher_{IS} и Publisher_{GS} (468) не является наивысшим среди элементов для Publisher_{GS}. Однако он относительно больше других элементов для Publisher_{IS}.

Чтобы идентифицировать то, какая пара атрибутов соответствует, система согласования схем оценивает содержимое взаимной информации для пары атрибутов. Взаимную информацию называют также перекрестной энтропией и приростом информации. Система согласования схем предполагает, что каждая схема представляет разбиение базы данных web в соответствии с атрибутами схемы. Пары атрибутов из различных схем, разбиения которых перекрываются в большей степени, вероятно должны соответствовать. В одном варианте осуществления система согласования схем оценивает взаимную информацию между парой атрибутов в соответствии с нижеследующим уравнением:

$$EMI(S_{1i}, S_{2j}) = \frac{m_{ij}}{M} \log \frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M} * \frac{m_{+j}}{M}} \quad (1)$$

в котором EMI является оценкой взаимной информации между i -м атрибутом схемы S_{1i} и j -м атрибутом S_{2j} , M является $\sum_{ij} m_{ij}$, m_{i+} является $\sum_i m_{ij}$ и m_{+j} является $\sum_j m_{ij}$. Матрица EMI для матрицы вхождений согласно Таблицы 1 показана в Таблице 2.

Таблица 2				
	Title _{GS}	Author _{GS}	Publisher _{GS}	ISBN _{GS}
Author _{IS}	-0,04	[0,11]	0,06	0,00
Title _{IS}	[0,19]	-0,03	-0,01	0,00
Publisher _{IS}	-0,03	-0,02	[0,14]	-0,01
Keyword _{IS}	-0,01	0,01	-0,07	0,17
ISBN _{IS}	0,00	0,00	0,00	[0,32]

Система согласования схем определяет соответствие между атрибутами в случае, когда один элемент матрицы ЕМІ больше других элементов для этого же интерфейсного атрибута (то есть в этой же строке), а также больше других элементов для этого же глобального атрибута (то есть в этом же столбце). Соответствующие атрибуты имеют большее перекрытие в содержимом информации (смысле информации) между собой, чем их перекрытие с другими атрибутами противоположной схемы, как показано посредством квадратных скобок. Например, элемент матрицы ЕМІ для Author_{IS} и Author_{GS} (то есть 0,11) является наибольшим как для интерфейсных атрибутов «автор», так и глобальных атрибутов «автор», и он является корректным соответствием. Соответствие атрибутов представлено посредством нижеследующего уравнения:

$$\text{MAP}(S_{1,i}, S_{2,j}) = \text{соответствие, если } e_{ij} \geq e_{ik} \mid k \neq j \text{ и } e_{ij} \geq e_{ik} \mid k \neq i \quad (2)$$

при этом MAP показывает, соответствует ли i-й атрибут схемы S₁ j-му атрибуту схемы S₂, и e_{ij} является элементом матрицы ЕМІ для i-го атрибута схемы S₁ и j-го атрибута схемы S₂.

В одном варианте осуществления система согласования схем идентифицирует соответствия между атрибутами различных баз данных web. Система согласования схем идентифицирует соответствия на основании подобия векторов из соответствующих матриц вхождений для баз данных web. Например, Таблица 3 представляет таблицу вхождений глобальный-интерфейсный для схемы S₁, и Таблица 4 представляет таблицу вхождений глобальный-интерфейсный для схемы S₂. Глобальной схемой GS является {Title, Author, Publisher, ISBN}, интерфейсной схемой IS₁ для web-сайта 1 является {Author1, Title1, Publisher1, Keyword1, ISBN1} и интерфейсной схемой IS₂ для web-сайта 2 является {Title2, Author2, ISBN2}.

Таблица 3				
	T _G	A _G	P _G	I _G
A ₁	93	498	534	0
T ₁	451	345	501	0
P ₁	62	184	468	2
K ₁	120	248	143	275
I ₁	0	0	0	258

Таблица 4				
	T _G	A _G	P _G	I _G
T ₂	166	177	118	0
A ₂ (P)	39	331	406	0
I ₂	0	0	0	18

Атрибут A1 представлен вектором первой строки Таблицы 3, и атрибут A2

представлен вектором второй строки Таблицы 4. Система согласования схем вычисляет сходство двух атрибутов, используя нижеследующее уравнение:

$$EVS(S_{1,i}, S_{2,j}) = \frac{\sum_k a_{ik} b_{jk}}{\sqrt{\sum_k a_{ik}^2} * \sqrt{\sum_k b_{jk}^2}} \quad (3)$$

в котором EVS является оценкой подобия векторов между i-м атрибутом схемы S_1 и j-м атрибутом схемы S_2 , a_{ik} представляет значения матрицы вхождений для схемы S_1 , и b_{jk} представляет значения матрицы вхождений для схемы S_2 .

В Таблице 5 представлены оценки подобия векторов, выведенные на основании Таблицы 3 и Таблицы 4.

Таблица 5			
	T_2	$A_2(P)$	I_2
A_1	0,84	[0,99]	0
T_1	[0,96]	0,84	0
P_1	0,71	0,95	0,01
K_1	0,72	0,67	0,66
I_1	0	0	[1,00]

Система согласования схем определяет соответствие между атрибутами в случае, когда один элемент матрицы EVS больше других элементов для того же интерфейсного атрибута одного web-сайта, а также больше других элементов для того же интерфейсного атрибута другого web-сайта. Квадратными скобками по Таблице 5 помечены наибольшие значения подобия и в ее строке, и в ее столбце, что также показывает корректное соответствие. Несмотря на то, что второй атрибут из IS_2 , $Author_2$, некорректно сопоставлен с $Publisher_2$ из GS , система согласования схем использует межсайтовое согласование, чтобы внести исправление в соответствие.

В одном варианте осуществления система согласования схем перекрестно проверяет соответствие глобальный-интерфейсный, соответствие глобальный-результатирующий, соответствие интерфейсний-результатирующий, соответствие интерфейсний-интерфейсный и соответствие результирующий-результатирующий, чтобы идентифицировать и исправить соответствия, которые могут быть некорректными. Система согласования схем кластеризует интерфейсные атрибуты (и подобным образом результирующие атрибуты) в множество кластеров на основании глобальных атрибутов, которым они соответствуют. Например, атрибуты различных баз данных web, для которых было установлено соответствие некоторому глобальному атрибуту, представляют один кластер. Эта кластеризация основана на внутрисайтовом согласовании. Межсайтовое согласование также может быть использовано, чтобы осуществить перекрестную проверку кластеров. Если внутрисайтовое и межсайтовое согласование были полностью корректны, то каждый атрибут базы данных web будет отображаться только на те атрибуты других баз данных web, которые находятся внутри этого же кластера. Другими словами, атрибуты из баз данных web будут согласованно отображаться друг на друга и на глобальные атрибуты. В одном варианте осуществления система согласования схем представляет атрибуты схем базы данных web в виде вершин и межсайтовое согласование в виде ребер между вершинами. Система согласования схем разбивает вершины так, чтобы срез ребер был минимизирован. Срез ребер является суммой весов всех ребер (например, каждое ребро имеет одинаковый вес) между элементами разбиения. Посредством минимизации среза ребер система согласования схем

минимизирует количество ребер между вершинами для различных кластеров.

В одном варианте осуществления система согласования схем аппроксимирует минимизацию среза ребер посредством использования начальных кластеров в качестве начального разбиения и перемещения вершин из одного кластера в другой до тех пор, пока количество сечений не уменьшится. Обычно вершину перемещают в кластер, в котором находятся большинство ее соседей. Соседние вершины между собой имеют ребро. Поскольку вершина должна быть перемещена, если перемещают многих из ее соседей, система согласования схем может использовать многократные проходы так, чтобы срез ребер сходил к локальному оптимуму. Если срез ребер сходится, система согласования схем разрешает межкластерное соответствие между атрибутами A_i web-сайта S_1 и B_j web-сайта S_2 , содержащимися в двух кластерах C_1 и C_2 , отбрасывая межкластерное соответствие и повторно устанавливая соответствие A_i с атрибутом B_k web-сайта S_2 , который находится в кластере C_1 , или наоборот.

На фиг.3 проиллюстрирован один проход разбиения системы согласования схемы в одном варианте осуществления. В этом примере глобальная схема содержит два атрибута { Author, Publisher }, и пять баз данных web содержат атрибуты $IS_1=\{A_a\}$, $IS_2=\{B_a, B_p\}$, $IS_3=\{C_a, C_p\}$, $IS_4=\{D_a, D_p\}$ и $IS_5=\{E_a, E_p\}$. Кластеры 301 и 302 иллюстрируют начальные кластеры атрибутов (представленные посредством вершин) на основе того, какому глобальному атрибуту они соответствуют (по внутрисайтовому согласованию), и ребра между парами атрибутов означают, что атрибуты соответствуют (по межсайтовому согласованию). В начальном состоянии атрибут A_a некорректно сопоставлен глобальному атрибуту Publisher и также некорректно сопоставлен B_p , в то же время он был корректно сопоставлен трем другим атрибутам в категории Author. Поэтому система согласования схем для уменьшения количества ребер перемещает A_a между кластерами из 3 в 1.

Перемещение исправляет согласующий атрибут A_a с Publisher на глобальный атрибут Author. После перемещения система согласования схем удаляет ребро между A_a и B_p и добавляет новое ребро между A_a и B_a (атрибут web-сайта 2, который сопоставлен глобальному атрибуту Author). Кластеры 311 и 312 представляют исправленные соответствия.

Глобальные схемы, интерфейсные схемы и результирующие схемы могут быть идентифицированы с использованием различных способов. Некоторые способы идентификации глобальных схем основываются на именах атрибутов и структуре элементов. (См. S. Castano, V. Antonellis and S. Vimercati. Global Viewing of Heterogeneous Data Sources. IEEE Trans. Data and Knowledge Eng., vol. 13, no. 2, 2001; и B. He, C.C. Chang. Statistical Schema Matching across Web Query Interfaces. Proc. ACM SIGMOD Conf, 2003, которые тем самым включены в настоящее описание путем ссылки.) Другие способы основываются на формальных онтологиях. (См. B. He, C.C. Chang. Statistical Schema Matching across Web Query Interfaces. Proc. ACM SIGMOD Conf, 2003; и F. Hakimpour, A. Geppert. Global Schema Generation Using Formal Ontologies. Proc. 21st Conf. on Conceptual Modeling, 2002, которые тем самым включены в настоящее описание путем ссылки.) Типовые значения глобальных атрибутов могут быть собраны из различных типовых баз данных web или сформированы вручную. Интерфейсная схема базы данных web может быть идентифицирована на основании относящихся к вводу данных тегов (неотображаемых элементов разметки), web-страницы запроса, как определено спецификацией языка гипертекстовой разметки (HTML). Некоторые способы идентификации результирующей схемы формируют объекты-оболочки, чтобы извлекать содержимое вложенных полуструктурированных данных из динамических,

формируемых по шаблону web-страниц. (См. A. Arasu, H. Garcia-Molina. Extracting Structured Data from Web Pages. Proc. ACM SIGMOD Conf., 2003; C.H. Chang, S.C. Lui. IEPAD: Information Extraction based on Pattern Discovery. Proc. 10th World Wide Web Conf, 681-688, 2001; V. Crescenzi, G. Mecca P. Merialdo. ROADRUNNER: Towards Automatic Data
 5 Extraction from Large Web Sites. Proc. 27th VLDB. Conf, 109-118, 2001; и J. Wang, F. Lochovsky. Data Extraction and Label Assignment for Web Databases. Proc. 12th World Wide Web Conf, 187-196, 2003, которые тем самым включены в настоящее описание путем ссылки). Один способ формирует объект-оболочку регулярного выражения на
 10 основании вложенного обнаружения повторяемого образа в страницах HTML. (См. J. Wang, F. Lochovsky. Data Extraction and Label Assignment for Web Databases. Proc. 12th World Wide Web Conf., 187-196, 2003, которое тем самым включено в настоящее описание путем ссылки). Специалист в данной области техники оценит тот факт, что
 15 каждая из этих схем также может быть идентифицирована вручную или комбинацией ручного и автоматизированного средств.

На фиг.4 показана блок-схема, на которой в одном варианте осуществления проиллюстрированы компоненты системы согласования схем. Система 410 согласования схем соединена с различными сайтами 401 баз данных web через линию
 20 связи 402. Система согласования схем включает в себя компонент 411 внутрисайтового согласования, компонент 412 межсайтового согласования, компонент 413 перекрестной проверки, компонент 414 формирования куба, компонент 415 проецирования куба, компонент 416 вычисления ЕМІ и компонент 417
 25 формирования матрицы соответствий. Система согласования схем также включает в себя хранилище 421 куба, хранилище 422 проекций, хранилище 423 ЕМІ и хранилище 424 соответствий. Компонент внутрисайтового согласования вызывает компонент формирования куба, чтобы сформировать куб вхождений. и вызывает
 30 компонент проецирования куба, чтобы сформировать матрицы вхождений глобальный-интерфейсный, глобальный-результатирующий и интерфейсный-результатирующий. Компонент внутрисайтового согласования вызывает также компонент вычисления ЕМІ, чтобы вычислить оценку взаимной информации на основании матриц вхождений, и вызывает компонент формирования матрицы
 35 соответствий, чтобы идентифицировать, какие пары атрибутов соответствуют. Компонент межсайтового согласования использует матрицы вхождений, чтобы вычислить оценку подобия векторов, и вызывает компонент формирования матрицы соответствий, чтобы идентифицировать соответствия. Компонент перекрестной
 40 проверки изменяет соответствие для атрибутов, которые выявляются согласованными некорректно. Хранилище куба содержит кубы вхождений, хранилище проекций содержит матрицы вхождений, хранилище ЕМІ содержит матрицы ЕМІ и хранилище соответствий содержит матрицы соответствий.

Вычислительное устройство, на котором реализована система согласования схем, может включать в себя центральный процессор, память, устройства ввода данных
 45 (например, клавиатуру и указательные устройства), устройства вывода (например, устройства вывода на экран) и устройства хранения данных (например, дисководы). Память и устройства хранения данных являются машиночитаемыми носителями, которые могут содержать команды, реализующие систему согласования схем. Кроме
 50 того, структуры данных и структуры сообщений могут храниться или передаваться средой передачи данных, такой как сигнал на линии связи. Могут использоваться различные линии связи, такие как сеть Интернет, локальная сеть, глобальная сеть, или двухточечное соединение по коммутируемым телефонным каналам.

Система согласования схем может быть реализована в различных операционных средах, которые включают в себя персональные компьютеры, серверные компьютеры, карманные или портативные устройства, многопроцессорные системы, микропроцессорные системы, программируемую бытовую электронику, сетевые ПК, миникомпьютеры, универсальные ЭВМ, распределенные вычислительные среды, которые включают в себя любые из вышеупомянутых систем или устройств, и подобное.

Система согласования схем может быть описана в общем контексте машиноисполнимых команд, например, программных модулей, исполняемых одним или несколькими компьютерами или другими устройствами. Обычно программные модули включают в себя процедуры, программы, объекты, компоненты, структуры данных и т.д., которые выполняют конкретные задачи или реализуют определенные абстрактные типы данных. Обычно функциональные возможности программных модулей могут быть объединенными или распределенными, как требуется в различных вариантах осуществления.

На фиг.5 показана блок-схема последовательности операций, на которой проиллюстрирована работа для компонента внутрисайтового согласования в одном варианте осуществления. Компонент идентифицирует соответствия глобальный-интерфейсный, глобальный-результатирующий и интерфейсный-результатирующий для базы данных web. На этапе 501 компонент вызывает компонент формирования куба, чтобы сформировать куб вхождений. На этапах 502-506 компонент осуществляет в цикле выбор пары схем (то есть глобальной и интерфейсной, глобальной и результирующей и интерфейсной и результирующей) и формирует матрицу соответствий, представляющую соответствие каждой пары. На этапе 502 компонент выбирает следующую пару схем. На этапе 503 ветвления, если все пары схем были уже выбраны, компонент завершает работу, иначе компонент продолжает работу на этапе 504. На этапе 504 компонент вызывает компонент проецирования куба, чтобы сформировать матрицу вхождений для выбранной пары схем. На этапе 505 компонент вызывает компонент вычисления ЕМІ, чтобы оценить взаимную информацию между парами атрибутов для выбранной пары схем. На этапе 506 компонент вызывает компонент формирования матрицы соответствий, чтобы сформировать матрицу соответствий, указывающую соответствия атрибутов для выбранной пары схем. Компонент затем переходит в цикле на этап 502, чтобы выбрать следующую пару схем.

На фиг.6 показана блок-схема последовательности операций, на которой проиллюстрирована работа компонента формирования куба в одном варианте осуществления. Компонент формирует куб вхождений для базы данных web на основе глобальной схемы, интерфейсной схемы и результирующей схемы. Куб вхождений является трехмерной матрицей, которая отображает в значение счета каждую комбинацию глобального атрибута, интерфейсного атрибута и результирующего атрибута. Значение счета представляет собой количество раз, которое элемент результата для запроса с этим интерфейсным атрибутом, установленным в глобальное значение атрибута этого глобального атрибута, имел это значение глобального атрибута в этом результирующем атрибуте. На этапе 601 компонент выбирает следующий глобальный атрибут. На этапе 602 ветвления, если все глобальные атрибуты уже выбраны, то компонент осуществляет возврат, иначе компонент продолжает работу на этапе 603. На этапе 603 компонент выбирает следующее значение глобального атрибута для выбранного глобального атрибута. На этапе 604

ветвления, если все глобальные значения атрибута для выбранного глобального атрибута были уже выбраны, то компонент переходит в цикле на этап 601, чтобы выбрать следующий глобальный атрибут, иначе компонент продолжает работу на этапе 605. На этапах 605-609 компонент осуществляет цикл выбора каждого интерфейсного атрибута и подачи запроса с этим интерфейсным атрибутом, установленным в выбранные значения глобального атрибута. Специалист в данной области техники оценит тот факт, что область значений для некоторых интерфейсных атрибутов может быть ограничена. Например, если интерфейсный атрибут представлен посредством элемента SELECT (выбор) языка HTML, то область его значений может быть ограничена значениями в ассоциированном элементе OPTION (вариант выбора). В таком случае компонент может подавать запросы только для значений глобального атрибута, которые являются "подобными" значению варианта выбора. Значение глобального атрибута может считаться подобным, если оно содержит значение варианта выбора. Специалист в данной области техники оценит тот факт, что могут использоваться другие меры подобия. Запросы для элементов CHECKBOX и RADIOBOX могут быть обработаны таким же образом. Так как область значений для TEXTBOX может быть неизвестной, компонент может исчерпывающе подавать запросы, используя все значения глобального атрибута для интерфейсного атрибута, представленного посредством TEXTBOX. В одном варианте осуществления компонент устанавливает значение только для одного интерфейсного атрибута для каждого запроса. Значения других интерфейсных атрибутов могут содержать значение по умолчанию, как определено web-сайтом. На этапе 605 компонент выбирает следующий интерфейсный атрибут. На этапе 606 ветвления, если все интерфейсные атрибуты были уже выбраны, компонент переходит в цикле на этап 603, чтобы выбрать следующее глобальное значение атрибута для выбранного глобального атрибута. На этапе 607 компонент формулирует запрос, используя выбранный интерфейсный атрибут и выбранное значение глобального атрибута. На этапе 608 компонент подает сформулированный запрос на web-сайт. На этапе 609 компонент обновляет куб вхождений на основании результата запроса и затем осуществляет переход в цикле на этап 605, чтобы выбрать следующий интерфейсный атрибут.

На фиг.7 показана блок-схема последовательности операций, иллюстрирующая работу компонента обновления куба в одном варианте осуществления. Компоненту передают индикатор глобального атрибута, значение глобального атрибута и интерфейсный атрибут и результат запроса. На этапе 701 компонент выбирает следующий элемент или строку результата. На этапе 702 ветвления, если все элементы результата уже выбраны, то компонент осуществляет возврат, иначе компонент продолжает работу на этапе 703. На этапе 703 компонент выбирает следующий результирующий атрибут или столбец. На этапе 704 ветвления, если все результирующие атрибуты уже выбраны, то компонент переходит в цикле на этап 701, чтобы выбрать следующий элемент результата, иначе компонент продолжает работу на этапе 705. На этапе 705, если значение глобального атрибута равно значению выбранного результирующего атрибута для выбранного элемента, то компонент продолжает работу на этапе 706, иначе компонент переходит в цикле на этап 703, чтобы выбрать следующий результирующий атрибут для выбранного элемента. На этапе 706 компонент увеличивает значение счета в кубе вхождений для переданного глобального атрибута, переданного интерфейсного атрибута и выбранного результирующего атрибута. Затем компонент переходит в цикле на этап 703, чтобы

выбрать следующий результирующий атрибут для выбранного элемента.

На фиг.8 показана схема последовательности операций, иллюстрирующая работу компонента проецирования куба в одном варианте осуществления. В этом варианте осуществления компонент формирует матрицу вхождений для соответствия 5 глобальный-интерфейсный. Система согласования схем может формировать матрицы вхождений для соответствия глобальный-результирующий и соответствия интерфейсный-результирующий таким же образом. В этом варианте осуществления компонент суммирует значения счета результирующих атрибутов для пары 10 глобального атрибута и интерфейсного атрибута, чтобы выполнить проецирование трехмерного куба вхождений в двумерную матрицу соответствий. Специалист в данной области техники оценит тот факт, что могут использоваться способы проецирования, отличные от прямого суммирования. Например, компонент может использовать взвешенное суммирование, в котором веса основаны на достоверности, 15 выведенной в течение автоматической идентификации результирующей схемы. На этапе 801 компонент выбирает следующий глобальный атрибут. На этапе 802 ветвления, если все глобальные атрибуты уже выбраны, то компонент осуществляет возврат, иначе компонент продолжает работу на этапе 803. На этапе 803 компонент 20 выбирает следующий интерфейсный атрибут. На этапе 804 ветвления, если все интерфейсные атрибуты уже выбраны, то компонент переходит в цикле на этап 801, чтобы выбрать следующий глобальный атрибут, иначе компонент продолжает работу на этапе 805. На этапе 805 компонент выбирает следующий результирующий атрибут. На этапе 806 ветвления, если все результирующие атрибуты уже выбраны, то 25 компонент переходит в цикле на этап 803, чтобы выбрать следующий интерфейсный атрибут, иначе компонент продолжает работу на этапе 807. На этапе 807 компонент увеличивает значение счета в матрице вхождений для выбранного интерфейсного атрибута и глобального атрибута на значение счета из куба вхождений для 30 выбранного глобального атрибута, интерфейсного атрибута и результирующего атрибута. Компонент затем переходит в цикле на этап 805, чтобы выбрать следующий результирующий атрибут.

На фиг.9 показана схема последовательности операций, иллюстрирующая работу компонента вычисления ЕМІ в одном варианте осуществления. Этот компонент 35 использует Уравнение 1, чтобы оценить взаимную информацию для пар атрибутов в матрице вхождений. Специалист в данной области техники оценит тот факт, что могут использоваться различные способы, чтобы оценивать вероятность соответствия пар атрибутов. Компоненту передают матрицу вхождений и он возвращает таблицу ЕМІ. На этапе 901 компонент вычисляет сумму всех значений счетчиков в таблице 40 вхождений. На этапе 902 компонент вычисляет сумму значений счета в каждой строке матрицы вхождений. На этапе 903 компонент вычисляет сумму значений счета в каждом столбце матрицы вхождений. На этапах 904-908 компонент осуществляет цикл выбора каждой пары атрибутов из матрицы вхождений и определяет вероятность 45 соответствия атрибутов. На этапе 904 компонент выбирает следующую строку матрицы вхождений. На этапе 905 ветвления, если все строки матрицы вхождений уже выбраны, компонент осуществляет возврат, иначе компонент продолжает работу на этапе 906. На этапе 906 компонент выбирает следующий столбец матрицы вхождений. 50 На этапе 907 ветвления, если все столбцы матрицы вхождений уже выбраны, то компонент переходит в цикле на этап 904, чтобы выбрать следующую строку матрицы вхождений, иначе компонент продолжает работу на этапе 908. На этапе 908 компонент вычисляет оценку взаимной информации для атрибутов, представленных

выбранной строкой и столбцом. Компонент затем переходит в цикле на этап 906, чтобы выбрать следующий столбец.

На фиг.10 показана схема последовательности операций, иллюстрирующая работу компонента формирования матрицы соответствий в одном варианте осуществления.

Компоненту передают матрицу, например матрицу ЕМІ, которая указывает вероятность соответствия пар атрибутов. Если вероятность для пары атрибутов является наивысшей вероятностью для обоих атрибутов (например, наивысшая в строке, представляющей один атрибут, и наивысшая в столбце, представляющем другой атрибут), компонент решает, что атрибуты соответствуют. На этапе 1001 компонент выбирает следующую строку переданной матрицы. На этапе 1002 ветвления, если все строки переданной таблицы уже выбраны, то компонент осуществляет возврат, иначе компонент продолжает работу на этапе 1003. На этапе 1003 компонент выбирает следующий столбец переданной матрицы. На этапе 1004 ветвления, если все столбцы переданной матрицы уже выбраны, то компонент переходит в цикле на этап 1001, чтобы выбрать следующую строку переданной матрицы, иначе компонент продолжает работу на этапе 1005. На этапе 1005 ветвления, если значение для выбранной строки и столбца является наивысшим в этой строке, то компонент продолжает работу на этапе 1006, иначе компонент переходит в цикле на этап 1003, чтобы выбрать следующий столбец. На этапе 1006 ветвления, если значение для выбранной строки и столбца является наивысшим внутри этого столбца, то компонент продолжает работу на этапе 1007, иначе компонент переходит в цикле на этап 1003, чтобы выбрать следующий столбец. На этапе 1007 компонент устанавливает значение матрицы соответствия для выбранной строки и столбца, чтобы указать соответствие, и затем переходит в цикле на этап 1003, чтобы выбрать следующий столбец выбранной строки.

На фиг.11 показана блок-схема последовательности операций, иллюстрирующая работу компонента межсайтового согласования в одном варианте осуществления. Компонент идентифицирует, какие атрибуты (интерфейсные и результирующие) одного web-сайта каким атрибутам другого web-сайта соответствуют. Компонент использует матрицу вхождений для соответствия глобальный-интерфейсный web-сайтов, чтобы идентифицировать согласования для схем интерфейса и таблицы вхождений для соответствия глобальный-результирующий web-сайтов, чтобы идентифицировать согласования для результирующих схем. На этапе 1101 компонент вызывает компонент формирования куба, чтобы сформировать куб вхождений для web-сайта А. На этапе 1102 компонент вызывает компонент проецирования куба, чтобы сформировать матрицы вхождений для web-сайта А. На этапе 1103 компонент вызывает компонент формирования куба, чтобы сформировать куб вхождений для web-сайта В. На этапе 1104 компонент вызывает компонент проецирования куба, чтобы сформировать матрицы вхождений для web-сайта В. На этапе 1105 компонент вызывает компонент вычисления оценки подобия векторов для интерфейсных атрибутов, чтобы сформировать вероятность того, что пары интерфейсных атрибутов для web-сайта А и web-сайта В соответствуют. Специалист в данной области техники оценит тот факт, что могут использоваться многие различные способы, чтобы оценивать эту вероятность, и что подобие векторов является лишь одним примером. На этапе 1106 компонент вызывает компонент формирования матрицы соответствий, передавая матрицу оценок подобия векторов для интерфейсных атрибутов, для того, чтобы сформировать матрицу, указывающую, какие пары интерфейсных атрибутов соответствуют. На этапе 1107 компонент вызывает компонент вычисления оценки

подобия векторов, чтобы сформировать матрицу оценок подобия векторов для
результатирующих атрибутов. На этапе 1108 компонент вызывает компонент
формирования матрицы соответствий, чтобы сформировать матрицу, указывающую,
какие пары результирующих атрибутов соответствуют. Затем компонент завершает
5 работу.

На фиг.12 показана блок-схема последовательности операций, иллюстрирующая
работу компонента вычисления оценки подобия векторов в одном варианте
осуществления. Компоненту передают матрицу вхождений для соответствия
10 интерфейсный-интерфейсный или соответствия результирующий-результатирующий и он
определяет вероятность соответствия каждой пары атрибутов. На этапе 1201
компонент выбирает следующий атрибут web-сайта А. На этапе 1202 ветвления, если
все атрибуты web-сайта уже были выбраны, то компонент осуществляет возврат,
иначе компонент продолжает работу на этапе 1203. На этапе 1203 компонент
15 выбирает следующий атрибут web-сайта В. На этапе 1204 ветвления, если все
атрибуты web-сайта В уже выбраны, то компонент переходит в цикле на этап 1201,
чтобы выбрать следующий атрибут web-сайта А, иначе компонент продолжает работу
на этапе 1205. На этапе 1205 компонент вычисляет в соответствии с Уравнением 3
оценку подобия векторов для выбранных атрибутов и затем переходит в цикле на
20 этап 1203, чтобы выбрать следующий атрибут web-сайта В.

На фиг.13 показана блок-схема последовательности операций, иллюстрирующая
работу компонента перекрестной проверки в одном варианте осуществления. В
случае, когда межсайтовые соответствия указывают, что внутрисайтовое соответствие
25 является некорректным, компонент изменяет соответствие атрибутов. На этапе 1301
компонент выбирает следующий глобальный атрибут. На этапе 1302 ветвления, если
все глобальные атрибуты уже выбраны, то компонент завершает работу, иначе
компонент продолжает работу на этапе 1303. На этапе 1303 компонент выбирает
30 следующий web-сайт. На этапе 1304 ветвления, если все web-сайты уже выбраны, то
компонент переходит в цикле на этап 1301, чтобы выбрать следующий глобальный
атрибут, иначе компонент продолжает работу на этапе 1305. На этапе 1305 ветвления,
если выбранный web-сайт имеет атрибут, соответствующий выбранному глобальному
атрибуту, то компонент продолжает работу на этапе 1306, иначе компонент
35 переходит в цикле на этап 1303, чтобы выбрать следующий web-сайт. На этапе 1306
ветвления, если выбранный атрибут следует переместить в другой глобальный
атрибут, то компонент продолжает работу на этапе 1307, иначе компонент переходит
в цикле на этап 1303, чтобы выбрать следующий web-сайт. На этапе 1307 компонент
40 изменяет выбранный атрибут так, чтобы он соответствовал другому глобальному
атрибуту. На этапе 1308 компонент изменяет внутрисайтовые соответствия
выбранного атрибута. Затем компонент переходит в цикле на этап 1303, чтобы
выбрать следующий web-сайт.

Специалисту в данной области техники должно быть понятно, что хотя конкретные
45 варианты осуществления системы согласования схем были описаны в документе с
целью иллюстрации, могут быть выполнены различные изменения без выхода за
пределы существа и объема изобретения. Соответственно, изобретение не является
ограниченным за исключением оговоренного прилагаемой формулой изобретения.

Формула изобретения

1. Реализуемый в компьютерной системе способ формирования куба вхождений,
содержащий этапы, на которых

для каждого глобального атрибута домена базы данных, для каждого интерфейсного атрибута базы данных подают запросы к базе данных, причем каждый запрос имеет значение упомянутого интерфейсного атрибута базы данных, установленного в значение глобального атрибута упомянутого глобального атрибута домена базы данных;

для каждого результата каждого поданного запроса осуществляют подсчет количества вхождений значения глобального атрибута в каждый результирующий атрибут упомянутого результата;

для каждой комбинации глобального атрибута, интерфейсного атрибута и результирующего атрибута сохраняют в качестве элемента куба вхождений сумму значений подсчета количества вхождений значения упомянутого глобального атрибута в каждый результирующий атрибут результата запроса, поданного с упомянутым интерфейсным атрибутом, установленным в значение глобального атрибута упомянутого глобального атрибута домена базы данных, при этом сохраненные элементы образуют куб вхождений.

2. Способ по п.1, содержащий этап, на котором формируют матрицу вхождений, ассоциированную с глобальными атрибутами и интерфейсными атрибутами, на основе куба вхождений.

3. Способ по п.1, содержащий этап, на котором формируют матрицу вхождений, ассоциированную с глобальными атрибутами и результирующими атрибутами, на основе куба вхождений.

4. Способ по п.1, содержащий этап, на котором формируют матрицу вхождений, ассоциированную с интерфейсными атрибутами и результирующими атрибутами, на основе куба вхождений.

5. Способ по п.1, в котором запрос подают для каждой комбинации значения глобального атрибута и интерфейсного атрибута.

6. Способ по п.1, в котором куб вхождений включает в себя значение подсчета для каждой комбинации глобального атрибута, интерфейсного атрибута и результирующего атрибута.

7. Реализуемый в компьютерной системе способ идентификации атрибутов базы данных внутри домена, содержащий этапы, на которых

обеспечивают значения подсчета вхождений, ассоциированных с глобальными атрибутами глобальной схемы домена и интерфейсными атрибутами интерфейсной схемы и результирующими атрибутами результирующей схемы базы данных, причем каждое значение подсчета представляет собой для каждой комбинации глобального атрибута, интерфейсного атрибута и результирующего атрибута количество вхождений значения глобального атрибута упомянутого глобального атрибута в качестве значения упомянутого результирующего атрибута результата запроса, поданного к базе данных с упомянутым интерфейсным атрибутом, установленным в упомянутое значение глобального атрибута;

оценивают взаимную информацию между парами схем на основе предоставленных значений подсчета; и

на основе оценки взаимной информации идентифицируют, какие атрибуты соответствуют; и

сохраняют указание соответствующих атрибутов.

8. Способ по п.7, в котором обеспечение значений подсчета включает в себя этап, на котором выполняют проецирование куба вхождений, обеспечивающего значение подсчета вхождений, ассоциированных с глобальными атрибутами, интерфейсными

атрибутами и результирующими атрибутами, в матрицу, ассоциированную с парами схем.

9. Способ по п.8, включающий в себя этап, на котором формируют куб вхождений посредством подачи запросов к базе данных со значениями интерфейсных атрибутов, установленными в значения глобальных атрибутов упомянутых глобальных атрибутов.

10. Способ по п.9, в котором внутри куба вхождений значение подсчета вхождений представляет количество раз, которое значения глобального атрибута глобального атрибута, используемого в качестве значения интерфейсного атрибута в запросе, встречаются в результирующем атрибуте результата запроса.

11. Способ по п.7, в котором интерфейсные атрибуты идентифицируются на основе элементов языка HTML, относящихся к вводу данных.

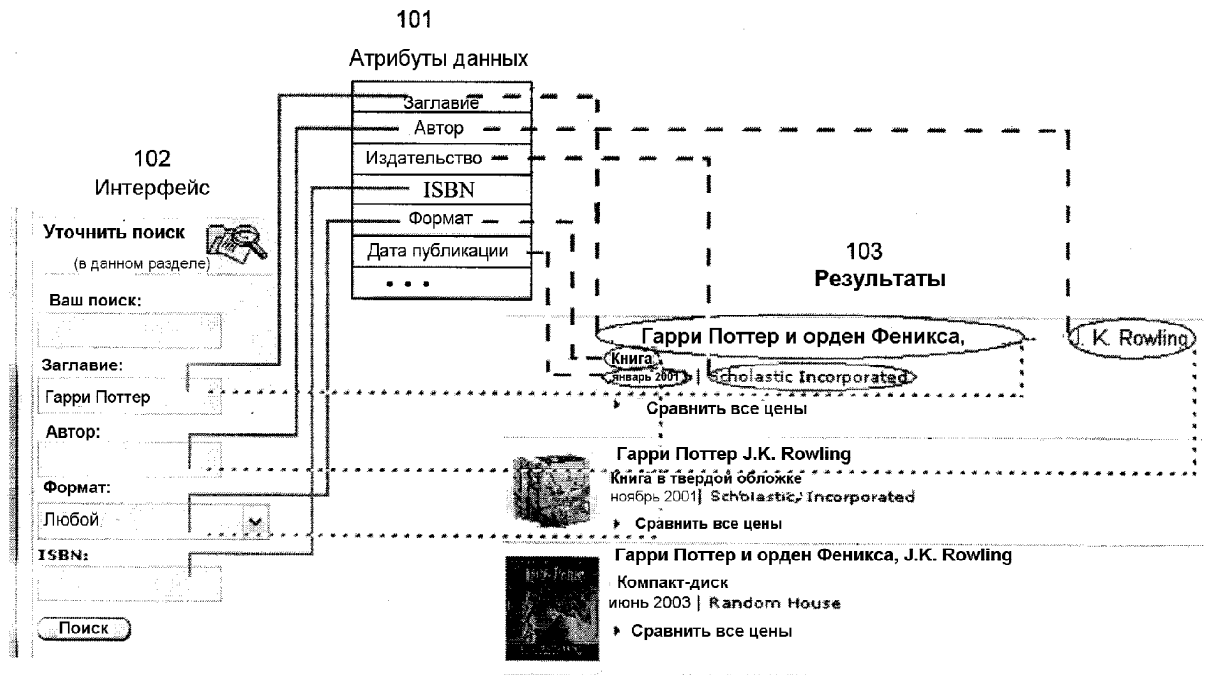
12. Способ по п.7, в котором результирующие атрибуты идентифицируются с использованием объекта-оболочки регулярного выражения.

13. Способ по п.7, в котором значения подсчета вхождений обеспечиваются посредством подачи запросов к базе данных со значениями интерфейсных атрибутов, установленных в значения глобальных атрибутов упомянутых глобальных атрибутов.

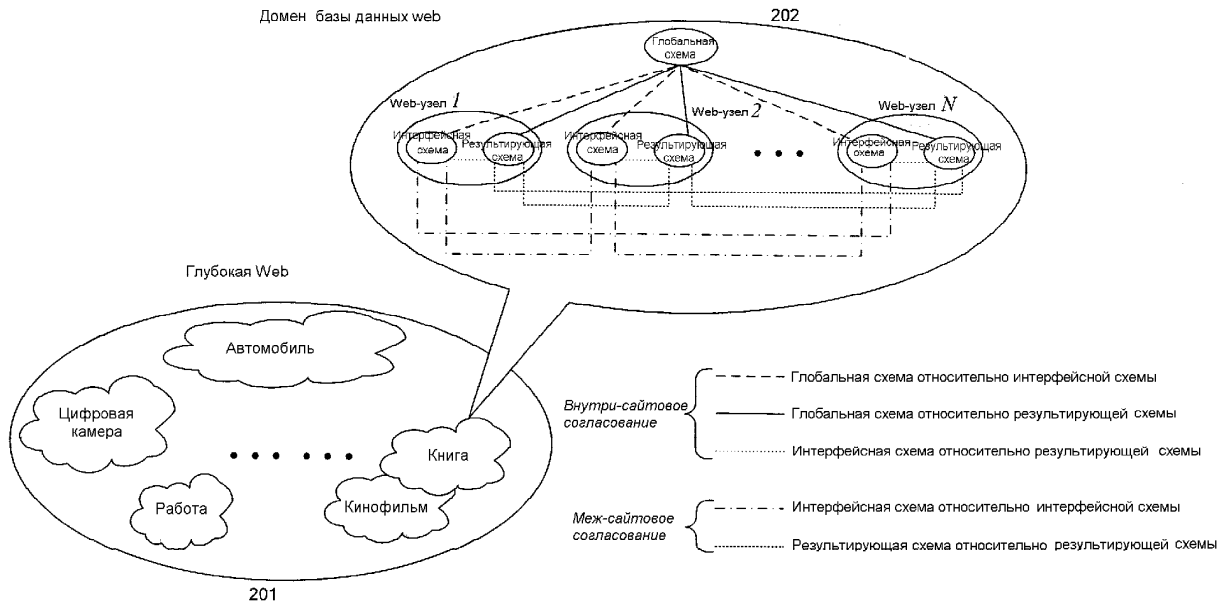
14. Способ по п.7, в котором взаимную информацию оценивают нижеследующим образом:

$$EMI(S_{1i}, S_{2j}) = \frac{m_{ij}}{M} \log \frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M} \cdot \frac{m_{+j}}{M}}.$$

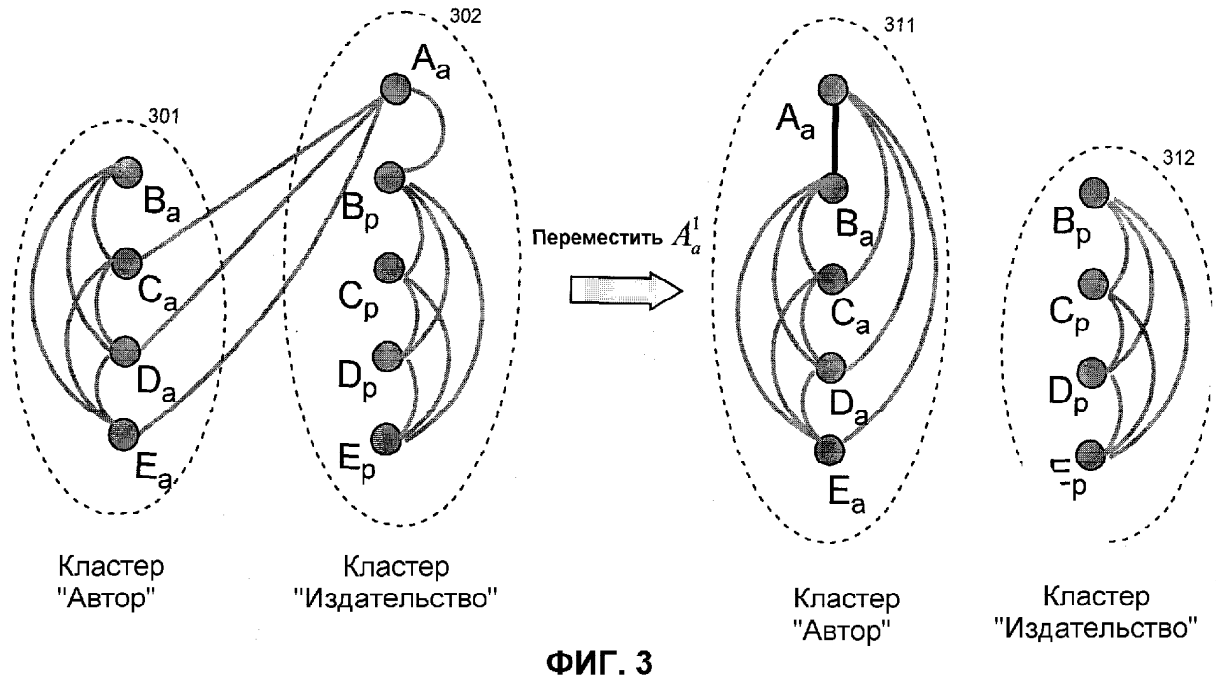
15. Способ по п.7, в котором соответствие между атрибутами в паре схем идентифицируют в случае, когда атрибут одной схемы, имеющий наивысшую оценку взаимной информации для атрибута другой схемы, не имеет более высокой оценки взаимной информации для другого атрибута упомянутой другой схемы.

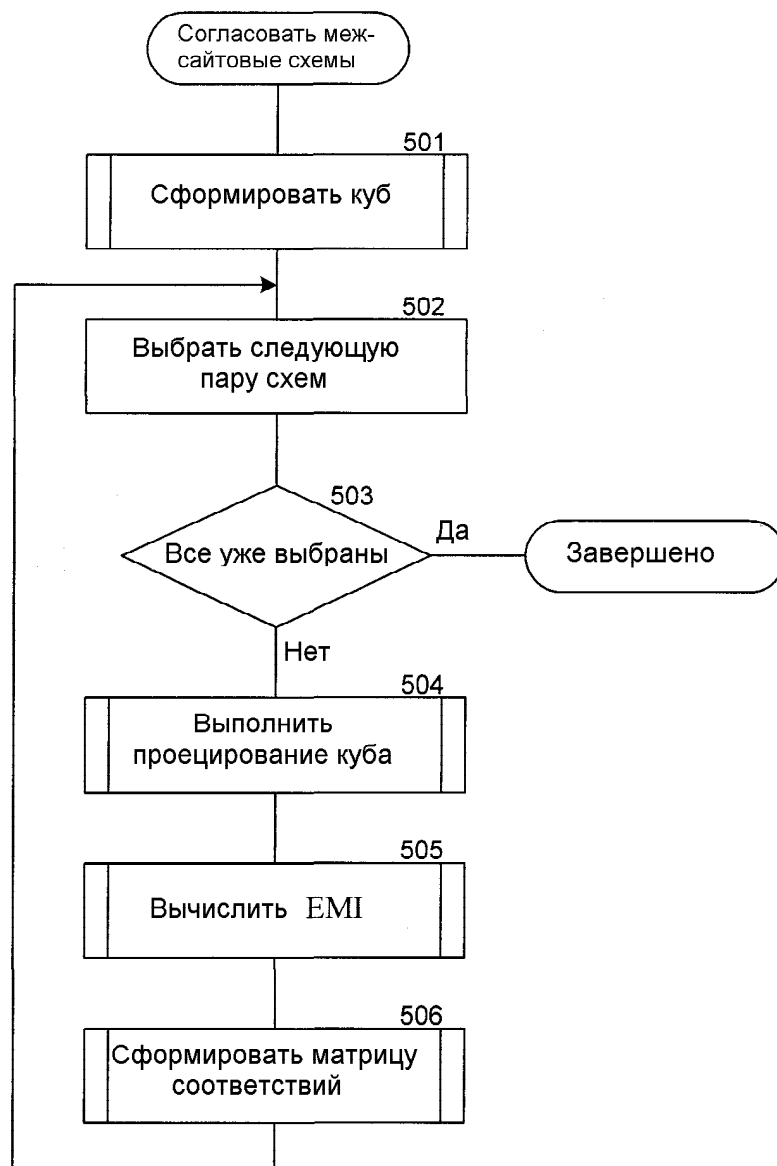


ФИГ. 1

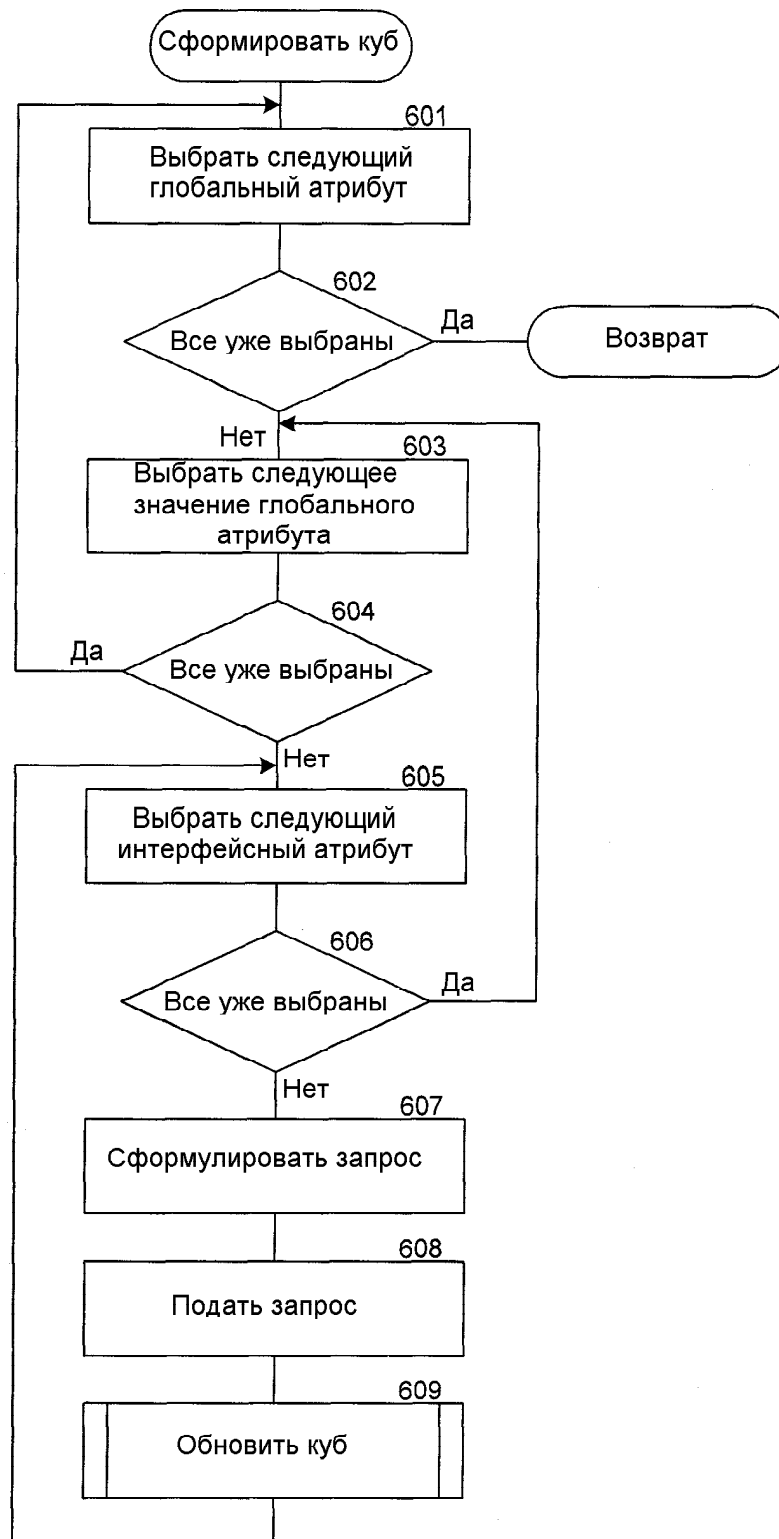


ФИГ. 2

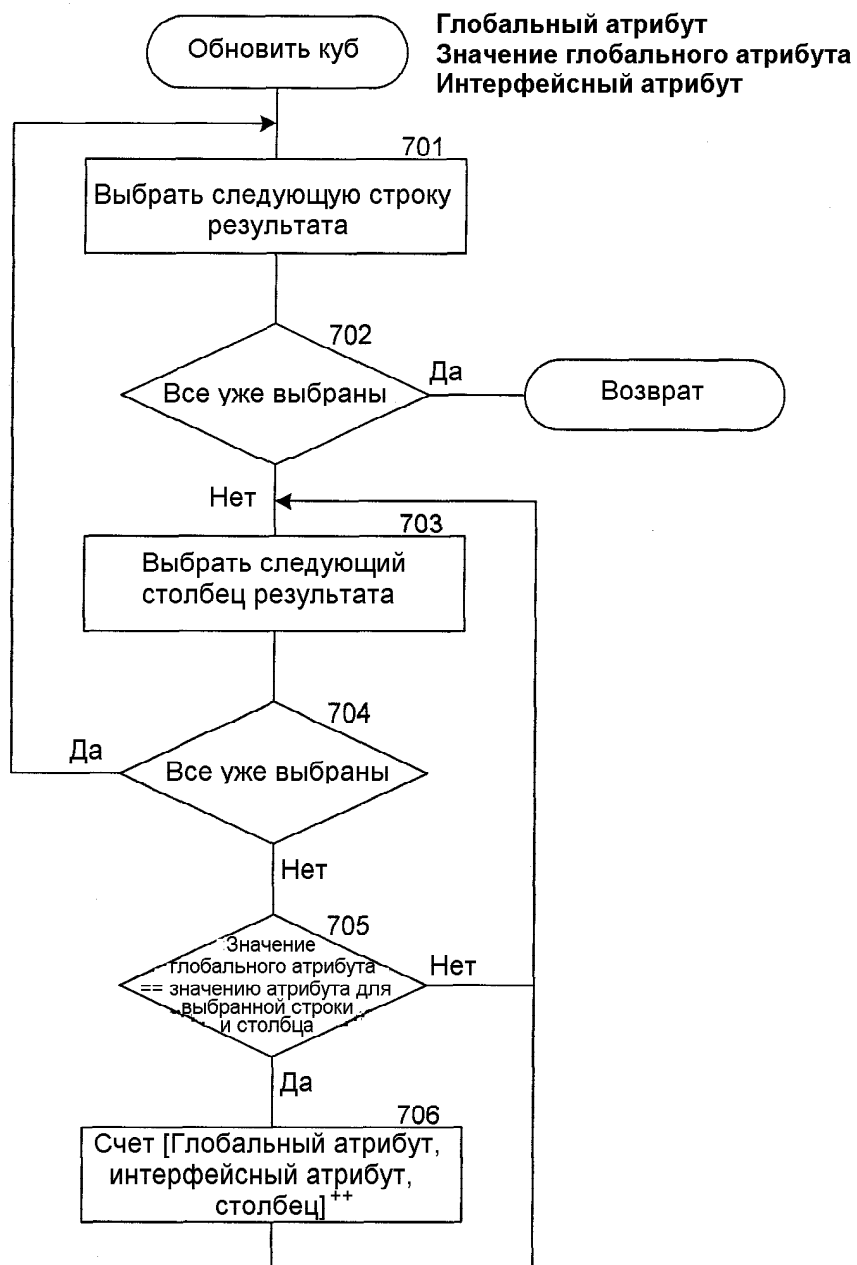




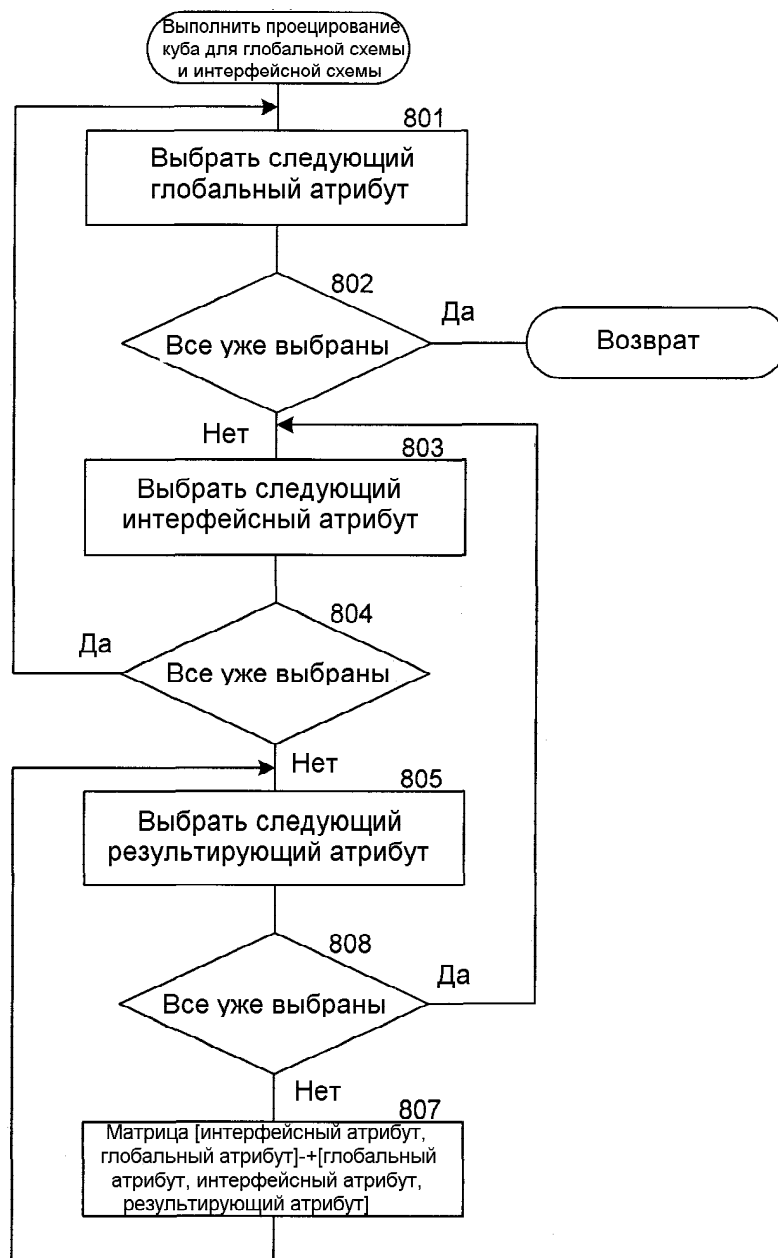
ФИГ. 5



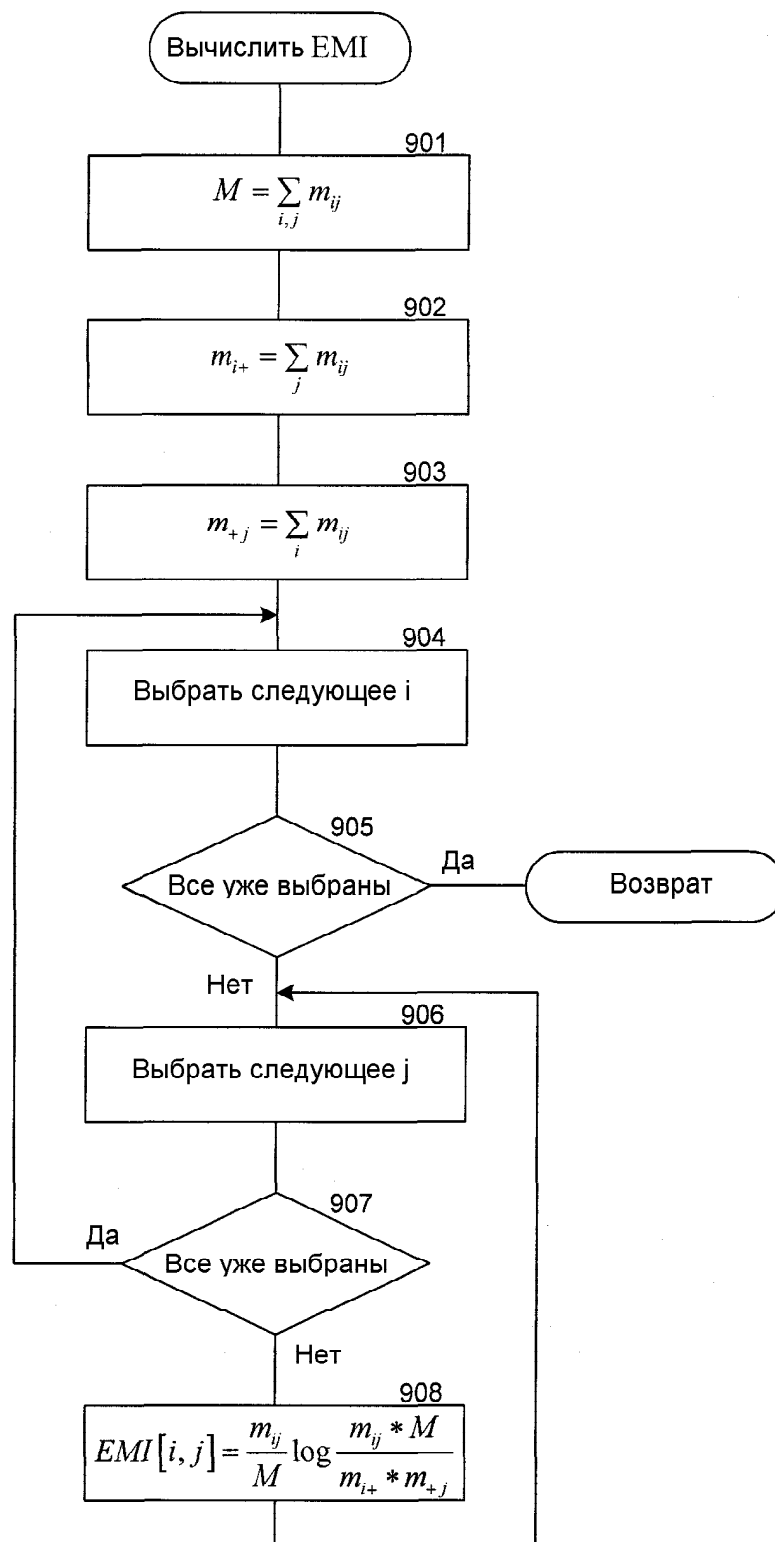
ФИГ. 6



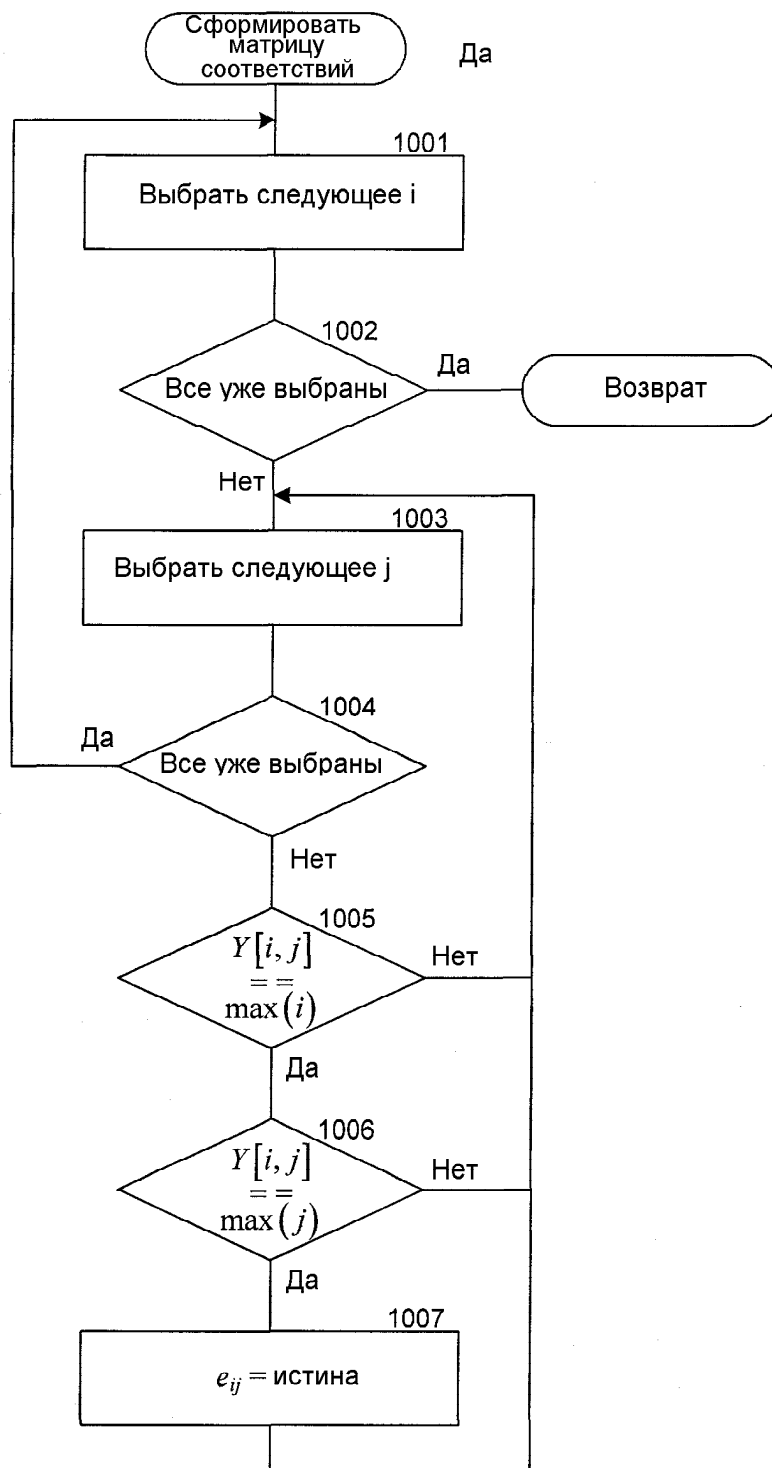
ФИГ. 7



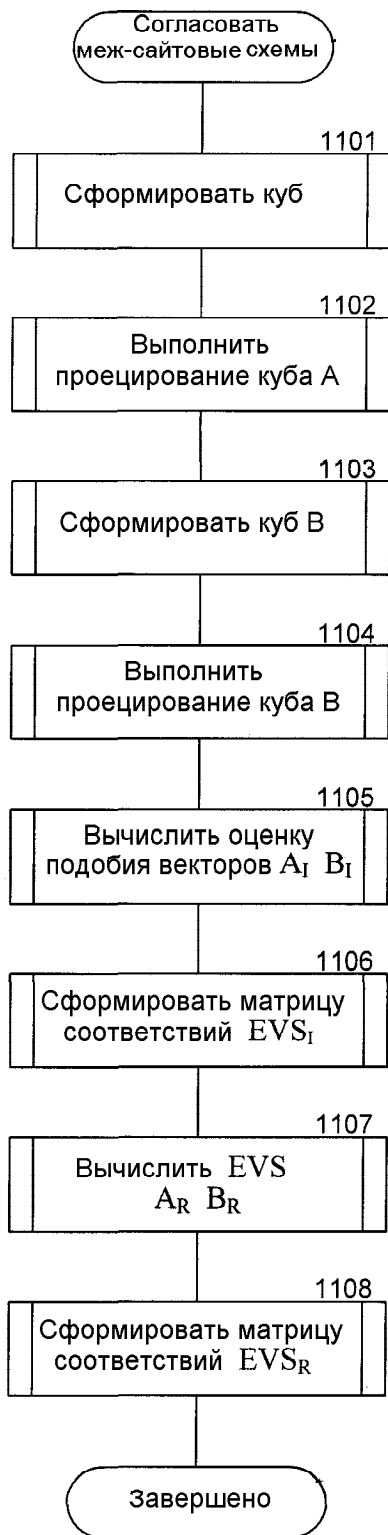
ФИГ. 8



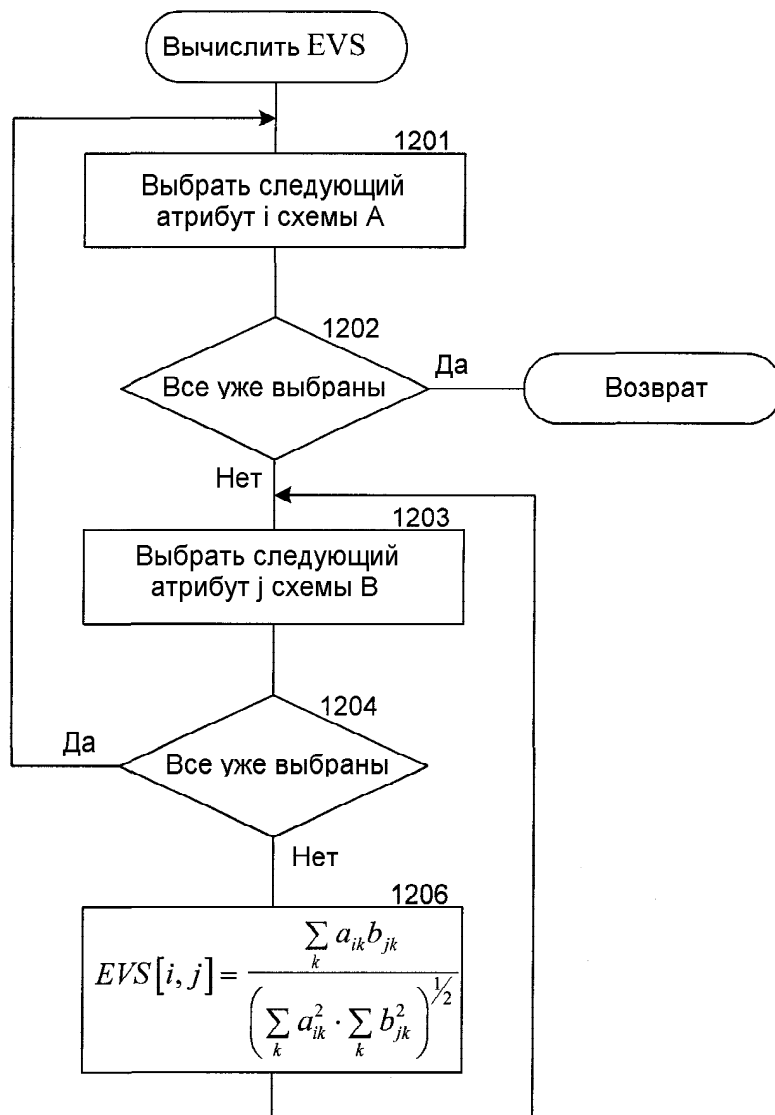
ФИГ. 9



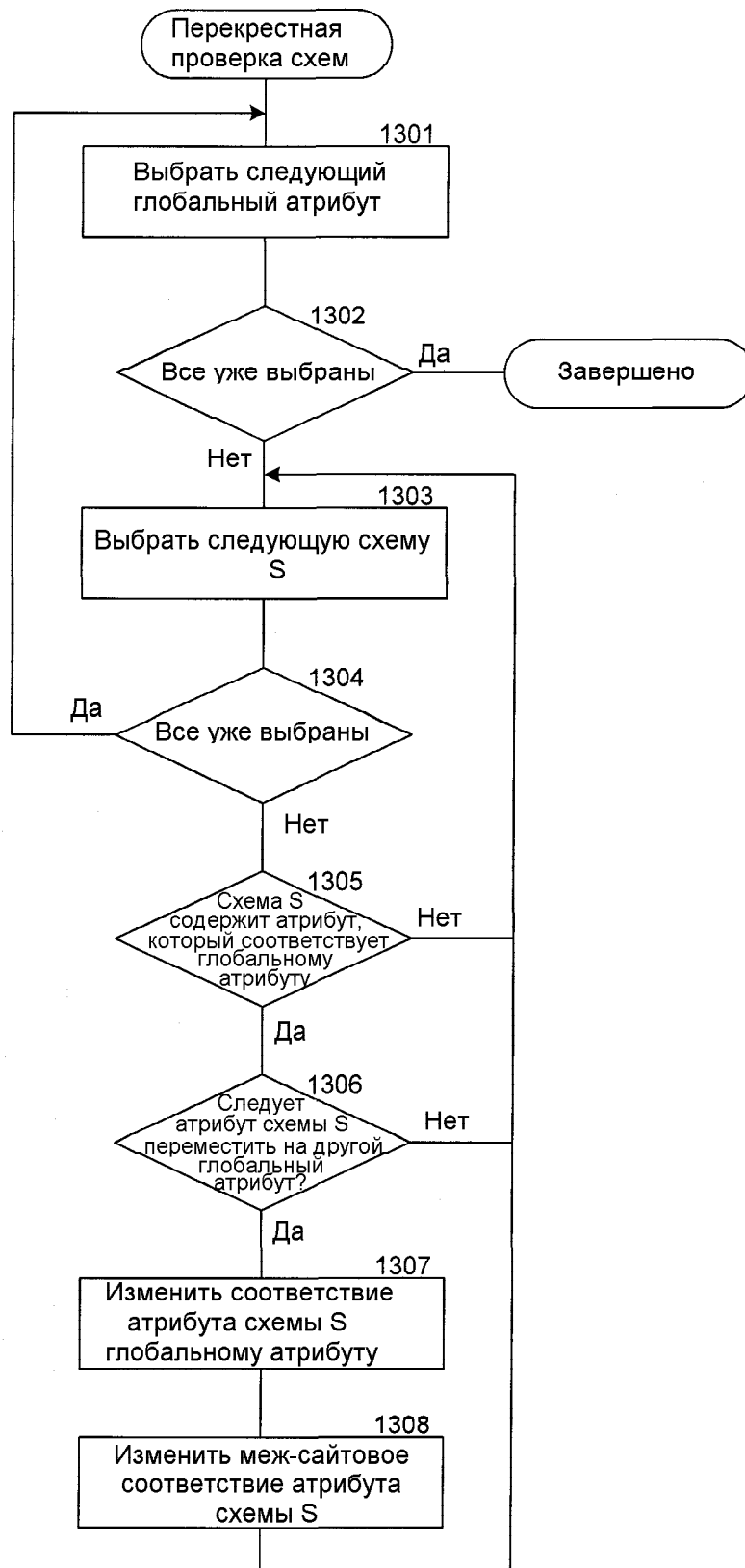
ФИГ. 10



ФИГ. 11



ФИГ. 12



ФИГ. 13