РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 419 857** (13) **C2**

(51) МПК
G06F 17/30 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21)(22) Заявка: **2005114666/08, 13.05.2005**

(24) Дата начала отсчета срока действия патента:
13.05.2005

Приоритет(ы):
(30) Конвенционный приоритет:
14.05.2004 US 10/846,949

(43) Дата публикации заявки: **20.11.2006** Бюл. № 32

(45) Опубликовано: **27.05.2011** Бюл. № **15**

(56) Список документов, цитированных в отчете о поиске: **SU 305479 A1, 01.01.1971. WEN J-R ET AL: "QUERY CLUSTERING USING USER LOGS" от 01.01.2001 найден по ссылке в Интернете URL: http: //research.microsoft.com/en-us/um/people/jrwen/jrwen_files/publications/qc-tois.pdf. MILLER J C, RAE G, SCHAEFER F "MODIFICATIONS OF KLEINBERG'S HITS ALGORITHM ISING MATRIX EXPONENTIATION AND WEB LOG RECORDS" от** (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):
ЧЖАН Бэньюй (US),
СЮЭ Гуй-Жун (US),
ЦЗЭН Хуа-Цзюнь (US),
МА Вэй-Ин (US),
ЧЭНЬ Чжэн (US)

(73) Патентообладатель(и):
МАЙКРОСОФТ КОРПОРЕЙШН (US)

(54) СПОСОБ И СИСТЕМА ДЛЯ ОПРЕДЕЛЕНИЯ ПОДОБИЯ ОБЪЕКТОВ НА ОСНОВАНИИ ГЕТЕРОГЕННЫХ СВЯЗЕЙ

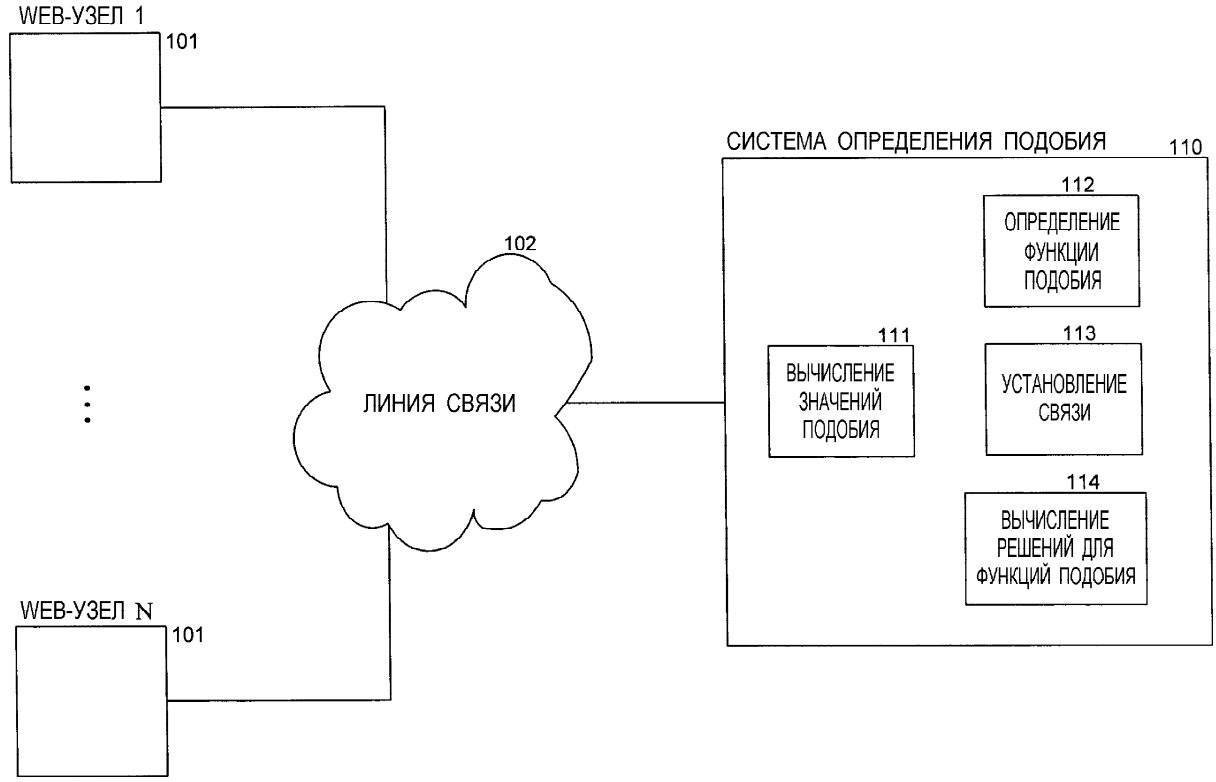
(57) Реферат:

Изобретение относится к способам определения подобия объектов и, в частности, к определению подобия на основании связей между объектами. Техническим результатом является увеличение точности поиска за счет вычисления межтипового подобия. Способ включает: для каждого типа, если подобие этого типа основано на внутритиповой связи, предоставление функции внутритипового подобия для каждой такой связи, которая измеряет подобие между объектами этого типа; если подобие этого типа основано на межтиповой связи, предоставление функции межтипового подобия для каждой такой связи, которая измеряет подобие между объектами этого типа; и предоставление функции подобия, которая измеряет подобие между объектами этого типа, основываясь на любых функциях внутритипового подобия и любых функциях межтипового подобия для этого типа; и для каждой связи предоставление

RU 2419857 C2

RU 2419857 C2

данных, которые определяют эту связь между объектами, ассоциированными с этой связью; одновременно решение предоставленных функций подобия на основании связей, определяемых обеспеченными данными; и сохраняют подобия, основанные на одновременном решении предоставленных функций подобия. 3 н. и 26 з.п. ф-лы, 5 ил.



Фиг. 1

(56) (продолжение):
12.09.2001 найден по ссылке в Интернете URL: http://cnls.lanl.gov/~jomiller/publications/hits_exp.pdf. TAHER H. HAVELIWALA «Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search» от 15.07.2002, найден по ссылке в Интернете URL: http://ilpubs.stanford.edu:8090/750/1/2003-29.pdf.

RU 2419857 C2

RU 2419857 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(19) **RU** (11) **2 419 857** (13) **C2**

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

(12) ABSTRACT OF INVENTION

(21)(22) Application: **2005114666/08, 13.05.2005**

(24) Effective date for property rights:
13.05.2005

Priority:
(30) Priority:
14.05.2004 US 10/846,949

(43) Application published: **20.11.2006** Bull. 32

(45) Date of publication: **27.05.2011 Bull. 15**

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):
ChZhAN Behn'juj (US),
SJuEh Guj-Zhun (US),
TsZEhN Khua-Tszjun' (US),
MA Vehj-In (US),
ChEhN' Chzhehn (US)

(73) Proprietor(s):
MAJKROSOFT KORPOREJShN (US)

(54) METHOD AND SYSTEM FOR DETERMINING SIMILARITY OF DOCUMENTS BASED ON HETEROGENEOUS CONNECTIONS

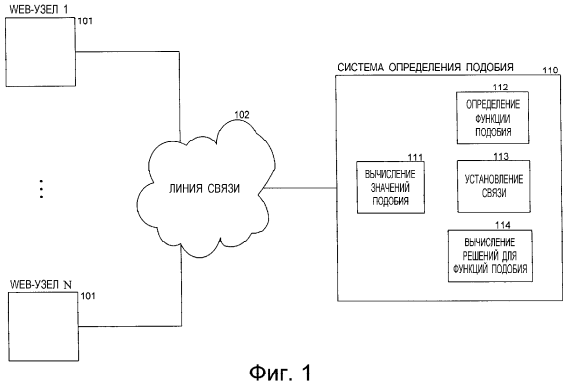
(57) Abstract:

FIELD: information technology.

SUBSTANCE: method involves the following: for each type, if similarity of that type is based on an intra-type connection, providing an intra-type similarity function for every such connection which measures similarity between objects of that type; if similarity of that type is based on an inter-type connection, providing an inter-type similarity function for every such connection which measures similarity between objects of that type; and providing a similarity function which measures similarity between objects of that type based on any intra-type similarity functions and any inter-type similarity functions for that type; and for each connection, providing data which determine that connection between objects associated with that connection; simultaneously solving the provided similarity functions based on connections determined by the provided data; and storing similarities based on simultaneous solving of the provided similarity functions.

EFFECT: high accuracy of search owing to calculation of inter-type similarity.

29 cl, 5 dwg



Фиг. 1

RU 2419857 C2

RU 2419857 C2

ОБЛАСТЬ ТЕХНИКИ

Описанная технология относится, в общем случае, к определению подобия объектов и, в частности, к определению подобия на основании связей между объектами.

ПРЕДПОСЫЛКИ СОЗДАНИЯ ИЗОБРЕТЕНИЯ

Многие поисковые службы, такие как Google и Overture, обеспечивают поиск информации, доступ к которой может быть осуществлен через сеть Интернет. Эти поисковые службы предоставляют пользователям возможность поиска отображаемых на дисплее страниц, например web-страниц, которые могут представлять интерес для пользователей. После того как пользователь подает поисковый запрос (также именуемый "запросом"), содержащий условия поиска, поисковая служба распознает web-страницы, которые могут иметь отношение к этим условиям поиска. Для быстрого распознавания соответствующих web-страниц поисковая служба может обеспечивать поддержку соответствия ключевых слов web-страницам. Поисковая служба может осуществлять генерацию этого соответствия путем поиска во "всемирной паутине" (то есть в сети "всемирная паутина") поисковым агентом для извлечения ключевых слов каждой web-страницы. Для поиска во "всемирной паутине" поисковым агентом поисковая служба может использовать список корневых web-страниц и распознавать все web-страницы, доступ к которым может быть осуществлен через эти корневые web-страницы. Ключевые слова любой конкретной web-страницы могут быть извлечены с использованием различных известных способов информационного поиска, например, путем распознавания слов заголовка, слов, имеющихся в метаданных web-страницы, выделенных слов и т.д. Поисковая служба может вычислять оценку релевантности, которая указывает, насколько каждая web-страница соответствует поисковому запросу, на основании близости каждого соответствия, популярности web-страницы (например, ранга популярности страницы (PageRank) службы Google) и т.д. Затем поисковая служба отображает пользователю ссылки на эти web-страницы в порядке, основанном на степени их соответствия. Механизмы поиска могут обеспечивать более широкий поиск информации в любой совокупности документов. Например, совокупности документов могут содержать все патенты США, все решения федерального суда, все архивные документы компании и т.д.

Поисковым службам могут потребоваться измерения подобия между различными объектами, например web-страницами или запросами. Например, в поисковой службе может быть предусмотрено интерактивное расширение запроса, для которого требуется вычисление подобия между терминами запроса и иными терминами. В качестве другого примера поисковая служба может захотеть сгруппировать web-страницы в кластеры сходных web-страниц для содействия пользователю при навигации по web-страницам. В типовых алгоритмах определения подобия объектов обычно используют вектор признаков, связанный с объектами, а затем вычисляют расстояние между векторами признаков, служащее в качестве показателя подобия. Например, web-страницы могут иметь признаки, содержащие ключевые слова, информационное содержимое и т.д., которые используют для вычисления подобия. Определение подобия в большинстве алгоритмов основано исключительно на признаках, связанных с объектами. Например, подобие между web-страницами может быть основано исключительно на содержимом web-страниц. Однако ряд алгоритмов учитывают признаки, основанные на разнородных объектах. Например, в одном из алгоритмов используются данные, выбираемые щелчком манипулятора типа "мышь",

в которых запросы являются подобными в том случае, если они содержат те же самые термины или приводят к выбору той же самой web-страницы пользователями. Таким образом, вектор признаков для таких запросов содержит информацию о тех web-страницах результата запроса, которые были выбраны пользователями.

Однако эти способы при вычислении подобия между объектами одного типа не учитывают подобие между объектами другого типа, которые могут быть связаны с ними. То есть измерения подобия для объектов одного типа могут быть связаны с измерениями подобия для объектов другого типа. Например, запрос может быть подобным другому запросу, отчасти, на основании подобия между web-страницами результатов, которые пользователи выбирают непосредственно или щелчком манипулятора типа "мышь". В ином случае, web-страницы могут являться подобными другим web-страницам, отчасти, на основании подобия между запросами, которые возвращают web-страницы в их результатах. Желательно иметь способ измерения подобия объектов, учитывающий связи между разнородными объектами.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

Предложены способ и система для измерения подобия объектов на основании связей с объектами одного и того же типа и различных типов и подобия этих объектов другим объектам. В одном из вариантов осуществления изобретения система определения подобия определяет функции внутритипового и межтипового подобия для каждого типа объекта. Система определения подобия может объединять функции внутритипового и межтипового подобия для определенного типа в функцию общего подобия для этого типа. После определения функций подобия система определения подобия производит сбор значений атрибутов для объектов, которые могут содержать данные о связях между объектами одного и того же типа, именуемых внутритиповыми связями, и о связях между объектами различных типов, именуемых межтиповыми связями. После сбора значений атрибутов для объектов система определения подобия вычисляет решения для функций внутритипового и межтипового подобия путем итерационного вычисления значений подобия для объектов до тех пор, пока не будет получено решение за счет сходимости значений подобия.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

На Фиг. 1 изображена блок-схема, на которой показаны компоненты системы определения подобия в одном из вариантов осуществления изобретения.

На Фиг. 2 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом определения функций подобия в одном из вариантов осуществления изобретения.

На Фиг. 3 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом установления связей в одном из вариантов осуществления изобретения.

На Фиг. 4 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом вычисления решения для функций подобия в одном из вариантов осуществления изобретения.

На Фиг. 5 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом вычисления функций подобия в одном из вариантов осуществления изобретения.

ПОДРОБНОЕ ОПИСАНИЕ

Предложены способ и система для измерения подобия объектов на основании связей с объектами одного и того же типа и различных типов и подобия этих объектов другим объектам. В одном из вариантов осуществления изобретения система

определения подобия определяет функции внутритипового и межтипового подобия для каждого типа объекта. Функция внутритипового подобия является мерой подобия между объектами одного и того же типа. Например, функция внутритипового подобия между запросами может быть основана на том, насколько близко совпадают условия поиска в запросах, на основании атрибута пользователей, подающих запросы. Внутритиповое подобие между объектами может также зависеть от подобия других объектов того же самого типа. Например, два запроса могут иметь более высокую степень подобия друг другу в том случае, если каждый из них имеет высокую степень подобия третьему запросу. Такое внутритиповое подобие между объектами, основанное на подобии между другими объектами, определяет рекурсивную функцию. Функция межтипового подобия является мерой подобия между двумя объектами одного типа на основании атрибутов объектов другого типа, в том числе, их подобия. Например, два запроса могут иметь более высокую степень подобия в том случае, если web-страница результата одного запроса, выбираемая пользователями путем щелчка манипулятора типа "мышь", является подобной web-странице результата другого запроса, выбираемой пользователями путем указания "мышью". Подобие объектов другого типа может также зависеть от подобия объектов первого типа. Кроме того, поскольку подобие объектов одного типа может зависеть от подобия объектов другого типа и наоборот, функции межтипового подобия являются рекурсивными между различными типами.

Тип объекта может иметь различные определения подобия, определяемые для его объектов на основании различных атрибутов объектов. Например, web-страница может иметь внутритиповое подобие, основанное на содержании web-страниц, и иное внутритиповое подобие, основанное на ссылках между web-страницами. Система определения подобия может объединять функции внутритипового и межтипового подобия для определенного типа в функцию общего подобия для этого типа. В одном из вариантов осуществления изобретения система определения подобия объединяет функции внутритипового и межтипового подобия посредством линейного уравнения с весовыми коэффициентами, присвоенными каждой функции внутритипового и межтипового подобия на основании ее воспринимаемой точности при отображении общего подобия между объектами этого типа. Например, функции внутритипового подобия, имеющей высокую точность, может быть присвоен высокий весовой коэффициент, а функции внутритипового подобия, имеющей низкую точность, может быть присвоен низкий весовой коэффициент.

После определения функций подобия система определения подобия производит сбор значений атрибутов для объектов, которые могут содержать данные о связях между объектами одного и того же типа, именуемых внутритиповыми связями, и о связях между объектами различных типов, именуемых межтиповыми связями. Например, web-страница может иметь значения атрибута, не основанные на связях, которые соответствуют ключевым словам web-страницы. Web-страница может также иметь внутритиповую связь, основанную на входящих и исходящих ссылках между web-страницами. Web-страница может иметь межтиповую связь с запросами, основанными на выборе web-страниц из результатов запроса щелчками мыши.

После сбора значений атрибутов для объектов система определения подобия вычисляет решения для функций внутритипового и межтипового подобия путем итерационного вычисления значений подобия для объектов до тех пор, пока не будет получено решение за счет сходимости значений подобия. Система определения подобия использует итерационный подход вследствие рекурсивного характера

функций подобия. Система определения подобия начинает с инициализации подобий, а затем вычисляет функцию подобия для каждого типа объекта на основании исходных подобий для получения новых подобий. Система определения подобия измеряет разность между новыми подобиями и старыми подобиями для определения того, было ли получено решение за счет сходимости значений подобия. Если это так, то новые подобия представляют собой решение. В противном случае система определения подобия повторяет процедуру с новыми подобиями, которые становятся старыми подобиями. Таким образом, система определения подобия вычисляет подобия объектов одного типа на основании подобий объектов другого типа и на основании связей между объектами различных типов.

Ниже приведен пример обработки, выполняемой в системе определения подобия, применительно к механизму поиска. Система определения подобия моделирует объекты (например, web-страницы и запросы) и связи (например, входящие ссылки и варианты выбора щелчком мыши), используемые механизмом поиска, в виде ориентированного графа G = (V,E), где узлы V отображают объекты механизма поиска, а ребра E отображают связи между объектами. Узлы V могут быть разделены на два подмножества $Q = \{q^1,q^2,\ldots,q^m\}$ и $P = \{p^1,p^2,\ldots,p^n\}$, где Q обозначает запросы, а P обозначает web-страницы. Связи между этими web-страницами и запросами могут содержать связь по входящей ссылке (OL - outgoing link), связь по исходящей ссылке (IS - incoming link) и связь щелчком мыши (CT - click-through). Для узла ν в графе $M_R(\nu)$ отображает набор соседних узлов, которые имеют связь R с узлом ν. Например, $M_{IL}(\nu)$ отображает набор web-страниц, которые являются источником входящих ссылок на web-страницу ν. $M_R^i(\nu)$ обозначает i-ю web-страницу в множестве. Для отображения подобия между объектами система определения подобия использует матрицу S подобия, и S[a,b] отображает подобие между объектами a и b.

Система определения подобия основана на том принципе, что объекты одного типа являются подобными, отчасти, на основании подобия взаимосвязанных объектов другого типа. Если два объекта одного типа имеют связь с одним и тем же объектом другого типа, то эти два объекта являются подобными в некоторой степени. Кроме того, если два объекта одного и того же типа имеют связь с двумя различными, но подобными объектами другого типа, то эти два объекта являются подобными в некоторой степени. Система определения подобия отображает этот принцип следующими уравнениями:

$$S_{O_1}[a,b] = \frac{C}{|M_R(a)||M_R(b)|} \sum_{i=1}^{|M_R(a)|} \sum_{j=1}^{|M_R(b)|} S_{O_2}\left[M_R^i(a), M_R^j(b)\right], \quad (1)$$

где S_{O_1} представляет собой подобие между объектами a и b типа O_1, S_{O_2} представляет собой подобие между объектами i и j другого типа, R отображает межтиповую связь, на которой основано подобие, а C - весовой коэффициент. Если a равен b, то $S_{O_1}[a,b]$ определяется равным 1, то есть подобие объекта самого с собой дает максимальное подобие, равное 1. Если оба объекта a и b связаны с одним и тем же объектом A в O_2, то $S_{O_2}[A,A]$ равно 1, что дает максимальный вклад в $S_{O_1}[a,b]$. Если любой из объектов a или b не имеет каких-либо соседей, то есть связь с объектом в O_2 отсутствует, то $|M_R(a)|$ или $|M_R(b)|$ равно нулю. В этом случае система определения подобия устанавливает $S_{O_1}[a,b]$ равным нулю, предотвращая деление на нуль. Например, предположим, что O_1 содержит объекты a и b, что O_2 содержит объекты A, B и C и что a связан с A и B, а b связан с B и C. Если $S_{O_2}[A,B]$ равно 0,7,

$S_{Q_1}[B,C]$ равно 0,7 и $S_{Q_2}[B,C]$ равно 0,49, а весовой коэффициент равен 0,7, то с использованием уравнения (1) $S_{Q_1}[a,b]$ равно 0,5 (например, 0,7/4·(0,7 + 0,49 + 1,0 + 0,7)).

Система определения подобия определяет общее подобие типа объектов на основании комбинации подобий, полученных из функций внутритипового подобия и функций межтипового подобия. В одном из вариантов осуществления изобретения система определения подобия использует линейную комбинацию подобий функций внутритипового подобия и функций межтипового подобия, представленную следующим уравнением:

$$S[a,b] = \alpha S_{внутр}[a,b] + \beta S_{меж}[a,b] \quad , (2)$$

где $S_{внутр}$ и $S_{меж}$ представляют собой подобия, полученные из функций внутритипового подобия и функций межтипового подобия, а α и β представляют собой весовые коэффициенты для подобий, причем $\alpha + \beta = 1$. Путем присвоения различных значений α и β система определения подобия может регулировать вклады различных функций подобия в общее подобие. Как описано выше, уравнение 2 может быть задано в рекурсивном виде, поскольку подобие одного объекта может быть определено на основании подобия другого объекта, которое, в свою очередь, может быть определено на основании подобия этого одного объекта. В одном из вариантов осуществления изобретения система определения подобия вычисляет решение для функций подобия путем итерационного вычисления значений подобия до достижения их сходимости (то есть, $\left\| S^i - S^{i-1} \right\| < \varepsilon$, где ε - пороговое значение разности).

Применительно к механизму поиска, для определения функции внутритипового подобия система определения подобия может использовать только содержимое запроса. Функция внутритипового подобия, основанная на информационном содержимом, может быть определена следующим уравнением:

$$S_{QC}[a,b] = \frac{\left| Ключевое\_слово(a) \cap Ключевое\_слово(b) \right|}{\left| Ключево\_слово(a) \cup Ключевое\_слово(b) \right|} \quad ,(3)$$

где a и b - запросы, а S_{QC} - матрица подобия информационного содержимого запросов на основании информационного содержимого. Например, когда запросы a и b имеют два условия (или ключевых слова) для поиска, при этом одно из ключевых слов является общим, то значение их подобия равно 0,33 (то есть 1/3). Система определения подобия может определить функцию межтипового подобия для запроса на основании связи с web-страницами указанием мышью посредством следующего уравнения:

$$S_{QCT}[a,b] = \frac{C_{CT}}{\left| M_{CT}(a) \right\| M_{CT}(b) \right|} \sum_{i=1}^{|M_{CT}(a)|} \sum_{j=1}^{|M_{CT}(b)|} S_{PCT}\left[M^i_{CT}(a), M^j_{CT}(b) \right] \quad , (4)$$

где S_{QCT} представляет собой матрицу подобия запросов на основании указания мышью, S_{PCT} представляет собой матрицу подобия web-страниц на основании указания мышью, $M_{CT}(a)$ обозначены указания мышью из запроса a на ссылках на web-страницы, распознанные из журналов регистрации запросов, а C_{CT} - весовой коэффициент. Система определения подобия объединяет уравнения (3) и (4) в функцию общего подобия для запросов, представленную в виде следующего уравнения:

$$S_Q[a,b] = \alpha S_{QC}[a,b] + \beta S_{QCT}[a,b] \quad , (5)$$

где S_Q представляет собой матрицу общего подобия запросов.

Система определения подобия отображает подобие web-страниц на основании внутритиповых связей входящих ссылок и исходящих ссылок и межтиповой связи с запросами, приводящими к обращению к web-страницам указанием мышью. Система определения подобия определяет функцию внутритипового подобия на основании связи по входящей ссылке для отражения того, что две web-страницы могут являться подобными в том случае, когда на них имеется ссылка на одной и той же web-странице (или на подобных друг другу web-страницах). Система определения подобия также определяет функцию межтипового подобия на основании связи по исходящей ссылке для отражения того, что две web-страницы могут быть подобными в том случае, когда они имеют ссылки на одну и ту же web-страницу (или на подобные друг другу web-страницы). Система определения подобия отображает функции внутритипового подобия для web-страниц на основании связей по исходящей и по входящей ссылке следующими уравнениями:

$$S_{OL}[A,B] = \frac{C_{OL}}{|M_{OL}(A)||M_{OL}(B)|} \sum_{i=1}^{|M_{OL}(A)||M_{OL}(B)} \sum_{j=1} S_{IL}\left[M_{OL}^{i}(A), M_{OL}^{j}(B)\right], \quad (6)$$

$$S_{IL}[A,B] = \frac{C_{IL}}{|M_{IL}(A)||M_{IL}(B)|} \sum_{i=1}^{|M_{IL}(A)||M_{IL}(B)} \sum_{j=1} S_{IL}\left[M_{IL}^{i}(A), M_{IL}^{j}(B)\right], \quad (7)$$

где A и B представляют web-страницы, C_{OL} и C_{IL} - весовые коэффициенты, S_{OL} и S_{IL} - матрицы подобия на основании исходящих и входящих ссылок, $M_{OL}(A)$ - целевые web-страницы исходящих ссылок из web-страницы A, а $M_{IL}(A)$ - источник входящих ссылок из web-страниц на web-страницу A. Система определения подобия отображает функцию межтипового подобия для web-страниц на основании связи указанием мышью следующим уравнением:

$$S_{PCT}[A,B] = \frac{C_{CT}}{|M_{CT}(A)||M_{CT}(B)|} \sum_{i=1}^{|M_{CT}(A)||M_{CT}(B)} \sum_{j=1} S_{QCT}\left[M_{CT}^{i}(A), M_{CT}^{j}(B)\right], \quad (8)$$

где $M_{CT}(A)$ представлены запросы, на которых пользователи производят указание мышью для доступа к web-странице A. Так как уравнение (8) определено через уравнение (4) (то есть через S_{QCT}) и наоборот, то эта пара уравнений определяет рекурсивную функцию. Система определения подобия определяет функцию общего подобия для web-страниц в виде линейной комбинации функций внутритипового подобия и функций межтипового подобия, которая представлена следующим уравнением:

$$S_P[A,B] = \alpha' S_{OL}[A,B] + \beta' S_{OL}[A,B] + \gamma' S_{PCT}[A,B], \quad (9)$$

где S_P представляет собой матрицу подобия для web-страниц, а α', β' и γ' - весовые коэффициенты, при этом $\alpha' + \beta' + \gamma' = 1$.

Следовательно, система определения подобия использует унифицированную структуру для объединения разнородных объектов и их межтиповых связей. Так как функции общего подобия являются рекурсивными, система определения подобия вычисляет решения для функций подобия одновременно итерационным способом. Функции подобия выражены следующими вышеупомянутыми уравнениями:

$$S_{QC}[a,b] = \frac{|Ключевое\_слово(a) \cap Ключевое\_слово(b)|}{|Ключево\_слово(a) \cup Ключевое\_слово(b)|}$$

$$S_{QCT}[a,b] = \frac{C_{CT}}{|M_{CT}(a)||M_{CT}(b)|} \sum_{i=1}^{|M_{CT}(a)||M_{CT}(b)} \sum_{j=1} S_P\left[M_{CL}^{i}(a), M_{CL}^{j}(b)\right]$$

$$S_Q[a,b] = \alpha S_{QC}[a,b] + \beta S_{QCT}[a,b]$$

$$S_{OL}[A,B] = \frac{C_{PC}}{|M_{OL}(A)||M_{OL}(B)|} \sum_{i=1}^{|M_{OL}(A)|} \sum_{j=1}^{|M_{OL}(B)|} S_P\left[M_{OL}^i(A), M_{OL}^j(B)\right] \quad (10)$$

$$S_{IL}[A,B] = \frac{C_{PR}}{|M_{IL}(A)||M_{IL}(B)|} \sum_{i=1}^{|M_{IL}(A)|} \sum_{j=1}^{|M_{IL}(B)|} S_P\left[M_{IL}^i(A), M_{IL}^j(B)\right]$$

$$S_{PCT}[A,B] = \frac{C_{CT}}{|M_{CT}(A)||M_{CT}(B)|} \sum_{i=1}^{|M_{CT}(A)|} \sum_{j=1}^{|M_{CT}(B)|} S_Q\left[M_{CT}^i(A), M_{CT}^j(B)\right]$$

$$S_P[A,B] = \alpha' S_{OL}[A,B] + \beta' S_{OL}[A,B] + \gamma' S_{PCT}[A,B].$$

Из уравнений (10) можно заметить, что на межтиповое подобие между любыми двумя запросами оказывает воздействие подобие web-страниц, как внутритиповое подобие, так и межтиповое подобие. Так как на межтиповое подобие между web-страницами оказывает воздействие подобие запросов как внутритиповых, так и межтиповых, уравнения (10) определяют рекурсивные связи. Таким образом, подобия web-страниц и запросов взаимно распространяются одно в другое и имеют сходимость к устойчивому состоянию.

На блок-схеме из Фиг. 1 показаны компоненты системы определения подобия в одном из вариантов осуществления изобретения. Web-узлы 101 соединены через линию связи 102 с системой 110 определения подобия. Система определения подобия содержит компонент 111 вычисления подобий, компонент 112 определения функций подобия, компонент 113 установления связей и компонент 114 вычисления решения для функций подобия. Компонент вычисления подобий вычисляет подобия между объектами на основании межтиповых связей и подобия объектов других типов. Компонент вычисления подобий вызывает компонент определения функций подобия, компонент установления связей и компонент вычисления решения для функций подобия. Компонент определения функций подобия может взаимодействовать с пользователем для определения типов объектов, связей между объектами и различных функций подобия для объекта каждого типа. Компонент установления связей осуществляет генерацию данных о связях на основании собранных данных. Например, собранные данные могут содержать запросы, web-страницы результатов запроса и журналы регистрации запросов. Компонент вычисления решения для функций подобия производит итерационное вычисление определенных функций подобия для генерации обновленных матриц подобия до тех пор, пока не будет получено решение за счет сходимости значений подобия матриц подобия.

Вычислительное устройство, в котором реализована система определения подобия, может содержать центральный процессор, память, устройства ввода данных (например, клавиатуру и координатно-указательные устройства), устройства вывода (например, устройства отображения) и запоминающие устройства (например, дисковые накопители). Память и запоминающие устройства представляют собой считываемые посредством компьютера среды, которые могут содержать команды, обеспечивающие реализацию системы определения подобия. Кроме того, структуры данных и структуры сообщений могут быть запомнены или переданы через среду передачи данных, например, в виде сигнала по линии связи. Могут быть использованы различные линии связи, например сеть Интернет, локальная сеть, глобальная сеть или прямое соединение по коммутируемой телефонной линии.

Система определения подобия может быть реализована в различных операционных средах. Различными известными вычислительными системами, средами и

конфигурациями, которые могут быть пригодными для использования, являются, в том числе, персональные компьютеры, серверы, карманные или портативные компьютерные устройства, многопроцессорные системы, системы на основе микропроцессоров, программируемые бытовые электронные устройства, сетевые персональные компьютеры (ПК), миникомпьютеры, большие универсальные вычислительные машины, распределенные вычислительные среды, содержащие любую из вышеупомянутых систем или любое из вышеупомянутых устройств, и т.п.

Система определения подобия может быть описана в общем контексте исполняемых посредством компьютера команд, например программных модулей, выполняемых одним или большим количеством компьютеров или иных устройств. Программные модули обычно содержат подпрограммы, программы, объекты, компоненты, структуры данных и т.д., выполняющие конкретные задачи или реализующие конкретные абстрактные типы данных. Как правило, функциональные возможности программных модулей в различных вариантах осуществления изобретения могут быть объединены или распределены желательным образом.

На Фиг. 2 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом определения функций подобия в одном из вариантов осуществления изобретения. В блоках 201-209 компонент производит циклический выбор каждого типа объекта и определяет функции внутритипового и межтипового подобия для объектов этого типа. В одном из вариантов осуществления изобретения компонент может взаимодействовать с пользователем для определения внутритиповых и межтиповых связей между объектами. Компонент может также определять функции подобия, которые не имеют рекурсивности, основанной на подобии между объектами, например, условиями поиска в запросе, основанными на подобии. В блоке 201 компонент производит выбор следующего типа объекта. В блоке 202 ветвления в том случае, если все типы объектов уже были выбраны, компонент производит возврат, а в противном случае компонент продолжает процедуру, переходя к блоку 203. В блоке 203 компонент производит выбор следующей внутритиповой связи для выбранного типа. В блоке 204 ветвления в том случае, если все внутритиповые связи уже были выбраны, компонент продолжает процедуру, переходя к блоку 206, а в противном случае компонент продолжает процедуру, переходя к блоку 205. В блоке 205 компонент определяет функцию внутритипового подобия для выбранного типа и выбранной связи. Затем компонент производит возврат в начало цикла в блок 203 для выбора следующей внутритиповой связи. В блоке 206 компонент производит выбор следующей межтиповой связи для выбранного типа. В блоке 207 ветвления в том случае, если все межтиповые связи уже были выбраны, компонент продолжает процедуру, переходя к блоку 209, а в противном случае компонент продолжает процедуру, переходя к блоку 208. В блоке 208 компонент определяет функцию межтипового подобия для выбранного типа и выбранной связи. Затем компонент производит возврат в начало цикла в блок 206 для выбора следующей межтиповой связи. В блоке 209 компонент определяет функцию общего подобия путем объединения определенных функций внутритипового подобия и функций межтипового подобия для выбранного типа. Компонент может присвоить каждой из объединенных функций подобия весовые коэффициенты. Затем компонент производит возврат в начало цикла в блок 201 для выбора следующего типа объекта.

На Фиг. 3 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом установления связей в одном из вариантов

осуществления изобретения. Компонент осуществляет обработку собранных данных и осуществляет генерацию данных о связях. В блоках 301-308 компонент производит циклический выбор каждого типа объекта и осуществляет генерацию данных о связях для этого типа объекта. В блоке 301 компонент производит выбор следующего типа объекта. В блоке 302 ветвления в том случае, если все типы уже были выбраны, компонент производит возврат, а в противном случае компонент продолжает процедуру, переходя к блоку 303. В блоке 303 компонент производит выбор следующей внутритиповой связи для выбранного типа. В блоке 304 ветвления в том случае, если все внутритиповые связи уже были выбраны, компонент продолжает процедуру, переходя к блоку 306, а в противном случае компонент продолжает процедуру, переходя к блоку 305. В блоке 305 компонент задает элементы данных о связях для выбранного типа и для выбранной внутритиповой связи. Затем компонент производит возврат в начало цикла в блок 303 для выбора следующей внутритиповой связи. В блоке 306 компонент производит выбор следующей межтиповой связи для выбранного типа. В блоке 307 ветвления в том случае, если все межтиповые связи уже были выбраны, компонент производит возврат в начало цикла в блок 301 для выбора следующего типа объекта, а в противном случае компонент продолжает процедуру, переходя к блоку 308. В блоке 308 компонент задает элементы данных о связях для выбранного типа и для выбранной межтиповой связи. Затем компонент производит возврат в начало цикла в блок 306 для выбора следующей межтиповой связи для выбранного типа.

На Фиг. 4 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом вычисления решения для функций подобия в одном из вариантов осуществления изобретения. В блоке 401 компонент инициализирует матрицы подобия. Например, компонент может установить значения подобия на диагоналях равными единице для указания максимального значения подобия, а другие значения подобия установить равными случайным числам. В блоке 402 компонент устанавливает значение разности равным очень большой величине для того, чтобы обеспечить выполнение, по меньшей мере, одной итерации. В блоках 403-408 компонент производит циклическое вычисление функций общего подобия с множеством итераций для обновления матрицы подобия до тех пор, пока не будет получено решение за счет сходимости значений подобия. В блоке 403 компонент производит выбор следующей итерации. В блоке 404 ветвления в том случае, если сумма значений разности для подобий типов является меньшей, чем пороговая разность, то получено решение за счет сходимости, и компонент производит возврат, а в противном случае компонент продолжает процедуру, переходя к блоку 405. В блоке 405 компонент производит выбор следующего типа объекта. В блоке 406 ветвления в том случае, если все типы уже были выбраны, компонент продолжает процедуру, переходя к блоку 408, а в противном случае компонент продолжает процедуру, переходя к блоку 407. В блоке 407 компонент вычисляет функцию подобия для выбранного типа, обновляя матрицу подобия для выбранного типа, а затем производит возврат в начало цикла в блок 405 для выбора следующего типа. В блоке 408 компонент вычисляет разность между значениями подобия при этой итерации и значениями подобия при предыдущей итерации для выбранного типа. Затем компонент производит возврат в начало цикла в блок 403, начиная следующую итерацию.

На Фиг. 5 изображена схема последовательности операций, на которой показана обработка, выполняемая компонентом вычисления подобий в одном из вариантов

осуществления изобретения. Через компонент пропускается тип объекта, и компонент обновляет матрицы подобия для этого типа. В блоке 501 компонент производит выбор следующей функции внутритипового подобия для пропускаемого типа. В блоке 502 ветвления в том случае, если все функции внутритипового подобия уже были выбраны, компонент продолжает процедуру, переходя к блоку 504, а в противном случае компонент продолжает процедуру, переходя к блоку 503. В блоке 503 компонент вычисляет новое значение подобия для каждого объекта пропускаемого типа. Затем компонент производит возврат в начало цикла в блок 501 для выбора следующей функции внутритипового подобия. В блоке 504 компонент производит выбор следующей функции межтипового подобия для пропускаемого типа объекта. В блоке 505 ветвления в том случае, если все функции межтипового подобия уже были выбраны, компонент продолжает процедуру, переходя к блоку 507, а в противном случае компонент продолжает процедуру, переходя к блоку 506. В блоке 506 компонент вычисляет новые значения подобия для каждого объекта пропускаемого типа с использованием выбранной функции межтипового подобия. Затем компонент производит возврат в начало цикла в блок 504 для выбора следующей функции межтипового подобия. В блоке 507 компонент объединяет матрицы с использованием весовых коэффициентов, осуществляя генерацию общего подобия для пропускаемого типа для текущей итерации. Затем компонент производит возврат.

Специалисту в данной области техники понятно, что хотя описание конкретных вариантов осуществления системы определения подобия приведено в иллюстративных целях, могут быть реализованы различные ее модификации, не выходя за пределы сущности и объема настоящего изобретения. Следовательно, настоящее изобретение ограничено только приложенной формулой изобретения.

Формула изобретения

1. Способ генерации измерений подобия между объектами, осуществляемый в компьютерной системе, причем каждый объект имеет один из множества типов, тип имеет внутритиповую связь, а пара типов имеет межтиповую связь, содержащий следующие операции:

для каждого типа,

если подобие этого типа основано на внутритиповой связи, предоставляют функцию внутритипового подобия для каждой такой связи, которая измеряет подобие между объектами этого типа;

если подобие этого типа основано на межтиповой связи, предоставляют функцию межтипового подобия для каждой такой связи, которая измеряет подобие между объектами этого типа, основываясь на подобии объектов другого типа, причем функция межтипового подобия для типа определяется рекурсивно на основании функции подобия другого типа, при этом функция межтипового подобия генерирует подобие для первого и второго объектов этого типа, которое является взвешенным средним подобия между парами объектов другого типа с одним объектом пары, имеющим связь с первым объектом, и другим объектом пары, имеющим связь с вторым объектом; и

предоставляют функцию подобия, которая измеряет подобие между объектами этого типа, основываясь на любых функциях внутритипового подобия и любых функциях межтипового подобия для этого типа; и

для каждой связи предоставляют данные, которые определяют эту связь между объектами, ассоциированными с этой связью;

одновременно решают предоставленные функции подобия на основании связей, определяемых обеспеченными данными; и

сохраняют подобия, основанные на одновременном решении предоставленных функций подобия.

2. Способ по п.1, в котором функция внутритипового подобия для типа определяется рекурсивно на основании функции подобия этого типа.

3. Способ по п.1, в котором функция подобия для типа представляет собой линейную комбинацию функций внутритипового и межтипового подобия для этого типа.

4. Способ по п.3, в котором каждой функции внутритипового подобия и межтипового подобия присвоен весовой коэффициент.

5. Способ по п.4, в котором сумма весовых коэффициентов функций внутритипового и межтипового подобия для упомянутого типа равна единице.

6. Способ по п.1, в котором функции подобия решаются путем итерационного вычисления подобий для объектов на основании функций подобия.

7. Способ по п.6, в котором функции подобия решаются, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, сходится.

8. Способ по п.6, в котором функции подобия решаются, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, меньше, чем пороговое различие.

9. Способ по п.1, в котором типы объектов включают в себя web-страницы и запросы, а межтиповая связь между запросом и web-страницей основана на указаниях мышью из запроса на web-страницу.

10. Способ по п.1, в котором типы объектов включают в себя web-страницы и запросы, внутритиповые связи для web-страниц основаны на входящих и исходящих ссылках, а межтиповая связь между web-страницей и запросом основана на указаниях мышью из запроса на web-страницу.

11. Машиночитаемый носитель, содержащий команды для управления компьютерной системой для генерации измерений подобия между объектами, причем каждый объект имеет один из множества типов, способом, содержащим

предоставление для каждого типа функции подобия, которая измеряет подобие между объектами этого типа на основании внутритиповых подобий между объектами этого типа, когда для этого типа определено внутритиповое подобие, и межтипового подобия между объектами этого типа на основании подобия объектов другого типа, когда для этого типа определено межтиповое подобие, причем межтиповое подобие для типа определяется рекурсивно на основании подобия объектов другого типа, при этом межтиповое подобие для первого и второго объектов этого типа является взвешенным средним внутритипового подобия между парами объектов другого типа с одним объектом пары, имеющим связь с первым объектом, и другим объектом пары, имеющим связь с вторым объектом;

для каждой связи, предоставление данных, которые определяют эту связь между объектами, ассоциированными с этой связью;

решение предоставленных функций подобия, основанных на связях, определенных предоставленными данными; и

сохранение подобий, основанных на решенных функциях подобия.

12. Машиночитаемый носитель по п.11, в котором функции подобия определяют систему линейных уравнений.

13. Машиночитаемый носитель по п.11, в котором функция подобия определяется

рекурсивно на основании подобий объектов этого типа для различных связей.

14. Машиночитаемый носитель по п.11, в котором функция подобия для типа представляет собой линейную комбинацию внутритиповых и межтиповых подобий для этого типа.

15. Машиночитаемый носитель по п.14, в котором каждому внутритиповому и межтиповому подобию присвоен весовой коэффициент.

16. Машиночитаемый носитель по п.15, в котором сумма весовых коэффициентов для внутритиповых и межтиповых подобий для упомянутого типа равна единице.

17. Машиночитаемый носитель по п.11, в котором функции подобия решаются путем итерационного вычисления подобий для объектов на основании функций подобия.

18. Машиночитаемый носитель по п.17, в котором функции подобия решаются, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, сходится.

19. Машиночитаемый носитель по п.17, в котором функции подобия решаются, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, меньше, чем пороговое различие.

20. Вычислительное устройство для вычисления оценки подобия для объектов, причем каждый объект имеет один из множества типов, и каждый тип имеет межтиповую связь с другим типом, причем вычислительное устройство содержит:

компонент для каждого типа, реализующий функцию подобия, которая предоставляет оценку подобия для пар объектов этого типа на основании межтипового подобия между объектами этого типа и объектами другого типа, при этом межтиповое подобие определяется рекурсивно на основании оценки подобия для пар объектов другого типа и межтиповых связей между парами объектов, при этом межтиповое подобие для первого и второго объектов этого типа является взвешенным средним внутритипового подобия между парами объектов другого типа с одним объектом пары, имеющим связь с первым объектом, и другим объектом пары, имеющим связь с вторым объектом;

компонент, который решает функции подобия на основании межтиповых связей, определенных для набора объектов, путем итерационного вызова компонентов, реализующих функцию подобия, до тех пор, пока не будет получена сходимость оценок подобия; и

компонент, который сохраняет подобия, основанные на решенных функциях подобия.

21. Вычислительное устройство по п.20, в котором функция подобия определяет систему линейных уравнений.

22. Вычислительное устройство по п.20, в котором тип имеет внутритиповую связь между объектами этого типа, и в котором функция подобия дополнительно основана на внутритиповых подобиях между объектами этого типа на основании внутритиповых связей между объектами.

23. Вычислительное устройство по п.22, в котором внутритиповое подобие для типа основано на межтиповых подобиях между объектами этого типа и другого типа.

24. Вычислительное устройство по п.22, в котором функция подобия определяется рекурсивно на основании внутритиповых подобий объектов для различных внутритиповых связей.

25. Вычислительное устройство по п.22, в котором функция подобия для типа представляет собой линейную комбинацию функций внутритипового и межтипового

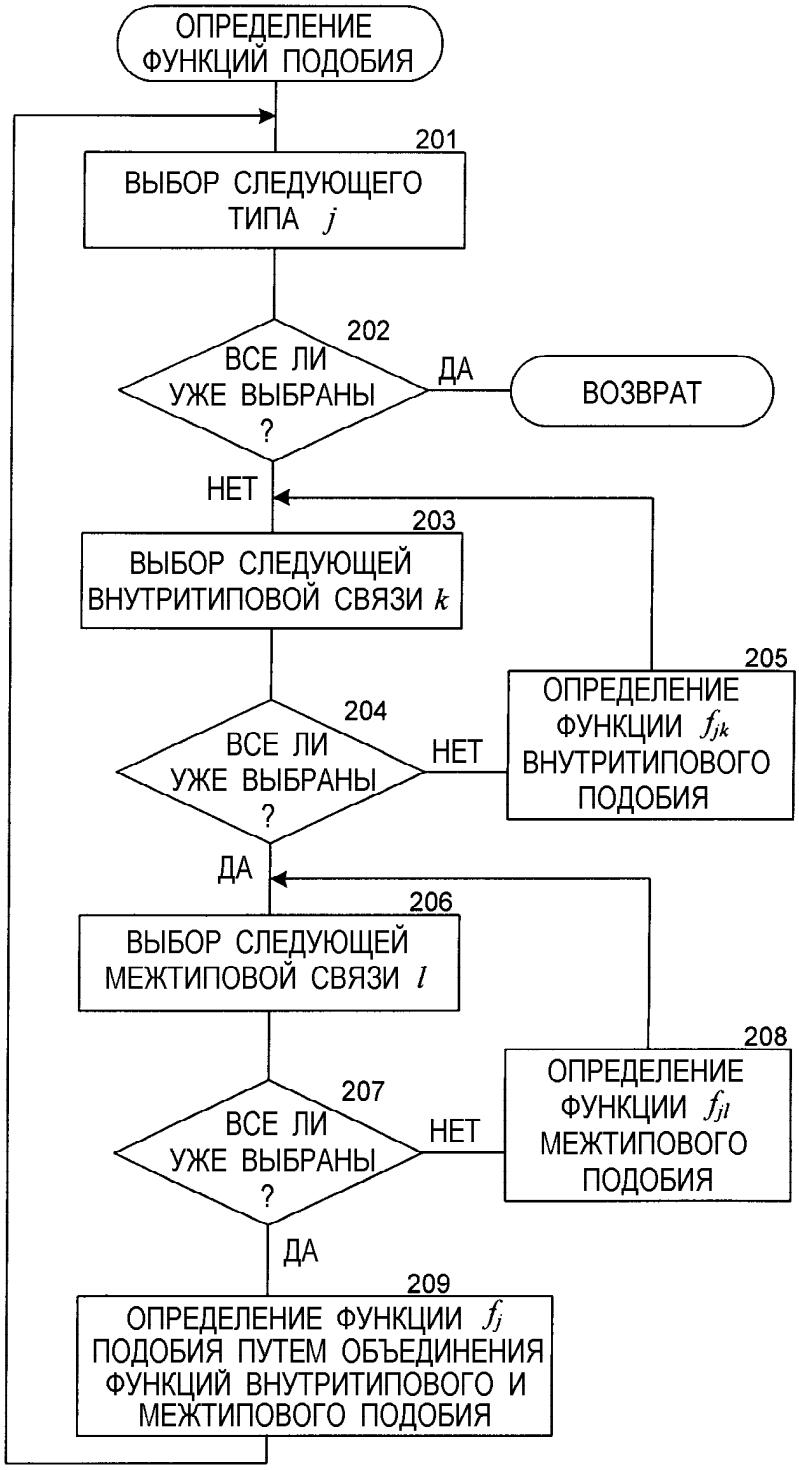
подобия для этого типа.

26. Вычислительное устройство по п.25, в котором каждой функции внутритипового и межтипового подобия присвоен весовой коэффициент.

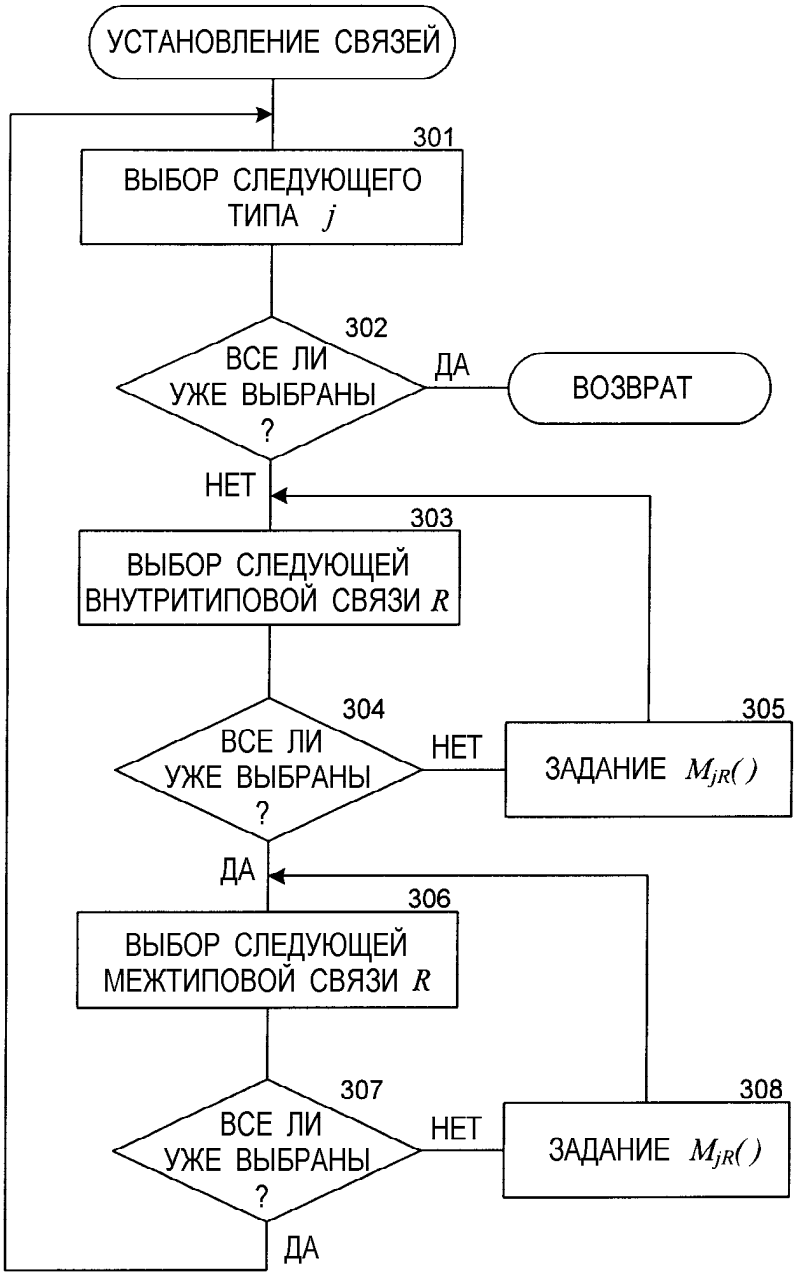
27. Вычислительное устройство по п.26, в котором сумма весовых коэффициентов функций внутритипового и межтипового подобия для упомянутого типа равна единице.

28. Вычислительное устройство по п.20, в котором оценки подобия сходятся, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, меньше, чем пороговое различие.

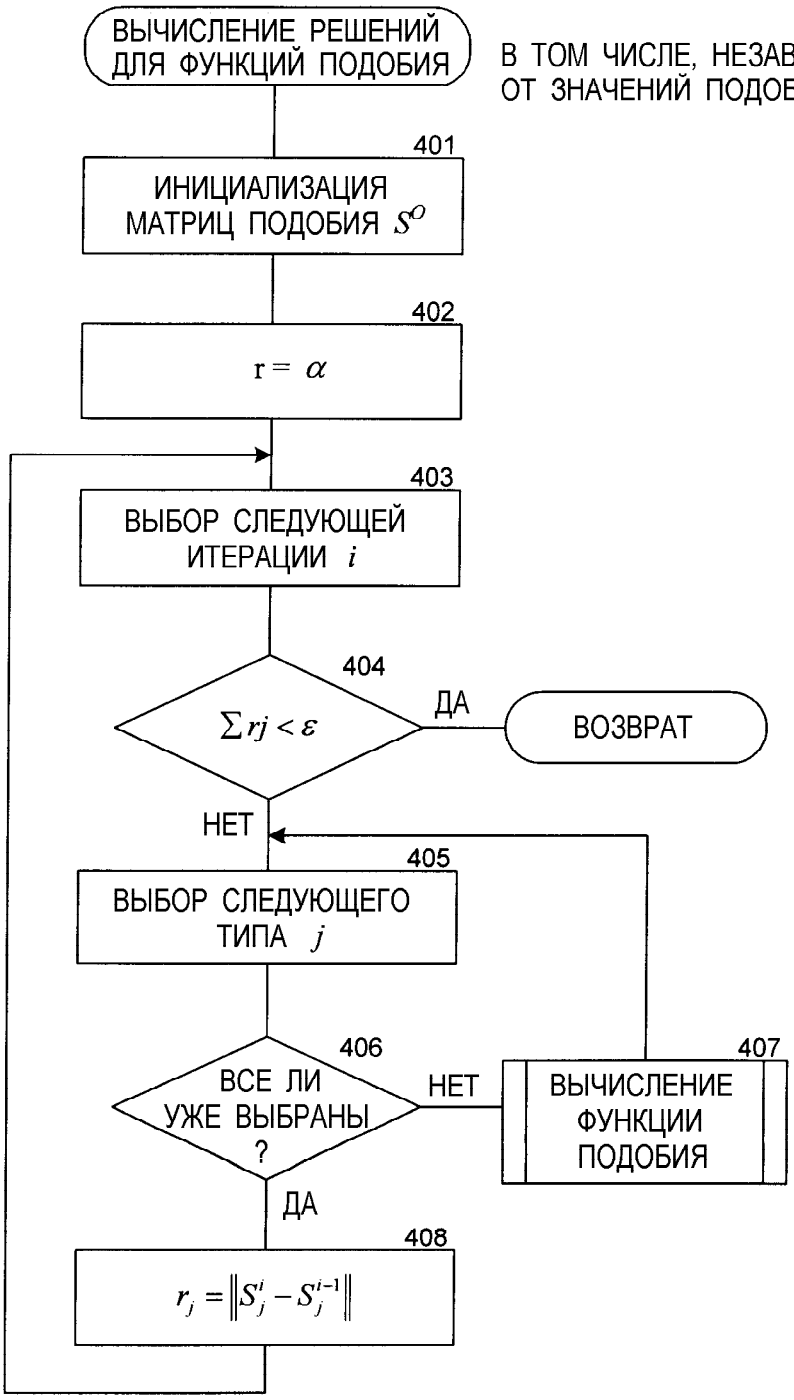
29. Вычислительное устройство по п.20, в котором функции подобия решаются, когда мера различия, основанная на подобиях от одной итерации к следующей итерации, меньше, чем пороговое различие.


ОПРЕДЕЛЕНИЕ ФУНКЦИЙ ПОДОБИЯ
201
ВЫБОР СЛЕДУЮЩЕГО ТИПА j
202
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
ДА
ВОЗВРАТ
НЕТ
203
ВЫБОР СЛЕДУЮЩЕЙ ВНУТРИТИПОВОЙ СВЯЗИ k
205
ОПРЕДЕЛЕНИЕ ФУНКЦИИ f_{jk} ВНУТРИТИПОВОГО ПОДОБИЯ
204
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
ДА
206
ВЫБОР СЛЕДУЮЩЕЙ МЕЖТИПОВОЙ СВЯЗИ l
208
ОПРЕДЕЛЕНИЕ ФУНКЦИИ f_{jl} МЕЖТИПОВОГО ПОДОБИЯ
207
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
ДА
209
ОПРЕДЕЛЕНИЕ ФУНКЦИИ f_j ПОДОБИЯ ПУТЕМ ОБЪЕДИНЕНИЯ ФУНКЦИЙ ВНУТРИТИПОВОГО И МЕЖТИПОВОГО ПОДОБИЯ

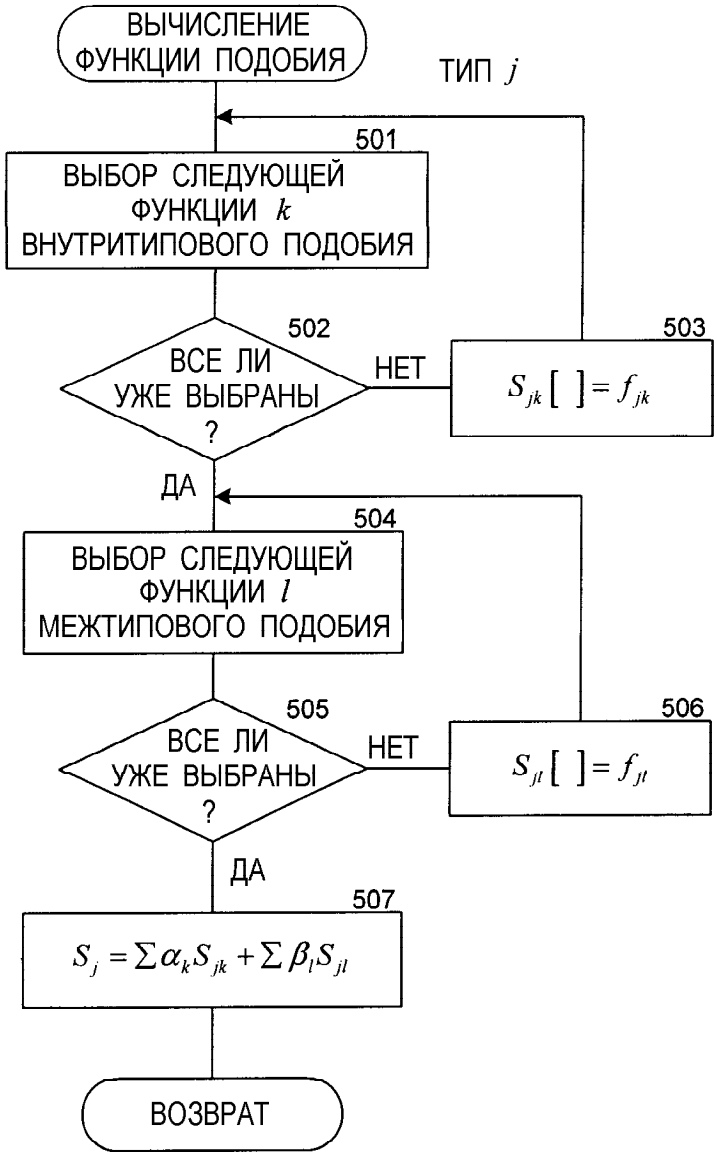
Фиг. 2


УСТАНОВЛЕНИЕ СВЯЗЕЙ
301
ВЫБОР СЛЕДУЮЩЕГО ТИПА j
302
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
ДА
ВОЗВРАТ
НЕТ
303
ВЫБОР СЛЕДУЮЩЕЙ ВНУТРИТИПОВОЙ СВЯЗИ R
304
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
305
ЗАДАНИЕ $M_{jR}()$
ДА
306
ВЫБОР СЛЕДУЮЩЕЙ МЕЖТИПОВОЙ СВЯЗИ R
307
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
308
ЗАДАНИЕ $M_{jR}()$
ДА

Фиг. 3


ВЫЧИСЛЕНИЕ РЕШЕНИЙ ДЛЯ ФУНКЦИЙ ПОДОБИЯ
В ТОМ ЧИСЛЕ, НЕЗАВИСИМО ОТ ЗНАЧЕНИЙ ПОДОБИЯ
401
ИНИЦИАЛИЗАЦИЯ МАТРИЦ ПОДОБИЯ S^0
402
$r = \alpha$
403
ВЫБОР СЛЕДУЮЩЕЙ ИТЕРАЦИИ i
404
$\sum rj < \varepsilon$
ДА
ВОЗВРАТ
НЕТ
405
ВЫБОР СЛЕДУЮЩЕГО ТИПА j
406
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
407
ВЫЧИСЛЕНИЕ ФУНКЦИИ ПОДОБИЯ
ДА
408
$r_j = \|S_j^i - S_j^{i-1}\|$

Фиг. 4


ВЫЧИСЛЕНИЕ ФУНКЦИИ ПОДОБИЯ
ТИП j
501
ВЫБОР СЛЕДУЮЩЕЙ ФУНКЦИИ k ВНУТРИТИПОВОГО ПОДОБИЯ
502
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
503
$S_{jk}[\ ] = f_{jk}$
ДА
504
ВЫБОР СЛЕДУЮЩЕЙ ФУНКЦИИ l МЕЖТИПОВОГО ПОДОБИЯ
505
ВСЕ ЛИ УЖЕ ВЫБРАНЫ ?
НЕТ
506
$S_{jl}[\ ] = f_{jl}$
ДА
507
$S_j = \Sigma \alpha_k S_{jk} + \Sigma \beta_l S_{jl}$
ВОЗВРАТ

Фиг. 5