РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 377 645** (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: **2005113190/09, 29.04.2005**

(24) Дата начала отсчета срока действия патента:
29.04.2005

(30) Конвенционный приоритет:
30.04.2004 US 10/836,319

(43) Дата публикации заявки: **10.11.2006**

(45) Опубликовано: **27.12.2009** Бюл. № **36**

(56) Список документов, цитированных в отчете о поиске: **JP 2003122608 A, 25.04.2003. RU 2210809 C2, 20.08.2003. US 6466940 B1, 15.10.2002. JP 4027271 A1, 30.01.1992. CHUE W L et al, "SVD: a novel content - based representation technique for web documents", INFORMATION COMMUNICATION AND SIGNAL PROCESSING, 2003 AND FOURTH PACIFIC RIMM CONFERENCE ON MULTIMEDIA PTROCEEDING OF THE 2003 JOINT CONFERENCE OF** (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),
ШЭНЬ До (US),
ЦЗЭН Хуа-Цзюнь (US),
МА Вэй-Ин (US),
ЧЭНЬ Чжэн (US)

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

(54) СПОСОБ И СИСТЕМА ДЛЯ КЛАССИФИКАЦИИ ДИСПЛЕЙНЫХ СТРАНИЦ С ПОМОЩЬЮ РЕФЕРАТОВ

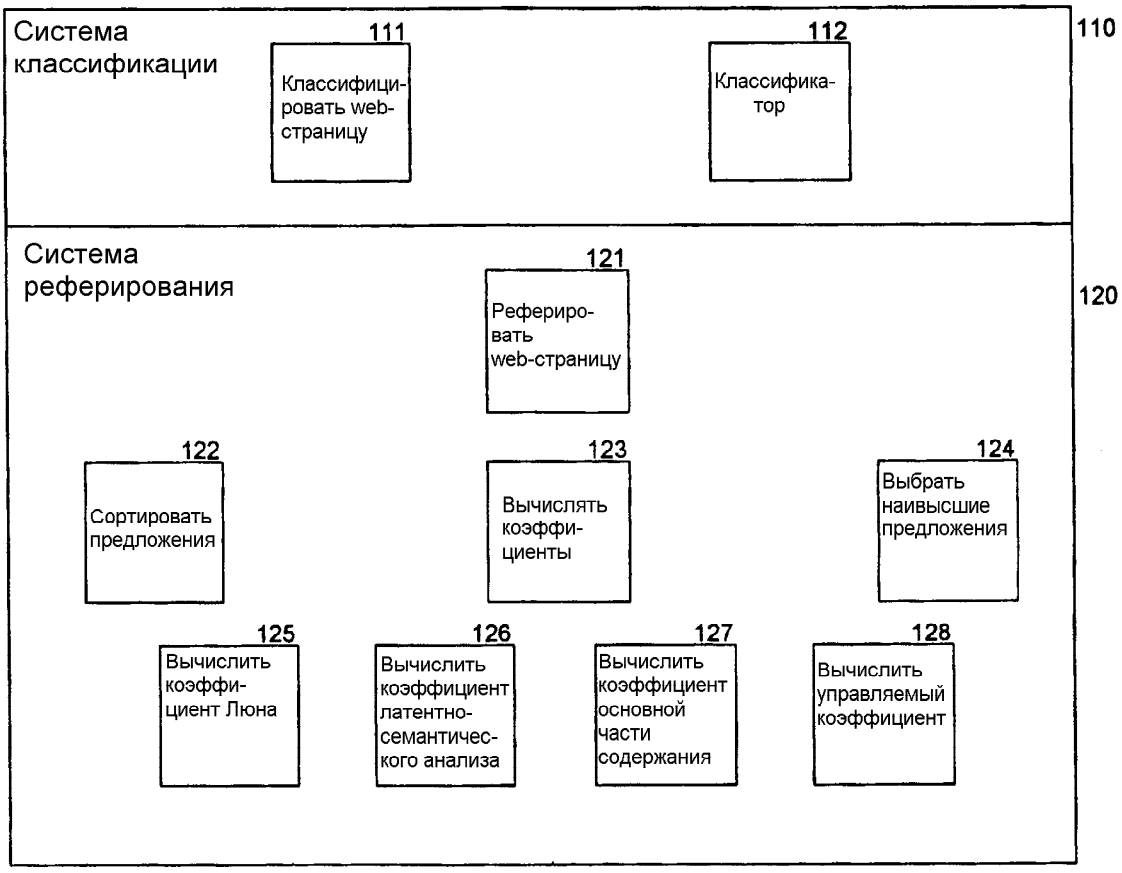
(57) Реферат:

Изобретение относится к средствам обеспечения классификации информации. Техническим результатом является повышение достоверности обрабатываемой информации. Система классификации web-страниц использует систему реферирования web-страниц для выработки рефератов web-страниц. Реферат web-страницы может включать в себя предложения web-страницы, которые являются наиболее тесно связанными с главной темой web-страницы. Система реферирования может объединять преимущества множества методов реферирования, чтобы выявлять предложения web-страницы, которые представляют главную тему web-страницы. Когда реферат выработан, система классификации может применить традиционные методы классификации к реферату, чтобы классифицировать web-страницу. Система классификации может использовать традиционные методы классификации, такие как упрощенный байесовский классификатор или метод опорных векторов, чтобы выявить

RU 2377645 C2

RU 2377645 C2

классификации web-страницы на основании реферата, выработанного системой реферирования. 7 н. и 59 з.п. ф-лы, 9 ил., 1 табл.



ФИГ.1

(56) (продолжение):
THE FOURTH INTERNATIONAL CONFERENCE ON SINGAPORE 15-18 DEC.2003, PISCATAWAY, NJ, USA, IEEE, VOL,3, 15 December 2003, pages 1840-1844.

RU 2377645 C2

RU 2377645 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(19) **RU** (11) **2 377 645** (13) **C2**

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

(12) ABSTRACT OF INVENTION

(21), (22) Application: **2005113190/09, 29.04.2005**

(24) Effective date for property rights:
29.04.2005

(30) Priority:
30.04.2004 US 10/836,319

(43) Application published: **10.11.2006**

(45) Date of publication: **27.12.2009 Bull. 36**

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),
ShEhN' Do (US),
TsZEhN Khua-Tszjun' (US),
MA Vehj-In (US),
ChEhN' Chzhehn (US)

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

(54) METHOD AND SYSTEM FOR CLASSIFYING DISPLAY PAGES USING SUMMARIES

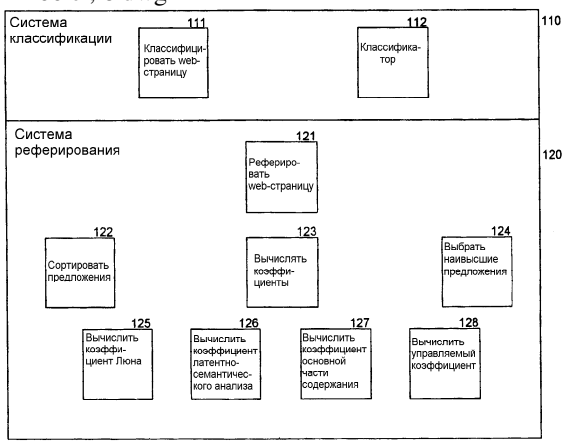
(57) Abstract:

FIELD: physics; computer engineering.

SUBSTANCE: invention relates to means of classifying information. A system for classifying web pages uses a web page summarisation system to generate summaries of web pages. The summary of a web page may include sentences of the web page which are most closely related to the primary topic of the web page. The summarisation system may combine advantages of multiple summarisation techniques to identify sentences of a web page that represent the primary topic of the web page. Once the summary is generated, the classification system can apply conventional classification techniques to the summary to classify the web page. The classification system can use conventional classification techniques such as a simplified Bayesian classifier or a support vector technique to identify the classifications of a web page based on the summary generated by the summarisation system.

EFFECT: increased reliability of processed information.

66 cl, 8 dwg



ФИГ.1

RU 2377645 C2

RU 2377645 C2

Область техники, к которой относится изобретение

Описанная технология относится в общем к автоматической классификации информации.

Предшествующий уровень техники

Многие услуги поисковых средств, такие как Google и Overture, обеспечивают поиск информации, которая доступна через Интернет. Эти услуги поисковых механизмов позволяют пользователям искать дисплейные страницы, такие как web-страницы, которые могут интересовать пользователей. После того как пользователь представляет запрос, который включает в себя поисковые термины, услуга поискового средства выявляет web-страницы, которые могут относиться к этим поисковым терминам. Чтобы быстро выявить релевантные web-страницы, услуги поисковых средств могут поддерживать отображение кодовых слов в web-страницы. Это отображение может вырабатываться путем «ползания» по сети (т.е. по Всемирной паутине), чтобы выявить ключевые слова каждой web-страницы. Для осуществления ползания по сети услуга поискового средства может использовать список корневых web-страниц для выявления всех web-страниц, которые доступны через эти корневые web-страницы. Ключевые слова любой конкретной web-страницы могут быть выявлены с помощью различных общеизвестных информационных поисковых методов, таких как выявление слов заголовка, слов, введенных в метаданные web-страницы, слов, которые выделяются, и т.д. Услуга поискового средства может вырабатывать коэффициент релевантности, чтобы указывать, насколько релевантной может быть информация web-страницы для поискового запроса, на основании близости каждого совпадения, популярности web-страницы (к примеру, ранг страницы (PageRank) Google) и т.д. Услуга поискового средства затем отображает пользователю ссылки на эти web-страницы в порядке, который базируется на их ранжировании.

Хотя услуги поисковых средств могут выдавать много web-страниц в качестве результата поиска, представление web-страниц в ранговом порядке может затруднить пользователю реальное нахождение тех web-страниц, которые особенно интересны пользователю. Поскольку web-страницы, которые представлены первыми, могут быть направлены на популярные темы, пользователю, интересующемуся неясной темой, может понадобиться просмотреть многие страницы результата поиска, чтобы найти интересующую его web-страницу. Чтобы облегчить пользователю нахождение интересующих его web-страниц, web-страницы результата поиска могут быть представлены в иерархической организации на основании какого-нибудь распределения по классам или категориям web-страниц. Например, если пользователь представил поисковый запрос на «court battles» («битвы на корте» или «судебные схватки»), результат поиска может содержать web-страницы, которые можно классифицировать как относящиеся к спорту или относящиеся к праву. Пользователь может предпочесть, чтобы ему сначала представили перечень классификаций web-страниц, так что пользователь сможет выбрать классификацию web-страниц, которая его интересует. К примеру, пользователю сначала может быть представлено указание, что web-страницы результата поиска классифицированы как относящиеся к спорту и относящиеся к праву. Пользователь затем может выбрать относящуюся к праву классификацию, чтобы просмотреть web-страницы, которые относятся к праву. В противоположность этому, поскольку спортивные web-страницы более популярны, чем правовые web-страницы, пользователю может потребоваться просмотреть много страниц, чтобы найти относящиеся к праву web-страницы, если наиболее

популярные web-страницы представлены первыми.

Было бы непрактично вручную классифицировать миллионы web-страниц, которые доступны в настоящее время. Хотя для классификации основанного на тексте содержания использованы методы автоматической классификации, эти методы не применимы в общем случае к классификации web-страниц. Web-страницы имеют организацию, которая включает в себя шумовое содержание, такое как реклама или навигационная панель, которое не относится напрямую к главной теме web-страницы. В силу того что традиционные методы основанной на тексте классификации будут использовать такое шумовое содержание при классификации web-страниц, эти методы будут иметь тенденцию вырабатывать неверные классификации web-страниц.

Желательно иметь метод классификации для web-страниц, который базировал бы классификацию web-страниц на главной теме web-страницы и придавал мало значения шумовому содержанию web-страницы.

Раскрытие изобретения

Система классификации и обобщения классифицирует дисплейные страницы, такие как web-страницы, на основании автоматически вырабатываемых рефератов дисплейных страниц. В одном варианте осуществления система классификации web-страниц использует систему реферирования web-страниц, чтобы вырабатывать рефераты web-страниц. Реферат web-страницы может включать в себя предложения этой web-страницы, которые наиболее тесно связаны с главной темой web-страницы. Система реферирования может сочетать преимущества многих методов реферирования, чтобы выявлять предложения web-страницы, которые представляют главную тему web-страницы. Когда реферат вырабатывается, система классификации может применить традиционные методы классификации к реферату, чтобы классифицировать web-страницу.

Краткое описание чертежей

Фиг.1 - блок-схема, которая иллюстрирует компоненты системы классификации и системы реферирования в одном варианте осуществления.

Фиг.2 - блок-схема алгоритма, которая иллюстрирует работу компонента классификации web-страницы в одном варианте осуществления.

Фиг.3 - блок-схема алгоритма, которая иллюстрирует работу компонента реферирования web-страницы в одном варианте осуществления.

Фиг.4 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления коэффициентов в одном варианте осуществления.

Фиг.5 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления коэффициентов Люна (Luhn) в одном варианте осуществления.

Фиг.6 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления коэффициента латентно-семантического анализа в одном варианте осуществления.

Фиг.7 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления коэффициента основной части содержания в одном варианте осуществления.

Фиг.8 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления управляемого коэффициента в одном варианте осуществления.

Фиг.9 - блок-схема алгоритма, которая иллюстрирует работу компонента вычисления объединенного коэффициента в одном варианте осуществления.

Подробное описание

Предлагаются способ и система для классификации дисплейных страниц на

основании автоматически вырабатываемых рефератов дисплейных страниц. В одном варианте осуществления система классификации web-страниц использует систему реферирования web-страниц, чтобы вырабатывать рефераты web-страниц. Реферат web-страницы может включать в себя предложения этой web-страницы, которые наиболее тесно связаны с главной темой web-страницы. После того как реферат выработан, система классификации может применить традиционные методы классификации к реферату, чтобы классифицировать web-страницу. Система реферирования может сочетать преимущества многих методов реферирования, чтобы выявлять предложения web-страницы, которые представляют главную тему web-страницы. В одном варианте осуществления система реферирования использует метод реферирования Люна, метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования либо по отдельности, либо в сочетании, чтобы вырабатывать реферат. Система реферирования использует каждый из методов реферирования, чтобы вырабатывать специфичный для конкретного метода реферирования коэффициент для каждого предложения web-страницы. Затем система реферирования комбинирует специфичные для конкретного метода реферирования коэффициенты для предложения, чтобы выработать общий коэффициент для этого предложения. Система реферирования выбирает предложения web-страницы с наивысшими общими коэффициентами, чтобы сформировать реферат web-страницы. Система классификации может использовать традиционные методы классификации, такие как упрощенный байесовский классификатор или метод опорных векторов, чтобы выявлять классификации web-страницы на основании реферата, выработанного системой реферирования. При этом web-страницы могут автоматически классифицироваться на основании автоматической выработки рефератов web-страниц.

В одном варианте осуществления система реферирования использует модифицированную версию метода реферирования Люна, чтобы вырабатывать коэффициент Люна для каждого предложения web-страницы. Согласно методу реферирования Люна вырабатывается коэффициент для предложения, который базируется на «значимых словах», имеющихся в этом предложении. Чтобы выработать коэффициент для предложения, согласно методу реферирования Люна выявляется часть предложения, заключенная между значимыми словами, которые разнесены не более чем на определенное число незначащих слов. Согласно методу реферирования Люна коэффициент предложения вычисляется как отношение квадрата числа значащих слов, содержащихся в упомянутой заключенной между значащими словами части, к числу слов в этой заключенной между значащими словами части. (См. H.P. Luhn. *The Automatic Creation of Literature Abstracts* [Автоматическое создание литературных рефератов], 2 IBM J. Of Res. & Dev. No. 2,

159-65 (April 1958).) Система реферирования модифицирует метод реферирования Люна путем определения совокупности значащих слов для каждой классификации. К примеру, относящаяся к спорту классификация может иметь совокупность значащих слов, которая включает в себя «корт» («court»), «баскетбол» и «спорт», тогда как относящаяся к праву классификация может иметь совокупность значащих слов, которая включает в себя «суд» («court»), «адвокат» и «преступник». Система реферирования может выявлять совокупности значащих слов на основании обучающего набора web-страниц, которые классифицированы заранее. Система реферирования может выбирать наиболее часто используемые слова на web-страницах с определенной классификацией в качестве совокупности значащих слов для

классификации. Система реферирования может также удалять из совокупности некоторые стоп-слова, которые могут представлять шумовое содержание. При подсчете коэффициента предложения web-страницы согласно модифицированному методу реферирования Люна вычисляется коэффициент для каждой классификации. Метод реферирования затем усредняет коэффициенты для каждой классификации, которые находятся над пороговым уровнем, чтобы выдать комбинированный коэффициент Люна для предложения. Система реферирования может выбирать предложения с наивысшими коэффициентами Люна для формирования реферата.

В одном варианте осуществления система реферирования использует метод реферирования на основе латентно-семантического анализа, чтобы вырабатывать коэффициент латентно-семантического анализа для каждого предложения web-страницы. Метод реферирования на основе латентно-семантического анализа использует декомпозицию по сингулярным значениям, чтобы вырабатывать коэффициент для каждого предложения. Система реферирования вырабатывает матрицу слово-предложение для web-страницы, которая содержит взвешенное значение термин-частота для каждой комбинации слово-предложение. Эта матрица может быть представлена следующим образом:

$$A=U\Sigma V^T \quad (1)$$

где A представляет матрицу слово-предложение, U является матрицей ортонормированных столбцов, столбцы которой являются левыми сингулярными векторами (представляет собой диагональную матрицу, диагональные элементы которой являются неотрицательными сингулярными значениями, рассортированными в убывающем порядке), а V является ортонормированной матрицей, столбцы которой являются правыми сингулярными векторами. После декомпозиции матрицы на U (и V) система реферирования использует правые сингулярные векторы для выработки коэффициентов для предложений. (См. Y.H. Gong & X. Liu. *Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis* [Реферирование общего текста с помощью меры релевантности и латентно-семантического анализа] в Proc. Of the 24th Annual International ACM SIGIR, New Orleans, Louisiana, 19-25 (2001).) Система реферирования может выбирать первый правый сингулярный вектор и выбирать предложение, которое имеет наивысшее значение индекса в этом векторе. Система реферирования далее задает этому предложению наивысший коэффициент. Система реферирования затем выбирает второй правый сингулярный вектор и задает предложению, которое имеет наивысшее значение индекса в этом векторе, второй наивысший коэффициент. Система реферирования далее продолжает таким же образом вырабатывать коэффициенты для остальных предложений. Система реферирования может выбирать предложения с наивысшими коэффициентами, чтобы сформировать реферат web-страницы.

В одном варианте осуществления система реферирования использует метод реферирования основной части содержания, чтобы вырабатывать коэффициент основной части содержания для каждого предложения web-страницы. Метод реферирования основной части содержания выявляет основную часть содержания web-страницы и задает наивысший коэффициент предложениям в этой основной части содержания. Чтобы выявить основную часть содержания web-страницы, метод реферирования основной части содержания выявляет базовые объекты и составные объекты web-страницы. Базовый объект представляет собой наименьшую информационную область, которую нельзя разделить дальше. Например, в HTML

(языке гипертекстовой разметки) базовым объектом является неделимый элемент внутри двух тегов (неотображаемых элементов разметки) или внедренный объект. Составным объектом является набор базовых объектов или иных составных объектов, которые скомбинированы для выполнения некоторой функции. После выявления объектов система реферирования разделяет объекты на категории, такие как информация, навигация, взаимодействие, украшение или специальная функция. Категория информации служит для объектов, которые представляют содержательную информацию, категория навигации служит для объектов, которые представляют руководство по навигации, категория взаимодействия служит для объектов, которые представляют пользовательское взаимодействие (к примеру, поле ввода), категория украшения служит для объектов, которые представляют украшения, а категория специальной функции служит для объектов, которые представляют такую информацию, как правовая информация, контактная информация, информация логотипа и т.д. (См. J.L. Chen, et al. *Function-based Object Model Towards Website Adaptation* [Основанная на функции объектная модель для адаптации web-сайта], Proc. Of WWW10, Hong Kong, China (2001).) В одном варианте осуществления система реферирования строит частоту появления термина инвертированным индексом частоты документа (т.е. TF*IDF) для каждого объекта. Затем система реферирования вычисляет подобие между парами объектов с помощью вычисления подобия, такого как косинусное подобие. Если подобие между объектами пары больше, чем пороговый уровень, система реферирования связывает объекты пары. Далее система реферирования идентифицирует объект, который имеет наибольшее число связей к нему, в качестве сердцевинного объекта, который представляет главную тему web-страницы. Основная часть содержания web-страницы является сердцевинным объектом вместе с каждым объектом, который имеет связь с этим сердцевинным объектом. Система реферирования выдает высокий коэффициент каждому предложению основной части содержания и низкий коэффициент каждому иному предложению web-страницы. Система реферирования может выбирать предложения с высоким коэффициентом, чтобы сформировать реферат web-страницы.

В одном варианте осуществления система реферирования использует метод управляемого реферирования, чтобы вырабатывать управляемый коэффициент для каждого предложения web-страницы. Метод управляемого реферирования использует обучающие данные для обучения функции реферирования, которая выявляет, следует ли выбирать предложение как часть реферата. Метод управляемого реферирования представляет одно предложение вектором признаков. В одном варианте осуществления метод управляемого реферирования использует признаки, определенные в Таблице 1, где f_{ij} представляет значение *i*-го признака в предложении *j*.

Таблица 1

Признак	Описание
f_{i1}	Позиция предложения S_i в содержащем его абзаце.
f_{i2}	Длина предложения S_i, которая является числом слов в S_i.
f_{i3}	(TF_W*SF_W, которая учитывает не только число слов W, но также его распределение по предложениям, где TF_W есть число появлений слова W на целевой web-странице и где SF_W есть число предложений, включающих в себя слово W на целевой web-странице.
f_{i4}	Подобие между S_i и заглавием, что можно вычислить как скалярное произведение между предложением и заглавием.
f_{i5}	Косинусное подобие между S_i и всем текстом на web-странице.
f_{i6}	Косинусное подобие между S_i и метаданными web-страницы.

f_{i7}	Число появлений слова из специального набора слов, которые имеются в S_i. Специальный набор слов можно построить собиранием на web-странице слов, которые выделяются (например, курсивом, жирным шрифтом или подчеркиванием).
f_{i8}	Средний размер шрифта слов в S_i. В общем, чем больше размер шрифта на web-странице, тем выше важность.

Система реферирования может использовать упрощенный байесовский классификатор для обучения функции реферирования. Функция реферирования может быть представлена следующим уравнением:

$$p(s \in S \mid f_1, f_2 \ldots f_8) = \frac{\prod_{j=1}^{8} p(f_j \mid s \in S)\, p(s \in S)}{\prod_{j=1}^{8} p(f_j)} \qquad (2)$$

где $p(s \in S)$ означает степень сжатия реферата (которая может быть заранее определена для различных приложений), $p(f_j)$ есть вероятность каждого признака j, а $p(f_j \mid s \in S)$ есть условная вероятность каждого признака j. Два последних фактора можно оценить из обучающего набора.

В одном варианте осуществления система реферирования комбинирует коэффициенты метода реферирования Люна, метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования, чтобы выработать общий коэффициент. Коэффициенты могут комбинироваться следующим образом:

$$S = S_{luhn} + S_{lsa} + S_{cb} + S_{sup} \qquad (3)$$

где S представляет скомбинированный коэффициент, S_{luhn} представляет коэффициент Люна, S_{lsa} представляет коэффициент латентно-семантического анализа, S_{cb} представляет коэффициент основной части содержания, а S_{sup} представляет управляемый коэффициент. Альтернативно система реферирования может применять весовой фактор для коэффициента каждого метода реферирования, так чтобы коэффициенты не всех методов реферирования были взвешены одинаково. Например, считается, что коэффициент Люна более точно отражает соотношение предложения с главной темой web-страницы, тогда весовой множитель для коэффициента Люна может быть 0,7, а весовые множители для остальных коэффициентов могут быть 0,1 для каждого. Если весовой множитель для метода реферирования установлен на нуль, то система реферирования не использует этот метод реферирования. Специалисту в данной области техники должно быть понятно, что любое число методов реферирования может иметь свои веса, установленные на нуль. К примеру, если весовой множитель 1 используется для коэффициента Люна и нуль для остальных коэффициентов, то «скомбинированный» коэффициент будет просто коэффициентом Люна. Кроме этого система реферирования может нормировать каждый из коэффициентов методов реферирования. Система реферирования может также использовать нелинейную комбинацию коэффициентов методов реферирования. Система реферирования может выбирать предложения с наивысшими скомбинированными коэффициентами, чтобы сформировать реферат web-страницы.

В одном варианте осуществления система классификации использует упрощенный байесовский классификатор, чтобы классифицировать web-страницу на основании реферата. Упрощенный байесовский классификатор использует правило Байеса, которое можно определить следующим образом:

$$P(c_j \mid d_i;\hat{\theta}) = \frac{P(c_j \mid \hat{\theta}) \prod_{k=1}^{n} P(w_k \mid c_j;\hat{\theta})^{N(w_k,d_i)}}{\sum_{r=1}^{|C|} P(c_r \mid \hat{\theta}) \prod_{k=1}^{n} P(w_k \mid c_r;\hat{\theta})^{N(w_k,d_i)}} \quad (4)$$

где $P(c_j \mid d_i;\hat{\theta})$ можно вычислить путем подсчета частоты каждой категории c_j, появляющейся в обучающих данных, $|C|$ есть число категорий, $p(w_i/c_j)$ есть вероятность того, что слово w_i появляется в классе c_i, $N(w_k,d_i)$ есть число появлений слова w_k в d_i, а n есть число слов в обучающих данных. (См. A. McCallum & K. Nigam. *A Comparison of Event Models for Naive Bayes Text Classification* [Сравнение моделей событий для упрощенной байесовской классификации текстов] в AAAI-98 Workshop on Learning for Text Categorization (1998).) Поскольку w_i может быть малым в обучающих данных, для оценки его значения можно использовать сглаживание Лапласа.

В альтернативном варианте осуществления система классификации использует метод опорных векторов для классификации web-страницы на основании ее реферата. Метод опорных векторов работает путем нахождения гиперповерхности в пространстве возможных входных данных. Гиперповерхность пытается отделить положительные примеры от отрицательных примеров путем максимизации расстояния между ближайшими к гиперплоскости из положительного и отрицательного примеров. Это обеспечивает правильную классификацию данных, которые подобны, но не идентичны обучающим данным. Можно использовать различные методы для обучения метода опорных векторов. Одна методика использует алгоритм последовательной минимальной оптимизации, который разбивает большую задачу квадратичного программирования на ряд малых задач квадратичного программирования, которые можно разрешить аналитически. (См. Sequential Minimal Optimization [Последовательная минимальная оптимизация] на http: //research.microsoft.com/~jplatt/smo.html.)

Фиг.1 является блок-схемой, которая иллюстрирует компоненты системы классификации и системы реферирования в одном варианте осуществления. Система 110 классификации включает в себя компонент 111 классификации web-страницы и компонент-классификатор 112. Система 120 реферирования включает в себя компонент 121 реферирования web-страницы, компонент 122 сортировки предложений, компонент 123 вычисления коэффициентов и компонент 124 выбора наивысших предложений. Компонент классификации web-страницы использует компонент реферирования web-страницы, чтобы вырабатывать реферат web-страницы, а затем использует компонент-классификатор, чтобы классифицировать web-страницу на основании реферата. Компонент реферирования web-страницы использует компонент вычисления коэффициентов, чтобы вычислять коэффициент для каждого предложения web-страницы. Компонент реферирования web-страницы затем использует компонент сортировки предложений, чтобы сортировать предложения web-страницы на основании их коэффициентов, и компонент выбора наивысших предложений, чтобы выбирать предложения с наивысшими коэффициентами, для формирования реферата web-страницы. Компонент вычисления коэффициентов использует компонент 125 вычисления коэффициента Люна, компонент 126 вычисления коэффициента латентно-семантического анализа, компонент 127 вычисления основной части содержания и компонент 128 вычисления управляемого коэффициента, чтобы вырабатывать коэффициенты для разных

методов реферирования. Затем компонент вычисления коэффициентов комбинирует коэффициенты для методов реферирования, чтобы выдать общий коэффициент для каждого предложения.

Вычислительное устройство, на котором реализуется система реферирования, может включать в себя центральный процессор, память, устройства ввода (к примеру, клавиатуру и координатно-указательные устройства), устройства вывода (к примеру, устройства визуального отображения) и устройства хранения (к примеру, дисководы). Память и устройства хранения являются машиночитаемыми носителями, которые могут содержать команды, которые воплощают систему реферирования. Кроме этого структуры данных и структуры сообщений могут храниться или передаваться носителем передачи данных, таким как сигнал на линии связи. Можно использовать различные линии связи, такие как Интернет, локальная сеть, региональная сеть или коммутируемое соединение от точки к точке.

Система реферирования может быть реализована в различных операционных средах. Описанная здесь операционная среда является лишь одним примером подходящей операционной среды и не предназначена налагать какое-либо ограничение в отношении объема использования или функций системы реферирования. Иные общеизвестные вычислительные системы, среды и конфигурации, которые могут быть пригодны для использования, включают в себя персональные компьютеры, серверные компьютеры, ручные или портативные устройства, многопроцессорные системы, основанные на микропроцессоре системы, программируемую бытовую электронику, сетевые ПК, мини-компьютеры, универсальные компьютеры, распределенные вычислительные среды, которые включают в себя любые из вышеуказанных систем или устройств, и т.п.

Система реферирования может быть описана в общем контексте машиноисполняемых команд, таких как программные модули, исполняемые одним или более компьютерами или иными устройствами. В общем программные модули включают в себя процедуры, программы, объекты, компоненты, структуры данных и т.д., которые выполняют конкретные задачи или воплощают определенные абстрактные типы данных. Как правило, функции программных модулей могут комбинироваться или распределяться в различных вариантах выполнения, как желательно.

Фиг.2 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента классификации web-страницы в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он выдает ее классификации. На этапе 201 компонент вызывает компонент реферирования web-страницы, чтобы выработать реферат для этой web-страницы. На этапе 202 компонент классифицирует web-страницу на основании реферата web-страницы с помощью классификатора, такого как упрощенный байесовский классификатор или метод опорных векторов. Затем этот компонент завершает работу.

Фиг.3 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента реферирования web-страницы в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет коэффициент для каждого предложения web-страницы и выбирает предложения с наивысшими коэффициентами, чтобы сформировать реферат web-страницы. На этапе 301 этот компонент вызывает компонент вычисления коэффициентов для вычисления коэффициента для каждого предложения. На этапе 302 компонент сортирует предложения на основании вычисленных коэффициентов. На этапе 303

компонент выбирает предложения с наивысшими коэффициентами, чтобы сформировать реферат для web-страницы. Затем компонент выдает реферат.

Фиг.4 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента вычисления коэффициентов в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет коэффициенты разных методов реферирования для предложений web-страницы и вычисляет комбинированный коэффициент для каждого предложения на основании этих коэффициентов методов реферирования. Компонент может альтернативно вычислять коэффициент с помощью только одного метода реферирования или различных комбинаций методов реферирования. На этапе 401 компонент вызывает компонент вычисления коэффициента Люна, чтобы вычислить коэффициент Люна для каждого предложения web-страницы. На этапе 402 компонент вызывает компонент вычисления коэффициента латентно-семантического анализа, чтобы вычислить коэффициент латентно-семантического анализа для каждого предложения web-страницы. На этапе 403 компонент вызывает компонент вычисления коэффициента основной части содержания, чтобы вычислить коэффициент основной части содержания для каждого предложения web-страницы. На этапе 404 компонент вызывает компонент вычисления управляемого коэффициента, чтобы вычислить управляемый коэффициент для каждого предложения web-страницы. На этапе 405 компонент вызывает компонент комбинации коэффициентов, чтобы вычислить скомбинированный коэффициент для каждого предложения web-страницы. Затем компонент выдает скомбинированные коэффициенты.

Фиг.5 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента вычисления коэффициента Люна в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет коэффициент Люна для каждого предложения переданной web-страницы. На этапе 501 компонент выбирает следующее предложение web-страницы. На этапе 502 ветвления, если все предложения web-страницы уже выбраны, компонент выдает коэффициенты Люна, иначе компонент продолжает работу на этапе 503. На этапах 503-509 компонент работает в цикле, вырабатывая коэффициент класса для выбранного предложения для каждой классификации. На этапе 503 компонент выбирает следующую классификацию. На этапе 504 ветвления, если все классификации уже выбраны, компонент переходит к этапу 510, иначе компонент переходит к этапу 505. На этапе 505 компонент выявляет слова выбранного предложения, которые заключены значащими словами выбранной классификации. На этапе 506 ветвления, если значащие слова выявлены, компонент переходит к этапу 507, иначе компонент возвращается к этапу 503 для выбора следующей классификации. На этапе 507 компонент подсчитывает значащие слова внутри заключенной между значащими словами части выбранного предложения. На этапе 508 компонент подсчитывает слова внутри заключенной между значащими словами части выбранного предложения. На этапе 509 компонент вычисляет коэффициент для классификации как квадрат числа значащих слов, деленный на число слов. Затем компонент возвращается к этапу 503 для выбора следующей классификации. На этапе 510 компонент вычисляет коэффициент Люна для выбранного предложения как сумму коэффициентов классификации, поделенную на число классификаций, для которых была выявлена заключенная между значащими словами часть выбранного предложения (т.е. среднее коэффициентов классификации, которые вычислялись). Затем компонент возвращается к этапу 501 для выбора следующего предложения.

Фиг.6 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента вычисления коэффициента латентно-семантического анализа в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет коэффициент латентно-семантического анализа для каждого предложения переданной web-страницы. На этапах 601-603 компонент работает в цикле, конструируя вектор термин-на-вес для каждого предложения web-страницы. На этапе 601 компонент выбирает следующее предложение web-страницы. На этапе 602 ветвления, если все предложения web-страницы уже выбраны, компонент переходит к этапу 604, иначе компонент переходит к этапу 603. На этапе 603 компонент конструирует вектор термин-на-вес для выбранного предложения, а затем возвращается к этапу 601 для выбора следующего предложения. Векторы термин-на-вес для предложений образуют матрицу, в отношении которой выполняют декомпозицию, чтобы выдать матрицу правых сингулярных векторов. На этапе 604 компонент выполняет декомпозицию этой матрицы по сингулярным значениям, чтобы выработать правые сингулярные векторы. На этапах 605-607 компонент работает в цикле, устанавливая коэффициент для каждого предложения на основании правых сингулярных векторов. На этапе 605 компонент выбирает следующий правый сингулярный вектор. На этапе 606 ветвления, если все правые сингулярные векторы уже выбраны, компонент выдает коэффициенты в качестве коэффициентов латентно-семантического анализа, иначе компонент переходит к этапу 607. На этапе 607 компонент устанавливает коэффициент предложения с наивысшим значением индекса выбранного правого сингулярного вектора, а затем возвращается к этапу 605 для выбора следующего правого сингулярного вектора.

Фиг.7 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента вычисления коэффициента основной части содержания в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет коэффициент основной части содержания для каждого предложения переданной web-страницы. На этапе 701 компонент выявляет базовые объекты web-страницы. На этапе 702 компонент выявляет составные объекты web-страницы. На этапах 703-705 компонент работает в цикле, вырабатывая вектор частота термина/инвертированная частота документа для каждого объекта. На этапе 703 компонент выбирает следующий объект. На этапе 704 ветвления, если все объекты уже выбраны, компонент переходит к этапу 706, иначе компонент переходит к этапу 705. На этапе 705 компонент вырабатывает вектор частота термина/инвертированная частота документа для выбранного объекта, а затем возвращается к этапу 703 для выбора следующего объекта. На этапах 706-710 компонент работает в цикле, вычисляя подобие между парами объектов. На этапе 706 компонент выбирает следующую пару объектов. На этапе 707 ветвления, если все пары объектов уже выбраны, компонент переходит к этапу 711, иначе компонент переходит к этапу 708. На этапе 708 компонент вычисляет подобие между выбранной парой объектов. На этапе 709 ветвления, если подобие выше, чем пороговый уровень подобия, компонент переходит к этапу 710, иначе компонент возвращается к этапу 706 для выбора следующей пары объектов. На этапе 710 компонент добавляет связь между выбранной парой объектов, а затем возвращается к этапу 706 для выбора следующей пары объектов. На этапах 711-715 компонент выявляет основную часть содержания web-страницы путем выявления сердцевинного объекта и всех объектов со связями к этому сердцевинному объекту. На этапе 711 компонент выявляет сердцевинный объект как объект с наибольшим числом связей к нему. На этапе 712

компонент выбирает следующее предложение web-страницы. На этапе 713 ветвления, если все предложения уже выбраны, компонент выдает коэффициенты основной части содержания, иначе компонент переходит к этапу 714. На этапе 714 ветвления, если предложение находится внутри объекта, который связан с сердцевинным объектом, это предложение находится в основной части содержания, и компонент переходит к этапу 715, иначе компонент устанавливает коэффициент выбранного предложения на нуль и возвращается к этапу 712 для выбора следующего предложения. На этапе 715 компонент устанавливает коэффициент выбранного предложения на высокий коэффициент, а затем возвращается к этапу 712 для выбора следующего предложения.

Фиг.8 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента вычисления управляемого коэффициента в одном варианте осуществления. Этому компоненту web-страница передается в качестве аргумента, и он вычисляет управляемый коэффициент для каждого предложения переданной web-страницы. На этапе 801 компонент выбирает следующее предложение web-страницы. На этапе 802 ветвления, если все предложения web-страницы уже выбраны, компонент выдает управляемые коэффициенты, иначе компонент переходит к этапу 803. На этапе 803 компонент вырабатывает вектор признаков для выбранного предложения. На этапе 804 компонент вычисляет коэффициент выбранного предложения с помощью выработанного вектора признаков и обученной функции реферирования. Затем компонент возвращается к этапу 801 для выбора следующего предложения.

Фиг.9 представляет собой блок-схему алгоритма, которая иллюстрирует работу компонента комбинирования коэффициентов в одном варианте осуществления. Этот компонент вырабатывает комбинированный коэффициент для каждого предложения web-страницы на основании коэффициента Люна, коэффициента латентно-семантического анализа, коэффициента основной части содержания и управляемого коэффициента. На этапе 901 компонент выбирает следующее предложение web-страницы. На этапе 902 ветвления, если все предложения web-страницы уже выбраны, компонент выдает комбинированные коэффициенты, иначе компонент переходит к этапу 903. На этапе 903 компонент комбинирует коэффициенты для выбранного предложения, а затем возвращается к этапу 901 для выбора следующего предложения.

Специалист поймет, что, хотя здесь для целей иллюстрации описаны конкретные варианты осуществления системы реферирования, могут быть сделаны различные модификации без отхода от сущности и объема изобретения. Специалист поймет, что классификация относится к процессу выявления класса или категории, связанных с дисплейной страницей. Классы можно определить заранее. Атрибуты дисплейной страницы, подлежащей классификации, могут сравниваться с атрибутами, выделенными из других дисплейных страниц, которые уже классифицированы (например, обучающий набор). На основании этого сравнения дисплейная страница классифицируется в класс, атрибуты дисплейных страниц которого подобны атрибутам классифицируемой дисплейной страницы. В противоположность этому кластеризация относится к процессу выявления из набора дисплейных страниц групп дисплейных страниц, которые подобны друг другу. Соответственно, изобретение не ограничивается ничем, кроме приложенной формулы изобретения.

Формула изобретения

1. Реализуемый в компьютерной системе способ классификации web-страниц,

содержащий этапы, на которых извлекают web-страницу;

осуществляют автоматическую выработку реферата извлеченной web-страницы посредством

идентификации объектов web-страницы, причем эти объекты имеют предложения,

получения произведения частоты термина на индекс инвертированной частоты документа для каждого объекта,

вычисления подобия между парами объектов на основе произведения частоты термина на индексы инвертированной частоты документа этих объектов,

связывания, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу,

выбора объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта web-страницы,

назначения высоких коэффициентов предложениям сердцевинного объекта и объектов со связями к сердцевинному объекту и низких коэффициентов всем остальным предложениям,

выбора предложений для формирования реферата web-страницы на основе назначенных коэффициентов; и определяют классификацию для извлеченной web-страницы на основании автоматически выработанного реферата.

2. Способ по п.1, в котором при автоматической выработке реферата вычисляют коэффициент для каждого предложения web-страницы с помощью множества методов реферирования.

3. Способ по п.2, в котором коэффициент для каждого предложения является линейной комбинацией коэффициентов множества методов реферирования.

4. Способ по п.1, в котором предложения с наивысшими коэффициентами выбираются для формирования реферата.

5. Способ по п.2, в котором методы реферирования включают в себя метод реферирования Люна (Luhn), метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования.

6. Способ по п.2, в котором методы реферирования включают в себя любые два или более из набора методов реферирования, состоящего из метода реферирования Люна (Luhn), метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования.

7. Способ по п.1, в котором при определении классификации используют упрощенный байесовский классификатор.

8. Способ по п.1, в котором при определении классификации используют метод опорных векторов.

9. Способ по п.1, в котором при автоматической выработке реферата используют метод реферирования Люна (Luhn).

10. Способ по п.1, в котором при автоматической выработке реферата используют метод реферирования на основе латентно-семантического анализа.

11. Способ по п.1, в котором при автоматической выработке реферата используют метод реферирования основной части содержания.

12. Способ по п.1, в котором при автоматической выработке реферата используют метод управляемого реферирования.

13. Реализуемый в компьютерной системе способ реферирования web-страницы,

содержащий этапы, на которых извлекают web-страницу;

для каждого предложения извлеченной web-страницы,

назначают коэффициент предложению на основе множества методов реферирования, причем при назначении коэффициента согласно одному из этих методов реферирования осуществляют

идентификацию объектов web-страницы, причем эти объекты имеют предложения,

получение произведения частоты термина на индекс инвертированной частоты документа для каждого объекта,

вычисление подобия между парами объектов на основе произведения частоты термина на индексы инвертированной частоты документа этих объектов,

связывание, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу,

выбор объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта web-страницы,

назначение высокого коэффициента предложениям сердцевинного объекта и объектов со связями к сердцевинному объекту и низкого коэффициента всем остальным предложениям, и

комбинируют коэффициенты, назначенные предложению, для выработки скомбинированного коэффициента для этого предложения; и выбирают предложения с наивысшими скомбинированными коэффициентами для формирования реферата извлеченной web-страницы.

14. Способ по п.13, в котором скомбинированный коэффициент для каждого предложения является линейной комбинацией назначенных коэффициентов.

15. Способ по п.14, в котором назначенные коэффициенты множества методов реферирования взвешиваются по-разному при комбинировании.

16. Способ по п.13, в котором методы реферирования включают в себя метод реферирования Люна (Luhn), метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования.

17. Способ по п.13, в котором методы реферирования включают в себя любые два или более из набора методов реферирования, состоящего из метода реферирования Люна (Luhn), метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования.

18. Способ по п.13, в котором метод реферирования является методом реферирования Люна (Luhn), где классификация имеет совокупность значащих слов.

19. Способ по п.18, в котором шумовые слова отбрасываются из совокупности.

20. Способ по п.13, в котором метод реферирования является методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признак, базирующийся на подобии между предложением и метаданными web-страницы.

21. Способ по п.13, в котором метод реферирования является методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признаки, основанные на словах предложения, которые выделяются на web-странице.

22. Способ по п.13, в котором метод реферирования является методом управляемого реферирования, при этом предложение представляется набором

признаков, который включает в себя признак, базирующийся на размере шрифта слов в этом предложении.

23. Способ по п.13, включающий в себя выявление классификации для извлеченной web-страницы на основании реферата извлеченной web-страницы.

24. Способ по п.23, в котором при выявлении классификации используется упрощенный байесовский классификатор.

25. Способ по п.23, в котором при выявлении классификации используется метод опорных векторов.

26. Машиночитаемый носитель, содержащий команды, предписывающие компьютерной системе вырабатывать реферат для дисплейной страницы способом, содержащим для каждого предложения дисплейной страницы выработку коэффициента, который базируется на множестве методов реферирования, причем один из этих методов реферирования подразумевает

вычисление подобия между парами объектов дисплейной страницы, причем эти объекты имеют предложения,

связывание, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу,

выбор объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта дисплейной страницы,

назначение высокого коэффициента предложениям сердцевинного объекта и объектов со связями к сердцевинному объекту и низких коэффициентов всем остальным предложениям и; выбор предложений с наивысшими выработанными коэффициентами, чтобы сформировать реферат дисплейной страницы.

27. Машиночитаемый носитель по п.26, в котором вырабатываемый коэффициент для каждого предложения является комбинацией коэффициента для каждого из множества методов реферирования.

28. Машиночитаемый носитель по п.27, в котором коэффициенты множества методов реферирования взвешиваются по-разному.

29. Машиночитаемый носитель по п.26, в котором методы реферирования включают в себя метод реферирования Люна (Luhn), метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования.

30. Машиночитаемый носитель по п.26, в котором методы реферирования включают в себя любые два или более из набора методов реферирования, состоящего из метода реферирования Люна (Luhn), метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования.

31. Машиночитаемый носитель по п.26, в котором метод реферирования является методом реферирования Люна (Luhn), где классификация имеет совокупность значащих слов.

32. Машиночитаемый носитель по п.31, в котором шумовые слова отбрасываются из совокупности.

33. Машиночитаемый носитель по п.26, в котором метод реферирования является методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признак, базирующийся на подобии между предложением и метаданными дисплейной страницы.

34. Машиночитаемый носитель по п.26, в котором метод реферирования является

методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признаки, основанные на словах предложения, которые выделены на дисплейной странице.

35. Машиночитаемый носитель по п.26, в котором метод реферирования является методом управляемого реферирования, в котором предложение представляется набором признаков, который включает в себя признак, базирующийся на размере шрифта слов в этом предложении.

36. Машиночитаемый носитель по п.26, включающий в себя выявление классификации для дисплейной страницы на основании реферата дисплейной страницы.

37. Компьютерная система для классификации дисплейных страниц, содержащая средство для автоматической выработки реферата дисплейной страницы посредством вычисления подобия между парами объектов дисплейной страницы, причем эти объекты имеют предложения, связывания, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу, выбора объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта дисплейной страницы, выбора предложений сердцевинного объекта и объектов со связями к сердцевинному объекту для формирования реферата дисплейной страницы и;

средство для выявления классификации для дисплейной страницы на основании автоматически выработанного реферата.

38. Компьютерная система по п.37, в которой средство для автоматической выработки реферата вычисляет коэффициент для каждого предложения дисплейной страницы с помощью множества методов реферирования.

39. Компьютерная система по п.38, в которой коэффициент для каждого предложения является линейной комбинацией коэффициентов множества методов реферирования.

40. Компьютерная система по п.37, в которой методы реферирования включают в себя метод реферирования Люна (Luhn), метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования.

41. Компьютерная система по п.37, в которой методы реферирования включают в себя любые два или более из набора методов реферирования, состоящего из метода реферирования Люна (Luhn), метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования.

42. Компьютерная система по п.41, в которой каждому предложению дисплейной страницы назначается коэффициент, который является комбинацией коэффициентов множества методов реферирования.

43. Реализуемый в компьютерной системе способ идентификации сердцевинного объекта web-страницы, содержащий этапы, на которых

идентифицируют объекты web-страницы, причем эти объекты имеют предложения,

получают произведение частоты термина на индекс инвертированной частоты документа для каждого объекта,

вычисляют подобие между парами объектов на основе произведения частоты термина на индексы инвертированной частоты документа этих объектов, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, связывают объекты этой пары, чтобы показать, что объекты удовлетворяют данному

порогу, и выбирают объект, который имеет наибольшее количество связей, в качестве сердцевинного объекта web-страницы.

44. Способ по п.43, дополнительно содержащий этапы, на которых назначают высокие коэффициенты предложениям сердцевинного объекта и объектов со связями к сердцевинному объекту и низкие коэффициенты всем остальным предложениям,

выбирают предложения для формирования реферата web-страницы на основе назначенных коэффициентов.

45. Способ по п.44, в котором для формирования реферата выбирают предложения с наивысшими коэффициентами.

46. Способ по п.44, в котором при формировании реферата вычисляют коэффициент для каждого предложения web-страницы с помощью множества методов реферирования.

47. Способ по п.46, в котором коэффициент для каждого предложения является линейной комбинацией коэффициентов множества методов реферирования.

48. Способ по п.46, в котором методы реферирования включают в себя метод реферирования Люна (Luhn), метод реферирования на основе латентно-семантического анализа, метод реферирования основной части содержания и метод управляемого реферирования.

49. Способ по п.46, в котором методы реферирования включают в себя любые два или более из набора методов реферирования, состоящего из метода реферирования Люна (Luhn), метода реферирования на основе латентно-семантического анализа, метода реферирования основной части содержания и метода управляемого реферирования.

50. Способ по п.44, в котором при формировании реферата используют метод реферирования Люна (Luhn).

51. Способ по п.44, в котором при формировании реферата используют метод реферирования на основе латентно-семантического анализа.

52. Способ по п.44, в котором при формировании реферата используют метод реферирования основной части содержания.

53. Способ по п.44, в котором при формировании реферата используют метод управляемого реферирования.

54. Способ по п.44, дополнительно содержащий этап, на котором определяют классификацию для извлеченной web-страницы на основании сформированного реферата.

55. Способ по п.54, в котором при определении классификации используют упрощенный байесовский классификатор.

56. Способ по п.54, в котором при определении классификации используют метод опорных векторов.

57. Машиночитаемый носитель, содержащий команды, предписывающие компьютерной системе идентифицировать сердцевинный объект для дисплейной страницы способом, содержащим

вычисление подобия между парами объектов дисплейной страницы, причем эти объекты имеют предложения,

связывание, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу, и выбор объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта дисплейной страницы.

58. Машиночитаемый носитель по п.57, дополнительно содержащий

назначение высокого коэффициента предложениям сердцевинного объекта и объектов со связями к сердцевинному объекту и низких коэффициентов всем остальным предложениям и выбор предложений с наивысшими выработанными коэффициентами, чтобы сформировать реферат дисплейной страницы.

59. Машиночитаемый носитель по п.58, в котором вырабатываемый коэффициент для каждого предложения является комбинацией коэффициента для каждого из множества методов реферирования.

60. Машиночитаемый носитель по п.59, в котором коэффициенты множества методов реферирования взвешиваются по-разному.

61. Машиночитаемый носитель по п.58, в котором метод реферирования является методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признак, базирующийся на подобии между предложением и метаданными дисплейной страницы.

62. Машиночитаемый носитель по п.58, в котором метод реферирования является методом управляемого реферирования, где предложение представляется набором признаков, который включает в себя признаки, основанные на словах предложения, которые выделены на дисплейной странице.

63. Машиночитаемый носитель по п.58, в котором метод реферирования является методом управляемого реферирования, в котором предложение представляется набором признаков, который включает в себя признак, базирующийся на размере шрифта слов в этом предложении.

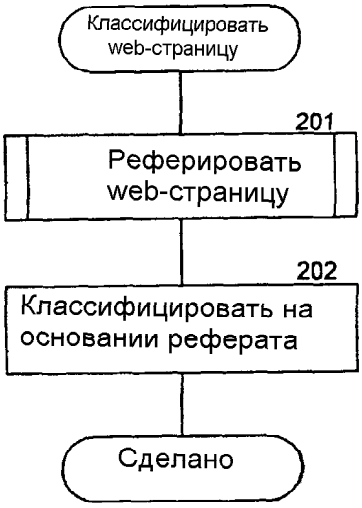
64. Машиночитаемый носитель по п.58, включающий в себя выявление классификации для дисплейной страницы на основании реферата дисплейной страницы.

65. Компьютерная система для идентификации сердцевинных объектов дисплейных страниц, содержащая средства для вычисления подобия между парами объектов дисплейной страницы, причем эти объекты имеют предложения,

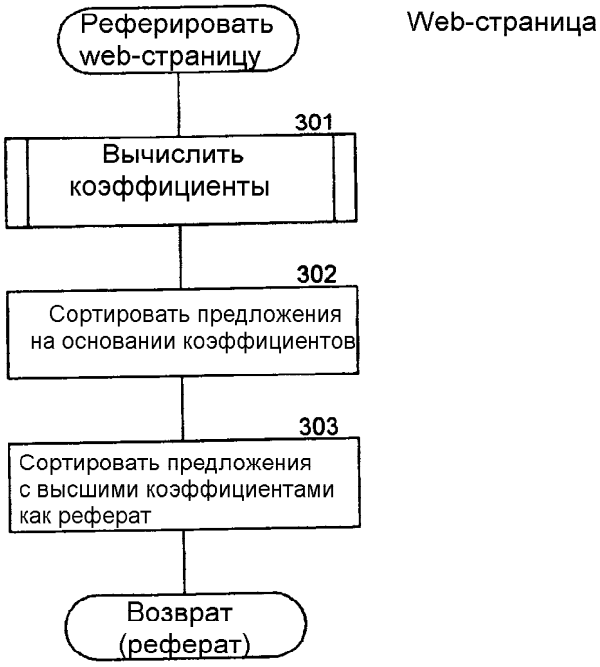
связывания, если вычисленное подобие между парой объектов удовлетворяет порогу подобия, объектов этой пары, чтобы показать, что объекты удовлетворяют данному порогу, и выбора объекта, который имеет наибольшее количество связей, в качестве сердцевинного объекта дисплейной страницы.

66. Система по п.65, дополнительно содержащая

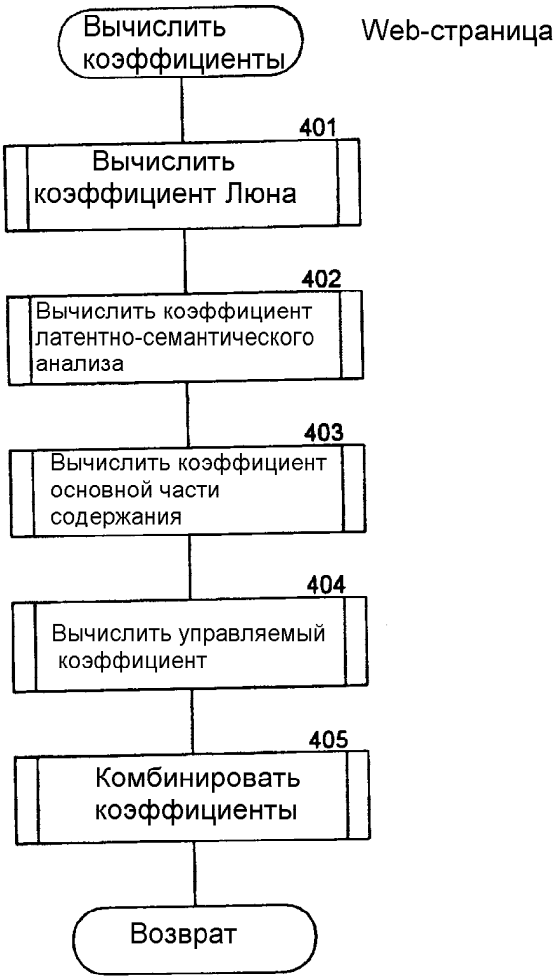
средство для выбора предложений сердцевинного объекта и объектов со связями к сердцевинному объекту для формирования реферата дисплейной страницы и средство для выявления классификации для дисплейной страницы на основании сформированного реферата.


Классифицировать web-страницу
201
Реферировать web-страницу
202
Классифицировать на основании реферата
Сделано

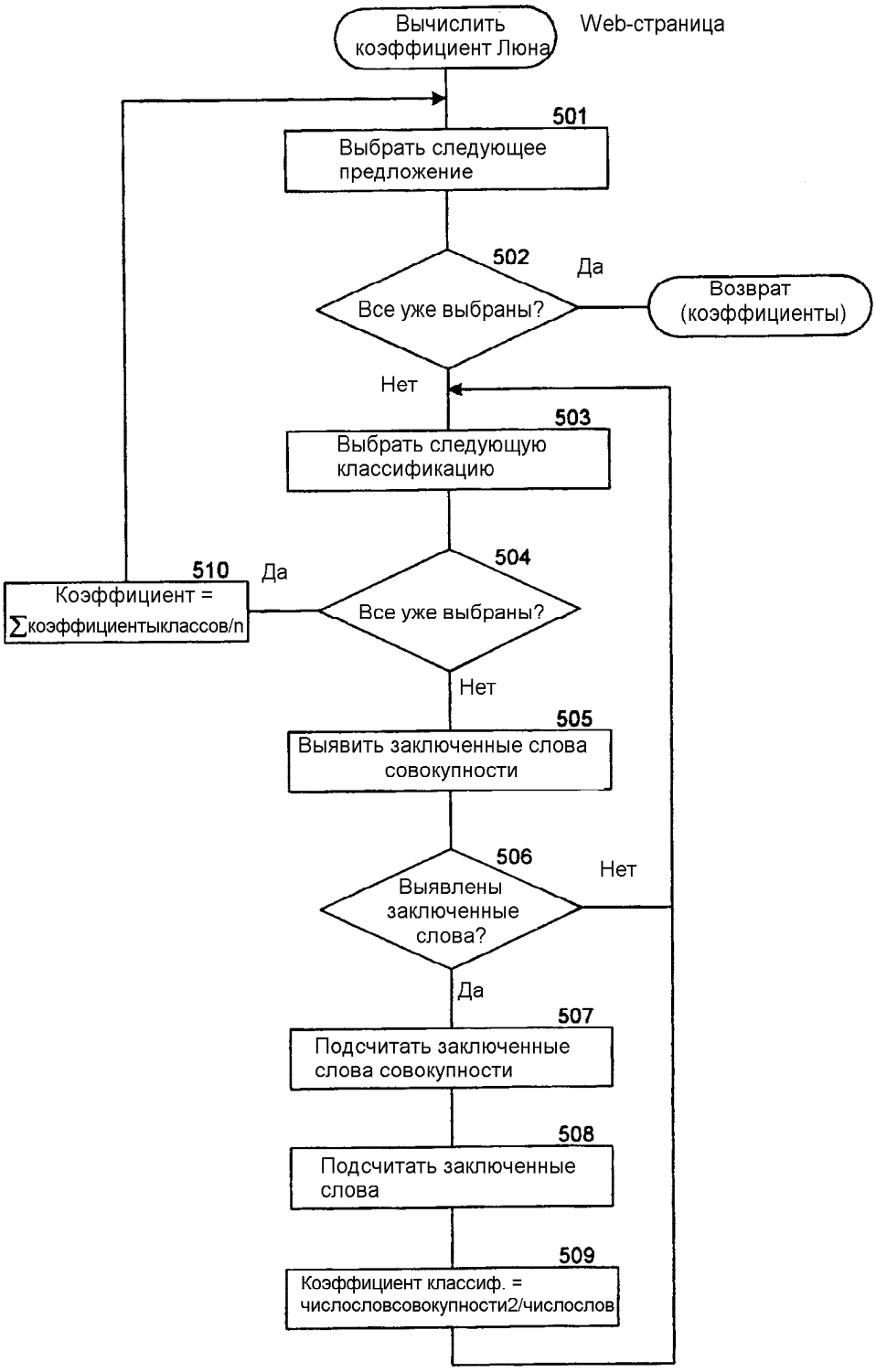
ФИГ.2


Реферировать web-страницу
Web-страница
301
Вычислить коэффициенты
302
Сортировать предложения на основании коэффициентов
303
Сортировать предложения с высшими коэффициентами как реферат
Возврат (реферат)

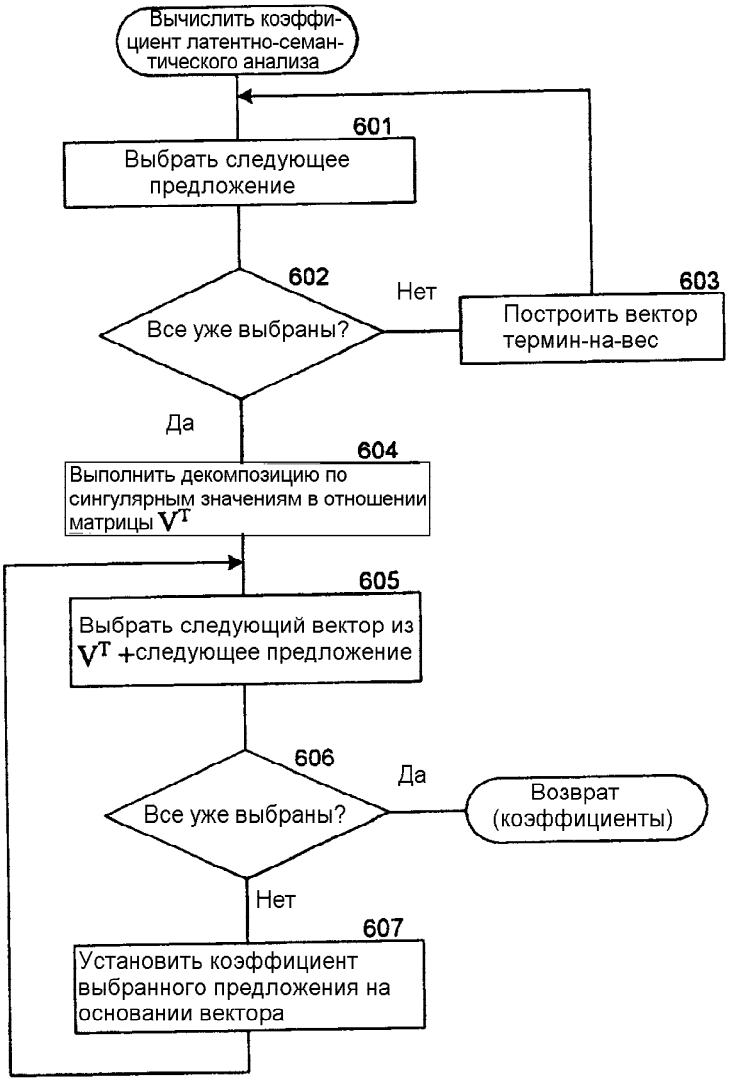
ФИГ.3


Вычислить коэффициенты
Web-страница
401
Вычислить коэффициент Люна
402
Вычислить коэффициент латентно-семантического анализа
403
Вычислить коэффициент основной части содержания
404
Вычислить управляемый коэффициент
405
Комбинировать коэффициенты
Возврат

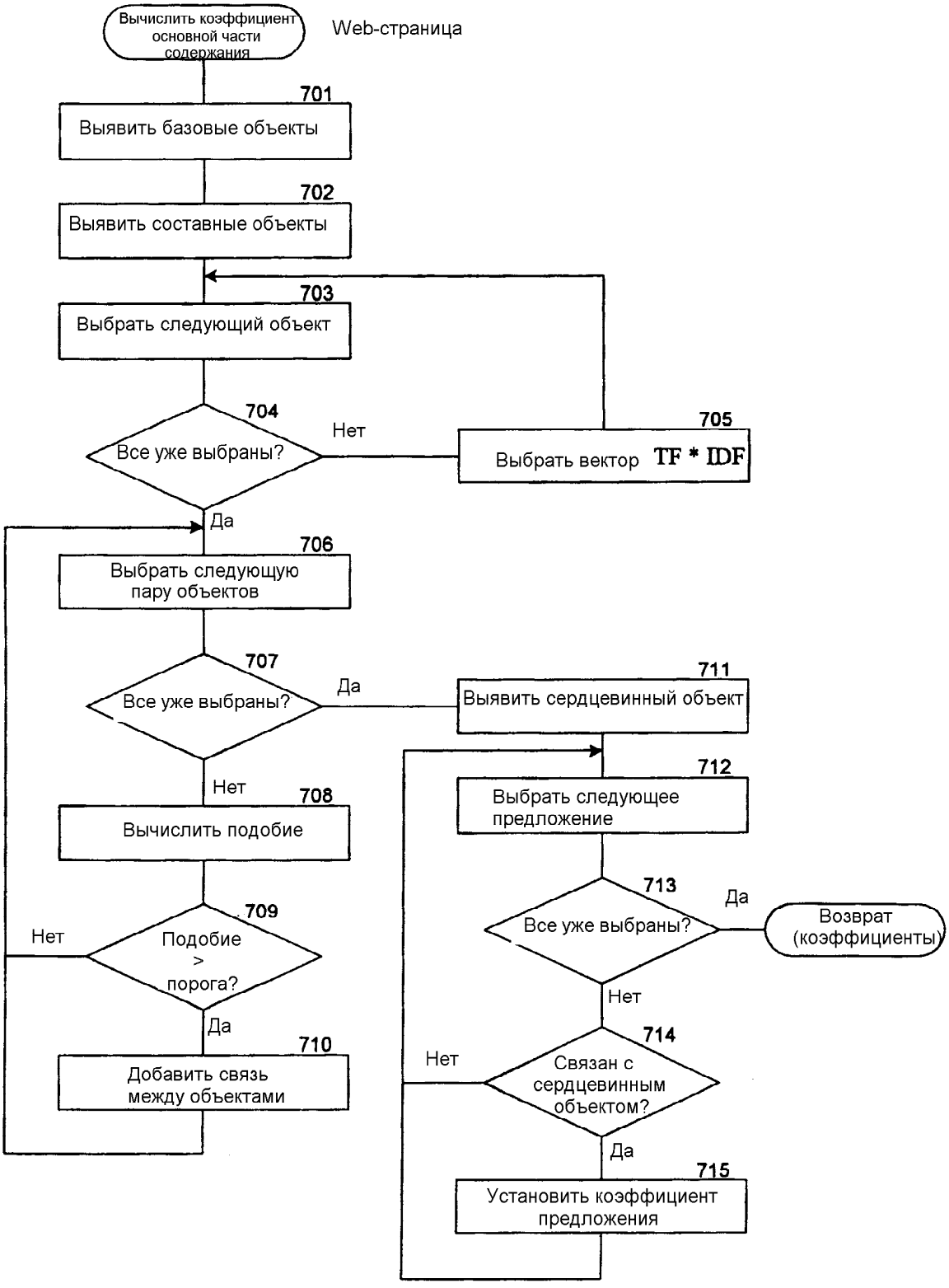
ФИГ.4


Вычислить коэффициент Люна
Web-страница
501
Выбрать следующее предложение
502
Все уже выбраны?
Да
Возврат (коэффициенты)
Нет
503
Выбрать следующую классификацию
504
Все уже выбраны?
Да
510
Коэффициент = ∑коэффициентыклассов/n
Нет
505
Выявить заключенные слова совокупности
506
Выявлены заключенные слова?
Нет
Да
507
Подсчитать заключенные слова совокупности
508
Подсчитать заключенные слова
509
Коэффициент классиф. = числословсовокупности2/числослов

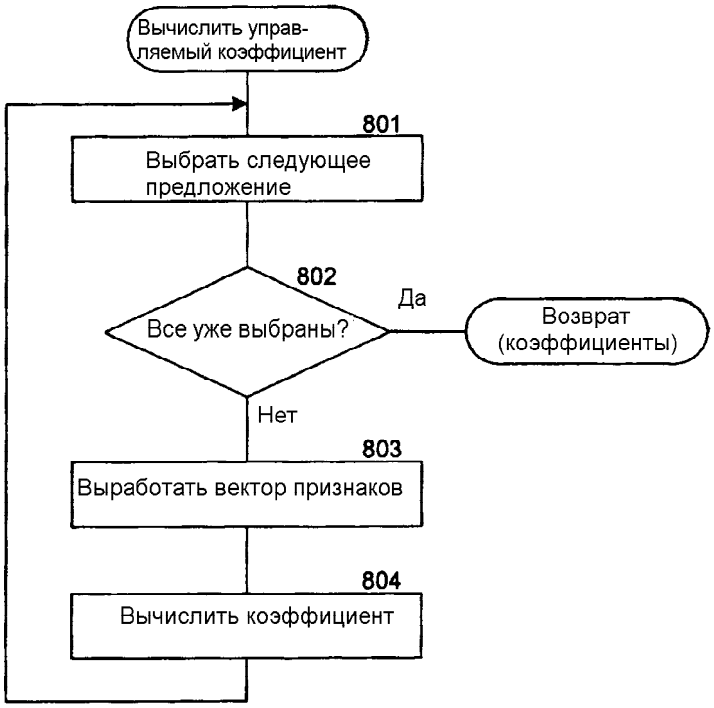
ФИГ.5


Вычислить коэффициент латентно-семантического анализа
601
Выбрать следующее предложение
602
Все уже выбраны?
Нет
603
Построить вектор термин-на-вес
Да
604
Выполнить декомпозицию по сингулярным значениям в отношении матрицы V^T
605
Выбрать следующий вектор из V^T +следующее предложение
606
Все уже выбраны?
Да
Возврат (коэффициенты)
Нет
607
Установить коэффициент выбранного предложения на основании вектора

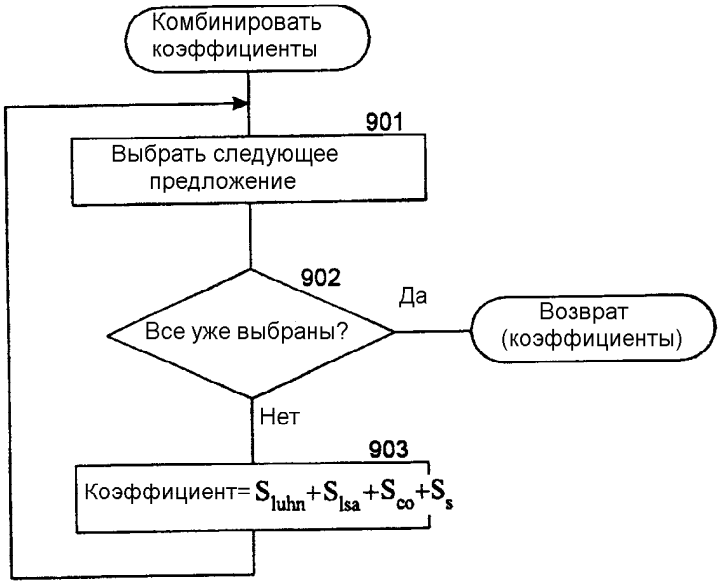
ФИГ.6


Вычислить коэффициент основной части содержания
Web-страница
701
Выявить базовые объекты
702
Выявить составные объекты
703
Выбрать следующий объект
704
Все уже выбраны?
Нет
705
Выбрать вектор TF * IDF
Да
706
Выбрать следующую пару объектов
707
Все уже выбраны?
Да
711
Выявить сердцевинный объект
Нет
708
Вычислить подобие
712
Выбрать следующее предложение
709
Подобие > порога?
Нет
713
Все уже выбраны?
Да
Возврат (коэффициенты)
Да
Нет
710
Добавить связь между объектами
714
Связан с сердцевинным объектом?
Нет
Да
715
Установить коэффициент предложения

ФИГ.7


Вычислить управляемый коэффициент
801
Выбрать следующее предложение
802
Все уже выбраны?
Да
Возврат (коэффициенты)
Нет
803
Выработать вектор признаков
804
Вычислить коэффициент

ФИГ.8


Комбинировать коэффициенты
901
Выбрать следующее предложение
902
Все уже выбраны?
Да
Возврат (коэффициенты)
Нет
903
Коэффициент= $S_{luhn}+S_{lsa}+S_{co}+S_s$

ФИГ.9