(19) Евразийское патентное ведомство

(11) 043657 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.06.08

(21) Номер заявки
202200059

(22) Дата подачи заявки
2021.09.21

(51) Int. Cl. ***G06F 17/00*** (2019.01)
G06F 16/21 (2019.01)
G06F 16/9538 (2019.01)

(54) ДОМЕННО-ОРИЕНТИРОВАННАЯ ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА (ДИПС) (СПОСОБ ЕЕ СОЗДАНИЯ)

(43) **2023.03.31**
(96) **2021/025 (AZ) 2021.09.21**

(71)(73) Заявитель и патентовладелец:
ИНСТИТУТ СИСТЕМ УПРАВЛЕНИЯ НАНА (AZ)

(72) Изобретатель:
Аббасов Али Мамед оглы, Касумов Вагиф Алиджавад оглы (AZ)

(56) US-B2-10817568
US-A1-20080091633
US-A1-20020129015
US-B1-7299222

(57) Изобретение относится к информационно-поисковым системам. Сущность изобретения состоит в создании системы для оптимизации информационного поиска в интернете, позволяющей относительно быстро обнаружить наиболее релевантные результаты. Технический эффект заявляемого изобретения позволяет повысить уровень эффективности информационного поиска путем систематизации Web-пространства, сужения области поиска и организации тематически направленного поиска, оставляя при этом без изменения программно-технические параметры поисковых систем.

043657 B1

043657 B1

Изобретение относится к информационно-поисковым системам.

Известно, что пользователь, выполняя веб-поиск с использованием поисковой системы, преследует две основных цели. Ему нужно, чтобы поисковая система выдавала наиболее релевантные результаты и чтобы это было выполнено относительно быстро. Однако анализ проблем информационного пространства Интернет, а также исследований в области поиска информации показывает, что в современных условиях астрономического роста объема информации в среде Интернет эффективность средств поиска информации явно отстает от желаемого уровня. Главной причиной данной проблемы в основном является неорганизованность информационных ресурсов, в том числе Web-документов в среде Интернет. Данное обстоятельство связано со следующими факторами:

информационные ресурсы, в том числе Web-документы Интернет, создаются независимо друг от друга в произвольной форме по желанию авторов или владельцев;

каждый информационный ресурс характеризуется только определенным набором признаков в лучшем случае и их частотными характеристиками;

ссылки между информационными ресурсами ставятся без учета их содержания в целом, в результате чего логическая связь между информационными ресурсами существует только между соседними, в лучшем случае через две-три ссылки;

существующие поисковые системы в основном дают возможность квазислучайного (малонаправленного) поиска по паутине Web, эффективность которого зависит от библиотечной подготовленности пользователей.

В настоящее время для решения указанной проблемы предлагаются различные способы и системы отбора информации в сети Интернет.

Известна система и способ [1] для получения информации от распределенного с использованием поискового агента, работающей в сети, по "всемирной паутине". Система предназначена для обеспечения клиентов сервера быстрыми и точными данными поиска веб-агентом и основана на анализе данных полученной информации от работающих в сети систем с использованием распределенного поискового веб-агента. Собранную поисковым веб-агентом сервера информацию сравнивают с данными полученной информации, распространенной через страницу результатов поискового механизма, либо посредством клиентских словарей, исходящих от сервера, которые обобщают данные поискового веб-агента. Предполагается, что техническим результатом является обеспечение клиентов сервера быстрыми и точными данными поиска веб-агентом.

Известен способ [2] отбора информации в сети интернет и использования этой информации в разделяемом веб-сайте, имеющем некоторое доменное имя, согласно которому с помощью интернет-робота осуществляют поиск информации в сети Интернет о субъектах. Используемый критерий поиска в виде списка ключевых слов, соответствующих доменному имени разделяемого веб-сайта, позволяет отобрать информацию, включающую средства индивидуализации субъектов, содержащие в себе соответствующую указанному доменному имени символьную последовательность. Указанный способ не всегда может обеспечить корректный отбор информации, из-за того что соответствие информации документу оценивается путем интернет-робота, что не исключает ошибок в определении тематики документа.

Задача изобретения состоит в создании поисковой системы для оптимизации информационного поиска в интернете, позволяющей относительно быстро обнаружить наиболее релевантные результаты.

Сущность изобретения состоит в способе создания доменно-ориентированной информационно-поисковой системы (ДИПС). Способ, предложенный в рамках информационной модели доменов (IM_D), включает в себя разделение виртуального информационного пространства на семантически слабосвязанные домены, являющиеся зонами охвата отдельных поисковых систем, создание многоуровневой иерархической структуры, на каждом уровне которой формируются доменные области со своими атрибутами, имеющие предметно-содержательную связь, формирование пользовательского запроса (UQ), организацию поиска информации с использованием модулей "MAPP2" (Mapping 2), "MAPP1" (Mapping 1), оценки релевантности и ранжирование результатов.

Информационная модель доменов (IM_D) представляет собой информационное пространство, организованное как иерархическая структура, на уровнях которой определяются Web-области, содержащие объекты со своими атрибутами и состоящие из доменных областей, включающие в себя множества объектов доменов, в том числе серверов Web-сайтов и других информационных сервисов. Каждый объект описывается множеством атрибутов, дескрипторов, ключевых слов и терминов. Определяются принципы размещения объектов на уровнях, отношения между объектами, между объектами и их атрибутами, а также между атрибутами объектов.

Предложенная для описания ресурсов информационного пространства Интернета информационная модель доменов (IM_D) позволяет также построить модель виртуального Web-окружения (Virtual Web Environment - VWE) и модель пользовательского запроса (User Query - UQ), а также разработать методы отображения одной модели в другую, определения релевантности отображения и ранжирования результатов.

Информационную модель доменов (IM_D) в общем виде можно представить в следующем виде:

$$IM_D \rightarrow \{N, E, H(E), H(E, E^*), A, R(E, A), R(A, A^*), R(E, E^*)\},$$

где N - максимальное количество уровней в иерархии доменов Web-пространства;

$E=\{e_i\}_n$ - множество объектов (сущностей) в домене;

$A=\{a_j\}_m$ - множество атрибутов, используемых для описания объектов домена;

H(E) - матрица (или вектор) размещения объектов (категорий) на уровнях иерархии доменов, где $H(e_i)=1$, если объект e_i размещен на домене, и $H(e_i)=0$ в обратном случае, $i = \overline{1,n}$;

H(E,E*) - матрица отношений категоризации (вложенности на subtype и supertype) между объектами, где $H(e_i,e_j^*)=1$, если объект e_i имеет отношение вложенности на subtype или supertype с объектом e_j^*, и $H(e_i,e_j^*)=0$ в обратном случае, $i,j = \overline{1,n}$;

R(E,A) - матрица отношений между объектами и атрибутами доменов, где $R(e_i,a_i)=1$, если объект e_i имеет отношение релевантности с атрибутом a_j, и $R(e_i,a_j)=0$ в обратном случае, $i = \overline{1,n}, j = \overline{1,m}$;

R(E,E*) - матрица отношений между объектами доменов где $R(e_i,e_j^*)=1$, если объект e_i имеет отношение с объектом e_j^*, и $R(e_i,a_j)=0$ в обратном случае, $i,j = \overline{1,n}$.

Исходя из вышеприведенной модели IM_D информационную модель виртуального Web-окружения (VWE) можно представить как

$$VWE \rightarrow \{URL, K, R(URL, K), R(K, K), R(URL, URL)\},$$

где URL - множество связанных (релевантных) Web-документов Интернета;

K - множество дескрипторов (ключевых слов), используемых для описания Web-документов;

R(URL,K) - матрица отношений между Web-документами и ключевыми словами;

R(K,K) - матрица взаимосвязей между ключевыми словами внутри Web-документов;

R(URL,URL) - матрица связей между Web-документами.

VWE является частью Web-окружения Интернета и относится к (релевантна) рассматриваемому конкретному домену. VWE содержит информацию о доменах Интернета в форме, подобной IM, и использует модель Web-документа.

Множество Web-документов URL также является объектами домена, поэтому можно писать $URL \subset E$. Тогда множество дескрипторов (ключевых слов) K также является подмножеством множества атрибутов, т.е. $K \subset A$. Учитывая это, матрицы отношений R(URL,K), R(K,K) и R(URL,URL) можно рассматривать как $R(K^E,K^A)$, $R(K^A,K^A)$ и $R(K^E,K^E*)$. Здесь K^E - подмножество дескрипторов (ключевых слов), описывающих объект E, K^A - множество дескрипторов (ключевых слов), описывающих Web-документы, т.е. $K^E \in E$ и $K^A \in A$.

Матрицы взаимосвязей модели VWE представляются следующим образом.

$R(K^E,K^A)$ - определяет Web-документы аналогично отношениям между объектами и их атрибутами, в виде

$$V_1 = \left\| v_{ij}^1 \right\|_{n \times m},$$

где

$$v_{ij}^1 = \begin{cases} URL, & \text{если } R(K^E, K^A) = 1 \\ 0, & \text{в обратном случае} \end{cases}$$

$R(K^E,K^E)$ - определяет Web-документы согласно отношениям между объектами в виде

$$V_2 = \left\| v_{ij}^2 \right\|_{n \times n},$$

где

$$v_{ij}^2 = \begin{cases} URL, & \text{если } R(K^E, K^E) = 1 \\ 0, & \text{в обратном случае} \end{cases}$$

$R(K^A,K^A)$ определяет взаимосвязь между дескрипторами (ключевыми словами) Web-документов в виде

$$V_3 = \left\| v_{ij}^3 \right\|_{m \times m},$$

где

$$v_{ij}^3 = \begin{cases} 1, \text{если ключевые слова } i \text{ и } j \text{ имеют взаимосвязь} \\ 0, \text{в обратном случае} \end{cases}.$$

Последняя матрица предоставляет возможность использовать взаимосвязи между ключевыми словами для повышения эффективности результата поиска путем учета семантической связанности ключевых слов, их синонимов и ассоциативных слов.

Формирование пользовательского запроса UQ дает возможность пользователю выражать свои требования на иерархии информационной модели и, используя существующий браузер, отправлять UQ через модуль отображения "MAPP2" (Mapping 2) в модель VWE. Формируя свой запрос в нижеследующем виде, пользователь ищет Web-документы из информационного пространства Интернета, которые как обычно содержат ключевые слова, связи из информационной модели доменов (IM_D) и виртуального Web-окружения (VWE):

$$UQ \rightarrow \{E, A, R(K^E, K^A), R(K^E, K^{E^*})\},$$

где

$$R(K^E, K^A), R\left(K^E, K^{E^*}\right) \in \{0,1\}.$$

В этом случае веса ключевых слов и связи из других Web-документов могут быть использованы для фильтрации результатов.

Модуль отображения "MAPP1" (Mapping 1) позволяет создать виртуальное Web-окружение (VWE) с помощью существующих Web-браузеров, представляющих возможность получения информации о доменах. Для создания виртуального Web-окружения (VWE) необходим переход от информационной модели виртуального Web-окружения к информационной модели заданного домена системы.

Учитывая, что множества ключевых слов взяты из множества дескрипторов $K \in E \cup A$, то преобразование модели VWE в модель IM_D можно рассматривать как

$$K \in K^E \cup K^A,$$

где $K^E \in K$ и $K^A \in A$.

Для заданных объектов не трудно выделять (классифицировать) K^E из E и K^A из A.

Таким образом, отношения между объектами и атрибутами $R(K^E,K^A)$ в множестве ключевых слов для VWE будет

$$R(K^E, K^A) = \begin{cases} 1, \text{если } R(K^E, K^A) = 1 \\ 1, \text{если } R(URL(K^E), URL(K^A)) = 1 \\ 0, \text{ иначе} \end{cases}$$

Выражение $R(K^E,K^A)=1$ означает, что внутри Web-документов существуют связи между K^E и K^A, а выражение $R(URL(K^E),URL(K^A))$ означает, что существуют связи между Web-документами, содержащие K^E и K^A. Аналогично выражению $R(K^E,K^A)$ определяются отношения между объектами $R(K^E,K^E)$ с учетом их ключевых слов для VWE:

$$R(K^E, K^A) = \begin{cases} 1, \text{если } R(K^E, K^E) = 1 \\ 1, \text{если } R(URL(K^E), URL(K^E)) = 1 \\ 0, \text{иначе} \end{cases}$$

Модуль MAPP2 осуществляет процесс поиска информации в VWE на основе UQ, определяя соответствие между их элементами, выполняет механизм отображения между UQ и VWE, реализует оптимизацию просмотра (поиска) результата, определяет, какое направление и какой маршрут просмотра будут наилучшими для планирования UQ. Элементами пользовательского запроса и виртуального Web-окружения являются:

множество объектов пользовательского запроса E отображается на множество ключевых слов (дескрипторов) K^A, описывающих эти объекты VWE, т.е. $E \rightarrow K^E$;

множество атрибутов A объектов домена отображается на множество ключевых слов (дескрипторов), описывающих эти атрибуты, т.е. $A \rightarrow K^A$;

так как множества E и A отображаются соответственно на множества K^E и K^A, то отношения между объектами и атрибутами, как и между самими объектами, будут отображаться на следующие отношения:

$$R(E,A) \rightarrow R(K^E, K^A) \quad \text{и} \quad R(E,E^*) \rightarrow R(K^E, K^{E^*}).$$

Согласно этому отображению выбор (поиск) Web-документов из VWE по пользовательскому запросу реализуется по одной из следующих строк.

№	$E \rightarrow K$	$A \rightarrow K^A$	$R(E,A) \rightarrow R(K^E,K^A)$	$R(E,E^*) \rightarrow R(K^E,K^{E^*})$	Результаты (Web-документы)
1.	T	T	T	T	$v_1 \cap v_2$
2.	T	T	T	F	v_1
3.	T	T	F	T	v_2
4.	T	T	F	F	$v_1 \cup v_2$
5.	T	F	F	T	v_2
6.	F	T	F	F	v_1
7.	F	F	F	F	-

Принципиальная схема ДИПС проиллюстрирована на чертеже.

Структура ДИПС включает в себя

модуль пользователя - Web-браузер пользователя;

пользовательский интерфейс (USER INHERFASE), состоящий из модуля формирования запроса (UQ forming) и модуля определения релевантности (Relevance definition);

информационную модель доменов (domen information model) IMD для описания информационных ресурсов интернета (Internet);

модули отображения Mapping-1 и Mapping-2;

модель виртуального Web-окружения (Virtual Web Environment - VWE);

модель Web-документа (Web document model);

модуль адаптации (ADAPTANION);

базу знаний (DB).

Работа системы осуществляется следующим образом.

Модуль интерфейса пользователя (UI) ДИПС выполняет функции формирования поискового запроса пользователя UQ и определения релевантности документов к пользовательским запросам. Пользователь, используя существующий Web-браузер, выражает свои информационные потребности и формирует необходимый поисковый запрос UQ, далее отправляет его через блок отображения "MAPP2" в модуль виртуального Web-окружения (VWE). Блок MAPP2, имеющий знание о UQ и VWE, выполняет механизм отображения между UQ и VWE. MAPP2 также реализует оптимизацию просмотра (поиска), используя базу данных и определяя наилучшее направление и маршрут процесса поиска для сформированного UQ. Когда пользователь получает список Web-документов из VWE, модуль оценки релевантности результатов поиска осуществляет ранжирование найденных Web-документов на основе критерия релевантности.

Технический эффект заявляемого изобретения позволяет повысить уровень эффективности информационного поиска в Интернете путем систематизации Web-пространства, сужения области поиска и организации тематически направленного поиска, оставляя при этом без изменения программно-технические параметры поисковых систем.

Литература.

1. Патент РФ № 2383920 "Система и способ для клиент-обоснованного поиска веб-агентом".

2. Патент РФ № 2413278 "Способ отбора информации в сети интернет и использования этой информации в разделяемом веб-сайте и компьютерный сервер для реализации этого способа".

ФОРМУЛА ИЗОБРЕТЕНИЯ

Способ создания доменно-ориентированной поисковой системы (ДОПС) включает

разделение виртуального информационного пространства IM на семантически слабосвязанные домены, которые являются зонами охвата отдельных поисковых систем;

создание многослойной иерархической структуры, на каждом уровне которой формируются доменные области со своими атрибутами, имеющие предметно-содержательную связь;

формирование пользовательского запроса UQ, поиска информации с использованием "MAPP2" (Mapping 2), "MAPP1" (Mapping 1), оценки релевантности; и

ранжирование результатов,

при этом множество отношений между объектами и их атрибутами на всех уровнях структуры описываются следующим выражением:

$$IM \to \{N, E, H(E), H(E, E^*), A, R(E, A), R(A, A^*), R(E, E^*)\}$$

где H(E) - матрица (или вектор) размещения объектов (категорий) на уровнях иерархии доменов, где $H(e_i)=1$, если объект e_i размешен на домене, и $H(e_i)=0$ в обратном случае, $i = \overline{1,n}$;

H(E,E*) - матрица отношений категоризации (вложенности на subtype и supertype) между объектами, где $H(e_i,e_j^*)=1$, если объект e_i имеет отношение вложенности на subtype или supertype с объектом e_j^*, и $H(e_i,e_j^*)=0$ в обратном случае, $i,j = \overline{1,n}$;

R(E,A) - матрица отношений между объектами и атрибутами доменов, где $R(e_i,a_i)=1$, если объект e_i имеет отношение релевантности с атрибутом a_j, и $R(e_i,a_j)=0$ в обратном случае, $i = \overline{1,n}, j = \overline{1,m}$;

R(E,E*) - матрица отношений между объектами доменов, где $R(e_i,e_j^*)=1$, если объект e_i имеет отношение с объектом e_j^*, и $R(e_i,a_j)=0$ в обратном случае, $i,j = \overline{1,n}$.

Общая структура ДИПС