УДК 681.3

Ю.В. Рогушина, канд. фіз.-мат. наук

МОДЕЛІ ПОДАННЯ ЗНАНЬ ПРО СТАЛІ ІНФОРМАЦІЙНІ ІНТЕРЕСИ КОРИСТУВАЧІВ В ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ

Інститут програмних систем НАН України, e-mail: jjj@cybergal.com

Розглянуто онтологічний підхід до інформаційного пошуку в Інтернеті, який через збільшення обсягу інформації і розосередження її джерел стає все більш складним. Показано, що критичним є не стільки час пошуку, скільки добір інформації, яка релевантна запиту користувача.

Вступ

Для того, щоб підвищити ефективність інформаційного пошуку в розподіленому середовищі Інтернету, потрібно враховувати відомості про область інтересів конкретного користувача. Для цього треба мати адекватні засоби формалізованого опису предметних областей, придатні для використання користувачами, що не є спеціалістами в галузі інформаційних технологій. У даній роботі пропонується використовувати для цього онтологічне подання знань про предметну область (ПрО).

Онтологічний підхід дозволяє користувачеві не тільки використовувати готові онтології, але і самостійно створювати і модифікувати їх. Така вимога призводить до спрощення подання онтологічних систем, але дозволяє значно розширити коло потенційних користувачів.

Аналіз досліджень інформаційно-пошукових систем

Інформаційний пошук являє собою процес зіставлення запиту користувача з відомостями про інформаційні ресурси (ІР), що відомі інформаційно-пошуковій системі (ІПС), до якої надійшов цей запит.

Запит користувача – це опис інформації, доступ до якої він хоче одержати. Релевантність результатів пошуку в даній роботі оцінюється з погляду користувача.

Документи, що за будь-якими параметрами, у т. ч. і не зазначеними явно в запиті, не задовольнили користувача, вважаються не релевантними. Чим складніше форма подання запиту, тим вище релевантність пошуку. Але ускладнення форми запиту призводить до ускладнення процедури його обробки і, отже, до збільшення часу пошуку.

Традиційні механізми пошуку в Інтернеті, як правило, розглядають запити користувача на пошук інформації ізольовано один від одного і не враховують отримані раніше результати.

Значно підвищити ефективність пошуку дозволяє його персоніфікація, тобто використання відомостей про попередні запити конкретного користувача і сферу його інформаційних інтересів.

Ураховуючи контекст пошуку – інформацію про користувача, його інтереси і виконані раніше запити, – можна отримувати більш релевантні результати. Існує кілька підходів до формалізації такого контексту.

Наприклад, у проекті Inquirus NEC Research Institute контекстна інформація задається явно у вигляді категорії даних, які хоче знайти користувач [1].

Ця інформація використовується для вибору механізмів пошуку, яким передається запит, для модифікації запитів і для визначення принципів упорядкування отриманих документів.

Деякі засоби здатні визначити контекст пошуку автоматично. Наприклад, система Watson моделює контекст інформації, що потрібна користувачу, на основі змісту документів, які користувач раніше редагував засобами Microsoft Word чи переглядав в Інтернеті Explorer.

Ці документи аналізуються за допомогою евристичного алгоритму, який виявляє слова, характерні для змісту документів. Потім знайдені слова автоматично додаються до запиту користувача.

Крім того, Watson у фоновому режимі шукає в Web-документи, пов’язані з матеріалами, що редагує та переглядає користувач.

Недоліком системи є непрозорість алгоритмів, що використовує система, для кінцевого користувача.

Remembrance Agent індексує визначені файли (повідомлення електронної пошти, наукові статті тощо) та, поки користувач працює з певним документом, веде пошук документів, пов’язаних з цим документом.

До аналогічних рішень можна віднести Autonomy’s Kenjin, агентів Fab, Letizia і WebWatcher, що вивчають область інтересів користувача для того, щоб запропонувати йому відповідні Web-сторінки.

Спільним недоліком всіх цих систем є те, що вони не використовують стандартизовані засоби для опису ПрО, яка цікавить користувача, внаслідок чого знання, здобуті однією системою, не можна застосовувати в іншій.

Постановка завдання

Для того, щоб підвищити ефективність інформаційного пошуку та враховувати контекст пошуку, потрібно розробити адекватну формальну модель подання знань про інформаційні інтереси користувачів Інтернет-ресурсів.

Онтології як засіб формалізації знань користувача про предметну область

Один з перспективних підходів до завдання контексту пошуку ґрунтується на онтологіях, що містять перелік основних термінів, зв'язки між ними і правила виведення.

Так, у проекті Semantic Web, спрямованому на аналіз семантики ІР, саме онтологічний підхід є основою для подання знань про різні ПрО. У рамках проекту Semantic Web задіяні передові інформаційні технології: агентно-орієнтований підхід у програмуванні – проект DAML+OIL (DARPA Agent Markup Language + The Ontology Inference Layer) [2], онтологічні системи [3], XML [4] тощо.

Консорціумом W3C розроблено ряд низькорівневих протоколів роботи Web-агентів, мова опису сервісів агентів WSDL (Web Service Description Language), протокол обміну інформацією між програмними агентами SOAP (Simple Object Access Protocol) та специфікація UDDI (Universal Description, Discovery and Integration), що пропонує користувачам уніфікований і систематизований засіб пошуку постачальників послуг через централізований реєстр Web-служб.

Для того, щоб формально відобразити в метаописі ІР знання про ПрО, до якої відноситься цей ІР, у проекті Semantic Web використовують онтологічні системи, що містять ієрархію концепцій ПрО й описують важливі властивості кожної концепції за допомогою механізму «атрибут – значення».

У Semantic Web машинна обробка змісту ІР здійснюється шляхом розмітки документів за допомогою онтологічних термінів.

Онтології дозволяють концептуалізувати домен фіксуванням сутностей і зв'язків між ними.

Для формального подання онтологій розроблено мови онтологій DAML+OIL та OWL. Ці мови засновані на розробках Консорціуму W3C, RDF (Resource Description Framework) [5] та RDF Schema.

DAML+OIL – це семантична мова розмітки Web-ресурсів, яка розширює стандарти RDF і RDF Schema більш повними примітивами моделювання. Онтологія DAML+OIL – це колекція RDF-трійок. Остання версія мови DAML+OIL забезпечує багатий набір конструкцій для створення онтологій і розмітки інформації таким чином, щоб їх могла читати і розуміти машина.

Класифікація онтологій DAML Ontology Library складається з груп онтологій, об'єднаних за такими параметрами:

– унікальним ідентифікатором ресурсу URI;
– датою подання;
– ключовими словами;
– каталогом Open Directory Category;
– класами;
– властивостями;
– просторами імен;
– джерелом фінансування (організацією);
– підлеглими організаціями (Ontology Language).

Мова подання онтологій OWL має розширити можливості XML, RDF, RDF Schema та DAML+OIL. Цей проект знаходиться на стадії розробки, але його розробники передбачають створення потужного механізму семантичного аналізу. Планується, що основні переваги OWL порівняно з DAML+OIL будуть полягати в усуненні деяких обмежень і тих конструкцій DAML+OIL, які не використовуються, а також у здатності прямо вказувати симетричність властивості.

Онтологія OWL є послідовністю аксіом, фактів і посилань на інші онтології. Онтології також містять компоненти для опису авторства і подібної інформації. Онтології OWL є документами Web. На них можна посилатися через URI.

Персоніфікація інформаційного пошуку

Проблема інформаційного пошуку ускладнюється тим, що різні групи людей, що займаються збором і пошуком інформації, застосовують для спілкування з ІПС як свої спеціальні терміни, так і терміни, широко використовувані іншими співтовариствами в іншому контексті.

Поряд із глобальними онтологіями, що описують досить широкі ПрО і для створення яких необхідні значні зусилля як експертів ПрО, так і інженерів, існують онтології, що дозволяють формально подати знання конкретного користувача про ПрО.

Такі онтології можуть створюватися і модифікуватися користувачами самостійно. Хоча, можливо, деякі знання користувача про ПрО є помилковими, але така онтологія описує ПрО, що відповідає інформаційним інтересам саме цього користувача.

Наприклад, якщо користувач помилково вважає дельфіна рибою і, запросивши зображення якої-небудь риби, отримає зображення дельфіна, тоді його інформаційна потреба буде задоволена.

Формальна модель онтології ХЕ “Формальна модель онтології [6] О – це упорядкована трійка

$$O = \langle X, \Re, \Phi \rangle,$$

де X – скінченна множина термінів ПрО (теми), яку описує онтологія О; $\Re$ – скінченна множина відношень між термінами заданої ПрО; Ф – скінченна множина функцій інтерпретації, заданих на термінах і/або відношеннях онтології О .

Для створення онтології, користувач має задати скінченну множину термінів ПрО X, скінченну множину відношень між цими термінами $\Re$ і скінченну множину функцій їх інтерпретації Ф, а потім указати, між якими саме термінами існують які відношення.

Онтологія ПрО може бути візуалізована у вигляді лісу орієнтованих графів з навантаженими дугами, в якому вершини відповідають термінам ПрО, а дуги – відношенням між ними.

Після цього користувач створює свій інформаційний запит і задає за онтологією контекст пошуку.

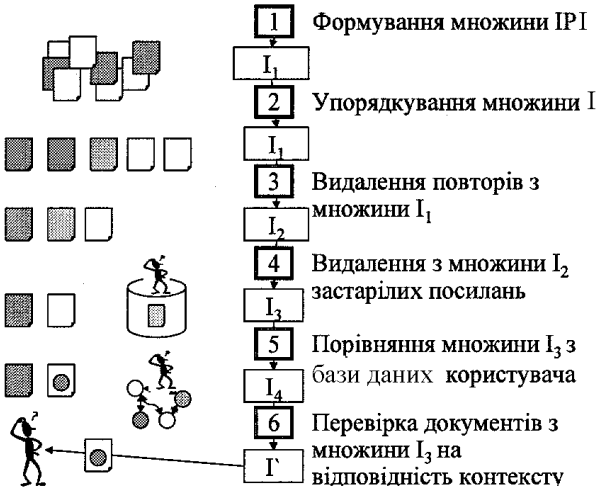
ІПС трансформує цей запит з урахуванням контексту пошуку, обирає потрібні користувачу ІР та передає йому відомості про ці ІР.

Спосіб виконання пошуку залежить від специфіки конкретних ІР.

Для того, щоб користувач мав можливість приступити до інформаційного пошуку, йому треба надати непорожню множину ІР Q, $Q = \langle Q_1, ... Q_n \rangle$, до яких він може звернутися.

Такими ІР можуть бути різні глобальні і локальні пошукові машини, окремі сайти, фіксовані документи тощо.

Етапи обробки результатів виконання запитів подані на рисунку.



Обробка результатів запитів

1. У результаті виконання інформаційного запиту користувача до Q за ключовими словами формується множина *I*: $I = \bigcup_{i=1}^{n} I_j$, де I_j – результат пошуку в ІР Q_j. Якщо є метаінформація про відповідний ІР (наприклад, у форматі RDF або MPEG7), то пошук здійснюється з її урахуванням. На жаль, більшість ІПС, що здійснюють пошук за ключовими словами, включають у *I* багато непотрібної користувачу інформації – повтори, нерелевантні і застарілі посилання, а також посилання на документи, уже відомі користувачу. Щоб позбавити користувача від необхідності переглядати вручну всі ці документи, потрібно здійснити їх фільтрацію, використовуючи відомості про попередні запити цього користувача і сфери його інформаційних інтересів.

2. Якщо множина *I* непорожня, то виконується її упорядкування за URL-адресами посилань. Інакше – завершення роботи.

3. Якщо отримана на кроці 2 підмножина *I* не порожня, то відфільтровуються посилання-“дзеркала”. Інакше – завершення роботи.

4. Відфільтровуються застарілі посилання.

5. Якщо отримана на кроці 3 підмножина *I* не порожня, то здійснюється перевірка за БД користувача, чи одержував він раніше кожне з посилань (якщо одержував, тоді рішення про те, чи залишати це посилання, залежить від того, як у минулому користувач обробив це посилання, а також від інших його інструкцій). Інакше – завершення роботи.

6. Якщо сформована на кроці 5 підмножина *I* не порожня, то виконується перевірка на відповідність документів $i_j, j = \overline{0,k}$ з множини $I`, I` \subseteq I$ контексту пошуку. Інакше – завершення роботи.

Саме на 6-му етапі використовується онтологія ПрО, створена раніше користувачем.

Застосування онтологій для завдання контексту пошуку, у першу чергу, орієнтовано на тих користувачів, що мають сталі інформаційні інтереси в мережі і потребують постійного надходження відповідної інформації. Запити таких користувачів можуть повторюватися від сеансу до сеансу чи змінюватися, але обмежена ПрО пошуку, в якій користувачі є експертами, практично не змінюються. Опис цих ПрО задається самими користувачами у вигляді онтологій, що містять перелік основних термінів, зв’язки між ними і правила виведення. Один користувач може створювати кілька онтологій, якщо він має кілька цікавлячих його прикладних областей, що не перетинаються.

В онтології користувач може відзначити терміни, наявність яких у шуканому документі є бажаною або небажаною, а також задати більш складні операції, наприклад, автоматично відзначити всі терміни, що знаходяться в певному відношенні з термінами, відзначеними раніше. Це дозволяє, зокрема, легко враховувати при пошуку синоніми чи близькі за значенням слова, а також здійснювати пошук відразу кількома мовами.

У результаті цього формується непорожня множина слів (або словосполучень) $W = \{w_1,...,w_m\}$, кожне з яких може мати свою позитивну або негативну вагу $v_k, k = \overline{1,m}$. Потім для кожного документу $i_j, j = \overline{0,k}$ з множини $I`, I` \subseteq I$ формується коефіцієнт відповідності контексту пошуку:

$$s_j, j = \overline{0,k}, s_j = \sum_{k=1}^{m} v_k * f(i_j, w_k), \quad (1)$$

де $f(i_j, w_k) = \begin{cases} 1, \text{ÿêùî} & w_k \in i_j; \\ 0, \text{ÿêùî} & w_k \notin i_j. \end{cases}$

Чим вище коефіцієнт (1), тим вище ймовірність того, що знайдений документ є релевантним запиту користувача. У деяких випадках замість рівняння (1) корисно використовувати більш складну формулу розрахунку коефіцієнта відповідності контексту пошуку:

$$s`_j, j = \overline{0,k}, s`_j = \sum_{k=1}^{m} v_k * f(i_j, w_k) * t_k,$$

де $t_k, k = \overline{1,m}$ – кількість входжень терміна $w_k, k = \overline{1,m}$ у документ $i_j, j = \overline{0,k}$.

Користувач може звертатися до онтологій, що створені іншими користувачами – переглядати їх, задавати за ними контекст пошуку, копіювати з них потрібні фрагменти, але не має права змінювати їх. ІПС має передбачати пошук онтологій, що містять уведені користувачем терміни, а також пошук онтологій, схожих на обрану користувачем онтологію. Це дозволяє створювати групи користувачів із спільними інформаційними інтересами і запобігти дублюванню у виконанні однакових багаторазових запитів різних користувачів. Адекватним засобом подання таких онтологій є мова OWL.

Висновки

Підвищення ефективності інформаційного пошуку потребує засобів формалізованого подання знань про предметну область, що цікавить користувача, які, з одного боку, мають достатню виразну потужність для відображення взаємозв'язків об'єктів, а з другого – можуть застосовуватися особами, що не є спеціалістами в галузі математики та інформаційних технологій.

Запропонований онтологічний підхід задовольняє ці вимоги та припускає автоматизовану обробку результатів інформаційних запитів користувачів.

Список літератури

1. *Glover E., Lawrence S., Birmingham W., Giles C.L.* Architecture of a metasearch engine that supports user information needs // In Eighth International Conference on Information and Knowledge Management, CIKM 99. – Kansas City, Missouri. – 1999. – Nov. – P.210–216.

2. *A Model-Theoretic* Semantics for DAML+OIL. – http://www.w3.org/TR/daml+oil-model .

3. *W3C* Web Ontology. – http://www.w3.org/2001/sw/WebOnt/.

4. *Distributed* XML: the role played by XML in the next-generation Web. – http://www.xml.com/pub/2000/09/06/distributed.html.

5. *RDF/XML* Syntax Specification. – http://www.w3.org/TR/rdf-syntax-grammar/.

6. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. – С.Пб.: Питер, 2001. – 312 с.

Стаття надійшла до редакції 21.06.04.

Ю.В. Рогушина

Модели представления знаний о постоянных информационных интересах пользователей в информационно-поисковых системах

Рассмотрен онтологический подход к информационному поиску в Интернете, который из-за увеличения объема информации и рассосредоточения ее источников становится все более сложным. Показано, что критическим является не столько время поиска, сколько отбор информации, релевантной запросу пользователя.

J.V. Rogushina

Models of constant user informational interests knowledge representation in informational retrieval systems

Ontological approach to sphere of user`s informational interests formalization for effectiveness increasing of informational retrieval results is proposed.