УДК 004.912

Борис Мороз, д.т.н., проф., Денис Костенко, Вікторія Костенко
Університет митної справи та фінансів

СЕМАНТИЧНА ТА ОНТОЛОГІЧНА КОМПОНЕНТИ МОДЕЛЕЙ ПОШУКУ

Розглядаються причини та проблеми необхідності вдосконалення процесу пошуку інформації. Пропонується використання семантичного та онтологічного підходів для вирішення багатьох проблем, які виникають в процесі пошуку інформації. Також ставиться питання про необхідність вирішення проблеми старіння інформації.
Ключові слова:дані, інформація, онтології, процес пошуку інформації, семантичний пошук.

Borys Moroz, Denys Kostenko, Victoriya Kostenko

SEMANTIC AND ONTOLOGICAL COMPONENTS OF THE SEARCH MODELS

Discusses the causes and problems necessity of improving the information search process. Propose the using semantic and ontological approaches to solve many problems arisingin the information search process. Also raises the question of the necessity of solving the aging information problems.
Keywords: data, information, ontologies, information retrieval process, semantic search.

Інформаційний пошук здійснюється засобами інформаційно-пошукової системи. Основними критеріями якості пошуку результатів інформаційного пошуку є: повнота, точність і оперативність. Пошук виконується в чотири етапи:

1) Визначення інформаційної потреби і власника інформаційного масиву;
2) Формулювання запиту;
3) Вилучення інформації з інформаційного масиву;
4) Ознайомлення з отриманою інформацією і оцінка результатів пошуку.

Семантичний пошук інформації – це процес пошуку документів за їх сенсовим ом. Метою семантичного пошуку є розширення стандартного словникового значення слова або фрази для того, щоб зрозуміти наміри користувача в рамках конкретного контексту. Інформація в процесі “зіставлення” повинна оброблятися з використанням знань (про користувача, ресурси і т.д.). При розробці онтологічної моделі інформаційної потреби користувача враховуються не тільки формальні відомості про запит, але і більш складноструктуровані знання про них. Наявність додаткових знань про те, що саме шукає користувач, дозволяє структурувати знайдену інформацію і надавати її користувачу в більш зручному вигляді. Концепція вирішення проблеми інформаційного пошуку на основі онтологічного походу передбачає використання декількох видів онтологій і онтологічних структур. Онтологія - це формальний опис результатів концептуального моделювання предметної області, представлена у формі, яка сприймається людиною і комп'ютерною системою.

В роботах Т. Грубера розглядалися різні аспекти взаємодії інтелектуальних систем між собою і з людиною. Інтелектуальними системами називаються програми, які моделюють деякі аспекти інтелектуальної діяльності людини. Звичайно, будь-яка програма займається таким моделюванням в тій чи іншій мірі, адже саме в цьому і полягає цінність комп'ютера для людини – комп'ютерна система дозволяє звільнити людину від виконання якоїсь однотипної діяльності. Ця діяльність може бути досить складною і вишуканою, але вона завжди однотипна. У цьому сенсі знання, які закладає в програму її творець (тобто алгоритм цієї програми), завжди статичні, вони не змінюються (звичайно, за винятком дуже конкретних знань, які ми називаємо “даними програми”). Інтелектуальна система в цьому сенсі більш універсальна – в ній знання про те, що треба робити в процесі виконання програми, не вшито в програму раз і назавжди, і може змінюватися. Якщо так, то ці знання необхідно передавати програмі як дані, тобто виникає необхідність їх опису та деталізації.

У ряді випадків онтологія тлумачиться як явна специфікація концептуалізації, тобто абстрактного представлення предметної області, спільне розуміння певної сфери зацікавленості. Це угода про спільне використання понять, що містить засоби подання предметних знань.

Не зайвим буде нагадати, що формально онтологія складається з термінів, організованих у таксономію, їх визначень і атрибутів, а також пов’язаних з ними аксіом та правил виведення. Формальна модель онтології (О) – це упорядкована трійка елементів

$$O = <T, R, F>,$$

де Т – скінченна множина термінів предметної області, яку описує онтологія О; R – скінченна множина відношень між термінами заданої предметної області; F – скінченна множина функцій інтерпретації, заданих на термінах і/або відношеннях онтології О.

На формальному рівні онтологія складається з наборів понять і тверджень про ці поняття, на основі яких можна будувати класи, об'єкти, відношення, функції та теорії. Онтологія як зразок домовленості про семантику предметної області сприяє встановленню коректних зв'язків між значеннями елементів такої області, створюючи умови для їх спільного використання. Цим займається один з засобів проектування онтологій – онтологічний інжинірінг.

Об’єктом дослідження є розробка покращеного алгоритму пошуку інформації у сукупності документів. Необхідність вдосконалення процесу пошуку обумовлена наступними причинами:

1) Неструктурований характер інформації більшості електронних документів. Неструктуровані дані становлять більшу частину інформації, з якою мають справу користувачі. Це – не менш 80-90% інформації, а 10-20% – це структуровані дані.

2) Необхідно зробити пошук динамічним і зручним для користувача.

3) Експонентне зростання кількості документів. При збільшенні простору пошуку пропорційно зростає і кількість документів у відгуку пошукової системи.

4) Відсутність стандартизованих механізмів семантичного індексування.

Для вирішення подібного роду проблем з документом необхідно пов'язати метадані, що дозволяють інтерпретувати й обробляти інформацію, яка зберігається в цьому документі, тобто включити в документ інформацію, яка описує структуру і семантику його змісту. Пропонується використання семантичного пошуку як одного з можливих варіантів вирішення проблем пошуку інформації. Онтологічна модель може бути використана для повнотекстового пошуку і для окремої класифікації. Для побудови онтології потрібно формальне декларативне подання чітко організованих конструкцій, які містять у собі словник термінів тематичної області, опис визначень цих термінів, існуючі взаємозв'язки між ними, і взагалі – теоретично можливі й неможливі взаємозв'язки. Обов’язковим у системі пошуку повинен бути процес “самонавчання” системи. Вважається, що цей процес дозволить ліквідувати ситуації з термінами або назвами, які записуються некоректно у базі даних.

Важливість дослідження авторів цих тез, пов’язаного з онтологіями, обумовлена також тим, що знання, яке не описане і не тиражоване, в кінцевому рахунку стає застарілим і непотрібним. А знання, яке поширюється, є генератором нових знань.

Звідси виникає ще один з майбутніх напрямків дослідження – вирішення проблеми старіння інформації.

Авторами здійснюється процес практичної реалізації програмних засобів семантичного пошуку, який би надавав результати не тільки за заданими словами з запиту, але й за еквівалентними за сенсом. При цьому приділяється увага засобам пошуку по тексту на основі шаблонів – регулярним виразам. Шаблон описує закономірність, якій повинні підкорятися шукані послідовності символів у тексті.

Проблема полягає в тому, щоб (використовуючи онтологічний підхід) зробити пошук динамічним, якісним та зручним для користувача. Для будь-якого типу запиту, що виникає в практичній діяльності, повинні бути знайдені адекватні знання в інформаційному просторі.

На даний момент розроблені деякі компоненти системи, функціональне призначення яких – надавати деякі функції можливості пошуку в неструктурованих текстових документах за різними критеріями. Також на стадії розробки знаходяться компоненти, які працюють з обробкою регулярних виразів.

Література

1. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учеб. пособие. - М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

2. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск: пер. с англ. – М: ООО “И.Д. Вильямс” – 2011. – 528 с.

3. Gruber T.R. The role of common ontology in achieving sharable, reusable knowledge bases // Principles of Knowledge Representation and Reasoning. Proceedings of the Second International Conference, 1991, pp. 601-602.