І. С. Кравчук, Д. В. Тарабанов

Автономно-пооб'єктна організація фактографічних систем

У багатьох випадках метою пошуку в лавиноподібно зростаючих масивах інформації є не документ, а конкретна інформація про певні об'єкти або явища. Інформаційні системи, призначені для такого пошуку, називаються фактографічними, а пошук в них – фактографічним пошуком. В залежності від форми вихідних даних слід розрізняти два види фак-

тографічного пошуку: 1) пошук у повнотекстових базах даних і 2) пошук у структурованих базах даних.

Відносно першого типу пошуку цілком слушною є оцінка Г. Селтона: «Не существует ни одной системы для семантической интерпретации даже относительно ограниченного множества объектов, связей и функций, и перспектива разработки таких систем в предвидимом будущем является сомнительной» [5:510]. Аналогічну думку з цього приводу висловлює і В.А. Широков: «Тим часом, принципово нових, проривних ідей у цих галузях (інтелектуальних мовно-інформаційних систем. – *І. К., Д.Т.*) ми поки що не спостерігаємо, і зовсім не тому, що в них немає потреби» [2:4].

Що стосується пошуку в структурованих базах даних, то він переважно розглядається в межах СКБД, які базуються на попередньому зображенні вихідних даних у формі об'єктно-характеристичних таблиць. У цьому випадку для віднаходження конкретного фактографічного запису необхідно спочатку відшукати потрібну таблицю, а потім вести пошук у межах знайденої таблиці. Для пошуку таблиці доводиться використовувати класифікаційні або дескрипторні мови з усіма їх недоліками, відомими з досвіду експлуатації документографічних ІПС.

Щоб уникнути першого етапу пошуку і забезпечити безпосередній доступ до фактографічних записів, нами пропонується автономно-пооб'єктна організація ІПС з використанням засобів семантики. Об'єктом зберігання при цьому є не об'єктно-характеристична таблиця, а фактографічний запис із зазначенням імені об'єкта, ознак об'єктів і значень ознак.

Така організація інформаційних систем має наступні переваги:

– завдяки автономності записів вона полегшує створення і ведення політематичних баз даних;

– припускає свободу в описуванні однотипних об'єктів;

– збільшує точність і повноту пошуку завдяки використанню парадигматичних семантичних ролей ознак об'єктів.

Опишемо прототип системи, побудованої на викладених засадах.

В основу бази даних розроблюваної фактографічної системи покладений фактографічний запис, що в уніфікованій формі описує той чи інший предмет дійсності. Дана система оптимізована для пошуку комерційних товарів. Таке тематичне обмеження накладено з двох причин: через необхідність конкретизувати зону пошуку та через існуючий серед користувачів попит на пошук такого характеру. Виходячи з цього, фактографічний запис має містити такий набір імен ознак і об'єктів, який дозволяє описувати будь-який існуючий товар. При цьому він не повинен збільшувати кількість імен ознак і об'єктів (а разом з ними – кількість полів у записі, більшість з яких може не використовуватися з причини гетерогенності предметів дійсності) та не повинен ускладнювати адміністрування системи. Проаналізувавши характеристичні таблиці з пошукової системи «Яндекс.Маркет» та комерційної системи «Amazon», проектувальники яких пішли шляхом створення структурованих баз даних при розробці фактографічних інформаційно-пошукових систем, ми виділили набір імен ознак і об'єктів, які відповідають описаним вище вимогам до фактографічного запису розроблюваної фактографічної системи:

Структура фактографічного запису

Параметр	**Значення параметру**
Назва	Унікальне
Категорія	Унікальне, фіксоване
Підкатегорія	Унікальне, фіксоване
Виробник	Множинне
Країна	Унікальне
Дата виготовлення	Унікальне, параметр
Ціна	Унікальне
Функція	Множинне, параметр
Властивість	Множинне, параметр
Компонент	Множинне, параметр
Зображення	Множинне
Опис	Унікальне, вільне
Гіперпосилання	Унікальне

Імена об'єктів, ознак і значень ознак записуються у так звані слоти, які утворюють фрейм (в теорії М. Мінського [3]) або фасет об'єкту (в теорії Ш. Ранганатана [4]). Обидві моделі є гомологічними, проте, щоб уникнути плутанини, ми будемо використовувати поняття «фрейму». «Фрейм є концептуальною структурою для декларативного подання знань про типізовану тематично єдину ситуацію, що містить слоти, зв'язані між собою певними семантичними відносинами. З метою наочності фрейм часто зображають у вигляді таблиці, рядки якої утворюють слоти. Кожний слот має своє ім'я і зміст (параметр і його значення – в інформаційних теоріях)» [1:20]. Така структурована модель запису дозволяє автоматичним системам обробляти інформацію про предмети дійсності без людського втручання. Людина у той самий

час теж має змогу сприймати ці записи, що полегшує їх створення та наповнення.

Параметри деяких слотів (поля) містять лише одне значення, інші містять одразу декілька значень, які вважаються однорідними.

Дамо стислий коментар щодо кожного з параметрів:

У полі «Назва» міститься ім'я об'єкта (товару). Так, наприклад, для об'єкта класу «книга» воно містить назву книги і/або прізвище автора, для об'єкта класу «праска» – назву моделі й/або фірми-виготовлювача.

Слоти «Категорія» і «Під категорія» є так званими ідентифікаторами класів – вони визначають родові відношення двох рівнів. Так для «пилососа» значенням слоту «категорія» буде «побутова техніка», а слоту «підкатегорія» – «пилосос».

У слоті «Виробник» міститься назва фірми-виготовлювача, якщо товар є продуктом кооперативного виробництва, або ім'я автора (видавництва, перекладача й тощо), якщо товар є результатом авторської діяльності.

У полі «Країна» міститься код країни за міжнародним стандартом ООН, у якій товар було вироблено. До слоту «Дата випуску» заноситься дата випуску товару за стандартом часу ISO. До слоту «Ціна» записується ціна товару, зазначена в одній з розповсюджених валют.

Слот «Характеристика» об'єднує три параметри: «Функція», «Властивість», «Компонент». Вони були виділені за лінгвістичними критеріями: кожний предмет дійсності розглядався, як такий, чий мовний референт може вступати у синтагматичні відношення з іншими референтами. Так під «функцією» розумілася дія (віддієслівний іменник, наприклад, «свердління», «розпилювання»); під «властивістю» – ознака (прикметник, наприклад, «синій»; числівник, наприклад, «120 *см*»), а під «компонентом» – елемент, що є складовою частиною об'єкта (наприклад, «блок живлення», «зарядний пристрій»). Подібний підхід раніше не використовувався у фактографічних системах, і його застосування може забезпечити описування семантичного рівня фактографічних даних.

Поле «Зображення» містить адресу графічного зображення продукту. У полі «Опис» міститься зроблений у вільній формі опис товару. Поле «Гіперпосилання» містить адресу сторінки з даним товаром в Інтернет-магазині, який його реалізує.

Фактографічний запис є синтаксично детермінованим, тобто його здійснено у термінах штучної мови розмітки. Це забезпечує уніфікований вигляд запису та полегшує подальше машинне опрацювання. У нашому випадку для зображення записів було використано мову XML (англ. *Extensible Markup Language* – розширювана мова розмітки).

XML, рекомендована Консорціумом Всесвітньої павутини мова розмітки, фактично є зведенням загальних синтаксичних правил. Ця «мова призначена для зберігання структурованих даних (замість існуючих файлів баз даних) та для обміну інформацією між програмами» [7]. Оскільки мова XML призначена також для створення на її основі більш спеціалізованих мов розмітки (таких, як XHTML, MathML, XSLT), у роботі [6: 48]) запропоновано спеціалізовану мову для описування фактографічних даних про комерційні товари – PTF (Product Trading Framework).

Як приклад оформлення фактографічного запису за допомогою синтаксису XML та простору імен PTF пропонуємо фактографічний референт об'єкту дійсності з класу «праска»:

```
<item id=«1»>
    <name>Roventa 200N</name>
    <parent-child>
        <type value=«побутова техніка»/>
        <subtype value=«праски»/>
        <series>праска</series>
    </parent-child>
    <universal-number/>
    <producer-info>
        <producer type=«manufacturer» url=«http://roventa.info»>Roventa</producer>
        </producer-info>
    <made-in>Тайвань</made-in>
    <date>2006</date>
    <price currency=«EUR»>27</price>
    <images>
        <image href=«http://dewr.com/irons/roventa200N_1.jpg»>Праска Roventa 200N з вертикальним струменем пари</image>
    </images>
    <characteristics>
        <feature type=«quality»>білий</feature>
        <feature type=«quality»>1,5 кг</feature>
    <feature type=«quality»>зручний</feature>
        <feature type=«function»>розприскувач</feature>
        <feature type=«function»>5 режимів нагрівання</feature>
        <feature type=«component»>очищувач поверхні</feature>
    </characteristics>
    <description>Roventa 200N – чудовий представник класу компактних прасок з усіма новими функціями, що роблять прасування ще більше простим і швидким</description>
    <url href=«http://dewr.com/irons/roventa200Nb.php»/>
</item>
```

Ми вважаємо, що фактографічні записи, подані таким чином, дещо покращать ефективність фактографічних систем за рахунок забезпечення незалежного доступу до кожного запису та зроблять пошуковий процес зручнішим завдяки порівняно уніфікованому оформленню запису.

На підставі викладених принципів було створено діючий прототип фактографічної пошукової системи «Findgood». Ця система є Інтернет-додатком, тобто розміщена на віддаленому сервері й може бути доступною з будь-якого комп'ютера, підключеного до Інтернет. Це повністю виправдано як призначенням системи, так і сформованою тенденцією до створення Інтернет-програм, які заміняють звичайні програми.

Фактографічна система складається з наступних програмних частин:

1. Агент-павук і валідатор.
2. Індекс (база даних про товари у форматі PTF).
3. Серверний модуль PHP: інтегрує парсер XML-PTF файлів, модуль пошуку, генератор документу формату XHTML.
4. Шаблонна сторінка XHTML із сервісом на основі JavaScript, AJAX, Session.

Принцип роботи розглядуваної системи такий самий, як і у більшості інших інформаційно-пошукових систем: включають два основні процеси – індексацію та пошук.

На етапі індексації агент-павук системи сканує масиви автономних фактографічних записів у форматі PTF з сайтів магазинів, а валідатор системи перевіряє їх на відповідність розглянутим вище правилам оформлення. Валідні документу записи копіюються на сервер системи і заносяться до індексу.

Другий етап – власне пошук, що ініціюється запитом користувача системи. Цей запит потрапляє у пошуковий модуль, який за допомогою автоматичного синтаксичного аналізатора здійснює аналіз фактографічних записів в індексі системи і знаходить ті з них, коефіцієнт релевантності яких є найбільшим. Ці записи подаються у форматі XHTML та відсилаються користувачеві.

Прототип системи було апробовано на конференції «Прикладна лінгвістика-2007» (Миколаїв). Наразі розроблення системи перебуває на стадії «тонкого налагодження».

Література

1. *Баранов А.Н.* Введение в прикладную лингвистику. – М.: Эдиториал УРСС, 2001. – 360 с. **2.** Корпусна лінгвістика / В.А. Широков та ін. – К.: Довіра, 2005. – 471 с. **3.** *Минский М.* Фреймы для представления знаний: Пер. с англ. / Под ред. Ф.М. Кулакова. – М.: Энергия, 1979. – 151 с. **4.** *Ранганатан Ш.Р.* Классификация двоеточием. Основная классификация. Пер. с англ. / Под. ред. Т.С. Гомолицкой. – М.: ГПНТБ СССР, 1970. – 422 с. **5.** *Сэлтон Г.* Автоматическая обработка, хранение и поиск информации. – М.: Советское радио, 1973. – 560 с. **6.** *Тарабанов Д.В., Пахомов М.М.* Метод оптимізації автоматичних пошукових систем на базі когнітивних фреймів // Прикладна лінгвістика 2007: проблеми і рішення. – Миколаїв: Видавництво НУК., 2007. – С. 58–59. **7.** W3C Recommendation: XML 1.1 W3C Working Draft. – 2001. – Available from: <http://www.w3.org/TR/2001/WD-xml11-20011213/>

АННОТАЦИЯ

В работе предлагается использовать в качестве единицы хранения в фактографических системах не таблицу, а самостоятельный объект таблицы. Рассмотрены унифицированные классы признаков таких объектов. Предложенная в статье организация информационной системы позволяет увеличить полноту и адекватность ее функционирования. Описана программная реализация и информационные потоки внутри системы.

SUMMARY

It is suggested in the article using not a table as a main storing unit in information systems, but a table self-dependent object. The article contains an overview of unified characteristic classes of such objects. The information system organization, described in the article, allows to increase the level of completeness and adequacy of it's functioning. It is also given an additional information about the technologies used in the prototype of the informational system.