УДК 004.023

К.т.н, доцент Білостоцький А.І., магістрант Філь А.О.

Національний технічний університет України
«Київський політехнічний інститут»

ДОСЛІДЖЕННЯ ІНТЕЛЕКТУАЛЬНИХ МЕТОДІВ ПОШУКУ ТА РОЗРОБКА ПОШУКОВИХ WEB-СЕРВІСІВ

Вступ

Довільний інформаційний WEB-ресурс як правило повинен надавати засоби пошуку потрібної інформації, яку розміщують на цьому WEB-ресурсі користувачі. Найбільш розповсюдженим зараз є пошук за ключовими словами, але останнім часом дуже стрімко почав розвиватись інтелектуальний пошук за смисловим змістом інформації [1, 2].

Проблемою, яку поставлена вирішити дана стаття, є неповнота пошуку за ключовими словами. Основними недоліками пошуку за ключовими словами, які не роблять його достатньо повним, є наступні:

1) пошук за ключовими словами не враховує числові параметри документа, за якими можна робити фільтри та сортування результатів пошуку;
2) простий пошук за ключовими словами не враховує категорії слів, які близько відносяться до предметної області інформації, по якій проводиться пошук. Такі категорії слів найбільш часто будуть зустрічатися в пошукових запитах. Наприклад, для предметної області бізнесу характерні такі категорії ключових слів: прибуток, інвестиції, окупність, виробництво, збут, склади тощо;
3) пошук за ключовими словами не враховує семантичні характеристики – семантичні відмінки та зв’язки;
4) пошук за ключовими словами орієнтований на предметну область, а не на користувача.

Особливістю інтелектуального пошуку є вирішення цих нестандартних задач пошукових систем. Їх вирішення було започатковано в [1, 2]. Це є актуальна наукова проблема в системах штучного інтелекту та інтелектуальних агентних системах. Це є також актуальна проблема, з практичної точки зору, в галузі створення інтелектуальних пошукових WEB-сервісів.

Проблема (2) вирішується частково, оскільки не враховує можливість заміни отриманої багатослівної синтаксеми однослівним синонімом. Також є невирішеною проблема (3).

Постановка задачі

Дослідити і реалізувати методи інтелектуального пошуку та інтегрувати їх в пошукові WEB-сервіси з метою досягнення більш повних та точних результатів за смисловим змістом інформації. Властивості інтелектуального пошуку повинні доповнювати існуючі в частині реалізації категорій, які включають в себе синоніми ключових слів, які найближче відносяться до предметної області бізнесу; реалізацію семантичних характеристик інтелектуального пошуку – семантичних відмінків та зв'язків.

Огляд існуючих рішень

Існуючі методи інтелектуального пошуку вирішують розглянуті недоліки пошуку за ключовими словами:

1) врахування числових параметрів документа вирішується за допомогою надання спеціального шаблону для створення документа. Окремі поля цього шаблону будуть відповідати числовим даним, і ці дані будуть зберігатися у внутрішній базі даних в числовому вигляді, а показуватися користувачеві у вигляді звичайного текстового документа;
2) врахування важливості ключових слів, близьких до предметної області, вирішується за допомогою створення в базі даних спеціальних груп синонімів для таких ключових слів і подальшого врахування їх пошуковою системою;
3) кожен користувач може створити персонального пошукового інтелектуального агента, і, таким чином, орієнтувати пошук не на предметну область, а на свої особисті дані.

Загальна вага пропозиції за статистичними і семантичними характеристиками підраховується як сума ваг всіх знайдених слів в даній бізнес-пропозиції.

Рішення

Агент семантичної фільтрації. Основними завданнями агента семантичної фільтрації є оцінка семантичної близькості запиту і документа (у відсотках) і впорядковування результуючого набору документів

відповідно до цієї оцінки (документи з великим відсотком семантичної релевантності показуються в першу чергу) [3]. Релевантність знайдених документів оцінюється по трьох параметрах:

1) семантичних відмінках (ролях). Тексти документів розрізають на фрагменти, що містять ключові слова запиту. Фрагменти передаються для обробки модулю семантичного аналізу, який будує їх пошуковий образ. Пошуковий образ – це індекс пар роль, іменна синтаксема; іменна синтаксема = прийменник + відмінок наступного іменника. Така ж процедура застосовується до запиту. Далі виконується порівняння семантичного образу запиту з семантичними образами фрагментів документа. Релевантність в цьому випадку є оцінка знайдених в образі документа пар з образу запиту;
2) семантичних зв'язках. Тексти документів розрізають на фрагменти, що містять ключові слова запиту. Фрагменти передаються для обробки модулю семантичного аналізу, який будує їх пошуковий образ. Пошуковий образ – це індекс трійок тип семантичного зв'язку, 1-а синтаксема, 2-а синтаксема. Далі порівнюється семантичний образ запиту з образами фрагментів документа. Релевантність по зв'язках є оцінка знайдених в образі документа трійок з образу запиту;

Модуль управління словником синонімів. Завданням модуля є управління словником синонімів і реалізація специфічних алгоритмів розширення запитів за рахунок використання словника синонімів. Алгоритм розширення запиту синонімами можна представити у вигляді рекурентної формули:

$$Query_i = (Word_i \,\&\, \mathrm{Syn}(Word_i)) \,\&\, Query_{i+1} \vee \mathrm{Syn}(NG),$$

де $Query_i$ – запит на наступному кроці, $Word_i$ – наступне слово, Syn() – функція додання синонімів, *NG* – іменована група для наступного слова.

Адекватність рішень

Розглянуті методи рішень роблять пошук більш повним в порівнянні з існуючими методами, оскільки:

1) агент семантичної фільтрації оцінює семантичну близькість запиту і документа (у відсотках) і впорядковує результуючий набір документів відповідно до цієї оцінки (документи з великим відсотком семантичної релевантності показуються в першу чергу);
2) модуль управління словником синонімів реалізує специфічні алгоритми розширення запитів за рахунок використання словника

синонімів і враховує можливість заміни отриманої багатослівної синтаксеми однослівним синонімом.

Приклади застосування

Розглянуті методи можна застосувати для розробки WEB-сервісу для пошуку бізнес-партнерів. Вони дають можливість врахувати фільтри та способи сортування результатів пошуку за числовими параметрами бізнес-пропозицій (прибуток, інвестицій, строки ведення бізнесу тощо); врахувати категорії ключових слів предметної області бізнесу (виробництво, збут, склади тощо); надати можливість створення власного інтелектуального параметризованого пошукового агента кожним користувачем.

Висновки

Розглянуті методи інтелектуального пошуку дозволяють робити пошук документів набагато повнішим, і мають наступні переваги перед простим пошуком за ключовими словами і існуючими методами інтелектуального пошуку:

1) враховуються числові параметри документа;
2) враховуються синоніми ключових слів, які мають близьке відношення до предметної області інформації WEB-сервісу;
3) враховуються семантичні відмінки та зв’язки;
4) орієнтація пошуку не на предметну область, а на користувача, за допомогою створення власного інтелектуального пошукового агента.

Література

1. *Козлов Е. Б., Метелкин А. В., Хорошевский В. Ф.* Мультиагентная система поиска информации в Интернет – М.: Физматлит, 2000. – С. 840–850.
2. *Куршев Е. П., Осипов Г. С., Рябков О. В., Самбу Е. И., Соловьева Н. В., Трофимов И. В.* Интеллектуальная метапоисковая система – М.: Наука, 2002. – С. 320–330.
3. *Кормалев Д. А., Куршев Е. П., Осипов Г. С., Сулейманова Е. А., Трофимов И. В.*: Препринт // Методы поиска и анализа информации. Автоматическое извлечение данных – Переславль-Залесский, ИПС РАН, 2003. – С. 260–263.