УДК 004.912

О. В. Бісікало
А. І. Лісовенко

АНАЛІЗ ХАРАКТЕРИСТИК СИСТЕМ ТИПУ «ЗАПИТАННЯ-ВІДПОВІДЬ»

Вінницький національний технічний університет

Анотація

В роботі проведено аналіз та характеристику систем типу «запитання-відповідь». Наведено основні структурні елементи системи та принципи їх взаємодії. На основі даної системи розроблено інформаційну технологію аналізу фахового вхідного контентую.

Ключові слова: аналіз тексту, система сипу «запитання-відповідь», база даних, релевантність відповіді, машинне навчання.

Abstract

In the article was spent the analysis of main characteristic of question-answering system. It's given the basic structural elements of QA system and the principles of their interaction. On the basis of this system was developed information technology of the analysis of specialized entrance content.

Keywords: expert model, QA system, database, relevance of the answer, machine learning.

Останнім часом інтереси дослідників все більше зміщуються у бік інтелектуального пошуку інформації. Значно підвищився інтерес до розробки інтелектуальних та нетрадиційних механізмів пошуку та отримання інформації. В наукових колах при оцінці методів інформаційного пошуку, що орієнтуються на роботу із англомовними інформаційними матеріалами, спостерігається постійна цікавість до розділу систем типу «запитання-відповідь» (СЗВ).

Головною особливістю систем типу «запитання-відповідь» є виконання пошуку відповіді такою системою на основі формування запитального речення природною мовою, а не шляхом підбору ключових слів. Відомі системи інформаційного пошуку дозволяють користувачу отримати список різного обсягу документів, які можуть містити інформацію, що нас цікавить, при цьому залишаючи користувачеві роботу по отриманню необхідних даних із документів, впорядкованих за рівнем релевантності запиту. На відміну від традиційних пошукових систем СЗВ забезпечує повернення короткої відповіді, а не переліку документів або посилань як у пошукових системах [1].

Розглянемо етапи роботи СЗВ – на першому з них виконується введення запитання природною мовою, початкова обробка та формалізація речення різноманітними аналізаторами (синтаксичним, морфологічним, семантичним), де визначаються відповідні йому атрибути для подальшого їх використання. Далі, на другому етапі відбувається пошук та аналіз документів: відбираються документи та їх фрагменти, в яких може міститись відповідь на вхідне запитання. На третьому етапі відбувається вилучення відповіді: система отримуючи текстові документи або їх фрагменти, вилучає із них слова, речення чи уривки тексту, які можуть стати відповіддю.

Слід відмітити, що важливу роль в отриманні результатів та розробці відіграє використання різноманітних словників-тезаурусів [2]. Застосування даних словників вирішує задачу визначення типів сутностей для виявлення відповідей, знаходження початкової форми слів для використання їх у пошукових запитах. Також дані словники використовуються для знаходження синонімів слів.

Для реалізації етапу аналізу введеного користувачем запитання, використовується модуль обробки запитання.

На вхід даного модулю подається запитання природною мовою, а його задачею є створення деякого представлення запитаної інформації. Тобто, модуль обробки запитання повинен:

- аналізувати запитання, щоб зрозуміти, яка основна інформація потрібна для відповіді на запитання;
- класифікувати тип запитання, щоб визначити тип очікуваної відповіді;
- переформулювати запитання, перетворити його в набір запитів для системи пошуку

інформації.

На виході модуля наявний набір запитів, які наступний модуль аналізу запитання може використовувати для пошуку інформації. Зазвичай для аналізу питання застосовуються шаблони (регулярні вирази, синтаксичні шаблони) для розпізнавання поширеного запитального обороту. Іншими відомими методами є метод автоматичного навчання статистичної моделі для визначення семантичного тегу.

Наступним кроком, який буде виконувати СЗВ є пошук інформації, який реалізовується на основі модуля обробки документів [3].

Даний модуль отримує на вході оброблене, переформульоване запитання, і на виході видає ранжований список релевантних документів, в яких може міститися відповідь на запитання. Модуль обробки документів зазвичай шукає інформацію за допомогою однієї або декількох пошукових систем, які майже завжди використовують Всесвітню Павутину як джерело документів. При пошуку використовується дещо інший підхід, ніж в популярних пошукових системах: для системи типу запитання-відповідь перш за все важливі документи, що містять всі ключові слова із запиту, тому що список ключових слів був детально відібраний модулем обробки запитань. Знайдені документи фільтруються та впорядковуються.

Релевантність знайденої інформації вимірюється по двом метрикам: точність та повнота [4].

Точність – це відношення кількості релевантних документів до загальної кількості знайдених документів.

Повнота – це відношення числа знайдених релевантних документів до загального числа релевантних документів у пошуковій базі.

Для пошуку документів у системі типу запитання-відповідь, повнота значно важливіша, ніж точність, тому що результати пошуку піддаються подальшій обробці.

Таким чином, головна мета модуля обробки документів – створити набір упорядкованих параграфів, що містять відповідь на запитання. Щоб досягти даної мети, необхідно здійснити:

-пошук релевантних запитанню документів;

-фільтрацію, для зменшення кількості документів і кількості тексту в кожному з них;

-впорядкування параграфів-кандидатів на вміст відповіді за степенем правдоподібності.

Зменшення об’єму документів до декількох параграфів виконується для прискорення роботи системи. Час реакції системи типу запитання-відповідь доволі суттєвий параметр, тому що система працює з користувачем інтерактивно.

Отже, шляхом аналізу характеристик СЗВ було визначено основні модулі системи, їх функціонал та принципи роботи. Основними затратами часу та ресурсів є ті, що виділяються на пошук та структурування інформації, яка найчастіше розміщується у мережі Інтернет. Так як застосування даної СЗВ в процесі розробки відповідної інформаційної технології супроводжується суттєвими недоліками, було прийняте рішення виконання пошуку у попередньо розробленій базі даних. Такі дані по замовчуванню є структурованими, що значно полегшує все етапи пошуку інформації та застосування СЗВ. Тому потребує розробки метод автоматизованої інтеграції даних текстової колекції у семантично орієнтовану базу даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Соловьев А.А. Синтаксические и семантические модели и алгоритмы в задаче вопросно-ответного поиска / А.А. Соловьев // Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2011, Воронеж, Россия, 2011. – сс. 201-210.
2. Vanitha Guda, Suresh Kumar Sanamrudi, I.Lakshmi Manyakamba Approaches for question answering // International Journal of Engineering Science and Technology. 2011. 3. №2. P. 990-995.
3. Широков В.А. Лінгвістичні та технологічні основи тлумачної лексикографії / В.А. Широков, В.М. Білоноженко, О.В. Бугаков та ін.. – К.: Довіра, 2010. – 295 с. – ISBN 978-966-507-283-6.
4. Агаєв Н.В. Исследование и разработка методов реализации вопросно-ответных систем / Н.В. Агаев // Курсова робота. Московський державний університет ім.М.В.Ломоносова. Факультет Обчислювальної матеиатики та кібернетики. Кафедра Системного програмування. – Москва, 2012. – 35 с.

Лісовенко Анна Ігорівна – асистент каф. АІВТ, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця, e-mail: alis@vntu.edu.ua.

Бісікало Олег Володимирович – д.т.н, проф., декан факультету комп'ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця.

Lisovenko Anna I. – assistant to department of Automatic Equipment And Information And Measuring Equipment, Faculty for Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: alis@vntu.edu.ua.

Bisikalo Oleg V. – Prof., DrSc, Dean of Faculty for Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia.