MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE IV INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2016

May 25 –27, 2016
Sumy, Ukraine

Development of a Method and Architecture of the Information System for Automated Collection of Thematic Information on the Internet

Ruslan Plaks, Natalia Fedotova
Sumy State University, Ukraine, lol.croatoan.lol@gmail.com

Abstract. The aim of my research is to develop the information system architecture collection of information on the Internet, which will allow to automate the process of information search. It must satisfy with setting the entire region searching and finding documents in accordance with it. Develop information system architecture model, which consists of indexing, searching service and virtual data warehouse that allows the researcher quickly obtain data on the topic of your search from various sources.

Keywords. Internet, Search System, Thematic Seach, Knowledge Bases.

ВСТУП

Останнім часом у зв'язку з швидким розвитком комп'ютерної техніки і телекомунікаційних технологій зросла проблема пошуку інформації в мережі Інтернет. На сьогоднішній день в електронному вигляді зберігається величезна кількість документів, описів, інструкцій, підручників, наукових статей та багато іншої неструктурованої інформації. Проблема знаходження серед такого обсягу потрібної інформації стає вкрай важливою і найчастіше важко розв'язуваної без використання спеціальних засобів. Таким чином, на сьогоднішній день існує потреба в опрацюванні цілого ряду аспектів, що стосуються функціонування систем інформаційного пошуку.

Завдяки цьому розвили свою популярність різні інформаційно–пошукові системи, пошукові роботи, мета – системи тощо. Але ці системи ефективні для пошуку популярної та релевантної інформації, проте вони не вирішують комплексних задач. Існуючі пошукові системи здатні знайти велику кількість інформації, частина якої тим чи іншим чином відноситься до запиту користувача, але більша частина інформації являє собою сміття. Це відбувається тому, що пошукові системи для пошуку інформації використовують ключові слова. Вони не відрізняють інформацію між собою. Для цього необхідно застосовувати фільтри, які б розділяли інформацію по предметній області та виконували пошук всередині цієї області.

На рис.1 показано спрощене схематичне зображення роботи моделі пошукової системи.

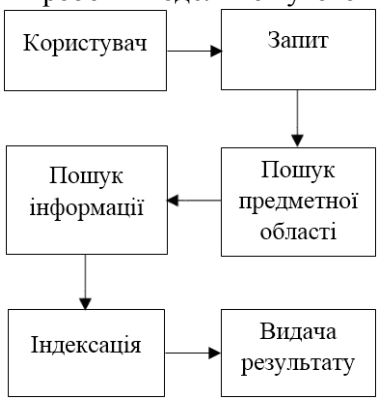

Рисунок 1 – Блок – схема моделі пошукової системи

Отже, необхідно розробити метод пошуку інформації, який би задовольнив поставлені завдання.

ЗАВДАННЯ ДОСЛІДЖЕНЯ

Для вирішення поставленої мети в поставленій роботі визначені наступні завдання дослідження: розробка методу розрахунку відповідності документа запиту; розробка алгоритму пошуку і збору даних; розробка критерію ефективності пошуку;

проектування архітектури інформаційно-пошукової системи.

Наукова новизна роботи полягає в тому, щоб вперше застосовати технологію семантико-ентропійного пошуку з використанням моделі контекстно-часової онтології для пошуку неформалізованих даних з різних джерел, використати обробку та аналіз семантичних даних в системах підтримки прийняття рішень з використанням контекстно-часової онтології, застосувати семантико-ентропійний пошук в мережі Internet, побудувати модель контекстно-часової онтології: ввести поняття фактора достовірності, що залежить від часу; запропонувати метод розрахунку оцінки невизначеності запиту з використанням ентропійної оцінки; запропонувати метод розрахунку оцінки релевантності документів з урахуванням коефіцієнтів достовірності, як розрахунок міри близькості графів, отриманих шляхом побудови семантичних мереж документа і запиту на підставі побудованої експертом контекстно-часової онтології.

Практична цінність роботи заключається в розробці архітектурної моделі ІС, що складається з індексуючого, пошукового сервісу і віртуального сховища даних та надає можливість дослідникові оперативно отримувати дані по темі свого пошуку з різних джерел.

ГІПОТЕЗА ДОСЛІДЖЕННЯ

Дослідження полягає в тому, щоб застосувати модель контекстно-часової онтології для пошуку неформалізованих даних з різних джерел і контролювати їх відповідність вимогам із забезпеченням функціонування ІС відповідно до вимог споживачів, яка дозволить підвищити ефективність функціонування інформаційної системи в цілому.

Припускається використовувати в роботі методи дослідження, які базуються на використанні різних методів теорії:

- теорії графів;
- теорії прийняття рішень;
- теорії інформації;
- нечіткої логіки;
- теорії ймовірності та математичної статистики;
- методів інформаційного пошуку;
- математичного моделювання;
- графовой кластеризації;
- модульного і об'єктно-орієнтованого програмування.

ВИСНОВКИ

В даний час робота з документами в інтернеті без застосування інформаційно – пошукових систем майже неможлива. Але навіть з ними це займає досить великої кількості часу. Тому дана робота була присвячена розробці архітектури та моделі інформаційної системи автоматичного збору тематичної інформації в мережі Internet.

Очікується, що розроблена модель пошукової системи значно підвищить точність отриманих результатів, за рахунок пошуку по заданій предметній області.

Були проаналізовані існуючі моделі пошукових систем, агентів, роботів. Застосовані методи теорії графів, теорії прийняття рішення, теорії інформації, нечіткої логіки, теорії ймовірності та математичної статистики, методів інформаційного пошуку, математичного моделювання, графової кластеризації, модульного і об’єктно – орієнтованого програмування тощо.

REFERENCES

[1] Cutting D., Pedersen J.O., Karger D., Tukey J. Scatter /Gather: A cluster-based approach to browsing large document collections. // Proceedings of SIGIR'92, Copenhagen, Denmark, June 21-24 1992, pp. 318-329.

[2] Craven M., DiPasquo D., Freitag D. et al. Learning to construct knowledge bases from the World Wide Web // Artificial Intelligence 118(1-2), pp. 69-113.

[3] Berendt B., Hotho A., Stumme G. Towards Semantic Web Mining // ISWC 2002, LNCS 2342, Springer-Verlag Berlin Heidelberg, 2002, pp. 264-278.

[4] Goldszmidt M., Sahami M. A probabilistic approach to full-text document clustering // SRI Technical Report ITAD-433-MS-98-044, 1997.

[5] Dhillon I.S., Fan J., Guan Y. Efficient clustering of very large document collections // Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishing, 2001, pp. 12-31.