ПОБУДОВА АЛГОРИТМУ АВТОМАТИЧНОЇ ІНДЕКАСАЦІЇ ПОВНОТЕКСТОВИХ ДОКУМЕНТІВ ЗА КЛЮЧОВИМИ СЛОВАМИ

Каніщева О.В, Ляхвацкая О.М
Науковий керівник – к.т.н., ст. викл. Каніщева О.В.
Національний технічний університет
«Харківський політехнічний інститут»
(61002, Харків, вул. Фрунзе, 21,
каф. інтелектуальних комп'ютерних систем, тел. (057) 707-63-60)

The given work is devoted working out information and the lingware for the automated indexing by keywords of the text-through document which is convenient and more useful in sphere of analysis of the text for the user.

In this article it is considered the analysis of a subject domain and statement of the task for automatic indexation of the document on keywords is considered, have analysed existing methods of automatic indexation of text-through documents. The review of modern industrial systems which realise automatic indexation by keywords is made. The algorithm of indexation text-through is shown the document.

Автоматизована індексація – індексація, технологія якої передбачає використовування формальних процедур, здійснюваних за допомогою обчислювальної техніки, і включає застосування інтелектуальних процедур при ухваленні основних рішень про склад пошукового зразку.

Індексація – в інформаційному пошуку – процес опису документів і запитів в термінах інформаційно-пошукової мови. У індексації кожному документу призначається набір ключових слів, що відображають його смисловий зміст.

Мета даної роботи – розробка алгоритму для автоматизованого індексування ключовими словами повнотекстового документу.

У роботі були проаналізовані наступні методи індексування повнотекстових документів:

1) Бінарне індексування – не залежить від мови документа по причині бінарної або словникової індексації. При бінарній індексації пошук ведеться на основі алгоритмів "нечіткого пошуку", тобто пошуку з помилками. В цьому випадку допускається неповний збіг слів з шаблоном.

2) Морфологічне індексування – виробляється с розрахунком морфології та семантики мови. При цьому методі індексації слова перетворяться в словоформи з відсіканням суфіксів і закінчень, що дозволяє шукати відміни і відмінювання шаблонів.

3) Індексування за «Ключовими»словами. Ключові слова – це слова, які визначають зміст документу, характеризують його значення. Але проблема в тому, що одні і ті ж слова в різному контексті можуть бути, а можуть і не бути "ключовими".

У роботі було зроблено огляд існуючих програмних систем, які реалізують функцію автоматичного індексування: Галактика-ZOOM, УІС Росія, Russian Context Optimizer (RCO).

Схема за якою буде створюватись програма автоматичної індексації повнотекстового документу за ключовими словами, представлена на рис. 1.

Рис. 1. Схема автоматичної індексації повнотекстового документу

У роботі була розглянута задача автоматичного індексування, як процесу автоматизованої обробки мови та побудовано алгоритм для написання програми. Проаналізована модель TF-IDF, за допомогою якої будуть вилучатися ключові слова для автоматичного індексування.

Список джерел інформації:

1. Автоматическая классификация текстовых документов с использованием алгоритмов и семантического анализа / Андреев А. М., Березкин Д. В., Морозов В. В., Симаков К. В. НПЦ «ИНТЕЛЛЕКТ ПЛЮС», 2003.

2. Модели и методы автоматической классификации текстовых документов / А.М. Андреев, Д.В. Березкин, В.В. Сюзев, В.И. Шабанов // Вестн. МГТУ. Сер. Приборостроение. – М.: Изд-во МГТУ. – 2003. – №3.