Дослідження методів повнотекстового пошуку та індексування документів електронної бібліотеки

А.А. Долженко, *студент,*
О.П. До́ренський, *викладач*
Кіровоградський національний технічний університет

Під час розробки програмного забезпечення електронних бібліотек постає задача реалізації повнотекстового пошуку інформації. Одним з найважливіших факторів, які впливають на якість пошуку, є метод внутрішнього преставлення документів.

Метою роботи є дослідження методів повнотекстового пошуку та індексування документів електронної бібліотеки.

Дослідження [1-4] показали, що застосування електронних бібліотек стало буденним явищем, а отже дослідження методів та методик їх програмної реалізації є актуальною та перспективною задачею, розв'язок якої дозволить запропонувати оптимальні шляхи програмної реалізації і впровадження сучасних бібліотечних інфорормаційних ресурсів.

Існує ряд методів, найбільш поширені з яких наступні:

а) лексичне індексування.

В основі лексичного індексування лежить булева модель. Запити користувача представляють собою деякий логічний вираз, в якому ключові слова поєднані операторами AND, NOT або ANDNOT. Під час використання цієї моделі індекс організується у вигляді інвертованого файлу, в якому для кожного терміна зі словника колекції зберігається список документів, в яких цей термін зустрічається.

Даний тип індексування досить добре поширений, але при цьому має істотні недоліки. Так як пошук ведеться за допомогою логічних об'єднань / перетинів документів, в яких наявні ключові слова, то результат пошуку є повністю безконтекстним, що значно знижує його релевантність.

Слід зазначити, що під релевантністю в даному випадку слід розуміти міру відповідності результатів пошуку завданню, поставленому в пошуковому запиті, тобто наскільки повно той або інший документ відповідає критеріям, вказаним в запиті користувача [4].

б) векторне індексування.

Більш досконалим і ефективним з погляду релевантності отриманих результатів є метод векторного індексування. У цій моделі запит користувача, також як і документи подаються у вигляді вектора в базисі слів словника. Найбільш релевантними вважаються ті документи, кути векторів яких з вектором запиту мінімальні.

в) ймовірнісне індексування.

Даний вид індексування зіставляє кожному слову його вагу в документі. Це призводить до значного підвищення якості пошуку в порівнянні з лексичним і ветокрним індексуванням.

г) приховане семантичне індексування.

Математичний апарат даного методу базується на економному сингулярному розкладанні матриць, яке дозволяє виявити приховані семантичні зв'язки при обробці великої колекції документів.

Теоретична ефективність методу набагато вище лексичного або векторного індексування, але через його високі вимоги до обчислювальних можливостей сервера застосування його ускладнене.

Існує досить багато інших методів внутрішнього представлення документів, але через їх складну формалізацію вони не отримали широкого поширення.

Процес індексування включає наступні етапи, які здійснюють у зазначеній нижче послідовності:

- аналіз і визначення змісту документа, як об'єкта індексування;
- вибір понять, які характеризують зміст документа;
- вибір термінів індексування для позначення понять;
- формування пошукового образу документа з термінів індексування.

Перераховані етапи можуть бути об'єднані в складі технологічних процедур за умови належного виконання кожного з етапів.

Пошуковий образ документа (ПОД) формують з обраних термінів індексування за допомогою граматичних засобів інформаційно-пошукової мови (ІПМ).

В процесі індексування не рекомендується описувати документ як фізичний об'єкт (з погляду його форми, об'єму тощо). Допускається відображати в ПОДі подібну інформацію, якщо вона дозволяє більш точно встановити відповідність документа інформаційної потреби користувача системи [5].

Метою автоматизації індексування є мінімізація матеріальних і людських ресурсів, що витрачаються на процедуру індексування, а також досягнення стабільності та однаковості її результатів. Автоматизоване індексування (АІ) здійснюють за:

- текстом первинного документа;
- заголовком та анотацією або рефератом документа.

АІ за текстом первинного документа повинно включати процедуру стиснення ПОД. З використанням обчислювальної техніки здійснюють такі змістовні етапи АІ:

- виявлення інформативних частин документа;
- ідентифікація слів тексту і приведення їх до нормалізованого вигляду (морфологічний аналіз і синтез);
- формування списку ключових слів вихідного тексту;
- підбір дескрипторів по тезаурусу;
- формування ПОД.

Продуктивність механізмів індексування і пошуку залежить від кількості і розміру індексованих документів, частоти виконання і складності запитів, а також від ресурсів комп'ютера.

Список літератури

1. Волкова А.Ю. Досвід створення віртуально-інформаційних бібліотечних центрів. / А.Ю. Волкова. // Матеріали IV Міжнар. НПК "Інформаційні ресурси: створення, використання, досвід". – Алушта: Центр інформаційних технологій Міжвузівського центру "Крим", 2007. – С. 24-27.
2. Дудченко С.В. Электронный документ как полнотекстовый, справочный и библиографический первоисточник / С.В. Дудченко // Культура народов Причерноморья. – Симферополь: Межвузовский центр "Крым". – 2004. – № 50, Т.3. – С. 59-64.
3. Закон України "Про Концепцію національної програми інформатизації" // Інформаційне законодавство: Збірник законодавчих актів у 6 томах / За заг. ред. Ю.С. Шемшученка, Т.С. Чижа. − Т. 1: Інформаційне законодавство України . − К.: "Видавництво "Юридична думка", 2005. − 416 с.
4. Доренський О.П. Мережні інформаційні технології / О.П. Доренський. – Кіровоград.: Вид-во "Код", 2010. – 234 с.
5. Захаров Д. Е. Разработка интеллектуальной нейросетевой поисковой системы «Нейропоиск». // Тезисы молодежной научно-технической конференции "Наукоемкие технологи и интеллектуальные системы ". – 2002. – С. 32-38.
6. Маршак М.Б., Информационно-поисковые системы Интернета и систем автоматизации библиотек: точки соприкосновения и принципиальные различия. Государственная публичная научно-техническая бибилиотека России. Москва, Россия.