ИСПОЛЬЗОВАНИЕ НОВЫХ МЕТОДОВ ПОИСКА ИНФОРМАЦИИ В ИНТЕРНЕТ - СРЕДЕ

Г.А. Мамедова

Институт Информационных Технологий Национальной Академии Наук Азербайджана, гор. Баку, тел: 38-05-89, e-mail: depart10@iti.ab.az

Аннотация: *В статье указывается, что в большинстве поисковых систем, при поиске информации в Интернете используется статистические методы, при котором, смысл отдельных слов или предложений не анализируется. Указывается на целесообразность использования вероятностно-семантических моделей поиска, при котором учитывается не только частота вхождения поисковых слов в документ, но и тематика документа, оценивается его смысловое содержание.*

Ключевые слова: *поиск информации, статистический поиск, индексирование страниц, семантический поиск.*

В настоящее время Интернет является одним из наиболее важных источников информации. Многие студенты, аспиранты и научные сотрудники уже сейчас значительную часть интересующей их информации получают через Интернет.

В Интернете поиск осуществляется в электронных документах, содержащихся на Web-страницах и в различных узлах сети Интернет. В электронных книгах возможен поиск по всему тексту документа, можно накапливать статистику и формировать различные поисковые *коллекции* документов. Полученные коллекции можно сравнивать между собой, что нельзя было сделать при традиционных методах поиска – в обычных книжных библиотеках.

Одной из основных проблем, с которыми сталкивается пользователь при поиске нужной ему информации является объем этой информации и ее качество. При работе с поисковой системой пользователь желает получить информацию, которая в полной мере удовлетворяла бы его потребностям, т.е. была бы *релевантной* его запросам.

В большинстве поисковых системах, при поиске информации в Интернете используется *статистические* методы [1, 2]. В таких поисковых системах документ рассматривается как последовательность слов и словосочетаний. При поиске информации учитывается заголовок документа, резюме, а также частота вхождения поисковых слов в документ. При этом способе поиска, смысл отдельных слов или предложений не анализируется

Но для получения пользователем более качественной и полной информации статистических методов поиска информации недостаточно. Необходимо анализировать смысловое содержание документа, его грамматику, выявить семантические связи между отдельными словами и словосочетаниями и оценить документ на релевантность. Ясно, что этот способ требует использования конкретного языка и осуществляется за счет применения различных словарей и тезаурусов. Такой способ анализа в литературе называется *семантическим*.

В последнее время появились новые методы поиска, которые делают попытку оценить смысловое содержание документа, использующие обратную связь от пользователя. Это происходит следующим образом: в ответ на запрос пользователя поисковая система выдает некоторое количество документов, в которых пользователь может пометить несколько документов, которые, как он считает, в полной мере соответствуют его запросам. Поисковая система использует полученную информацию для автоматического уточнения запроса, т.е. включает в запрос несколько новых ключевых слов (термов), выделенных в отмеченных документах. Процесс уточнения может повторятьсмя несколько раз.

В других системах поиска (AltaVista, Google и др.) при индексировании документов используют также и информацию о ссылках одних документов на другие, т.е. в алгоритмах поиска используют такую величину, как число ссылок из разных документов на «авторитетный» документ и используют эту информацию при *ранжировании* документов.

В различных системах в алгоритмах поиска используются методы линейной алгебры, векторного пространства, теорию нечетких множеств и другие. На наш взгляд, целесообразным

является использование вероятностно-семантических моделей поиска. При этом способе учитывается не только частота вхождения поисковых слов в документ, но и тематика документа, оценивается его смысловое содержание. Применение этих моделей в поисковых системах повысит эффективность поиска и качество получаемой пользователем информации.

Литература:

1. Хоникат Д. Интернет без проблем, М: Бином, 1996, 334 с.
2. Моррис Б. HTML в действии, СПб.: Питер, 1998, 256 с.
3. Berners-Lee T., Hendler J., Lassila O. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities // Scientific American. 2001. No. 5. P. 34–43.
4. Шумский С. Я. Интернет разумный // Открытые системы. 2001. № 3. С. 43–46.