ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МАТРИЦЫ ИНЦИДЕНТНОСТИ «ТЕРМИН-ДОКУМЕНТ» В ПОЛНОТЕКСТОВОМ ИНФОРМАЦИОННОМ ПОИСКЕ

Дашкевич Е.С.
Национальный технический университет
«Харьковский политехнический институт»
г.Харьков, ул. Пушкинская 79/2, тел. 0978314534
e-mail: esdashkevich@yandex.ua

В эпоху стремительного развития информационных систем, а также быстро растущих потребностей современного общества, остро встает вопрос о создании специфически-ориентированного программного обеспечения. В условиях современной глобальной информатизации, затрагивающей все сферы человеческой деятельности, каждый пользователь нуждается в мгновенных методах обработки различных видов информации. В первую очередь, это касается текстовых документов. С целью усовершенствования методов обработки информации необходимо рассмотреть существующие методы решения задачи построения матрицы инцидентности «термин-документ» для коллекции документов и разработать программное обеспечения для решения данной задачи.

Информационный поиск – обширная междисциплинарная область науки, стоящей на пересечений большого количества наук. Автоматические системы информационного поиска используются для уменьшения так называемой «информационной нагрузки». Наиболее известными примерами ИП являются методы, используемые в интернете. Большинство современных поисковых систем при организации поиска используют скрытое семантическое индексирование. Появление терминов в документе представляется при помощи матрицы «термин-документ».

Для сортировки текстов в коллекции по релевантности, необходимо определить, соответствует ли данный текст запросу, насколько высоко это соответствие. Для этого может быть использована матрица инцидентности «термин-документ» для коллекции документов.

Матрица инцидентности в широком понимании – одна из форм выражения графа, в которой обозначены святи между инцидентными элементами графа. Столбцы матрицы соответствуют ребрам, строки – вершинам графа. Ненулевое значение в ячейке на пересечении строки и столбца указывает на связь между вершиной и ребром – их инцидентность.

В частном случае (а именно в области информационного поиска) матрица инцидентности представляет собой таблицу, отображающую наличие термина в документе. Принцип заполнения матрицы следующий:

1) столбцами матрицы являются документы из предложенной коллекции документов, строками – термины из данных документов;

2) при наличии термина в тексте, в соответствующую ячейку на пересечении столбца и строки заносится единица; при его отсутствии – ноль.

На основе проведенного анализа был разработан следующий, наиболее рациональный, общий алгоритм решения данной задачи:

- создание массива структур;
- разбиение текстов на лексемы;
- исключение повторений терминов в рамках одного текста;
- анализ повторений терминов во всех текстах;
- заполнение матрицы инцидентности.

Преимущества рассматриваемого общего алгоритма:

- непосредственный доступ к файлу, в котором производится поиск;
- исключение повторений в рамках одного текста предоставляет возможность корректного заполнения матрицы;
- последовательный анализ повторений терминов во всей коллекции документов позволяет быстро заполнить матрицу инцидентности «термин-документ»;
- с помощью построенной таким образом матрицы инцидентности «термин-документ» становится возможным построение инвертированного списка для дальнейшего поиска.

Наряду с рассмотренными преимуществами, данный способ решения имеет и недостаток. Им является высокая сложность обработки большого количества документов коллекции.

После выполнения программы, пользователь получает выходную информацию на экран. Стоит заметить, что пользователь не принимает участие в ходе программы, что исключает появление некоторых ошибок. После получения результата работы программы, предоставляется возможность обрабатывать данный результат согласно потребностям пользователя.

Данная задача носит прикладной характер, так как при помощи матрицы инцидентности «термин-документ» становится возможным составлять инвертированные списки, которые являются основой полнотекстового информационного поиска. Таким образом, матрица инцидентности «термин-документ» является связующим звеном в процессе информационного поиска.