ПОЛНОТЕКСТОВЫЙ ПОИСК ПРИ ОБРАБОТКЕ ПЕРИОДИЧЕСКИХ ИЗДАНИЙ

Яковлев Б.С., Проскуряков Н.Е., Стрепков А.Н.
Тульский государственный университет

Abstract
The analysis of different in nature and applies of programs used for full-text search, was done. A number of factors influencing the search for articles and their parts, templates, queries and clarifications was identified. Examined the validity of the use of cloud storage services to quickly find of duplicates or attributes of articles in full-text search.
Keywords: *Full-text search, program, cloud storage services, quickly find*

Периодическим изданиям постоянно необходимо вести учет дубликатов текстовых материалов и проверять их на возможность выхода в прошлом или будущих номерах. Это одна из важнейших задач редакторской группы журналов, так как при постоянной текучке материалов есть большой шанс повторно занести одну и туже статью в разные номера. Также часто встречается ситуация, когда авторы присылают разные версии своих статей в течение недели или месяца, изменив, название файла или самой статьи. Все это действительно создает большие проблемы при обработке статей и подготовке журналов к выпуску.

Решая эту задачу редакторы обязаны ориентироваться не на свою память, а на анализ имеющихся у них архивных данных. Причем программное обеспечение, применяемое для этих целей должно быть надежным и осуществлять полнотекстовый поиск по многостраничным документам.

В силу заявленных требований можно предположить какими свойствами должно обладать программное обеспечение:

1. Поддерживать большое количество типов файлов, включая архивы.
2. Должен поддерживаться поиск слов, фраз и предложений.
3. На поиск не должна влиять структура построения текста в файле, кодировка.
4. Поддерживать морфологию слов в запросе, включая синонимы.

Определившись с необходимыми требованиями к ПО необходимо выявить их перечень и сферу применения. Это не тривиальная задача

из-за того, что программное обеспечение с подобными возможностями встречаются в самых разных отраслях. Поэтому было принято решение проанализировать наиболее популярные программные продукты из следующих типов программных продуктов: специализированные программы для работы с текстовыми данными и поиску в них текста, файловые менеджеры, сетевые или облачные хранилища, а также инструменты операционной системы.

Как говорилось ранее из исследуемых групп ПО были выявлены самые популярные из них: профессиональные программы для поиска – Archivarius 3000 [1]; файловый менеджеры – Total Commander 8.01; средства ОС – встроенное средство поиска из ОС Windows 7 [2]; облачные хранилища – Google Диск [3]. Последний был выбран по причине его широкого использования при подготовке выпуска периодической книжной продукции и наличием в нем контроля версий.

Изначально предполагалось, что наихудший результат дают инструменты поиска, встроенные в операционные системы и файловые менеджеры, так как они не предназначены изначально для полнотекстового поиска, и по большей части должны искать файлы на локальных дисках по их названию.

Таблица 1. Возможности поиска в различных форматов файлов

Формат	Total Commander	Встроенное средство поиска в ОС Windows 7	Google Диск	Archivarius 3000
1. DOC	+	+	+	+
2. DOCX	-	+	+	+
3. RTF	-	+	-	+
4. PDF, средствами MS Word 2007	-	-	+	+
5. PDF, средствами Adobe PDF (от разработчиков)	-	-	+	+
6. PDF, средствами PDF Creator	-	-	+	+
7. RAR (WinRar)	-	-	-	+
8. ZIP (WinRar)	-	-	+	+

Однако первое, что необходимо выявить – какие типы файлов поддерживаются поисковыми программами, существуют ли идеальные для поиска форматы. Для этого был проведен эксперимент с различными типами файлов. Для этого была выбрана статья авторов Б.С. Яковлева, Н.Е. Проскурякова, выполненная в формате DOCX, и в неё был добавлен текст «А.Н. Стрепков». После чего из данного файла были получены интересующие нас форматы (см. табл. 1).

Стоит отметить, что для файлового менеджера Total Commander была проведена специфическая настройка – выбран способ кодировки UTF16, т.к поисковая информация была на русском языке.

В качестве поискового запроса мы использовали неделимую надпись из описанного выше документа – «Б.С. Яковлев, Н.Е. Проскуряков, А.Н. Стрепков». Результаты приведены в табл. 1.

Проведенный эксперимент показал, что наиболее большой охват по форматам имеют ПО Archivarius 3000 и Google Диск. Однако, более важным выводом будет то, что все поисковые программы одинаково хорошо способны работать только с одним типом файла – старыми версиями MS Word 1997–2003 (DOC).

Поэтому на примере формата DOC мы можем проанализировать другие не менее важные проблемы поиска информации. Все последующие эксперименты будут увеличивать требования для поисковых программ, но это будет касаться только возможности выполнить поиск по филологическим методам, так как эти методы являются единственными вариантами при которых мы можем гарантировать, что будет найдена информация и по точному, и отрывочному запросу.

Исходя из опыта можно сказать, что существует несколько очень часто применяемых способов поиска информации:

– по фамилии автора или авторов,

– по названию статьи,

– по ключевым словам.

Первое, что можно проанализировать - это влияние написания слов с буквой «ё». Не секрет, что большинство авторов с фамилиями, имеющие данную букву пишут ее через «е». В результате, если ПО очень прямолинейно действует и производит жесткое сравнение без алгоритма, обрабатывающего фразообразование, то оно не решит поставленной перед ним задачи.

Для проверки мы использовали тот же файл что и в первом эксперименте, создав на его основе 4 файла и заменив фамилии авторов на Киселев, Киселёв, Перепелкин, Перепёлкин. Тип файла – DOC, т.к он обрабатывается всеми ПО. В табл. 2 приведены результаты.

Таблица 2. Возможности программ отличить буквы е и ё из запроса

Поисковый запрос	Total Commander	Встроенное средство поиска в ОС Windows 7	Google Диск	Archivarius 3000
	Количество найденный совпадений по запросу			
1. Киселев	1	2	2	2
2. Перепелкин	1	2	2	2
3. Киселёв	1	2	2	2
4. Перепёлкин	1	2	2	2

Результаты исследования поясним на примере. Фамилия Киселев была в одном файле, Киселёв в другом, соответственно 2 найденных файла программой – это максимальный и правильный результат поиска. Этот эксперимент показал, что буквы «е» и «ё» для большинства программ влияние на адекватность поиска не оказывают.

Проверку адекватности поиска по точному и неточному поисковому запросу было решено проводить по фамилиям авторов. Это сделано для того, чтобы исключить из эксперимента влияние структуры документа и его форматирование на конечный результат поиска.

Для эксперимента использовалась та же статья, что в табл. 1. Поисковым запросом были – «Стрепков» (неточный запрос) и «Б.С. Яковлев, Н.Е. Проскуряков, А.Н. Стрепков» (точный запрос). Результаты приведены в табл. 3.

Таблица 3. Влияние точности запроса на поиск информации

Поисковый запрос	Total Com-mander	Встроенное средство поиска в ОС Windows 7	Google Диск	Archivarius 3000
	Результат			
1. Стрепков	+	+	+	+
2. Б.С. Яковлев, Н.Е. , А.Н. Стрепков	+	+	+	+

Проверка влияния структуры документа и верстки текста на поиск также была проведена для описанного выше файла. В качества текста запроса используется название статьи – «Малозатратный подход к обеспечению устойчивой работы Интернет-ресурсов». В каждом файле были явно заданы три наиболее часто встречающихся ситуации, которые по нашему мнению могут привести к нежелательным последствиям для поиска:

– разрыв строки клавишей <Enter>;

– разрыв строки принудительным разрывом;

– неточное название статьи.

Для проверки последнего пункта в изначальный файл было внесено дополнительное изменение - название статьи стало «Малозатратный подход и обеспечение устойчивой работы Интернет-ресурсов», т.е. предлог «к» и окончание слова «обеспечению» были заменены и теперь не соответствуют поисковому запросу. Таким образом в исследовании теперь 4 файла и максимальный результат у программ – будет 4 найденных совпадения. Результаты исследований приведены в табл. 4.

Таблица 4. Влияние структуры и верстки документа на поиск информации по запросу «Малозатратный подход к обеспечению устойчивой работы Интернет-ресурсов»

Типы изменения структуры документа или фразы	Total Commander	Встроенное средство поиска в ОС Windows 7	Google Диск	Archivarius 3000
	Количество найденный совпадений по запросу			
1. Принудительный разрыв <Enter>	0	1	1	1
2. Принудительный разрыв строки	0	1	1	1
3. Измененное название статьи с разрыв клавишей <Enter>	0	0	1	0
4. Измененное название статьи с принудительным разрывом текстовой строки	0	0	1	0
Итого:	0	2	4	2

В ходе эксперимента было выявлено, что по запросу измененного названия статьи «Малозатратный подход и обеспечение устойчивой работы Интернет-ресурсов» результат работы программ стал немного иным. Поэтому приводим результаты исследований в табл. 5.

Таблица 5. Влияние структуры и верстки документа на поиск информации по запросу «Малозатратный подход и обеспечение устойчивой работы Интернет-ресурсов»

Типы изменения структуры документа или фразы	Total Commander	Встроенное средство поиска в ОС Windows 7	Google Диск	Archivarius 3000
	Количество найденный совпадений по запросу			
1. Принудительный разрыв <Enter>	0	1	1	1
2. Принудительный разрыв строки	0	1	1	1
3. Измененное название статьи с разрыв клавишей <Enter>	0	0	1	1
4. Измененное название статьи с принудительным разрывом текстовой строки	0	0	1	1
Итого:	0	2	4	2

Мы считаем, что такой парадокс связан с тем, что в Archivarius 3000 пока не предусмотрен алгоритм словообразования и морфологии, а реализован поиск по словам, встречающимся во фразе или предложении.

Проведенные эксперименты позволили сделать выводы:

1. Запрос по фамилиям даст более практичен т.к. данные запросы содержат уникальные слова, тогда как названия статей нередко содер-

жат термины и определения, устоявшиеся и часто использующиеся обороты речи, и они объемны по тексту.

2. На основе результатов экспериментов по частичному и точному запросам (табл. 3) и с учетом поддерживаемых форматов файлов (табл. 1) можно сделать вывод, что программы Google Диск и Archivarius 3000 дают более качественный результат поиска по сравнению со всеми остальными.

3. Детальное сравнение работы программного обеспечения (ПО) Google Диск и Archivarius 3000 (табл. 4, 5) показало, что на результат работы ПО Archivarius 3000 имеет большее влияние формат строки запроса. К сожалению в этом случае данное ПО не может гарантировать полное нахождение информации, особенно дублирующейся и измененной.

4. Технология Google Диск показала качественную работу и позволила найти текстовые файлы и часть архивных. Недостатком данной системы является то, что она не может работать с архивами типа RAR. Также стоит отметить, что технология дает очень большой набор результатов, поэтому его нужно сокращать, применяя «метки» папок, т.е. выполнять процедуру поиска в помеченных зонах. Однако это в целом снизит качество поиска.

5. Архив ZIP – идеальный формат для задач поиска и хранения архивных файлов в издательском деле. Он единственный архивный формат, который поддерживается ПО Google Диск и Archivarius 3000.

6. Можно рекомендовать использовать наборы текстовых файлов для хранения названий статей. Это сократит время поиска в архивах и даст возможность проводить поиск и анализировать информацию ПО Google Диск без загрузки на него всего архива.

Литература

1. Архивариус 3000. Мгновенный полнотекстовой поиск документов и e-mail сообщений на 18 языках. [Электронный ресурс]. URL: http://www.likasoft.com/ru/document-search/features.shtml (дата обращения: 03.03.2017).
2. Поиск по содержимому файлов в проводнике Windows 7. [Электронный ресурс]. URL: http://blog.depit.ru/poisk-po-soderjimomu-failov-v-windows-7 (дата обращения: 28.02.2017).
3. Официальная страница облачного сервиса Google Диск. [Электронный ресурс]. URL: https://www.google.com/intl/ru_ru/drive/using-drive/ (дата обращения: 20.02.2017).