

УДК 025.4.03

## **Современные подходы к организации поиска в автоматизированных библиотечно-информационных системах**

*Маркелов И.Н., студент*

*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,  
кафедра «Программное обеспечение ЭВМ и информационные технологии»*

*Кузнецов А.А., студент*

*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,  
кафедра «Программное обеспечение ЭВМ и информационные технологии»*

*Научный руководитель: Шиваров А.Е., к.т.н., доцент*

*Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,  
кафедра «Программное обеспечение ЭВМ и информационные технологии»*

[ashivarov@bmstu.com](mailto:ashivarov@bmstu.com)

### **Введение**

Автоматизированная библиотечно-информационная система библиотеки МГТУ им. Н.Э. Баумана «Яуза» – это экспериментальная разработка АБИС для крупного университета, проводимая совместно кафедрой программного обеспечения ЭВМ и информационных технологий (ИУ-7) и библиотекой МГТУ им. Н.Э. Баумана. Система внедрена и используется уже более 15 лет в библиотеке МГТУ[6].

В рамках статьи предстоит проанализировать имеющийся опыт эксплуатации электронного каталога АБИС «Яуза», оценить действующую организацию поиска в нём, выявить недостатки и предложить пути их решения, основываясь на современных тенденциях организации поиска в АБИС.

### **Анализ системы поиска**

В АБИС «Яуза» web-сайт предоставляет пользователям интерфейс для осуществления поисковых запросов к электронному каталогу библиографических записей АБИС. В системе предусмотрены два варианта поиска – простой и расширенный. Простой поиск позволяет пользователю ввести набор ключевых слов, которые могут содержаться в библиографическом описании или аннотации к книге (например, название или авторы) и осуществляет поиск по этим словам. В режиме расширенного поиска пользователь может задать значения конкретных полей – автор, заглавие, издательство, серия, различные

рубрики (УДК, ББК), ISBN, год издания и некоторые другие ограничения. Предполагается, что, используя расширенный поиск, пользователь имеет возможность сузить область поиска, исключив заранее нерелевантные результаты.

Любую систему организации поиска можно разделить на два этапа: построение поискового образа по данным, среди которых будет осуществляться поиск, и получение результатов поиска на основании поискового запроса.

В АБИС «Яуза» построение поискового образа осуществляется на базе библиографических описаний и иных сведений о книге (например, аннотация). Алгоритм формирования базы данных, являющейся поисковым образом электронного каталога, в том числе, повышает значимость слов, входящих в название книги или список её авторов, путём их многократного повторения в поисковом образе. Так в системе реализовано упорядочивание по релевантности. Стоит отметить, что для поддержания поискового образа в актуальном состоянии он регулярно полностью перестраивается, что приводит к временной недоступности системы поиска в электронном каталоге.

Получение результатов поиска в АБИС «Яуза» разделяется на 2 этапа: подготовка SQL запроса и выполнение SQL запроса на полнотекстовый поиск средствами инструментов MS SQL Server, то есть сама АБИС непосредственный поиск в базе данных не осуществляет, а выполняет следующие действия:

- формирует SQL запрос;
- передаёт его на выполнение в SQL Server;
- представляет пользователю результаты работы SQL Server.

Из этой схемы выполнения запроса следует, что влиять на результаты поиска можно либо изменением алгоритма формирования поискового индекса или путём изменения SQL запроса, что значительно ограничивает возможности поиска и сортировки результатов по релевантности.

### **Анализ эффективности поиска**

Вопрос оценки эффективности поисковых механизмов сложен в силу того, что опираясь на понятие релевантности, то есть семантического соответствия поискового запроса и поискового образа документа, которое является субъективным мнением пользователя, дать объективную оценку невозможно. Тем не менее, выработан ряд формальных признаков (так называемая формальная релевантность), позволяющих оценивать различные поисковые механизмы. Например, мера TF-IDF или функция

ранжирования BM25, используемые для оценки важности отдельных слов в контексте документа[1].

Эффективностью поиска предложено считать степень соответствия результатов поиска поисковому запросу. Для оценки такого соответствия разработаны следующие способы:

- Точность;
- Полнота;
- Выпадение;
- F-мера (мера Ван Ризбергена)[5].

Все указанные критерии опираются на соотношения между фактически найденными документами и документами, которые должны были быть найдены. Однако, как было отмечено выше, заранее не известно, какие документы действительно соответствуют поисковому запросу, что делает невозможным применение данных методов. При решении задач оценки эффективности поиска в крупных системах для этого привлекают экспертов, однако в рамках разработки АБИС университета данный подход представляется нереализуемым.

Тем не менее, ввиду необходимости оценки механизмов поиска, требуется предложить, возможно, более узкий и менее точный, но применимый к задаче поиска в электронном каталоге АБИС метод. Таким методом может стать формальное соответствие текстов частного набора поисковых запросов и результатов поиска, а именно:

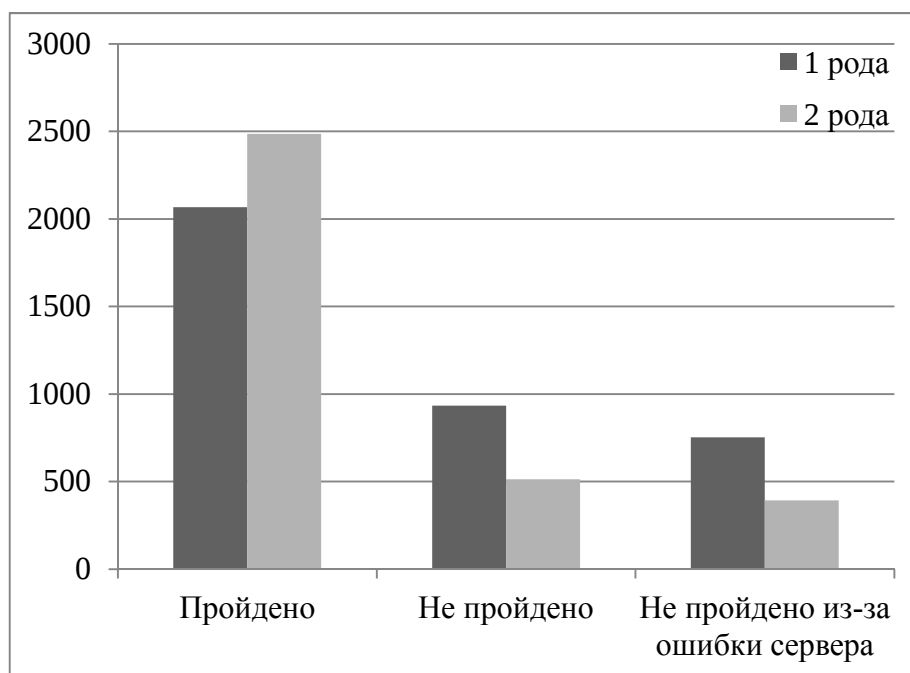
1. Если слово <S> имеется в поле <P> описания документа <D>, то при поиске по слову <S> в результатах поиска должен присутствовать документ <D>;
2. Если при поиске по слову <S> в результатах поиска выдаётся документ <D>, то слово <S> должно содержаться в одном из полей описания документа <D>.

Данные критерии базируются на определении ошибок первого и второго рода математической статистики. При этом ошибками первого рода являются ненайденные релевантные документы, содержащиеся в базе данных, а к ошибкам второго рода можно отнести документы, которые выдаются поиском, несмотря на их нерелевантность (несоответствие) поисковому запросу.

Таким образом, если автоматизировать тесты, проверяющие соответствие данных содержащихся в базе данных электронного каталога АБИС и в поисковой выдаче можно получить процентные показатели по ошибкам первого и второго рода.

Составленные тесты выполняют 3000 запросов к АБИС и SQL-серверу и сравнивают полученные результаты по методике, указанной выше. Итоги тестирования

представлены на рисунке. В результате тестирования ошибки первого рода составили 31,1%, а второго – 17,1%.



Рассмотрим подробно одну из ошибок поиска, чтобы понять её причины.

Например, пользователь ищет книги автора «Григорян Л.А». В результатах поиска на одной из первых позиций есть издание со следующим библиографическим описанием: «Философия в системе культуры : учеб. пособие для вузов / Ильин В. В., Базалин В. Г., Бушуева В. В. [и др.] ; ред. Ильин В. В. ; МГТУ им. Н. Э. Баумана. - Калуга : Полиграф-Информ, 2004. - 611 с. - ISBN 5-93999-103-3.». Как видно из цитаты, Григорян Л.А. не является его автором и не упоминается в библиографическом описании книги, т.е. данный результат не является релевантным, но выводится на седьмой позиции списка результатов поиска. При этом ниже указанного результата имеются книги, автором которых является искомый «Григорян Л.А». Таким образом, очевидно, нарушается ожидаемая сортировка по релевантности.

Рассмотрим запрос, который использует АБИС для поиска.

```
SELECT *
FROM [search_array].[dbo].[description]
WHERE CONTAINS (
    [description_keywords],
    'ISABOUT( "григорян л.а.", "григорян л.а.", "григорян л.а.",
        FORMSOF (INFLECTIONAL, "григорян л.а." ))');
```

Если выполнить этот запрос, то среди результатов поиска действительно будет указанный документ. Это связано с тем, что в поле, по которому ведётся поиск, имеется значение «Тер-Григорян Е. О.», указывающее одного из авторов этого документа. То есть в данном случае проблемы поиска обусловлены не только несовершенством самого поискового механизма, обеспечившего вывод на первые позиции не самых релевантных результатов. К ошибке приводит и неудовлетворительный алгоритм составления поискового индекса, задающего фамилию и инициалы автора как независимые друг от друга слова, что не является верным подходом – при поиске по запросу «Толстой Л.» не должны выводиться на первых позициях книги автора «Толстой А.», а ниже – автора «Толстой Л.».

Рассмотрим ещё один пример. Производим запрос по слову «математика». В результатах поиска на первой странице оказывается книга «История западной философии и ее связи с политическими и социальными условиями от Античности до наших дней: В 3 кн. / Рассел Б. - 4-е изд., стер. - М. : Академический Проект : Мир, 2004. - 1003 с. - (Концепции). - ISBN 5-8291-0389-3. - ISBN 5-902357-10-1.». Из библиографического описания не ясно, почему по запросу «математика» данный документ попадает в результаты. Более детальный анализ показал, что в аннотации к книге указано, что её автор – Рассел является крупным математиком. Так книга, никак не соответствующая тематике запроса, выводится на первых позициях в результатах поиска.

Таким образом, действующая *система поиска совершает ошибки двух типов* – отображает результаты, не являющиеся релевантными, и не выводит результаты, соответствующие поисковому запросу. Обе эти ошибки критичны для поисковых систем, так как первый тип ошибок влечёт увеличение количество результатов поиска и затрудняет выбор нужного, а второй изначально не предоставляет пользователю полной информации о фондах библиотеки. Причём ошибки возникают, в том числе из-за некорректного составления поискового индекса.

Кроме того, текущая версия поиска демонстрирует недостаточную производительность – время ожидания результатов редко оказывается менее 7 секунд, что приводит к ошибкам на сервере (см. рис. 1) и, соответственно, отказам в обслуживании пользователя.

Отметим также, что анализ журналов современных поисковых систем показывает, что пользователи делают ошибки в 10–15% поисковых запросов<sup>[3]</sup>. Поскольку поиск по искаженному запросу обычно приводит к нерелевантной выдаче, поисковые системы уделяют большое внимание проблеме исправления опечаток в запросе и стараются

исправлять их на всех этапах поискового процесса — начиная с корректировки запроса во время набора и заканчивая уточнением во время показа результатов поиска. В АБИС «Яуза» система, позволяющая выявлять и исправлять ошибки, отсутствует.

Многих ошибок при наборе поискового запроса можно избежать, если поисковая система выдаёт поисковые подсказки по мере ввода поискового запроса. Такая функция так же позволяет ускорить процесс ввода поискового запроса, а значит, сократить общее время на поиск. В АБИС «Яуза» поисковые подсказки не реализованы.

### **Выводы по результатам анализа**

Таким образом, на данный момент поисковый модуль АБИС «Яуза» требует существенной переработки для уменьшения ошибок обоих типов.

Выделим требования к улучшенному поисковому механизму АБИС «Яуза»:

- Отсутствие ошибок первого класса и минимизация ошибок второго класса;
- Наличие системы упорядочивания по релевантности, учитывающей значимость термов поискового запроса и поискового образа;
- Время ответа системы при обработке одного запроса менее 3 секунд (например, кеширование популярных запросов);
- Настраиваемые фильтры для обработки знаков препинания, окончаний слов и стоп-слов;
- Учёт рубрикаторов при поиске;
- Использование непрерывного обновления поискового индекса вместо его регулярного полного перестроения;
- Наличие опции, позволяющей учитывать количество книг для выдачи при сортировке результатов поиска;
- Наличие поисковых подсказок для режима простого поиска и для отдельных полей в режиме расширенного поиска;
- Уточнение поискового запроса с помощью исправления возможных ошибок ввода поискового запроса.

Очень полезным улучшением, которое необходимо реализовать, является интерактивный пользовательский интерфейс при поиске, создаваемый с использованием технологии AJAX. При вводе данных пользователю будут предложены подсказки в виде выпадающего списка из слов, первые буквы которых соответствуют уже введенным[4].

Также следует реализовать исправление ошибок в поисковом запросе. Наиболее распространённые функции коррекции запросов – подсказки и автозамена. Подсказка

представляет собой сообщение вида «Возможно, вы искали <...>» и ссылку, по которой пользователь может перейти на страницу с поисковой выдачей по данному запросу. В случае автозамены пользователю показывается поисковая выдача сразу по исправленному запросу, а под поисковой строкой помещается сообщение об автоисправлении.

Недостатком подсказки при исправлении ошибки является необходимость совершать дополнительные действия (нажать на кнопку). Автозамена экономит время, однако точность исправления должна быть намного выше, чем при использовании подсказок, так как неверные автоматические изменения поискового запроса будут приводить к выдаче неверных результатов поиска.

Для определения, какую из функций коррекции применять, используется классификатор надёжности, разделяющий исправления на надёжные (автозамены) и ненадёжные (подсказки). Однако, учитывая многообразие научных терминов, по которым нередко осуществляются запросы в АБИС «Яуза», целесообразно не принимать за пользователя окончательное решение, а предложить ему альтернативы предположительно ошибочному поисковому запросу и реализовать в АБИС «Яуза» исправление подсказками.

Ошибки в поисковых запросах можно отнести к одному из следующих классов:

- Ошибки в отдельных словах – пропуск, вставка, перестановка и замена букв в словах;
- Ошибки слитно-раздельного написания – пропуск или вставка пробела между словами;
- Неверная раскладка клавиатуры – использование английской раскладки для набора русских слов или наоборот.

Для реализации исправления ошибок необходимо взять за основу один из алгоритмов нечёткого поиска в тексте. Данные алгоритмы характеризуются метрикой – функцией расстояния между двумя словами, позволяющей оценить степень их сходства в данном контексте. Для данной задачи предлагается использовать расстояние Дамерау-Левенштейна, являющееся модификацией расстояния Левенштейна: помимо операций вставки, удаления и замены символов рассматривается также операция перестановки символов[2].

В качестве метода повышения качества поиска для исключения ошибок первого и минимизации ошибок второго рода, а так же повышения быстродействия предлагается осуществить интеграцию с полнотекстовой поисковой машиной *Sphinx*. Подобные системы позволяют обрабатывать имеющиеся данные, выстраивая собственный

поисковый образ с учётом значимости тех или иных полей и производить поиск по построенному образу. При этом важное значение имеет качественная и тонкая настройка поискового механизма и механизма индексирования, так как неверные параметры могут сильно ухудшить поисковую выдачу. Рассмотрим, какие недостатки текущей реализации поискового модуля можно решить, внедрив поисковую машину *Sphinx*.

- Низкое быстродействие поиска АБИС «Яуза» удастся улучшить благодаря высокой скорости индексации (до 10-15 МБ/сек на каждое процессорное ядро) и высокой скорости поиска (до 150—250 запросов в секунду на каждое процессорное ядро с 1 000 000 документов в поисковом образе) *Sphinx*;
- Несовершенство алгоритма упорядочивания по релевантности можно преодолеть благодаря поддержке нескольких полей полнотекстового поиска в документе (до 32), поддержке стоп-слов и поиска с учётом морфологии.

### **Заключение**

В данной статье проанализировано текущее состояние механизма поиска в АБИС «Яуза», указаны недостатки имеющегося решения, сформулированы требования к качеству поисковой системы АБИС, предложен новый подход к реализации поиска в каталоге АБИС.

### **Список литературы**

1. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. Киев: Вильямс, 2011. 528 с.
2. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. Академий Наук СССР, 1965. С. 845–848.
3. Панина М.Ф, Байтин А.В, Галинская И.Е. Автоматическое исправление опечаток в поисковых запросах без учета контекста – Яндекс, Москва. Режим доступа: <http://www.dialog-21.ru/digests/dialog2013/materials/pdf/PaninaMF.pdf> (дата обращения 01.04.2015)
4. Поисковые подсказки. Компания Яндекс. Режим доступа <https://company.yandex.ru/technologies/suggest/> (дата обращения 01.04.2015)
5. Черненький В. М., Гапанюк Ю. Е. «Методика идентификации пассажира по установочным данным» // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2012. Спец. вып. № 3 «Биометрические технологии». С. 30-39.

6. Шиваров А. Е., Абрамов Г. В., Пескова О. В., Белостоцкий Н. А. Автоматизированная библиотечно-информационная система технического университета // Вестник МГТУ им. Н. Э. Баумана. Сер. Приборостроение. 2007. № 4. С. 21-32.