Поиск, полнотекстовый поиск и поиск по фасетам в библиотеке прошлого, настоящего и будущего

К. О. Сбойчаков
Государственная публичная научно-техническая библиотека России, Москва, Россия

Статья обобщает результаты работ по развитию оригинальной методики построения и использования фасетов в системе автоматизации библиотек ИРБИС64.

В начале любого пути сортировка и поиск.
Дональд Э. Кнут

ИРБИС64+ предлагает составление Электронной Библиотеки на основе библиографических описаний полных текстов. В системе сочетаются два вида поиска: по библиографии и полнотекстовый. Поиск по библиографии может использовать любые виды полей библиографического описания документа: автора, заглавие, год издания и т.д. Разумеется, если пользователь, извиняюсь, читатель (в библиотеку приходят читатели) заранее знает книги, статьи какого автора он хочет найти или их заглавие, или хотя бы часть слов из заглавия, то библиографический поиск его полностью удовлетворит. И только в том случае, когда читатель из массы тех, кто “сам не знает, что хочет”, то есть он не может четко сформулировать свою потребность, тогда ему может помочь полнотекстовый поиск.

Полнотекстовый поиск подобен старому – не виртуальному, а живому – читателю, который ходил вдоль книжных полок и заглядывал то в одну, то в другую книгу, вычитывая из каждой, возможно, всего по паре строк, и выбирал на основании одного ему известного критерия какую книгу взять. Еще одна важная особенность полнотекстового поиска состоит в его самостоятельности: кроме слов текста он ни на что больше не опирается и может, в принципе работать с “чистым” документом, не имеющим никакого заранее составленного библиографического описания.

Понятно, что читатель не знает имеет ли искомый документ-текст библиографическое описание или нет. Поэтому он прежде всего формулирует свой запрос (REQUEST) на естественном языке, то есть у человека, по-простому говоря, есть вопрос. Он может на равных основаниях дополнить REQUESTбиблиографическими полями как некоторое уточнение, в результате которого найденные документы и тексты будут отсортированы согласно приказа, извиняюсь, согласно тому насколько часто в найденных текстах содержаться слова из REQUEST. Если документ не имеет полного текста и является лишь библиографическим описанием книги, статьи и т.д., он будет подан наверх списка найденных. В результате поиска читатель увидит сначала библиографические описания без полных текстов, затем описания с текстами, к которым прикреплены ссылки на найденные страницы. Чем больше слов из REQUEST нашлось на странице и чем ближе они друг к другу, тем раньше эта страница будет показана в списке найденных страниц.

Место текста в списке найденных, то есть близость соответствия запросу определяет его ранг, который рассчитывается на основании числовых критериев. Ниже мелким шрифтом приводится краткое описание этих критериев.

*В системе применяется классический алгоритм поиска с ранжированием, который расширен за счет учета расстояний между словами. Вес каждого слова в запросе оценивается следующим образом: TF*IDF, где TF – частотность слова в тексте. В системе этот параметр равен единице (достаточно наличия хотя бы одного слова).*
IDF – коэффициент уменьшения веса слова в зависимости от его распространенности в словаре базы данных.
*IDF = log2(MaxMFN/df+1) / log2(MaxMFN+1), где MaxMFN – число текстов в базе данных и df – число текстов, содержащих данное слово. При расчете ранга текста RANGиспользуются веса слов запроса TF*IDF и расстояния между ними в тексте. Для каждой пары слов берется минимальное расстояние между ними.*
*RANG = ∑∑ij (TFIDFi*TFIDFj / MIN(Rij)2), где Rij – расстояние между парой слов. Если слова рядом, Rij = 1.*

При поиске можно указать дополнительный параметр – максимальное расстояние между словами. В этом случае необходимым условием выдачи является наличие в тексте фрагментов, включающих все найденные слова из запроса. Если результат поиска нулевой, делается попытка применить менее строгий критерий отбора за счет учета неполных фрагментов. При расчете ранга текста используется минимальное расстояние между словами. Постулируется принцип – лучше один "хороший" фрагмент, чем несколько "плохих". Максимальный размер фрагмента (для включения текста в выдачу) составляет величину:
*F = NumWords*MaxDistance, где NumWords – число слов в запросе, MaxDistance – максимальное расстояние между словами.*

Вернемся к началу и представим себе именно такого читателя, который "в поисках того, не зная чего" бредет "туда, не зная куда". Очевидно, ему мало будет полнотекстового критерия выдачи из-за невозможности сформулировать REQUEST (нет слов!), также и библиографические данные ему не помогут по той простой причине, о которой часто говорит наше банальное: "А судьи кто? Почему именно этому автору я должен довериться в данном вопросе? Из-за его авторитета? Как будто мало ниспровергнутых авторитетов в прошлом!" Но оставим риторические вопросы читателю и перейдем к нашей теме – фасеты.

"Далее он расставил всех присутствующих по ... кругу (строго как попало)".

Льюис Кэролл

Фасеты – это такая штука, которая помогает читателю оценить как Электронная Библиотека в целом реагирует на его REQUEST, как соотносится то, что он ищет с тем, что имеется в наличии. Фасеты представляют собой (здесь должно было бы быть единственное число – фасет!) соотношение части и целого, взятое в ракурсе содержания части. Фасеты надо рассматривать как некое извлеченное из массива информации частное содержание, которое дает бросить на целое взгляд из угла, посмотреть на общую картинку через цветное стеклышко.

Представим себе, что в результате некоего поиска читатель получил довольно большой объем информации – список найденных им документов включает сотни или даже тысячи наименований книг, журналов, статей и т.д. При этом дополнительной информации для уточнения поиска у читателя нет. Как ему поступить? Именно здесь ему помогут фасеты, которые структурируют найденный материал и могут дать ключ от двери в волшебную страну, где живут ответы на вопросы. Каким образом фасеты умудряются структурировать материал, с которым имеет дело читатель лично? Как они "узнают" вопрос?

Подготовим заранее, например, список ключевых слов (или авторов, или рубрик и т.д.), которые участвуют во всех библиографических описаниях Электронной Библиотеки. Каждое слово из этого списка встретится во всех текстах (включая библиографические описания) Электронной Библиотеки определенное число раз, благодаря чему можно расставить их (ключевые слова) в строгом порядке (возрастания), по сути "как попало", потому что сам по себе такой порядок ничего не говорит. Но теперь если взять из этого сортированного списка только те слова, которые содержаться в найденных читателем документах (не забудем о порядке их следования!), то получатся как раз они – фасеты.

Вместе они составят "фасет ключевых слов". Этот фасет будет включать слова, которые наиболее часто встречаются в найденных документах. Если в тысячах найденных документах по запросу (ну, просто как пример): "Что станется с библиотеками в будущем?" – содержится некий фасет ключевых слов, то он (если мы сумеем его правильно прочесть!) будет нам подсказкой в каком направлении надо искать ответ.

Теперь дальше – для того, чтобы ответить на именно этот вопрос ("Что станется с библиотеками в будущем?"), нужно занять другой уровень осмысления. Вопрос о будущем библиотек риторический с технической точки зрения. Мир развивается, появляются новые виды коммуникаций, увеличивается скорость доступа к информации и т.д. Нас же интересует как меняется человек внутренне вместе с этим развитием, какую роль для человека играет библиотека.

Для начала достаточно слова. Затем появляется строчка, как стрела, обращенная к миру, ко всем.
После ставится точка.
Кто помыслит про-ч-есть получает вопрос, возникает причина, дальше все обстоятельства, и понеслось – развернулась пружина!
В золоченном оплете с тугим корешком Книга Жизни готова, и потерянный странник бредет с посошком, ищет нужное слово...

Цитата предлагает считать, что не странник-читатель потерял слово и ищет его, а сам он в некотором смысле “потерялся”. Раз он потерялся, он вынужден искать, искать дорогу домой, и на этом пути ему полагается слово-подсказка – указатель на пути. Этот указатель и есть библиотека.

Библиотека хранит опыт многих поколений людей, которые шли по этому пути домой и оставили свои впечатления в виде книг и воспоминаний, многие из которых со временем превратились в легенды и мифы. Другие люди оставили в библиотеке книги о том, что никакого пути нет, никакого дома не существует, а все, что происходит – просто происходит само собой. Что ж, это тоже опыт, тоже в некотором смысле путь.

Как только я включаю в рассуждение человека самого человека, оно претендует на хранение во Всемирной Идеальной Библиотеке как его личный жизненный вклад в совокупный общечеловеческий итог. Есть люди, жизнь которых стала легендой, их идеи – это живые книги, и они только частично хранятся в библиотеке. Когда мы читаем тексты таких книг нам кажется (да так и есть на самом деле), что мы общаемся с автором, как с нашим личным знакомым. Есть книги, которые написаны людьми знакомыми с автором идеи только понаслышке. Более того, со временем некоторым людям кажется, что они записали именно свои собственные мысли – так авторство становится коллективным делом. В любом случае Библиотека принимает на хранение любой опыт. Кажется, что ценится именно опыт прохождения пути, а не оригинальные мнения на этот счет. Например, если я захочу узнать что-нибудь о всемирном тяготении, я прочту Ньютона, а если о любви, то ... тоже прочту Ньютона!

Таким образом, вопрос о будущем библиотеки превратился в вопрос о будущем человека. Всемирная Идеальная Библиотека была, есть и будет, но сумеем ли мы прочесть в ней что-нибудь?