

# Информационный поиск в коллекции разнородных документов

© М.В. Губин

ИК Кодекс  
max@gubin.spb.ru

## 1 Введение

Современные электронные библиотеки редко содержат полностью однородные коллекции документов. Чаще система содержит массивы документов, которые поступили из различных источников или создавались различным способом. В случае библиотеки электронных книг это могут быть книги в различном формате или написанные в разные периоды времени. Для информационно-правовой системы, это, например, нормативные документы, книги и учебники по праву, словари, энциклопедии и т.д. Для системы локального поиска - файлы, почтовые сообщения, сообщения интернет-пейджера. Для глобальной поисковой системы в Интернет - сайты, новостные ленты, сообщения конференций, блоги и т.д.

Современные исследования в области качества информационного поиска обычно используют для оценки однородные коллекции документов [9, 1]. При этом результаты, полученные при таких экспериментах, не обязательно могут быть повторены на разнородной коллекции. Можно предположить, что следующие факторы могут оказывать отрицательное влияние на качество информационного поиска по разнородной коллекции:

1. Разные размеры документов. Как известно, нормализация по длине документа является задачей, которая решается при любом алгоритме поиска, но вносимая при этом систематическая ошибка может быть не заметна в однородной коллекции, и оказаться критичной для разнородной коллекций, где статистические характеристики распределения длин документов сильно отличаются. То, что данные характеристики могут заметно влиять на качество информационного поиска подтверждается в статье [14].
2. Различия в оформлении документов. Некоторые коллекции могут иметь не все параметры форматирования, или эти параметры по-разному отражают содержание докумен-

та. Например, в случае нормативной коллекции, выделенный заголовок почти всегда информативен, а для web коллекции, многие страницы имеют заголовки не связанные с текстом.

3. Разная структура гипертекстового графа документов коллекции. В некоторых коллекциях документы могут быть сильно связаны между собой, в других нет. Например, коллекция нормативных документов РОМИП [1] имеет сильно связанный граф связей, а web-коллекция, которая взята как подмножество интернета, имеет не связанный граф, с большим числом связей, направленных вне коллекции.

В данной статье рассматривается классическая задача информационного поиска, когда пользователь вводит запрос в виде нескольких слов или фразы, и получает результат поиска в виде списка документов, отсортированных в порядке убывания оценки их релевантности запросу. В дальнейшем, все множество документов доступных системе для поиска мы будем называть **коллекцией**, а каждое подмножество, которое выделяется в ней по какому-либо признаку и рассматривается нами как однородное - **массивом** документов.

В данной статье рассматриваются два варианта поиска по разнородной коллекции:

1. **„Сквозной“ поиск.** При таком подходе все массивы документов, с точки зрения поискового механизма, объединяются в один массив, то есть информация о принадлежности документа к некоторому массиву не учитывается.
2. **„Раздельный“ поиск.** При этом подходе поиск по каждой из массивов проводится отдельно, а результат представляет собой объединение списков, полученных поиском по каждому из массивов с помощью специального алгоритма объединения.

При реализации обоих вариантов поиска использовался один и тот-же базовый алгоритм, который в первом случае применялся ко всей коллекции, а во втором к каждому массиву докумен-

тов отдельно. Основной задачей, которая решается при реализации „раздельного“ поиска, является такое объединение результатов поиска по массивам, при котором происходит увеличение качества.

## 2 Базовый алгоритм

Используемый нами базовый алгоритм информационного поиска является вариантом классического и широко распространенного алгоритма TFIDF [15]. В отличие от классического варианта, дополнительный вес получают слова, которые встречаются в заголовках документов. Использовался так-же учет гипертекстовых связей между документами с помощью варианта алгоритма Local PageRank [2].

Этот алгоритм показал хорошее качество поиска при использовании наших внутренних оценок и оценок по методике и с использованием коллекций РОМИП [4, 5]. Поэтому, данную реализацию можно считать удачным базовым вариантом, который подходит для каждой из используемых коллекций отдельно.

## 3 Алгоритм объединения результатов

Задача алгоритма объединения состоит в том, чтобы:

1. Не ухудшить точность поиска, т.е. разместить документы, имеющие максимальную вероятность релевантности в начале списка.
2. Полно предоставить массивы документов, то есть не ухудшить показатели полноты.

Область информационного поиска достаточно активно развивается более 50 лет. В ней накопилось множество подходов и идей, которые можно использовать для решения задачи разработки алгоритма объединения результатов поисков по разным массивам документов. В качестве областей, которые могут быть полезны в данном случае, нами были рассмотрены:

1. Мета-поиск;
2. Распределенный поиск;
3. Выбор фрагментов.

### 3.1 Мета-поиск

Системы мета-поиска (meta-search) достаточно популярны в настоящее время в Интернете [6]. В основном они используют результаты глобальных интернет поисковых машин для увеличения полноты и точности результата поиска, а так же для предоставления более удобного пользовательского интерфейса для работы с результатами поиска.

При использовании подходов мета-поисковых машин для решения задачи поиска по разнородной коллекции можно назвать следующие особенности:

1. Мы не ограничены в информации о каждом из поисковых алгоритмов и коллекций. Если мета-поисковой машине в Интернет, как правило, доступны только сортированные списки документов, то в нашем случае возможно получить любую информацию о внутренних статистических характеристиках коллекций, другим отличием является то, что используется один алгоритм поиска по каждому из массивов.
2. Отсутствие повторений документов в результатах поисков. Большой класс мета-поисковых алгоритмов использует факт появления одного и того-же документа для упорядочивания документов в результате, в данном случае прямое применение данного подхода не возможно.

В связи с этим, особый интерес представляют алгоритмы мета-поиска, которые используют данные о коллекциях. Например, в работе [16], где для объединения используются данные о различии в частотах терминов запросов в каждом из массивов и в объединенной коллекции.

### 3.2 Распределенный поиск

Исследуемую нами задачу можно рассматривать как один из вариантов задачи распределенного поиска (distributed search). При этом каждый массив документов можно рассматривать как отдельный узел, где осуществляется поиск. В работах, посвященных распределенному поиску [7, 10, 13] предлагается ряд алгоритмов (CORI, bGIOS, LM, KM, ReDDE), которые учитывают распределение частот терминов и размеры каждого из массивов.

### 3.3 Выбор фрагментов

Другой активно исследуемый подход, методы которого можно использовать, является выбор фрагментов (passage retrieval). Каждый массив документов можно рассматривать как один большой документ, а входящие в массив отдельные документы как его фрагменты. В известных алгоритмах реализации данного подхода [12, 17, 8, 11] используется частота терминов запроса в коллекции и в данном документе, количество фрагментов, содержащих слова запроса, а так же распределение длин документов и выбираемых фрагментов в коллекции.

## 4 „Раздельный“ поиск

Мы считаем, что используемый алгоритм поиска для каждого из массива возвращает список документов с хорошими параметрами релевантности. Из рассмотренных подходов, можно выделить следующие общие характеристики, используемые алгоритмами объединения результатов поиска - это частоты терминов и количество документов, которые отобраны из каждого из массивов. Мы рассматривали следующие характеристики массивов:

1. Статистика распределения терминов запроса в коллекции. Вводится следующая величина для каждого из массивов:

$$Fr_i = \frac{F_i}{F},$$

где

$F$  - общая частота терминов в коллекции;

$F_i$  - частота термина в массиве  $i$ .

2. Количество документов, которые система вернула по запросу из каждого массива. Вводится следующая величина для каждого из массивов:

$$Rr_i = \frac{Sr_i}{S_i},$$

где

$Sr_i$  - количество документов, отобранных как релевантные из данного массива;

$S_i$  - размер  $i$ -го массива в документах.

Однако, используемый базовый алгоритм имел специальную обработку ситуации, когда не находилось документов, содержащих все термины запроса. При этом из запроса отбрасывались некоторые термины. В связи с этим, при проведении экспериментов оказалось, что для коллекций не содержащих релевантные документы система иногда формировала излишне большой список документов, что приводило к „перевешиванию“ при использовании этого коэффициента, поэтому данный параметр использовался только тогда, когда не производилось отбрасывание терминов.

Кроме этого, для каждого массива был введен коэффициент  $Qu_i$  - указывающий насколько данный массив оценен как „качественный“. При „раздельном“ поиске использовался следующий алгоритм:

1. Производится поиск по каждому из массивов;
2. получается оценка веса, каждого из массивов, на основании величин  $Fr_i$ ,  $Rr_i$  и  $Qu_i$ . Использовалась следующая простейшая формула:

$$W_i = Qu_i Rr_i \sum Fr_j$$

3. формируется объединенный список. При этом используется следующий алгоритм: Формируется массив, который заполняется вычисленными на предыдущем этапе весами. Далее выбирается элемент массива с наибольшим значением и в результирующий список перемещается первый документ из результата поиска соответствующего массива документов. После этого, ко всем элементам массива весов, кроме большего, прибавляется вес соответствующего массива документов и опять повторяется цикл отбора, пока не будут изъяты все документы из всех списков.

## 5 Экспериментальные результаты

Для оценки качества информационного поиска использовались следующие коллекции:

1. **Коллекции РОМИП** [1]. Массивами документов для данных экспериментов использовались две коллекции *legal* и *web*, которые представляли два массива и смешанный набор запросов. Данные эксперименты должны проверить влияние на качество предлагаемых методов, в случае если коллекции содержат разнородные документы, взятые из различных источников.
2. **Коллекции нормативных документов „Кодекс“**. Использовались массивы документов и наборы запросов:

(а) Российского федерального законодательства - 70 000 документов,

(б) законодательства Санкт-Петербурга - 36 000 документов.

Запросы для этой серии экспериментов были взяты из протоколов работы пользователей ИПС "Кодекс". Данные эксперименты должны проверить влияние на качество поиска предлагаемых методов, в случае если массивы содержат почти однородные документы и слабо отличаются.

Для каждой коллекции и набора запросов выполнялся поиск с использованием „сквозного“ и „раздельного“ поиска с использованием описанных алгоритмов.

К сожалению, к моменту подготовки этой статьи обработанных результатов РОМИП еще не было, поэтому невозможно провести детальный анализ.

Для „Коллекции нормативных документов „Кодекс““ была выполнена оценка качества поиска по внутренней методике, описанной в статье [3]. Экспертами, которые хорошо знакомы с коллекцией и предметной областью, были сформированы тестовые данные. Каждому из них было поставлено задание сформировать „идеальный“ ответ поисковой системы на запрос. Далее из результата выдачи системы отбиралось фиксированное количество документов, имеющих наибольший вес. Тем самым моделировалась типичная ситуация, когда пользователь просматривает только несколько первых документов выдачи, что совпадает с данными исследований по эргономике. Всего было обработано 25 запросов, которые были разбиты на три группы:

1. 12 запросов общей направленности, которым релевантны документы как Российского, так и Санкт-Петербургского законодательства, например, „медицинское страхование“.
2. 7 запросов, для которых не должно быть релевантных документов в массиве Санкт-Петербургского законодательства, то есть вопросы регулируемые на федеральном уровне, например, „договор комиссии“.

Запросы	Сквозной	Раздельный
1 группа	35	33
2 группа	14	13
3 группа	4	14
Сумма	53	60

Таблица 1: Качество информационного поиска.

3. 6 запросов, которые должны содержать в основном документы Санкт-Петербургского законодательства, например, „бюджет Санкт-Петербурга“ и „транспортный налог“.

Результаты, полученные по данной методике, приведены в таблице 1.

Из результатов видно, что при „сквозном“ поиске базовый алгоритм недооценивает документы из массива Санкт-Петербургских документов, что связано со структурой гипертекстового графа. Как показал наш анализ, эти документы имеют много гипертекстовых связей, указывающих на документы Российского законодательства, а обратных связей в документах нет.

При „раздельном“ поиске наблюдалась несколько обратная ситуация, система чаще предлагала документы массива Санкт-Петербургского законодательства, что привело к некоторому ухудшению качества поиска для 1 и 2 группы запросов, но значительно улучшило для 3-й группы.

## 6 Выводы

Использование предлагаемого „раздельного“ алгоритма поиска для исследованных разнородных коллекций позволяет увеличить качество информационного поиска для некоторых запросов. Для остальных запросов качество поиска изменилось не значительно.

## Список литературы

- [1] Российский Семинар по Оценке Методов Информационного Поиска. <http://romip.narod.ru>.
- [2] Google buys new algorithm. <http://www.search-marketing.info/newsletter/articles/google-continued.htm>.
- [3] Губин М.В. Исследование качества информационного поиска с использованием пар слов. In *Труды RCDL-2003*, pages 186–191, 2003.
- [4] Губин М.В. Опыт участия ИС „Кодекс“ в РОМИП 2003. In *Труды РОМИП'2003*, pages 31–42, 2003.
- [5] Губин М.В. Участие ИПС „Кодекс“ в семинаре РОМИП 2004. In *Труды РОМИП'2004*, pages 28–39, 2004.
- [6] Finding information on the internet: A tutorial. <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html>, 2004.
- [7] J. Callan. *Distributed information retrieval*. Kluwer Academic Publishers, 2000.
- [8] Hang Cui, Ji-Rong Wen, and Tat-Seng Chua. Hierarchical indexing and flexible element retrieval for structured document. In *ECIR*, pages 73–87, 2003.
- [9] Donna Harman. What we have learned, and not learned, from trec. In *Proc. of the BCS IRSG'2000*, pages 2–20.
- [10] Panagiotis G. Ipeirotis and Luis Gravano. When one sample is not enough: improving text database selection using shrinkage. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 767–778, New York, NY, USA, 2004. ACM Press.
- [11] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.
- [12] Gerard Salton, J. Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58, New York, NY, USA, 1993. ACM Press.
- [13] L. Si and J. Callan. The effect of database size distribution on resource selection algorithms. In *SIGIR 2003 Workshop on Distributed Information Retrieval*, 2003.
- [14] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Research and Development in Information Retrieval*, pages 21–29, 1996.
- [15] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, 1979.
- [16] Zonghuan Wu, Weiyi Meng, Clement Yu, and Zhuogang Li. Towards a highly-scalable and effective metasearch engine. In *WWW '01: Proceedings of the tenth international conference on World Wide Web*, pages 386–395, New York, NY, USA, 2001. ACM Press.
- [17] Justin Zobel, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. Efficient retrieval of

partial documents. In *TREC-2: Proceedings of the second conference on Text retrieval conference*, pages 361–377, Elmsford, NY, USA, 1995. Pergamon Press, Inc.

## **Information Retrieval in a Heterogeneous Document Collection**

Maxim Gubin

Modern digital libraries usually contain documents from different sources. For example, public Internet search engine search through cites, conferences, BLOGs etc., local search tools look for files, e-mail messages etc. These documents have different formatting, size and other attributes. Modern studies in the area usually evaluate quality of information retrieval over standard uniform collections.

We studied two methods of search in heterogeneous collections: "search through"when the search engine unites all documents in one set and "differential search"when the search engine performs separated searches in every set and then join results. The article contains discussion of development of the algorithm of the results join.

We perform a number of experiments using two heterogeneous collections:

1. legal documents and web documents;
2. legal documents from two different sources - federal legislation and local Saint-Petersburg legislation.

The experiments show that the second method slightly increase search quality.