Маршрутизация запросов в системах распределенного поиска

Игорь Некрестьянов
igor@meta.math.spbu.ru
Санкт-Петербургский Государственный Университет

Аннотация

В статье рассматривается задача выбора подмножества коллекций распределенной поисковой системы, которые будут обрабатывать запрос, т. е. задача маршрутизации запросов в системах распределенного поиска. Особое внимание уделяется системам, в которых применяются тематические коллекции. Такие системы обладают рядом существенных особенностей по сравнению с общей задачей распределенного поиска, учет которых позволяет повысить качество маршрутизации.

На основе стандартного тестового набора данных TREC-6 [19] сравнивается эффективность нескольких методов маршрутизации запросов при работе как с тематическими, так и с нетематическими коллекциями.

1 ВВЕДЕНИЕ

Огромный объем доступной в Интернет информации делает поисковые системы незаменимым инструментом. Количество существующих поисковых систем исчисляется сотнями и большинство из них принадлежат к одному из двух классов:

Многоцелевые системы: Такие системы (Altavista, Google, Yandex) предназначены для поиска информации по любым запросам. Для выполнения этой задачи они пытаются проиндексировать всю доступную в Интернет информацию[1].

Специализированные системы: В отличие от многоцелевых систем, такие системы предназначены для ответов на запросы, относящиеся к некоторой специализированной области. Например, информацию о популярной музыке и музыкантах можно искать на `www.allmusic.com`, прогноз погоды — на `www.weather.com`, а работу в США — на `www.bestjobsusa.com`. Каталог из 400 подобных специализированных систем доступен по адресу `www.invisibleweb.com`.

Специализированная поисковая система производит поиск по значительно меньшему количеству ресурсов, чем любая популярная многоцелевая поисковая система. Однако, этот факт имеет ряд положительных следствий для специализированных систем [12]:

- Информация, не относящаяся к специализации данной поисковой системы, не попадает в ее индекс.
- Возможно применение более вычислительно трудоемких методов поиска.
- Возможно привлечение экспертов в соответствующей области, а также поддержка возможности рекомендации ресурсов пользователями системы, что повышает качество и полноту коллекции.

Поэтому, зачастую поиск в соответствующей запросу специализированной поисковой системе быстрее и лучше удовлетворяет информационные потребности пользователя.

В то же время, из-за специализации таких поисковых систем, выбор конкретной системы для выполнения поиска является довольно сложной задачей. Для решения этой проблемы `www.invisibleweb.com`, например, предлагает возможность поиска по построенным вручную описаниям специализированных систем. Такой подход очень трудоемок и не всегда работает из-за ограниченности построенных вручную описаний. Автоматическое построение таких описаний является предметом современных исследований [17].

Из-за огромного объема индексируемой информации многоцелевые поисковые системы подразумевают использование очень мощных аппаратных ресурсов, и зачастую не могут использовать многие современные разработки в области информационного поиска из-за их вычислительной трудоемкости.

[1] Мы не выделяем в отдельный класс системы, которые ограничивают область индексирования по таким критериям, как используемая кодировка или домен.

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

Тем не менее, даже самые мощные поисковые системы не могут проиндексировать всю доступную в Интернет информацию. Так, по состоянию на февраль 2000 года в Интернете было опубликовано более миллиарда страниц[2], а результаты оценки[3] размеров крупнейших поисковых систем (июнь 2000) показывают, что максимальный размер индекса составляет 350 (560) миллионов страниц[4].

В целях повышения производительности и надежности большинство современных многоцелевых систем имеют уже не централизованную, а параллельную архитектуру [1].

В последние годы также активно исследовалась возможность применения распределенных архитектур к поисковым системам [2, 20, 1, 3, 16, 5]. В распределенных поисковых системах единый индекс разбивается на несколько отдельных частей (коллекций) по некоторому принципу, причем внутри разных коллекций могут использоваться разные методы выполнения поиска. При создании распределенной поисковой системы необходимо решить ряд вопросов [6, 1]:

Как маршрутизировать запросы? Для понижения нагрузки на сеть и повышения эффективности операция поиска выполняется не во всех коллекциях, а только в некотором их подмножестве. Этот процесс называется *маршрутизацией запросов* (query routing) [13, 16, 18].

Как объединять ответы? Процесс объединения ответов от отдельных частей индекса в единый ответ системы называется слиянием результатов (database fusion) [16, 14].

Какие протоколы использовать? Выбор протокола обмена данными в рамках системы влияет на открытость и потенциальную функциональность системы. Примерами существующих протоколов являются STARTS [10], Z39.50 [11].

В этой работе мы рассматриваем задачу *маршрутизации запросов*, продолжая наши предыдущие исследования в этой области [13].

В предыдущей работе [13] мы изучали маршрутизацию запросов в специфическом подклассе систем распределенного поиска — системах, в которых все коллекции тематические [2], т. е. все входящие в коллекцию документы имеют общую тематику. Поскольку запросу пользователя также соответствует некоторая тематика[5], то релевантные этому запросу документы неравномерно распределены по существующим коллекциям. В силу этой особенности почти по любому запросу можно выбрать небольшое подмножество доступных коллекций таким образом, что они содержат почти все релевантные запросу документы.

Рассматриваемая проблема похожа на задачу выбора специализированных коллекций по запросу [17], однако мы предполагаем, что тематические коллекции сами ответственны за построение собственных описаний требуемого вида [13].

В отличие от стандартной постановки задачи маршрутизации запросов, в рамках нашей модели на выполнение поиска по запросу может быть потрачено некоторое количество ресурсов[6], и стоимость обработки запроса в коллекции зависит от числа возвращаемых коллекцией результатов. Таким образом, распределение ресурсов, выделенных на поиск по отобранным коллекциям, также должно учитываться алгоритмом маршрутизации запросов.

В рамках этой работы на основе стандартного тестового набора данных TREC-6 [19] сравниваются несколько различных методов маршрутизации запросов при работе как с тематическими, так и с нетематическими коллекциями.

2 МАРШРУТИЗАЦИЯ ЗАПРОСОВ

Классическая задача маршрутизации запросов (query routing или collection selection problem) — это задача выбора подмножества из множества Θ доступных информационных ресурсов, т. е. коллекций, в которые будет направлен рассматриваемый запрос [3, 20, 16, 13]. При этом желательно, с одной стороны, минимизировать размер выбранного множества, с другой — включить в него как можно больше коллекций, которые содержат релевантные документы. Уменьшение числа коллекций, которые реально обрабатывают запрос, влечет повышение производительности системы. А отсутствие коллекции в выбранном множестве означает, что документы из этой коллекции, включая релевантные запросу, наверняка не будут возвращены пользователю.

В отличие от классической постановки задачи мы также рассматриваем вопрос распределения ресурсов, выделенных на выполнение поиска, между отобранными коллекциями. В качестве такого ресурса может выступать, например, общее число документов, которое мы хотим получить от коллекций[7].

2.1 Критерии оценки качества

Один из простейших критериев оценки качества маршрутизации запросов является $NetSurfRank$, предложенный в работе [4]. Для каждого запроса $q \in Q$ учитыва-

[2] `http://www.searchenginewatch.com/reports/sizes.html`

[3] `http://www.searchengineshowdown.com`

[4] Максимальный заявленный объем индекса составляет 560 миллионов страниц, однако независимые оценки реальных размеров индексов не превышают 350 миллионов.

[5] Отметим, что это утверждение не подразумевает ни существования коллекции с такой же тематикой, ни того, что все релевантные запросу документы содержатся в какой-либо одной коллекции.

[6] Это могут быть как аппаратные ресурсы (вычислительные ресурсы или сетевой траффик), так, возможно, и денежные.

[7] Отметим, что общее число полученных от коллекций документов — это вообще-то не то же самое, что и число документов возвращаемых пользователю, поскольку в рамках типичной системы распределенного поиска обычно присутствует дополнительный этап поиска среди всех возвращенных из коллекций результатов (так называемый database fusion или result merging).

ется только факт $isGood(q, n)$ (0 или 1) присутствия хотя бы одной коллекции с релевантными запросу документами среди первых n выбранных коллекций. Собственно $NetSurfRank$ по отношению к множеству запросов Q вычисляется следующим образом

$$NetSurfRank(Q, n) = \sum_{q \in Q} \frac{isGood(q, n)}{|Q|}$$

Такой подход имеет много слабых мест. Например, он не учитывает сколько релевантных коллекций было обнаружено и не делает разницы между коллекциями с большим и малым количеством релевантных документов. Тем не менее он все еще зачастую используется из-за своей простоты [17].

Попытка адаптировать стандартные в области информационного поиска критерии *точности* (precision) и *полноты* (recall) к задаче маршрутизации запросов была сделана в рамках проекта Gloss [9].

Точность маршрутизации определяется как

$$Precision(q) = \frac{|Chosen(q) \cap Ideal(q)|}{|Chosen(q)|},$$

где $Chosen(q)$ — это множество выбранный для маршрутизации коллекций, а $Ideal(q)$ — множество коллекций, которые содержат релевантные запросу q документы.

Полнота маршрутизации определяется аналогично:

$$Recall(q) = \frac{|Chosen(q) \cap Ideal(q)|}{|Ideal(q)|}$$

Поскольку первоочередной задачей маршрутизации запросов является предотвращение обработки запросов коллекциями, не содержащими релевантных документов, то на наш взгляд важным критерием оценки качества маршрутизации также является "чувствительность" ($Sensitivity$), т. е. отношение числа выявленных неподходящих коллекций к общему числу неподходящих коллекций:

$$Sensitivity(q) = \frac{|\Theta \setminus (Ideal(q) \cup Chosen(q))|}{|\Theta \setminus Ideal(q)|}$$

Однако, подход Gloss [9] также имеет недостатки. Так, например, все коллекции в множестве $Ideal(q)$ считаются одинаково важными, что не позволяет различать коллекции с большим и малым числом релевантных документов. Отметим еще один серьезный недостаток — никак не учитывается тот факт, что разные коллекции могут содержать одни и те же документы.

В работе [15] были предложены два критерия оценки качества, лишенные вышеупомянутых недостатков. Первый, $DocRecall$, — это отношение общего числа релевантных документов в n выбранных коллекциях по отношению к общему числу релевантных документов во всех коллекциях:

$$DocRecall(q, n) = \frac{\left|\bigcup_{C \in Chosen(q,n)} Answer(C, q)\right|}{\left|\bigcup_{C \in \Theta} Answer(C, q)\right|}$$

Здесь $Answer(C, q)$ обозначает множество всех релевантных запросу q документов в коллекции C.

Второй критерий, предложенный в [15], — $AvrgDocs$ — это среднее число релевантных документов в n выбранных коллекциях коллекциях, т. е.:

$$AvrgDocs(q, n) = \sum_{C \in Chosen(q,n)} \frac{|Answer(C, q)|}{n}$$

3 МЕТОДЫ МАРШРУТИЗАЦИИ ЗАПРОСОВ

В рамках этой работы мы рассматриваем маршрутизацию запросов в двух классах систем распределенного поиска — в системах, в которых коллекции имеют тематическую ориентацию, и в системах, в которых коллекции не имеют специфической тематической ориентации. Хотя, формально, описываемые ниже методы могут быть применены к обоим классам систем, некоторые из них неявно используют тот факт, что коллекции тематические.

Типичный метод маршрутизации запросов состоит из:

- Метода построения описаний коллекций.
- Метода выбора подмножества коллекций для выполнения поиска на основе построенных описаний коллекций.
- Метода распределения ресурсов между отобранными коллекциями.

3.1 Построение описаний коллекций

Описание коллекции используется для того, чтобы предсказать насколько полезна эта коллекция при ответе на данный запрос. Мы придерживаемся статистического подхода и, в качестве меры "полезности" коллекции запросу, мы рассматриваем оценку количества релевантных запросу документов в этой коллекции.

Прямолинейный подход — хранить в описании информацию о том в каких документах в рамках каждой коллекции встречается каждый терм. Действительно, в таком случае можно давать относительно точные предсказания, но размер такого описания будет сравним с суммарным размером всех коллекций. Поэтому такой подход практически не применим из-за крайне низкой масштабируемости.

Другая крайность — использование тематического рубрикатора для описания тематики коллекций. В этом случае запросу также (автоматически) сопоставляется множество рубрик. Этот подход использовался в работе [5], где в качестве тематического рубрикатора использовался каталог "Library of Congress". Хотя такие описания компактны, применимость этого подхода тоже ограничена, поскольку как построение таких описаний, так и контроль за соответствием содержимого коллекций их описаниям требуют огромного количества ручной работы.

Мы полагаем, что описания коллекций должны строиться полностью автоматически и отражать фактическое содержимое коллекции. В этой статье мы рассматриваем описания, имеющие форму взвешенного списка термов. Такая форма описаний часто используется в исследованиях в области маршрутизации запросов [8, 9, 7, 20].

3.1.1 *Полное описание*

Один из возможных способов построения описания коллекции C — это включить в него следующую информацию:

- Все термы, встречающиеся хотя бы в одном документе коллекции (основы слов, за исключением стоп-слов)
- для каждого терма t – $df_C(t)$ — число документов в коллекции C, содержащих терм t,
- общее число документов в коллекции N_C.

Отметим, что "терм" здесь понимается как "атомарная часть документа" и может быть как словом, так и основой слова в зависимости от параметров системы.

Такое описание коллекции мы далее будем называть *полным* описанием коллекции. Подобные описания рассматривались в ряде работ, посвященных распределенному поиску [18, 20].

3.1.2 *Тематическое описание*

Тематическое описание коллекции [13] — это сокращенная версия *полного* описания. Сокращение происходит за счет удаления информации о термах, которые в этой коллекции встречаются реже, чем в среднем по всем коллекциям. Иными словами информация о терме t присутствует в *тематическом* описании коллекции C, если:

$$\frac{df_C(t)}{N_C} \geq \frac{\sum\limits_{C\in\Theta} df_C(t)}{\sum\limits_{C\in\Theta} N_C}$$

где C есть множество всех коллекций.

В отличие от *полного* описания, *тематическое* описание коллекции учитывает не только информацию о содержимом этой коллекции, но также использует некоторую информацию (в виде полных описаний этих коллекций) о содержимом других коллекций. Таким образом, построение *тематического* описания не может быть осуществленно совершенно независимо и подразумевает доступ к некоторой разделяемой информации о системе в целом (например, к *полным* описаниям всех коллекций или к статистике об использовании термов во всей системе).

3.1.3 *Сокращенное описание*

При практическом использовании систем распределенного поиска размер описаний играет важную роль, поскольку напрямую влияет на общую масштабируемость системы. В работе [3] было показано, что можно выкидывать значительную часть термов из описания (иногда до 80%) без существенного ухудшения качества.

Для сокращения размера описания используется следующий подход. Из сокращаемого описания удаляются те термы, частота появления которых в данной коллекции не превосходит некоторое пороговое значение α:

$$\frac{df_C(t)}{N_C} < \alpha$$

Этот способ применим, как к полным, так и к тематическим описаниям.

3.2 Выбор коллекций

В рамках нашей модели мы не только выбираем множество коллекций, в которые будет направлен запрос, но также и распределяем некоторые ресурсы между этими коллекциями. Окончательное определение множества коллекции для маршрутизации запроса происходит уже на этапе распределения ресурсов — те коллекции, которым был назначен положительный ресурс, и составляют это множество.

Поэтому на этапе выбора коллекций мы ограничиваемся вычислением оценок близости $r(q, C)$ каждой из коллекций $C \in \Theta$ к запросу q. Соотношения между оценками дает возможность сравнивать относительную близость коллекций к запросу.

В рамках нашей модели запрос q представляет собой множество взвешенных термов. Вес $w_q(t)$ терма t определяет его важность. Далее мы также используем понятие "подзапроса": q' — подзапрос запроса q, если

$$\forall\, t \in q' \Rightarrow t \in q \;\&\; w_{q'}(t) = w_q(t)$$

Мы рассматриваем два способа вычисления оценок $r(q, c)$:

1. В первом подходе используется модель, предполагающая равномерное и независимое распределения термов запроса по документам.

   Тогда оценка $\overline{df}_1(\hat{q}, C)$ числа документов в коллекции C содержащих все термы запроса $\hat{q}$ может быть следующим образом:

   $$\overline{df}_1(\hat{q}, C) = \frac{\prod_{t\in\hat{q}} df_C(t)}{N_C^{|\hat{q}|}} N_C$$

   Собственно оценка релевантности коллекции C запросу q определяется как:

   $$r_1(q, C) = \max_{q'} \;\; \overline{df}_1(q', C) \cdot (2N_C)^{|\hat{q}|-|q|}$$

   Второй множитель здесь используется для того, чтобы подчеркнуть относительную значимость более длинных подзапросов.

2. В рамках этого подхода предполагается наличие корреляции вхождения термов в документы. Точнее, предполагается что, если

   $$df_C(t_i) \leq df_C(t_j), \quad t_i, t_j \in q,$$

   то каждый документ, содержащий терм t_i, содержит также и терм t_j.

   В этом случае оценка числа документов содержащих все термы запроса $\hat{q}$ есть

   $$\overline{df}_2(\hat{q}, C) = \min_{t\in\hat{q}} df_C(t)$$

   А оценка релевантности $r_2(q, C)$ вычисляется аналогично $r_1(q, C)$:

   $$r_2(q, C) = \max_{q'} \;\; \overline{df}_2(q', C) \cdot (2N_C)^{|\hat{q}|-|q|}$$

Набор данных	Полное описание	Тематическое описание
Тематические коллекции	48906	34835
Тематические коллекции, стемминг	38368	28127
Нетематические коллекции	33259	24357
Нетематические коллекции, стемминг	25503	19054

Таблица 1: Средние размеры (в термах) построенных описаний

3.3 Распределение ресурсов

Распределение ресурсов завершающий шаг при определении множества коллекций для маршрутизации запроса.

Мы ограничимся рассмотрением ситуации, когда необходимо решить сколько документов запросить из каждой коллекции, чтобы общее число полученных документов не превысило заданного N.

Для распределения ресурсов можно использовать одну из следующих стратегий:

- Поровну среди всех коллекций с ненулевыми оценками релевантности
- Пропорционально вычисленным оценкам релевантности, т. е. квота для коллекции C будет

$$N \cdot \frac{r(q, C)}{\sum_{C' \in \Theta} r(q, C')}$$

- Гибридный вариант: половину поровну среди коллекций с ненулевыми оценками релевантности, а остальное пропорционально оценкам
- Пропорционально положению в упорядоченном по убыванию вычисленных оценок списке коллекций

В общем случае, ограничение на ресурсы зачастую не одно, что влечет необходимость использования методов решения многокритериальных задач оптимизации. К сожалению, такие методы относительно трудоемки, что сказывается на общей эффективности маршрутизации запросов.

4 ЭКСПЕРИМЕНТЫ

Основная цель экспериментов — сравнение эффективности различных методов маршрутизации запросов на основе стандартного тестового набора данных.

4.1 Тестовый набор данных

В качестве исходного набора данных для экспериментов использовался набор тестовых данных TREC-6 [19]. Наборы данных TREC используются в рамках одноименной конференции[8] для проведения сравнительной оценки эффективности различных подходов к решению одних и тех же задач информационного поиска.

[8] `http://trec.nist.gov`

4.1.1 Построение тематических коллекций

При создании этих тестовых наборов используется труд экспертов для получения "идеальных" оценок релевантности. Однако, наборы данных TREC достаточно велики[9] и эксперты не могут вручную оценить релевантность всех документов для каждого запроса. Поэтому используется следующая процедура: для каждого запроса полуавтоматическим способом отбирается ограниченное множество документов, которые в дальнейшем анализируются экспертами на релевантность.

Для отбора этого множества документов обычно используются эвристики, основанные на происхождении используемых документов. В TREC обычно используются документы из источников подобных архивам газет. Для таких документов, зачастую, доступна некоторая метаинформация (например, рубрика в которой документ был опубликован). Так, например, ответ на вопрос про политические взаимоотношения России и Франции стоит искать среди документов, опубликованных в разделе внешняя политика, но не в разделе спорт. Отметим, что такой подход не гарантирует, что эксперты обнаружат все релевантные документы.

К сожалению, набор TREC-6 в явном виде не содержит данных для экспериментов с маршрутизацией запросов. Поэтому мы формирования коллекций мы использовали эвристический подход, основанный на используемой в TREC процедуре получения идеальных оценок релевантности. Мы полагали, что все документы, для которых присутствуют любые явные оценки релевантности для данного запроса, прошли стадию полуавтоматического отбора и, следовательно, относительно близки тематически. Поэтому, мы использовали множество таких документов в качестве содержимого соответствующей тематической коллекции.

Для формирования коллекций мы взяли тот же набор из 50 запросов, который использовался в TREC для экспериментов с маршрутизацией документов[10]. Используя описанную выше эвристику, мы получили 49 коллекций[11]. Среднее число документов в коллекции составило 1095, изменяясь от 715 до 1472. Содержимое коллекций довольно сильно пересекалось — всего в коллекциях со-

[9] Каждый набор содержит около 600 Mb упакованной текстовой информации.

[10] В отличии от задачи маршрутизации запросов, задача маршрутизация документов — это разновидность задачи автоматической классификации потоков документов.

[11] Каждому из 50 запросов соответствовали документы из нескольких разных наборов данных TREC. Для одного из запросов ни один из релевантных документов не входил в состав используемого нами набора TREC-6.

Параметры эксперимента: тип оценки и описания	Усредненные показатели					
	NetSurfRank(1)	Precision	DocRecall(1)	DocRecall(5)	DocRecall(10)	AvrgDocs(5)
Тематические коллекции						
r_1, полное	0.646	0.496	0.347	0.532	0.597	7.78
r_2, полное	0.673	0.496	0.336	0.540	0.607	6.74
r_1, тематическое	0.66	0.532	0.351	0.537	0.603	7.05
r_2, тематическое	0.69	0.529	0.36	0.551	0.621	7.38
Тематические коллекции, стемминг						
r_1, полное	0.7	0.505	0.371	0.581	0.642	8.96
r_2, полное	0.746	0.505	0.36	0.597	0.649	7.416
r_1, тематическое	0.66	0.532	0.354	0.585	0.544	7.63
r_2, тематическое	0.76	0.531	0.363	0.612	0.661	7.52
Нетематические коллекции						
r_1, полное	0.46	0.425	0.062	0.192	0.327	1.36
r_2, полное	0.473	0.425	0.056	0.193	0.334	1.41
r_1, тематическое	0.446	0.404	0.042	0.182	0.297	1.24
r_2, тематическое	0.44	0.407	0.057	0.182	0.311	1.62
Нетематические коллекции, стемминг						
r_1, полное	0.473	0.43	0.06	0.217	0.366	1.46
r_2, полное	0.48	0.43	0.081	0.224	0.363	1.585
r_1, тематическое	0.433	0.497	0.043	0.205	0.331	1.57
r_2, тематическое	0.36	0.496	0.059	0.211	0.347	1.37

Таблица 2: Сравнительная эффективность вычисления оценок релевантности на основе набора запросов Q_1.

держалось 26365 уникальных документов, т. е. каждый документ в среднем встречался в двух коллекциях.

4.1.2 Построение нетематических коллекций

Поскольку нашей основной целью является сравнение эффективности маршрутизации в рамках тематических и нетематических наборов коллекций, то мы хотели, чтобы наборы тестовых коллекций отличались как можно меньше.

Поэтому при построении нетематического набора коллекций использовались те же самые 26365 документов, что использовались в тематических коллекциях. Более того, размеры коллекций в нетематическом наборе совпадали с размерами коллекций из тематического набора. Распределение же документов по нетематическим коллекциям происходило случайным образом.

4.1.3 Множества тестовых запросов

В наших экспериментах мы использовали два множества тестовых запросов. Эти запросы, как и информация о множествах релевантных им документов, являются частью тестовых наборов TREC.

Первое множество запросов, Q_1, состояло из 150 запросов. Каждому из запросов в рассматриваемом наборе из 26365 документов считалось релевантными от 1 до 49 документов. Среди этих запросов были как запросы, ответы на которые были сосредоточены в одной-двух коллекциях, так и запросы, для которых нельзя было выделить небольшое подмножество доминирующих коллекций.

Второе множество, Q_2, представляло собой подмножество Q_1 и содержало только 45 запросов. Однако, для каждого запроса из этого множества существовала единственная коллекция, содержащая большинство релевантных запросу документов.

4.2 Экспериментальные результаты

Целью первой серии экспериментов являлась общая сравнительная оценка эффективности различных комбинаций типа описания и метода вычисления оценки релевантности. Для этой цели мы провели измерения как на наборе тематических, так и на наборе нетематических коллекций, используя множество запросов Q_1 . Варьируемым параметром экспериментов было использование стемминга при построении описаний. Полученные результаты (таблица 2) позволяют сделать следующие выводы:

- Некоторое преимущество имеет подход, использующий корреляционную модель распределения термов запроса по документам.
- Как и ожидалось, в случае нетематических коллекций, тематические описания менее эффективны, чем полные. Использование же тематических описаний в случае тематических коллекций оказывается более эффективным, хотя и не так значительно. Мы полагаем, что незначительность превышения объясняется тем фактом, что состав коллекций значительно пересекался, и это затруднило выделение особенностей словарного запаса коллекций.
- Использование стемминга влечет улучшение качества маршрутизации.
- Применение тематических описаний позволяет значительно (до 40%, смотри таблицу 1) сократить размер описания практически без потерь в качестве маршрутизации.

Одна из возможных причин лишь незначительного превосходства тематических описаний в случае тематических коллекций — отсутствие тематической специализации запросов, т. е. тот факт, что релевантные запросу документы сильно распределены по большому множеству коллекций. Для проверки этой гипотезы мы оценили качество маршрутизации, используя множество запросов Q_2. Как видно из результатов экспериментов в таблице 3 превосходство тематических описаний стало несколько более заметным.

Отметим, что, по сравнению с результатами в таблице 2, эффективность маршрутизации повысилась во всех случаях. Этот факт подтверждает естественную гипотезу, что легче маршрутизировать запросы, которым соответствует небольшое количество доминирующих коллекций.

4.2.1 Эффективность сокращенных описаний

Сокращение размеров описаний уменьшает объем разделяемой информации в системе и, как следствие, повышает ее масштабируемость.

Мы изучали влияние сокращения описаний на эффективность маршрутизации запросов. Ранее было показано [13, 3], что сокращение размера описаний методом, описанным в разделе 3.1.3, позволяет сократить размер описания до 80% без потерь в качестве маршрутизации.

Однако, наши эксперименты показали, что это не совсем так. Действительно, сокращение описаний почти не влияет на точность маршрутизации, но относительный порядок на множестве выбранных коллекций может значительно измениться. Так, например, при использовании полных описаний для тематических коллекций (рис. 1) сокращение описаний на 20% влечет падение *DocRecall*(5) на на 25%, хотя *Precision* практически не изменился.

В то же время, использование тематических описаний вместо полных дает 40% сокращение объема описаний без значительного ухудшения[12] качества поиска.

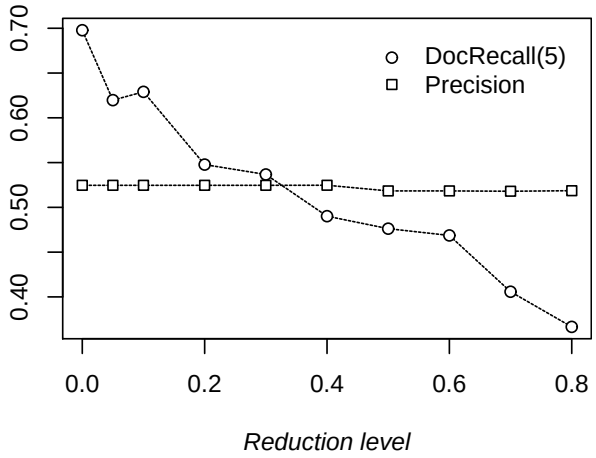

Рис. 1: Влияние процента сокращения описаний на качество маршрутизации

5 ЗАКЛЮЧЕНИЕ

В рамках этой работы мы рассматриваем проблему маршрутизации запросов в системах распределенного поиска, продолжая наши предыдущие исследования в этой области [13].

Сравнительный анализ эффективности маршрутизации, при использовании нескольких различных методов построения описаний и вычисления оценок релевантности, производился на основе стандартного тестового набора данных TREC-6. Результаты проведенных экспериментов позволяют сделать следующие основные выводы:

- Некоторое преимущество в качестве маршрутизации имеет подход, основанный на корреляционной модели распределения термов запроса по документам.

- Использование тематических описаний в случае тематических коллекций оказывается более эффективным, хотя и не так значительно. Мы полагаем, что незначительность превышения объясняется тем фактом, что состав коллекций значительно пересекался, и это затруднило выделение особенностей словарного запаса коллекций

- В случае нетематических коллекций, тематические описания несколько менее эффективны, чем полные. Однако, поскольку размер тематических описаний значительно (до 40%) меньше, чем размер полных описаний, снижение эффективности не велико, то применение тематических описаний может быть оправданным в целях повышения производительности системы.

- Использование стемминга влечет улучшение качества маршрутизации вне зависимости от типа описания и метода вычисления оценок релевантности.

- Применение прямолинейных процедур сокращения размеров описаний может значительно снизить на качество маршрутизации. Для сокращения размеров полных описаний предпочтительным решением является переход к использованию тематических описаний.

[12] В случае тематических коллекций использование тематических описаний может даже улучшить качество маршрутизации.

Список литературы

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.

[2] Mikhail Bessonov, Udo Heuser, and Igor Nekrestyanov. Open architecture for distributed search systems. In *Proc. of Sixth International Conference on Intelegence in Services and Networks (IS&N'99)*, Barcelona, Spain, April 1999.

[3] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proceedings of the SIGIR'95*, 1995.

Параметры эксперимента: тип оценки и описания	Усредненные показатели					
	NetSurfRank(1)	Precision	DocRecall(1)	DocRecall(5)	DocRecall(10)	AvrgDocs(5)
Тематические коллекции						
r_1, полное	0.672	0.503	0.362	0.547	0.607	18.47
r_2, полное	0.691	0.503	0.367	0.564	0.613	18.53
r_1, тематическое	0.692	0.590	0.383	0.611	0.657	16.24
r_2, тематическое	0.72	0.610	0.395	0.624	0.692	19.11
Тематические коллекции, стемминг						
r_1, полное	0.7	0.511	0.371	0.556	0.623	19.35
r_2, полное	0.746	0.516	0.373	0.571	0.625	19.94
r_1, тематическое	0.777	0.604	0.385	0.627	0.684	19.05
r_2, тематическое	0.812	0.615	0.461	0.631	0.701	21.47

Таблица 3: Сравнительная эффективность вычисления оценок релевантности на основе набора запросов Q_2.

[4] A. S. Chakravarthy and K. B. Haase. NetSerf: Using semantic knowledge to find internet information archives. In *Proc. of the SIGIR'95*, pages 4–11, 1995.

[5] Peter B. Danzig, Jongsuk Ahn, John Noll, and Katia Obraczka. Distributed indexing: A scalable mechanism for distributed information retrieval. In *Proc. of the SIGIR'91*, 1991.

[6] D. Drelinger and A. E. Howe. Expiriences wth selecting search engines using MetaSearch. *ACM Transactions on Information Systems*, 15(3):195–222, 1997.

[7] L. Gravano and H. Garcia-Molina. Generalizing gloss to vector-space databases and broker hierarchies. In *Proc. of the VLDB'95*, 1995.

[8] L. Gravano, H. Garcia-Molina, and A. Tomasic. The effectiveness of gloss for the text-database discovery problem. In *Proc. of the ACM SIGMOD'94*, 1994.

[9] L. Gravano, H. Garcia-Molina, and A. Tomasic. Precision and recall of gloss estimators for database discovery. In *Proc. of the 3rd International Conference on Parallel and Distributed Information Systems (PDIS'94)*, 1994.

[10] Gravano, Luis and Chang, Chen-Chuan K. and Garcia-Molina, Hector and Paepcke, Andreas. STARTS: Stanford Proposal for Internet Meta-Searching. In *Proceedings of the International Conference on Management of Data*, 1997.

[11] B. Kahle, H. Morris, J. Goldman, T. Erickson, and J. Curran. Interfaces for distributed systems of information servers. *Journal of the American Society for Information Science*, 44(8):453–485, 1993.

[12] David King. Specialized search engines: Alternatives to the big guys. 24(3), May 2000.

[13] Igor Kuralenok, Vladimir Dobrynin, Igor Nekrestyanov, Mikhail Bessonov, and Ahmed Patel. Distributed search in topic-oriented document collections. In *Proc. of World Multiconference on Systemics, Cybernetics and Informatics (SCI'99)*, volume 4, pages 377–383, August 1999.

[14] Anna Le Calve and Jacques Savoy. Database merging strategy based on logistic regression. *Information Processing and Management*, 36(3):341–359, May 2000.

[15] Z. Lu, J. Callan, and W. Croft. Measures in collection ranking evaluation. Technical Report TR96-39, University of Massachusetts, 1996.

[16] Jacques Savoy and Justin Picard. Report on the TREC-8 Experiment: Searching on the Web and in Distributed Collections. In *Proc. of the TREC'8*, 1999.

[17] Atsushi Sugiura and Oren Etzioni. Query routing for web search engines: Architecture and experiments. In *Proc. of the WWW-9*, May 2000.

[18] A. Tomasic, L. Gravano, C. Lue, P. Schwarz, and L. Haas. Data structures for efficient broker implementation. *ACM Transactions on Information Systems*, 15(2), April 1997.

[19] Elen Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC-6)*, 1997.

[20] Jinxi Xu and Jamie Callan. Effective retrieval with distributed collections. In *Proc. of the SIGIR'98*, 1998.