Исследование качества информационного поиска с использованием пар слов.

© Максим Губин

ИК Кодекс
max@gubin.spb.ru

1 Введение

Растущий размер коллекций документов делает все более важной задачей увеличения качества информационного поиска. Одним из методов, который потенциально может улучшить качества поиска является учет словосочетаний. Словосочетания являются устойчивыми лингвистическими объектами, которые имеют определенную семантику. Этим они очень похожи на отдельные термины. Поэтому один из способов их использовать, это обрабатывать при поиске подобным образом. То есть, кроме терминов, обработка осуществляется так же по словосочетаниям, которые рассматриваются как дополнительные термины запроса.

Еще одним аргументом для использования такого подхода является тот факт, что из изучения статистики пользовательских запросов известно, что большинство из них очень короткие и как раз содержат словосочетания [1].

Выделение словосочетаний является отдельной достаточно сложной проблемой, как с точки зрения лингвистики, так и компьютерной обработки естественного текста. Поэтому, практически применяют упрощенные методы [2], [3]. Простейшим случаем словосочетания является сочетание двух слов расположенных в непосредственной близости в тексте. Использование при информационном поиске таких контактных пар слов называется поиск по парам (word pair). При использовании такого подхода неявно исходят из гипотезы, что если есть устойчивое словосочетание, то оно обязательно в релевантном документе встретится в виде слов стоящих на небольшом расстоянии. Подобные упрощения достаточно распространены в области информационного поиска. Например, исторически первая и достаточно распространенная векторная модель исходит из предположения статистической независимости терминов, при этом дает не плохие результаты.

Одна из первых работ, посвященных подобному поиску в работе [4]. Однако в ней излагается только идея, авторам не удалось реализовать работающую систему и оценить качество поиска. Другая исследовательская группа получила противоречивый результат, получив улучшение на одной коллекции и ухудшение на другой [5]. Кроме этого, данный подход упоминается в работе [6], но анализа его влияния на качество не производится.

Наличие индекса по словосочетаниям позволяет увеличить скорость обработки запроса, т.к. не требуется пересекать индексы по положениям отдельных слов. Другим преимуществом данного подхода является то, что наличие информации об имеющихся в коллекции парах слов позволяет сформировать для пользователя подсказку по вводу следующего слова. В литературе имеется достаточно много статей, посвященных построению индекса для поиска по рядом расположенным терминам [7], [8], [9], [10]. Исследования подобных индексных структур проводил и автор данной статьи [11] и полученные результаты позволяли создать достаточно эффективную реализацию поиска по контактным парам.

Простота, возможность эффективной реализации делает этот метод очень привлекательным. Однако отсутствие однозначных результатов о том, приводит ли данный метод к улучшение качества поиска, явилось причиной выполнения данной работы.

2 Описание методики использование пар.

При данном подходе используется один из многочисленных алгоритмов информационного поиска, рассматривающий документ как множество терминов (bag of words). Но к поиску по отдельным терминам добавлен поиск по парам стоящих рядом в текстах и запросах терминов. Парами считаются любые два слова, которые отстоят друг от друга не далее чем через заданное расстояние, при этом шумовые слова не учитываются. Ограничения на размер индекса приводят к тому, что при использовании специальных поисковых структур расстояние между словами не превышает двух. Кроме этого, не учитываются пары, разделенные границами абзаца, так как они однозначно не являются словосочетаниями. Слова перед помещением в пару нормализуются по морфологии. Для пары учитываются те же веса и флаги, что для слова.

В процессе выполнения экспериментов, оказыва-

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

лось, что качество поиска ухудшается, поэтому алгоритм поиска несколько раз изменялся, так как делались попытки все-таки добиться устойчивого увеличения качества.
Первоначально предполагалось, что при поиске будут учитываться все пары слов, которые можно выделить из запроса, однако первые же эксперименты показали, что в этом случае качество поиска очень сильно ухудшается, вплоть до отсутствия релевантных документов в результатах большинства запросов, поэтому был добавлен алгоритм выделения "настоящих"словосочетаний. Данный алгоритм основывается на идее, что если данные слова являются устойчивым словосочетанием, то в большом количестве документов, которые содержат оба слова, должно встречаться и данная пара. То есть, если это устойчивое словосочетание, то вероятность того, что документ содержащий оба этих слова содержит и пару выше, чем в случае, если совместное появление данных слов случайно. Использовался следующий алгоритм:

1. Для каждой пары получают список документов, которые содержат эту пару S_{pair}. Для каждого входящего в пару слова также формируются списки вхождения S_{w1} и S_{w2}.

2. Формируется пересечение списков документов содержащих пару и списков для слов - $S_c = S_{w1} \bigcap S_{w2} \bigcap S_{w2}$.

3. Если Размер $|S_{pair}| * K > |S_c|$ то данная пара считается словосочетанием и оставляется, иначе отбрасывается. Параметр K определяется экспериментально.

Подобный алгоритм заметно улучшил качество поиска. Кроме того, заметно возросла скорость обработки запросов с большим количеством слов, т.к. уменьшилось количество обрабатываемых пар. Кратко алгоритм поиска системы выглядел следующим образом:

1. Из запроса выделяются все возможные пары слов. То есть при обработке запроса не учитывалось расстояние между словами, а формировался полный набор сочетаний по два слова из слов запроса.

2. Из сформированного массива пар с помощью описанного выше алгоритма отбираются "настоящие"словосочетания, которые добавляются к массиву слов запроса как отдельные термины.

3. Осуществляется поиск и взвешивание документов. Алгоритм поиска является вариантом классического TF*IDF [12]. Пары слов рассматривались точно так же, как и отдельные термины.

4. Результаты поиска выводились в виде списка документов, упорядоченных по убыванию веса.

Таким образом, изменения алгоритма поиска по сравнению с не использующим пары слов минимальны - только добавлен этап выделения пар слов и индекс по парам.
Однако, даже с помощью отбора „настоящих“ словосочетаний не удалось добиться улучшения качества поиска по сравнению со случаем не использования пар, поэтому было сделано предположение, что контактное появление является слишком сильным ограничением. Поэтому программное обеспечение было переработано таким образом, чтобы можно было исследовать влияние расстояния между словами в паре. Т.к. при расстояниях больше 2 количество пар при индексировании становится слишком большим, для данных экспериментов использовалось сканирование по текстам. Сначала с использованием индекса по отдельным словам отбирались документы, содержащие слова запроса, далее тексты этих документов сканировались и в них искались пары. В остальном алгоритм совпадал с описанным алгоритмом поиска по парам.

3 Описание методики оценки качества поиска

Целью проводимой работы было улучшение качества работы существующей коммерческой версии системы. Поэтому разрабатывалась методика, основанная на модели реального использования системой, соответственно ее результаты должны быть наиболее близки к требованиям пользователя. С другой стороны, методика должна быть достаточно проста, и требовать приемлемых затрат. Классическими критериями качества информационного поиска являются полнота (recall) и точность (precision) [13]. Однако применение данных характеристик для оценки качества на практике затруднительно по следующим причинам:

1. Для их оценки необходимо для каждого запроса получить список всех релевантных документов коллекции. Однако для реальной коллекции построение такого списка невозможно, либо требует слишком больших затрат на анализ.

2. Данные характеристики не обязательно совпадают с оценкой системы пользователем. Наблюдения за пользователями, работающими с системой, показали, что реально для пользователя важны только несколько первых документов, которые выдаются системой. Такое поведение достаточно типично при использовании поисковых систем.

Другим важным критерием качества работы системы является порядок в котором выдаются документы в результате. Практически все современные системы информационного поиска при выдаче результатов производят взвешивание полученных документов по мере релевантности документа запросу. Данная функция, ранее вспомогательная, с ростом размеров коллекций становится все

более важной. Данные эргономических исследований показывают, что пользователи, как правило, просматривают только несколько документов, имеющих наибольший вес. Однако, это не означает, что пользователю можно выдавать только первые несколько документов. Большой список, даже просмотренный поверхностно дает пользователю информацию:

1. О полноте базы данных (количестве документов) поисковой системы

2. Насколько широко представлены документы, относящиеся к запросу, в базе данных.

Современные системы, как правило, выдают максимально большие списки документов, применяя формальное упрощенные понимание релевантности. Например, популярный Internet поисковик Google (http://www.google.com) формирует список всех документов, которые содержат все слова запроса. Поэтому при выборе критерия оценки качества необходимо учитывать порядок выдачи документов, что не учитывается при классических критериях.
На основании этих соображении было принято решение использовать следующую методику оценки качества. 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 данный 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.
Первоначально планировалось применить достаточно сложную методику, аналогичную описанной в [14]. Однако данная методика имеет много эмпирических коэффициентов, значения которых очень сложно определить и обосновать. Поэтому был использован более простой вариант оценки. В качестве оценки эффективности использовалась величина, вычисляемая по следующей формуле:

$$Q = \sum \left| R_{sys} \bigcap R_{user} \right|$$

, где $\sum$ - сумма по всем запросам R_{sys}- множество документов в выдаче системы, находящихся на первых позициях, помещающихся на первом экране выдачи. R_{user} - множество документов, указанных пользователем как "идеальная"выдача системы. Легко доказать, что выбранная оценка является осмысленной с точки зрения критерия полноты/точность, если сделать допущение, что сформированная экспертами выборка документов для данного запроса содержит все релевантные. Количество отобранных экспертами документов не зависит от метода поиска, поэтому оно постоянно для всех экспериментов, $|R_{user}| = const$. В качестве результирующей выборки отбиралось некоторое постоянное количество первых документов из выдачи системы. При этом количество документов в результате у нас аналогично постоянно, то есть $|R_{sys}| = const$. Используя классическую формулу для полноты, и подставляя в нее оценку для одного эксперимента, получаем:

$$RECALL = \frac{|R_{sys} \bigcap R_{user}|}{|R_{user}|} = \frac{Q}{|R_{user}|} = \frac{Q}{Const}$$

Аналогично для точности:

$$PRECISION = \frac{|R_{sys} \bigcap R_{user}|}{|R_{sys}|} = \frac{Q}{Const}$$

Таким образом, используемая оценка в ограничениях используемой методики исследования и при допущении "идеальности"работы эксперта прямо пропорциональна точности и полноте. Можно показать, что подобную зависимость можно вывести и для средней точности и полноты серии экспериментов. Увеличение данной оценки означает, что данные изменения алгоритма поиска приводят к улучшению как полноты, так и точности поиска.

4 Экспериментальные результаты

В качестве коллекции документов была взята база действующих документов Российского законодательства. Она содержала около 60 тысяч документов, со средним размером документа 4 Кб. В качестве поисковой системы использовалась информационная система "Кодекс". Для запросов, использованных при экспериментах, из протоколов работы системы были взяты запросы реальных пользователей. Были отобраны запросы, удовлетворяющие следующим критериям:

1. Содержат более одного слова без орфографических ошибок. Запросы из одного слова не интересовали, т.к. не содержат словосочетаний.

2. Не содержат номеров документов и дат. Т.к. эти атрибуты обрабатываются специальным образом.

Из этих запросов случайно выбрали 30 запросов. По каждому из запросов экспертами были отобраны от 4 до 6 наиболее релевантных документов.
Первые же эксперименты показали, что качество поиска при использовании пар сильно ухудшилось, вплоть до того, что в результате не получалось ни одного релевантного документа. Было очевидно, что многие сформированные из запроса пары не являются словосочетаниями, и поэтому отбрасывается много релевантных документов. Поэтому надо было каким-то образом отобрать пары, являющиеся настоящими словосочетаниями. Для этого был использован алгоритм, который описан в разделе 2. При

Запросы	Качество без пар	Качество с парами
Все запросы	27	26
Двухсловные запросы	10	10

Таблица 1: Качество поиска при использовании контактных пар

Расстояние	2	15	20	30
Качество	26	44	42	38

Таблица 2: Качество поиска от расстояния между словами

выборе K равным 1.5 отбираемые пары слов стали осмысленными словосочетаниями. При K меньшем, в районе 1, словосочетания не выделялись, т.к. таких пар практически не было. При значении K больше 2, отбирались практически все варианты, т.е. алгоритм не обеспечивал нужной избирательности.

После этой доработки были произведены эксперименты и подсчитана оценка качества для случая поиска с учетом пар и без учета. Результаты приведены в таблице 1. Видно, что использование пар не только не привело к улучшению качества, но даже ухудшило его. Двухсловные запросы выделены отдельно, т.к. ожидалось, что для них изменение качества будет наиболее заметно, но и это не подтвердилось. Для прояснения сложившейся ситуации был проведен анализ запросов, по которым при учете двухсловных терминов эффективность понизилась. Для примера рассмотрим запрос "хранение оружия". По этому запросу найдено 202 документа, то есть в базе содержалось 202 документа содержащих контактное вхождение этих слов, при этом слово "оружие"содержалось в 1807 документах, а "хранение"в 6609 документах. Экспертами для рассматриваемого запроса выделено 5 релевантных документов. Среди первых 5 найденных документов только 1 релевантный. При поиске без учета словосочетаний среди первых 5 найденных было 2 релевантных документа. При учете словосочетаний в результат не попал приказ МВД России от 17.11.1999 N 938. Оказалось, что в данном документе нет ни одного появления пары "хранение оружия"с расстоянием между словами не больше 1. Поэтому документа нет не только среди первых 5, но и среди всех 202 отобранных. Между тем, в документе очень много разреженных появлений термина "хранение оружия"в виде фраз типа: "Правила хранения и ношения боевого ручного стрелкового оружия".

Еще один пример. По запросу "положение о разграничении Полномочий"найдено 255 документов, из запроса выделена пара "разграничении полномочий", содержащийся в 372 документах. Среди первых 6 документов содержится только 1 релевантный (всего релевантных документов 6). При поиске без учета терминов среди первых 6 содержится 3 релевантных документа. Таким образом, имеем 2 не попавших в результат релевантных документа. Первый из них: Конституция РФ от 12.12.93. Здесь нет ни одного появления термина "разграничении полномочий"с расстоянием между словами не больше 1, однако много появлений фразы: разграничение предметов ведения и полномочий. Второй документ, Указ Президента РФ от 12.03.96 N 370, выдан 63-м по порядку. Здесь имеется 2 появления термина "разграничении полномочий"с расстоянием между словами не больше 1 и очень много появлений фразы типа: "разграничение предметов ведения и полномочий".

Для исследования влияния расстояния между словами использовалась та же самая коллекция и такой же набор запросов. Проводился поиск для каждого запроса, при этом эксперименты отличались заданным расстоянием между словами в паре. Так как при этих экспериментах система производила сканирование текста документов, то прогоны занимали относительно много времени. Поэтому были проведены эксперименты только с некоторыми значениями расстояния. Результаты экспериментов приведены в таблице 2. Видно, что с ростом расстояния качество заметно улучшается. Однако, когда расстояние превышает 15 слов качество начинает ухудшаться, т.к. снижается избирательность. Таким образом, очевидно, что причиной ухудшения качества поиска с использованием пар является то, что ограничение на контактные появления слов является слишком сильным, необходимо рассматривать пары с расстоянием между словами около 10-15 слов.

5 Выводы

Использование контактных пар с небольшим расстоянием между словами для информационного поиска приводит к ухудшению качества информационного поиска. Причиной этого является то, что они накладывают слишком сильное ограничение, которое приводит к уменьшению количества релевантных документов в выдаче, тем самым снижается полнота и качество поиска. Экспериментальный данные показывают, что для того, чтобы не происходило ухудшение полноты, надо увеличить расстояние между словами при формировании пар. Устойчивое увеличение качества информационного поиска получается только при использовании достаточно большого расстояния между словами - 10-15. При таком расстоя-

нии количество пар слов становится неприемлемо большим для построения индекса, содержащего все пары.
Невозможность использования специального индекса делает поиск по парам намного менее привлекательным, т.к. при сканировании или хранении в индексе полной координатной информации можно использовать значительно более сложные алгоритмы выделения и обработки словосочетаний.

6 Перспективы

Значительное улучшение качества поиска при использовании пар слов с относительно большим расстоянием, говорит о том, что учет информации о взаимном положении слов в документах и запросе позволяет заметно улучшить поиск. Однако описываемый подход (пары слов) достаточно примитивен, возможно более сложный учет взаимного положения слов даст еще более значительный прирост качества, что требует дальнейшего изучения.

7 Благодарности

Автор выражает благодарность Харину Николаю Петровичу за большой объем проделанной работы и ценные советы и предложения.

Список литературы

[1] Gray. Запросы Рунета, 2003.

[2] Gael Dias, Sylvie Guillore, Jean-Claude Bassano, and Jose Gabriel Pereira Lopes. Combining linguistics with statistics for multiword term extraction: A fruitful association?

[3] Toru Takaki. Ntt data: Overview of system approach at trec-8 ad-hoc and question answering.

[4] James A. Danowski. Wordij: A word-pair approach to information retrieval.

[5] Santiago Garcia Ernest P.Chan and Salim Roukos. Trec-5 ad hoc retrieval using k nearest-neighbors re-scoring. April 1997.

[6] K. Sparck Jones, S. Walker, and S.E. Robertson. A probalistic model of information retrieval: development and status. August 1998.

[7] D. Bahle, H. E. Williams, and J. Zobel. Compaction techniques for nextword indexes. In *Proceedings of the SPIRE Conference on String Processing and Information Retrieval*, pages 33–45, San Rafael, Chile, 2001.

[8] D. Bahle, H. E. Williams, and J. Zobel. Efficient phrase querying with an auxiliary index. In K. Jarvelin, M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 215–221, Tampere, Finland, August 2002.

[9] D. Bahle, H.E. Williams, and J. Zobel. Compaction techniques for nextword indexes, November 2001.

[10] Hugh E. Williams, Justin Zobel, and Phil Anderson. What's next? - index structures for efficient phrase querying.

[11] Максим Губин. Изучение статистики встречаемости терминов и пар терминов в текстах для выбора методов сжатия инвертированного файла., 2002.

[12] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, 1979.

[13] Некрестьянов И.С. Кураленок И.Е. Оценка систем текстового поиска.

[14] И.Ашманов Н.Харин. Методика Н. Харина-И.Ашманова для оценки релевантности, 2000.

Studing Search Quality of A Word-pair Approach to Informational Retrieval

Maxim Gubin

Improving of a search quality is an important task of modern Information Retrieval. One possible approach is to use information about words collocations. The simplest variant of words cooccurence is word pairs. With this approach a system searches not only using bag of words document model, but including specific term pairs as if they were separate terms. The method has a number of advantages. First of all it is very easy to implement - modifications of a search engine are minimal. Second, it is possible to create an auxiliary next-word index when adjacency term pairs are used.
The paper is studying search quality of the method. Real user queries was used and a collection of documents of Russian legislation. A special evaluation method based on a real user behavior has been proposed.
The experiments show that the word-pair approach using contact collocation decrease search quality. It has been experimentally proved that usage of adjacency term pairs puts too strict restriction and therefore reduces recall. Tests show quality improvement only if it is used word pairs with distance 10-15 words between terms. But with this distance it is impossible to use a special index structure which makes the word-pair approach less attractive.