Расширение запросов с помощью вероятностного латентного семантического индексирования

В. Добрынин, И. Некрестьянов

Санкт-Петербургский Государственный Университет

vdobr@oasis.apmath.spbu.ru, igor@meta.math.spbu.ru

Аннотация

Целью данной работы является исследование эффективности применения метода вероятностного латентного семантического индексирования для решения проблемы расширения запроса пользователя при использовании обратной связи от пользователя. Для экспериментов используется стандартный набор тестовых данных TREC-5.

Введение

При использовании любой поисковой системы пользователь выражает свои информационные потребности в виде запроса, представленного на языке запросов данной системы [3]. Однако, анализ статистики запросов к различным поисковым системам показывает, что в 75% средняя длина запроса пользователя не превышает двух слов [12]. Не удивительно, что точность поиска по таким запросам обычно относительно невысока и зачастую не удовлетворяет пользователя.

Вообще говоря, для построения хорошего запроса мало хорошо понимать предметную область, в которой выполняется поиск. Необходимо также знать реальную статистику распределения слов по документам, проиндексированным этой системой. Такая информация дает возможность, например, не включать в запрос слова, которые в этой системе являются общеупотребительными и только “размывают” запрос. Обычно пользователь не имеет доступа и/или не обладает достаточной квалификацией, чтобы использовать эту информацию, и это является еще одной из причин, влекущей низкое качество поиска.

Известны два подхода, позволяющие расширить запрос пользователя дополнительными ключевыми словами, решая тем самым проблемы коротких запросов и необходимости учета распределения ключевых слов в коллекции:

- **Механизм обратной связи[1].**

  В рамках этого подхода поиск выполняется в две или более итераций [11,4,10,7,2,5,8]. На первом этапе производится поиск по исходному запросу пользователя. После получения результатов пользователь помечает некоторые из полученных ссылок как релевантные его информационным потребностям. Система использует эту информацию для автоматически выполняемого расширения запроса за счет включения в него ряда слов из отмеченных документов. Отметим, что такой подход требует дополнительной информации от пользователя.

- **Локальный контекстный анализ.**

  В этом случае поиск также выполняется в две или более итераций, но эти итерации прозрачны для пользователя [6,14]. На первой итерации используется исходный запрос. Полученные результаты поиска анализируется с целью выявления статистических отличий распределения слов в обнаруженных документах и распределения слов в системе в целом. Результаты этого анализа используются для расширения запроса. Отметим, что в отличие от механизма обратной связи, в этом случае от пользователя не требуется никакой дополнительной помощи.

Выбор конкретного подхода к расширению запроса зависит от многих факторов. Например, в случае, когда для запроса пользователя имеется достаточно много релевантных документов, более эффективным может

оказаться локальный контекстный анализ, т.к. в этом случае естественно ожидать, что среди первых возвращенных системой поиска результатов будет велика доля релевантных ссылок, и статистический анализ соответствующих документов позволит построить эффективное расширение исходного запроса. Если же документов, релевантных запросу пользователя, в данной коллекции мало, то предпочтительнее использовать механизм обратной связи. Немаловажным также является возможность дополнительной кооперации с пользователем, который, как показывает статистика использования поисковых систем [12], не очень предрасположен к поиску в несколько итераций.

Вне зависимости от выбранного подхода остается открытым вопрос о выборе ключевых слов для расширения запроса словами из выбранного множества документов. В этой работе исследуется возможность использования метода вероятностного латентного семантического индексирования [9] для выбора ключевых слов для расширения запроса.

Прямое применение вероятностного латентного семантического индексирования к коллекции большого объема требует использования значительных вычислительных ресурсов. Кроме того, необходимо регулярно повторять индексирование в связи с включением в коллекцию новых документов. В данной работе мы рассматриваем возможность индексирования лишь относительно небольшой доли коллекции и аппроксимации представление прочих документов в построенном пространстве факторов.

Экспериментальная проверка работоспособности предлагаемого подхода проводилась на основе стандартной тестовой коллекции TREC-5 [13]. Наблюдаемое в наших экспериментах повышение точности поиска демонстрируют работоспособность предлагаемого подхода.

Статья организована следующим образом: в следующем разделе кратко описаны основные понятия, связанные с вероятностным латентно-семантическим индексированием; далее рассматривается применение его к задаче расширения запросов, а в последнем разделе приведены результаты наших практических экспериментов на основе тестового набора данных TREC.

Вероятностное латентное семантическое индексирование

Метод вероятностного латентного семантического индексирования (PLSI) был предложен в работе [9], где и можно найти его подробное описание. В этом разделе мы лишь кратко опишем основные понятия и принципы, которые важны для понимания предлагаемого нами подхода.

Метод вероятностного латентного семантического индексирования ставит своей задачей выявление латентных, скрытых факторов (тем), присутствующих в коллекции и связанных с ее документами и словами. Именно, фиксировав число скрытых факторов r, с помощью метода PLSI можно оценить следующие величины

- $P(z_i)$ — вероятность того, что случайно выбранный из коллекции документ наиболее тесно связан с фактором (в наибольшей степени соответствует теме) z_i
- $P(d_j \mid z_i)$ — вероятность того, что наиболее тесно связанный с данным фактором z_i документ — это d_j
- $P(w_j \mid z_i)$ — вероятность того, что для данного фактора z_i наиболее тесно связано с ним слово — это w_j

[1] relevance feedback

Здесь $D = \{d_1, \ldots, d_n\}$ — множество всех проиндексированных документов, $W = \{w_1, \ldots, w_n\}$ — множество всех различных слов, встретившихся в проиндексированных документах, $Z = \{z_1, \ldots, z_r\}$ — множество латентных факторов.

Обозначим за $P(d, w)$ вероятность совместного наблюдения документа d и слова w, т.е. $P(d, w) = P(d)P(w \mid d)$. В рамках подхода, используемого при вероятностном латентном семантическом индексировании, величина $P(d, w)$ оценивается как $P(d, w) = \sum_{z \in Z} P(z)P(d \mid z)P(w \mid z)$.

В соответствии с принципом максимального правдоподобия, функции $P(z), P(d \mid z)$ и $P(w \mid z)$ определяются путем максимизации функции правдоподобия $L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)$, где $n(d, w)$ есть число вхождений слова w в документ d.

На этапе максимизации функции L мы использовали простейший из изложенных в [9] методов - стандартный метод *оценивания-максимизации*. На каждой итерации выполняются шаг *оценивания*

$$P(z \mid d, w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z'} P(z')P(d \mid z')P(w \mid z')}$$

и затем шаг *максимизации*

$$P(w \mid z) = \frac{\sum_d n(d, w)P(z \mid d, w)}{\sum_{d, w'} n(d, w')P(z \mid d, w')}, \; P(d \mid z) = \frac{\sum_w n(d, w)P(z \mid d, w)}{\sum_{d', w} n(d', w)P(z \mid d', w)},$$

$$P(z) = \frac{1}{R} \sum_{d, w} n(d, w)P(z \mid d, w), \; R = \sum_{d, w} n(d, w)$$

В наших экспериментах начальные значения для функций $P(z)$, $P(d \mid z)$ и $P(w \mid z)$ выбирались случайным образом, и сходимость наблюдалась после 50-70 итераций.

Расширение запросов на основе PLSI

Основная идея нашего подхода состоит в выявлении тематической принадлежности документов, отмеченных пользователем как релевантные на первом этапе поиска, и расширении запроса словами из отмеченных документов, представляющими выделенные темы. Однако, применение PLSI к коллекциям большого размера требует использования больших вычислительных ресурсов. Необходимо иметь возможность оценивать тематическую принадлежность любого документа коллекции на основе использования информации о тематической принадлежности документов из некоторого относительно небольшого множества случайно отобранных документов.

Аппроксимация образа нового документа в пространстве факторов

Будем далее называть вектор $(P(d \mid z_1), \ldots, P(d \mid z_r))$ образом документа d в пространстве факторов Z и, аналогично, вектор $(P(w \mid z_1), \ldots, P(w \mid z_r))$ образом слова w. Обозначим через D' некоторое множество документов, случайно отобранных из множества D всех документов, в котором выполняется поиск, а через W' — множество всех слов из документов вошедших в D'. Пусть также $d \in D - D'$.

Далее мы предполагаем, что множество D' может рассматриваться в качестве представительной выборки документов из коллекции D, и в ней затронуты все темы, отраженные в полной коллекции. Иными словами,

(произвольный) документ d содержит значительное число слов из W', при этом это подмножество W' достаточно полно отражает тематическую направленность d.

Рассмотрим систему линейных алгебраических уравнений

$$P(d,w) = \sum_{z \in Z} P(z)P(w \mid z)P(d \mid z)$$, где $w \in W'$, $n(d,w) > 0$

В качестве неизвестных рассматриваем величины $P(d \mid z)$, $z \in Z$, значения величин $P(z)$, $P(w \mid z)$, $z \in Z$, $w \in W'$ получены в результате применения PLSI к множеству документов D'. Величина $P(d \mid w)$ аппроксимируется по формуле $P(d,w) \approx \frac{1}{|D'|} \frac{n(d,w)}{length(d)}$, где $length(d)$ — количество слов из W' присутствующих в документе d.

Для приближенного решения данной системы мы строим псевдообратную матрицу для матрицы коэффициентов системы методом Гревилля (см., [1]). Полученное приближение является наилучшим (по методу наименьших квадратов), и может рассматриваться как хорошая аппроксимация образа документа d в пространстве факторов Z.

Расширение запроса пользователя

Как отмечалось во введении, в случае, когда коллекция документов, в которой выполняется поиск, содержит небольшое число документов релевантных запросу, расширение запроса, основанное на использовании обратной связи от пользователя, эффективнее использования локального контекстного анализа. Именно такая ситуация возникла в ходе проведения наших экспериментов, в связи с чем дальнейшее изложение учитывает использование только обратной связи от пользователя.

Расширение запроса пользователя на заданное число (q) слов на основе отмеченных пользователем документов (множество S) происходит следующим образом:

1. для всех слов w из W' таких, что они встречаются в документах из S вычисляется их вес $weight(w) = \sum_{d \in S, z \in Z} P(z)P(d \mid z)P(w \mid z)$.
2. множество слов w упорядочивается по убыванию весов $weight(w)$
3. из построенного списка выбираются первые q слов

При этом если документ d входит в множество D', то значения величин $P(z)$, $P(d \mid z)$ и $P(w \mid z)$ уже известны (вычислены при применении PLSI к множеству документов D'). В противном случае, неизвестные величины $P(d \mid z)$, $z \in Z$ оцениваются в соответствии с описанным в предыдущем разделе алгоритмом.

Наши эксперименты

На данный момент завершена только серия предварительных экспериментов, демонстрирующая общую перспективность подхода. Результаты дополнительных экспериментов и их анализ будет включен в полную версии статьи.

Тестовый набор данных

Тестирование проводилось на тестовой коллекции TREC-5 , для которой известны подготовленные экспертами запросы и списки релевантных документов для каждого запроса. Запросы в TREC-5 представляют собой достаточно длинные тексты, состоящие из нескольких разделов. В наших экспериментах мы были

заинтересованы в коротких запросах (поскольку именно короткие запросы доминируют в большинстве практических поисковых систем) и поэтому в качестве запросов мы выбирали краткое описание темы запроса (тег <topic>).

Коллекция TREC-5 содержит более 250 тысяч документов. Из раздела FBIS (foreign broadcast information system) были выбраны случайным образом 500 документов, множество которых (D') и было проиндексировано методом PLSI. Поиск выполнялся в коллекции, состоящей из первых 20000 документов (множество D) из FBIS.

Критерии оценки качества

Традиционным подходом к проблеме оценки качества поиска является построение усредненной по запросам кривой полнота-точность. Однако в наших экспериментах этот подход не приемлем в связи с широким разбросом числа релевантных документов среди проиндексированных 20000 для различных запросов (от одного до нескольких сотен). В связи с этим мы использовали для сравнения результатов поиска по исходному и расширенному запросу следующий подход. Обозначим через R_q множество рангов (порядковых номеров) релевантных документов для запроса q среди первых 200 документов, полученных в результате поиска. Пусть далее Q есть множество запросов, а $rel(q)$ - общее число документов в D, релевантных запросу q. Пусть

$$quality = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{(1+rel(q))}\sum_{i\in R_q}\frac{1}{i}.$$

Именно величину $quality$ мы будем использовать далее для оценки эффективности предлагаемого метода расширения запроса, вычисляя ее для исходных и расширенных запросов. Эта интегральная оценка учитывает как ранги всей совокупности релевантных документов, полученных в ответ запрос, так и общее число релевантных документов в коллекции. Ранги полученных релевантных документов для запросов с небольшим общим числом релевантных документов учитываются с большими весами, что соответствует интересам пользователя, желающего видеть релевантные документы среди первых возвращенных системой результатов, особенно в случае их малого числа.

Эксперименты на основе Open Muscat

Индексирование коллекции и поиск по исходным и расширенным запросом выполнялся с помощью системы Open Muscat[2] . Open Muscat является библиотекой с открытым кодом на C++, предназначенной для индексирования и поиска в коллекциях электронных документов. В свою очередь эта библиотека использует систему управления базами данных от SleepyCat. Open Muscat использует сочетание булевой и вероятностной моделей поиска. В наших экспериментах все запросы являлись дизъюнктами ключевых слов одинакового веса.

В экспериментах имитировался поиск информации в две итерации с обратной связью от пользователя. Первоначальный запрос передавался в поисковую систему Open Muscat и, затем, среди первых 200 документов результата выбирались первые три документа (или меньшее их число, если среди первых 200 документов не было трех релевантных), которые и использовались при расширении запроса пятью новыми ключевыми словами. Расширенный запрос вновь передавался в систему Open Muscat. Поиск проводился только среди 20000 документов из раздела FBIS, проиндексированных системой Open Muscat.

Результаты типичного эксперимента отражены в следующей таблице:

[2] `http://sourceforge.net/projects/openmuscat/`

Общее число запросов (для которых имеются релевантные документы в D)	27
Число запросов, для которых запрос не был расширен (среди первых 200 результатов нет релевантных документов)	4
Число запросов, для которых качество поиска для расширенного запроса выше качества поиска для исходного запроса (по критерию quality)	17
Среднее (по всем запросам) качество поиска по исходному запросу	0.061
Среднее (по всем запросам) качество поиска по расширенному запросу	0.125

Как видно среднее качества поиска значительно повышается (в два раза) по сравнению со случаем отсутствия расширения запросов. Заметим, что в тех случаях, когда качество поиска по расширенному запросу хуже качества поиска по исходному запросу, проигрыш при использовании расширенного запроса невелик, что и отражено в двукратном превышении среднего качества поиска по расширенному запросу по отношению к среднему качеству поиска по исходному запросу.

Для того, чтобы лучше понять насколько сложно хорошо расширить рассматриваемые запросы мы провели ряд экспериментов с расширением запросов на основе информации о синонимах, предоставляемой системой WordNet. Расширение производилось за счет тех слов, которые были синонимами к большему числу термов из запроса. Однако, таким образом мы смогли добиться только 20% улучшения точности (хотя информация в WordNet в значительной степени контролируется вручную). Кроме того, мы наблюдали стабильное снижение качества результатов при расширении запросов таким способом на 3 или более термов. Хотя это и объясняется тем, что короткие запросы несут мало информации, но это также иллюстрирует тот факт, что наивное расширение запросов может только ухудшить результат.

На данный момент еще многие вопросы требуют дополнительного изучения. Так, например, мы провели ряд экспериментов исследующих вопросы оптимального выбора эвристических параметров, определяющих получаемое расширение запроса. В частности, мы обнаружили, что на нашем тестовом наборе данных наилучшее качество достигается при расширении запроса на 7 слов и составляет 0.13308. Однако, этот результат скорее всего зависит от конкретного набора данных и запросов и выявление более общих закономерностей требует дополнительных экспериментов.

Заключение

В работе изучается применение вероятностного латентно-семантического анализа для расширения запросов. Для того, чтобы избежать проблемы с низкой вычислительной масштабируемостью метода предлагается строить пространство факторов по некоторому достаточно “представительному” подмножеству документов и аппроксимации образов прочих документов.

Практические эксперименты проводились на основе стандартного тестового набора данных TREC. Полученные результаты подтвердили общую перспективность рассматриваемого подхода и выявили ряд вопросов требующих дополнительного исследования. В частности, необходимо исследовать зависимость оптимального выбора эвристических параметров от наборов данных и запросов.

Библиография

1 Гантмахер. Теория матриц. Москва, 1967.

2 J. Allan. Relevance feedback with too much data. Technical Report UM-CS-1995-006, , 1995.

3 Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999.

4 C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In Proc. 17th Annu. Internat. ACM-SIGIR conference on Research and Development in Information Retrieval, pages 292-301. Springer-Verlag, 1994.

5 Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In Text REtrieval Conference, , 1994.

6 Bruce W. Croft and Jinxi Xu. Query expansion using local and global document analysis. In Proc. of the SIGIR'96, pages 4-11, 1996.

7 D. Haines and W.B. Croft. Relevance Feedback and Inference Networks. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2-11, 1993.

8 Kyoung-Soo Han, Dae-Ho Baek, and Hae-Chang Rim. Automatic text summarization based on relevance feedback with query splitting. In Proc. of the Fifth International Workshop on Information Retrieval with Asian Languages, pages 201-210, 2000.

9 Thomas Hofmann. Probabilistic latent semantic indexing. In Proc. of the SIGIR'99, pages 50-57, 1999.

10 Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In Research and Development in Information Retrieval, pages 206-214, 1998.

11 G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4):182-188, 1990.

12 C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. Technical Report 1998-014, COMPAQ System Research Center, October 1998.

13 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.

14 Jinxi Xu. Improving the effectiveness of informational retrieval with local context analysis.

QUERY EXPANSION WITH USAGE OF THE PROBABILISTIC LATENT SEMANTIC INDEXING.

V.Dobrynin, I. Nekrestyanov
Saint-Petersburg State University
vdobr@oasis.apmath.spbu.ru, igor@meta.math.spbu.ru

The goal of the paper is to investigate the effectiveness of the query expansion based on the usage of the probabilistic latent semantic indexing and relevance feedback. Standard set of test data (TREC-5) was used for experiments.