Особенности поискового алгоритма и архитектуры Exactus

© Тихомиров И. А.

Институт системного анализа РАН
matandra@isa.ru

Аннотация

В статье описаны проблемы, возникающие при использовании лингвистических методов поиска. Приведены особенности поискового алгоритма и архитектуры Exactus. Проанализированы результаты участия Exactus в РОМИП'2007, сделаны выводы о перспективе использования лингвистических алгоритмов поиска и дальнейших направлений исследований.

1. Введение

Специалисты в области поисковых технологий сосредоточили свое внимание на методах поиска, основанных на статистических характеристиках документов (TF*IDF-веса термов, ссылочное ранжирование и т.д.). Становится очевидным, что, несмотря на относительную языковую независимость этих методов, простоту реализации и определенные плюсы с точки зрения вычислительных ресурсов дальнейшее развитие этих методов является малоперспективным, и значительного увеличения качества поиска добиться на этом пути не удастся [1].

С другой стороны, коллективами лингвистов разработаны новые методы поиска и анализа текстов [1,2]. Очевидно, что в перспективе эти методы могут принести существенный выигрыш в точности и полноте поиска. Основным препятствием к непосредственному использованию лингвистических методов в популярных поисковых машинах является отсутствие их четкой оценки. Неизвестно, насколько хорошо работает метод, пока он не проверен на больших объемах данных. На отсутствие оценки влияет и тот фактор, что коллективы лингвистов, как правило, не имеют хорошей аппаратной

базы и опыта реализации задач в области программирования. Доведение лингвистического алгоритма «до ума» и его проверка в рамках серьезного соревнования (например, РОМИП) весьма трудоемкая и неподъемная для лингвистов задача.

Серьезную проблему составляет отсутствие у лингвистов опыта в области математики, что приводит к непониманию статистических формул и методов, используемых подавляющим большинством поисковых машин. Это приводит к тому, что лингвистические алгоритмы никак не учитывают хорошо зарекомендовавшую себя статистическую составляющую алгоритмов поиска. В результате имеем только статистические алгоритмы поиска (с поддержкой морфологии в лучшем случае) и лингвистические алгоритмы поиска (не учитывающие статистику).

В последние несколько лет целью разработчиков Exactus является эффективное взаимодействие лингвистов, математиков, программистов на пути решения задачи объединения статистических и лингвистических методов поиска [2]. В результате на семинар РОМИП был представлен экспериментальный алгоритм поисковой машины Exactus, включающий как статистические критерии поиска, так и языковые особенности естественного языка (русский синтаксис и семантику).

2. Несколько слов об алгоритме поиска Exactus

Алгоритм поиска Exactus объединяет статистическую и лингвистическую составляющие. Из статистических характеристик текста Exactus учитывает TF*IDF веса термов и значимость фрагментов текстов (на основе HTML-разметки документов). Лингвистическая составляющая – значения синтаксем и их сочетаемость в конкретном предложении [4,5]. Это позволяет отбирать только те тексты, в которых значение синтаксемы совпадает с ее значением в запросе (что невозможно в обычных статистических методах). Кроме того, это позволяет обработать ситуацию, когда целевая синтаксема является элементом более сложной синтаксической конструкции (например, находится в отношении подчинения). Пример:

Запрос: «Кто выиграл выборы на Украине».

Документ1: «Выборы на Украине выиграл Янукович».

Документ2: «Выборы на Украине выиграла партия Януковича».

В результате для системы Exactus первый документ наиболее предпочтителен, так как во втором документе «Янукович»

находится в отношении подчинения слова «партия» и в другом семантическом значении [6]. Тут следует отметить, что вопросно-ответный поиск в Exactus реализуется естественным образом (вопросительным конструкциям автоматически производится сопоставление их заместителей из индекса в рамках того же предложения, где находятся другие слова запроса).

Поиск в Excatus может быть проведен только после предварительной индексации документов. На этапе индексации производится преобразование документов к внутреннему формату Exactus, обсчет TF*IDF весов термов с учетом морфологии русского языка. Параллельно этому производится синтаксический и семантический анализ текстов, что позволяет выявить подчинения синтаксем в тексте и их семантические значения. Полученные в результате анализа данные укладываются в линейные упорядоченные списки.

Очевидно, что уложить семантическую сеть текста, полученную после синтаксического и семантического анализа, в линейные списки не простая задача [6]. Разработчики Exactus в своем экспериментальном алгоритме пожертвовали сетевой структурой, трансформировали сеть в дерево и уже его отражали в списочную структуру.

Алгоритм поиска Exactus представляет собой слияние и переранжирование линейных упорядоченных списков, что опять же аналогично концепции большинства поисковых машин. Особенностью алгоритма являются весовые коэффициенты и алгоритм предварительной индексации текстов, которые позволяют учесть как статистические, так и семантические составляющие текстов.

3. Особенности архитектуры и программно-аппаратных средств Exactus

Современная архитектура Exactus имеет модульную структуру, модули расположены на узлах кластерной установки с возможностью параллельного выполнения задач [3]. Основным способом параллелизма является позадачное распараллеливание. Управление задачами осуществляется посредством PVM-машины (Parallel Virtual Machine). Модули можно разделить на два типа: основные (лингвистические процессоры, индексаторы и т.д.) и вспомогательные (агрегаторы, синхронизаторы и т.д.). Задачей основных модулей является решение конкретных задач поисковой машины. Задачей вспомогательных модулей является сервисная

составляющая: обеспечение масштабируемости системы, распределенное хранение индекса, объединение результатов поиска и много другое [3].

Система Exactus является кросс-платформенной и может функционировать на широком спектре Unix-подобных операционных систем. Версия, используемая для РОМИП'2007, функционирует на Linux Debian 4.0. Экспериментальная установка состоит из 8-и задействованных узлов кластера пиковой производительностью 100 Gigaflops. Особенностью Exactus является то, что в качестве вычислительных узлов используются обычные персональные компьютеры, объединенные в стойку (концепция, аналогичная Google). Узлы неравнозначны по своим аппаратным характеристикам, так, например, для хранения индекса нужны большие винчестеры и большой объем оперативной памяти, а для лингвистических процессоров – высокая производительность центрального процессора и большой объем оперативной памяти. Для взаимодействия узлов используется Gigabit Ethernet.

4. Краткий анализ результатов Exactus на РОМИП'2007

Одним из замечательных моментов является то, что представленный в РОМИП алгоритм Exactus не использует ссылочное ранжирование (в отличие от подавляющего большинства поисковых алгоритмов). По обеим дорожкам (LEGAL и BY) был проведен всего один прогон экспериментального алгоритма Exactus. Несмотря на указанное выше, удалось продемонстрировать неплохие результаты по коллекции LEGAL и очень хорошие результаты по коллекции BY.

Наилучшие результаты достигнуты в AND-оценке по точности, которая с точки зрения разработчиков Exactus является доминирующей. Хорошие результаты по AND-оценке объясняются тем, что в случае OR-оценки не удалось обеспечить согласованного мнения экспертов по поводу релевантности того или иного документа.

В предыдущих версиях Exactus использовались по большей части лингвистические критерии оценки релевантности, что на предыдущих семинарах РОМИП приводило к большому числу слаборелевантных ответов (в предыдущих семинарах РОМИП у Exactus оценки по OR относительно других участников были лучше, чем AND) [7].

Можно полагать, что при поиске по легальной коллекции лингвистическая составляющая для узкоспециализированных текстов не дает ощутимого выигрыша (без настройки семантических словарей на предметную область). Более качественные результаты на специфических тематических коллекциях можно получить с использованием долгой настройки и подгонки весовых коэффициентов при подсчете статистики.

5. Заключение

Полученные в ходе экспериментов РОМИП результаты показывают перспективность применения лингвистических алгоритмов анализа текстов и возможность их применения в реальных условиях, тем более что одно из препятствий (отсутствие достоверных оценок лингвистических алгоритмов) частично снято. Эксперименты показывают, что скорость поиска в Exactus сравнима по скорости с современными поисковыми машинами на больших объемах данных. Все проблемы поиска удалось перенести на задачу индексации, которая, по-прежнему, остается узким местом лингвистического анализа. Однако, за счет использования современных вычислительных систем и параллельных вычислений синтаксический и семантический анализ больших коллекций текстов становятся вполне разрешимыми задачами.

Среди ближайших направлений исследований – включение в алгоритм индексации ссылочного ранжирования и заранее составленного каталога ресурсов. Кроме того, существуют определенные соображения по методам трансформирования семантической сети текста в линейные упорядоченные списки для последующего использования при поиске, что также должно повысить точность последнего.

В перспективе, лаборатория интеллектуальных динамических систем ИСА РАН планирует расширить свое участие в дорожках РОМИП и проверить таким образом новые алгоритмы каталогизации и контекстно-зависимого аннотирования.

Литература

[1] Осипов Г.С., Завьялова О.С., Климовский А.А., Кузнецов И.А., Смирнов И.В., Тихомиров И.А. Проблемы обеспечения точности и полноты поиска: Пути решения в интеллектуальной метапоисковой системе "Сириус". //Труды мезденародной конференции Диалог'2005, с. 390-395, Москва, Наука, 2005.

[2] Osipov G. S., Smirnov I. V., Tikhomirov I. A., Vybornova O.V, Zavjalova O. S. Linguistic Knowledge for Search Relevance Improvement.// Papers of Joint conference on knowledge-based software engineering JCKBSE'06, IOS Press, 2006. - P. 294-302.
[3] Осипов Г.С., Тихомиров И.А., Смирнов И.В. Exactus – система интеллектуального метапоиска в сети Интернет. // Труды десятой национальной конференции по искусственному унтеллекту с международным участием КИИ-2006. М: Физматлит, 2006. т. 3. - С. 859-866.
[4] Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова, М. 2004 – 544 с.
[5] Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988 – 440 с.
[6] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997.
[7] Тихомиров И. А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus.//Труды четвертого российского семинара по оценке методов информационного поиска РОМИП'2006. Санкт-Петербург: НУ ЦСИ, 2006. - с. 80-85.