RCO на РОМИП 2003: отчет об участии в семинаре по оценке методов информационного поиска

© Плешко В.В., Ермаков А.Е., Митюнин В.А.

ООО «Гарант-Парк-Интернет»
rco@metric.ru

Аннотация

Настоящая работа является отчетом об экспериментах по поиску web-страниц и классификации web-сайтов, проведенных в рамках инициативы РОМИП. Главной целью работы была апробация методов оценки качества информационного поиска на русскоязычных текстовых корпусах.

1. Введение

Наличие свободно доступных текстовых корпусов для оценки качества методов информационного поиска, а также регулярный обмен опытом между исследователями и разработчиками в области построения информационно-поисковых систем (ИПС) являются важными факторами прогресса в данной области.

Такой опыт уже более 10 лет успешно практикуется в США в форме конференции Text REtrieval Conference (TREC) [1]. Хотя англоязычные текстовые корпуса и полученные на них результаты позволяют в какой-то мере использовать передовой опыт зарубежных коллег, у российских исследователей до сих пор не имелось в распоряжении масштабных русскоязычных текстовых корпусов, на которых можно было бы получить достоверные оценки качества создаваемых систем.

Этот пробел призван устранить Российский семинар по Оценке Методов Информационного Поиска (РОМИП) [2,3]. Участникам семинара предлагалось принять участие в экспериментах (дорожках) по решению двух задач: «поиск web-страниц» и «тематическая

классификация web-сайтов». Методика оценки результатов, использованная организаторами семинара, является общепринятой для задач информационного поиска и описана в [4].

Компания «Гарант-Парк-Интернет» поддержала эту инициативу и приняла участие в обеих дорожках. При проведении экспериментов использовалось «коробочное» программное обеспечение (ПО), производимое компанией. Описание использованного ПО доступно на сайте [5], посвященном семейству технологий Russian Context Optimizer (RCO).

2. Поиск web-страниц

В качестве исходных данных участникам были предоставлены подмножество сайтов, расположенных в домене narod.ru, размером около 7Gb и фрагмент журнала запросов поисковой машины Яндекс. Необходимо было построить поисковый индекс по всем страницам предложенной коллекции документов и выполнить все запросы, представив в качестве результата первые 100 страниц, содержащиеся в ответе системы и упорядоченные по убыванию соответствия запросу.

2.1 Описание системы

Для решения поставленной задачи была использована поисковая машина Russian Context Server, предназначенная для поиска по текстовым и реляционным данным. Ниже приводятся основные принципы работы системы с текстовыми данными.

Индексируемые документы обрабатывались следующим образом:

1. Текст разбивается на последовательность лексем. Лексемы, входящие в список стоп-слов, удаляются из последовательности.
2. Каждая лексема приводится к исходной грамматической форме. Последовательности символов кириллицы анализируются только средствами словарной морфологии. Слова, отсутствующие в словаре, оставляются без изменений. Из последовательности символов латиницы делается попытка выделения основы слова при помощи алгоритма Портера.
3. Информация о каждой лексеме и ее позиции в документе заносится в инвертированный список. Инвертированные списки хранятся на диске в упакованном виде с использованием арифметических кодов, аналогичных, изложенным в [6].

При поиске текст запроса подвергается такому же преобразованию. Затем для каждой лексемы из индекса извлекается инвертиро-

ванный список, и полученные списки обрабатываются в соответствии с условиями запроса (И, ИЛИ, НЕ, ФРАЗА).

Численное значение соответствия документа запросу рассчитывается по классической схеме $tf \times idf$, но с некоторыми упрощениями, позволяющими использовать целочисленную арифметику и повышающими локальность используемых данных.

Пусть $Q = \{q_1, q_2, \ldots, q_K\}$ — последовательность, возможно повторяющихся, искомых терминов (слов, словосочетаний), D — документ, который система вернула в ответ на запрос. Тогда степень соответствия документа запросу равна

$$r(Q,D) = 0.3 + 0.7\frac{1}{K}\sum_{i=1}^{K}\left[0.5 + 0.5\frac{tf_i}{tf_i + 2}\right]\frac{\log_2\left(N/df_i\right)}{\log_2\left(N\right)}, \qquad (1)$$

где K — число слов запроса, N — общее число документов в коллекции, tf_i — частота термина q_i в документе D, df_i — число документов коллекции, содержащих термин q_i.

2.2 Подготовка данных

Все данные были перенесены в файловую систему NTFS. При этом для каждого сайта был создан отдельный каталог, совпадающий с адресом. Все ресурсы сайта были преобразованы в имена файлов путем замены символа «/» (прямой слэш) на «_» (подчеркивания). Попутно была создана таблица соответствий имен файлов ресурсам сайта с целью обратного преобразования при представлении результатов.

Перенос в файловую систему NTFS привел к потере нескольких страниц, названия ресурсов которых отличались только регистром. Также оказалось, что небольшое число файлов представлено в кодовой странице KOI, и еще некоторое количество файлов имело нулевой размер. Тем не менее, качество коллекции можно считать хорошим, так как число таких страниц составило доли процента и не могло повлиять на результат.

2.3 Настройка системы

Индекс был построен по содержимому HTML тегов BODY, TITLE, KEYWORDS, DESCRIPTION.

Поисковые запросы делались только к содержимому тега BODY. Слова запросов объединялись логической связкой ИЛИ. Результирующий список упорядочивался по убыванию соответствия страниц запросу.

Были также испробованы варианты с использованием дополнительно других полей. При анализе ответов системы на единичных запросах, полезность использования дополнительных полей показалась сомнительной. В результате было решено оставить для оценки только один прогон.

2.4 Результаты

При оценке результатов каждой паре <документ, запрос> выставлялась оценка «соответствует/не соответствует». В силу того, что большинству пар было выставлено по две оценки, обобщающие показатели считались двумя способами: «weak» и «strong». В первом случае считалось, что документ соответствует запросу, если была выставлена хотя бы одна положительная оценка. Во втором случае считалось, что документ не соответствует запросу, если была выставлена хотя бы одна отрицательная оценка.

В табл. 1 приведены показатели качества, оцененные двумя способами.

Таблица 1. Оценки качества поиска

	weak	strong
Recall	0.5927	0.6793
Precision(5)	0.0952	0.0333
Average precision	0.1027	0.0386
Precision(10)	0.1148	0.0500
R-precision	0.0952	0.1745
Precision	0.1068	0.0383

При сопоставлении способов подсчета заметный рост Recall и R-precision наряду с падением остальных показателей говорит о том, что основные разногласия у оценщиков были вызваны наименее релевантными документами (хвостом списка).

Зависимости точности результатов от полноты приведена на рис. 1. Полученные результаты тяжело интерпретировать без проведения дополнительных экспериментов.

Общий уровень результатов остальных участников (максимальная точность ~0.18) наводит на мысль, что, возможно, без привлечения дополнительной информации о гипертекстовой структуре web значительно повысить качество не получится.

Кроме того, отсутствие ярко выраженного пика при небольших значениях полноты свидетельствует о необходимости доработки формулы (1) вычисления степени соответствия документа запросу.

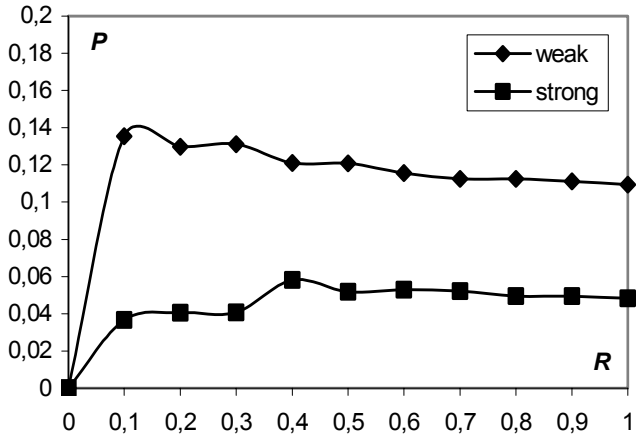

Рисунок 1. Зависимость точности от полноты

3. Классификация web-сайтов

Исходными данными для задачи классификации было то же подмножество сайтов домена narod.ru, что и для задачи поиска. В качестве набора классов было выбрано 164 рубрики каталога narod.ru. Каждой из отобранных рубрик было отнесено не менее 10 сайтов. Таким образом, исходное множество сайтов было разделено на обучающую (те, что вошли в каталог) и тестовую (остальные) выборки.

Задание состояло в том, чтобы на основе данных из обучающей выборки сопоставить каждому сайту из тестовой выборки набор рубрик, которым он соответствует. Допускалось не сопоставлять сайту ни одной рубрики.

3.1 Описание системы

Для выполнения задания была собрана система RCO Classifier, состоящая из двух готовых библиотек: RCO Semantic Network и RCO TopTree. Первая библиотека использовалась для построения терминологического вектора, отражающего значимость терминов анализируемого текста. Вторая библиотека использовалась для построения профилей рубрик и принятия решения о соответствии терминологического вектора профилю рубрики.

Построение терминологического вектора текста в библиотеке RCO Semantic Network состоит из следующих основных этапов:

1. Разбиение текста на предложения, являющиеся последовательностями лексем и разделителей, определение грамматических признаков каждой лексемы;
2. Объединение, в соответствии с базой правил, последовательностей лексем в сложные текстовые единицы [7];
3. Локальный синтаксический разбор каждого предложения. Дополнительно может быть проведен и полный разбор с использованием словаря моделей управления [8];
4. Синтез терминов (именных групп) и их ранжирование по тематическому весу (от 0 до 100). Максимальное число терминов и минимальный тематический вес определяются настройками;
5. Расчет ассоциативных связей между терминами.

Ассоциативные связи между терминами при формировании терминологических векторов не учитывались.

Библиотека RCO TopTree содержит набор алгоритмов классификации и кластерного анализа, абстрагированными от природы объектов и их признаков. Большинство реализованных в библиотеке алгоритмов приведено в [9]. В системе RCO Classifier был использован алгоритм, использующий понятие центра тяжести.

Пусть $T = \{t_i\}$ — множество всех различных терминов (слов, словосочетаний), выделенных к качестве значимых на текстах тестовой и обучающей выборок. Сопоставим каждому сайту s вектор $W_s = (w_i)^T$ размерности $|T|$, каждый элемент w_i которого равен тематическому весу термина t_i в контексте сайта. Обозначим через L количество рубрик, через $S_1, \ldots, S_L$ — множества сайтов, входящих в обучающую выборку и соответствующих каждой рубрике, и через S_0 — множество сайтов тестовой выборки.

В процессе обучения профиль каждой рубрики описывался центром масс обучающей выборки:

$$C_j = \frac{1}{|S_j|} \sum_{s \in S_j} W_s, \quad j = 1, \ldots, L . \tag{2}$$

Степень сходства сайта и рубрики вычислялась как корреляция соответствующих векторов:

$$sim(s, C_j) = \frac{W_s^T \cdot C_j}{\|W_s\| \cdot \|C_j\|} . \tag{3}$$

Кроме того, при обучении были рассчитаны минимальные степени сходства между каждым профилем рубрики и сайтами из соответствующей обучающей выборки:

$$h_j^0 = \min_{s \in S_j} sim(s, C_j) . \quad (4)$$

В процессе классификации для каждого сайта обучающей выборки сначала отбиралось подмножество классов R_s^0, степень сходства с которыми превышает рассчитанный при обучении порог:

$$R_s^0 = \{C_j : sim(s, C_j) \geq h_j^0, j = 1..L\}, \quad s \in S_0 . \quad (5)$$

Затем, в качестве окончательного множества классов выбирались только те, у которых относительное различие с наилучшем, в смысле сходства, классом не превышало заданный параметр δ. Таким образом, окончательное множество рубрик R_s, которые сопоставлялись сайту выглядит как

$$R_s = \{C_j : sim(s, C_j) \geq (1 - \delta) \max_{C \in R_s^0} sim(s, C), C_j \in R_s^0\} . \quad (6)$$

3.2 Подготовка данных

Каждый сайт был представлен в виде одного файла как конкатенация текстов всех его страниц в той последовательности, как они были расположены в файловой системе при выполнении задания по поиску web-страниц. При обработке для быстроты учитывались только первые 15Mb содержания сайта.

Для каждой рубрики был создан каталог, куда были скопированы файлы обучающей выборки. В процессе обучения в каждом из каталогов были созданы файлы с профилями рубрик, которые затем были использованы при обработке тестовой выборки.

3.3 Настройка системы

Построение терминологических векторов сайтов проводилось со следующими настройками: максимальное число терминов = 100, минимальный тематический вес термина = 5, максимальное число слов в термине = 5, синтаксический разбор = локальный.

Параметр δ, используемый для окончательного отбора рубрик, был взят равным 0.1, то есть в результат заносилась верхушка из 10% рубрик-кандидатов.

3.4 Результаты

Как и в случае поисковой дорожки, было получено два набора оценок: «weak» и «strong». Полученные оценки результатов представлены в табл. 2.

Таблица 2. Оценки качества классификации

	weak	strong
F1 (macro average)	0.2546	0.1433
Recall	0.2094	0.1805
Precision (macro average)	0.2510	0.0976
F1	0.1992	0.1153
Recall (macro average)	0.2582	0.2692
Precision	0.2807	0.1348

На обучающей выборке параметры Recall и Precision были равны соответственно 0.7265 и 0.9315. Более чем трехкратная деградация показателей при экстраполяции может быть вызвана описанием классов одним обобщенным терминологическим вектором, тогда как содержание сайтов в большинстве случаев является политематическим и должно описывается набором терминологических векторов.

4. Вычислительные ресурсы

Вычисления проводились на машине с процессором PIV-1.7GHz и объемом оперативной памяти 750Mb.

Построение поисковых индексов заняло 28 часов. Выполнение поисковых запросов заняло примерно 4 часа.

Построение профилей рубрик на обучающей выборке заняло 3 суток. Обработка тестовой выборки заняла около недели.

5. Заключение

В работе представлен отчет об участии компании «Гарант-Парк-Интернет» в первом годовом цикле семинара РОМИП. Основным результатом работы является выполнение заданий по поиску web-страниц и классификации web-сайтов именно с технической и методической точек зрения.

Несмотря на то, что специфика web нередко диктует использование упрощенных методов анализа документов, мы считаем, что многие из полученных участниками оценок не так уж плохи, о чем свидетельствует сравнение с аналогичными результатам, которые демонстрируют зарубежные системы на

подобных тестах. В целом можно сказать, что первый российский "блин" получился вполне удачным и вдохновляющим, а учет ошибок и обмен опытом позволят в следующем годичном цикле всем участникам семинара продемонстрировать более высокие результаты.

Литература

[1] Сайт Text REtrieval Conference (TREC) http://trec.nist.gov/

[2] Браславский П.И., Губин М.В., Добров Б.В., Добрынин В.Ю., Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., Сегалович И.В. Инициативный проект Российского семинара по Оценке Методов Информационного Поиска (РОМИП). *Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. – Москва, Наука, 2003.*

[3] Сайт Российского семинара по Оценке Методов Информационного Поиска (РОМИП) http://romip.narod.ru/

[4] Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска. *Программирование, 28(4): 226-242, 2002.*

[5] Сайт Russian Context Optimizer (RCO) http://www.rco.ru/

[6] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. – *New York, Von Nostrand Reinhold, 1994.*

[7] Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте. *Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. – Москва, 2003.*

[8] Ермаков А.Е. Неполный синтаксический анализ текста в информационно-поисковых системах. *Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2002. В двух томах. Т.2. "Прикладные проблемы". – Москва, Наука, 2002.*

[9] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – *Москва, Финансы и статистика, 1989.*

RCO at RIRES 2003: report on participation in information retrieval evaluation seminar

Pleshko V.V., Ermakov A.E., Mityunin V.A.

This article presents report on experiments in web page retrieval and web site classification tasks that were made as a part of RIRES initiative. Main goal of this article is to approbate information retrieval quality evaluation methods for Russian language corpora.