РОССИЙСКИЙ СЕМИНАР ПО ОЦЕНКЕ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА (РОМИП) В 2004 ГОДУ

Агеев М.С.[1,2], Губин М.В.[3], Добров Б.В.[1,2], Кураленок И.Е.[4], Некрестьянов И.С.[4], Плешко В.В.[5], Сегалович И.В.[6], Шабанов В.И.[7]

[1] Научно-исследовательский вычислительный центр МГУ
[2] АНО Центр информационных исследований
[3] ИК «Кодекс»
[4] Санкт-Петербургский государственный университет
[5] ООО «Гарант-Парк-Интернет», RCO Research Group
[6] ООО Яндекс
[7] ООО Рамблер Интернет Холдинг

romip@oasis.apmath.spbu.ru

Представлена деятельность РОМИП в 2004 году. Созданы следующие коллекции: Веб коллекция Narod.ru (Яндекс, 730000 док.), Веб коллекция DMOZ (Рамблер, 300000 док.), коллекция правовых актов РФ (ИК «Кодекс», 60000 док.). Программа РОМИП'2004 состояла из дорожек по поиску, классификации и фактографическому поиску (всего 5 дорожек, в 2003 году было 2 дорожки). Получено 34 варианта ответов от 9 участвующих коллективов (в 2003 году 14 вариантов от 7 команд).

1. Введение

Для сближения позиций различных исследователей в области информационного поиска уже сложилось несколько форумов, имеющих международный характер – например, американский TREC [1], европейский CLEF [2], японский NTCIR [3].

РОМИП, как представляется авторам, вбирает в себя лучшее из методологий проведения названных конференций (прежде всего TREC). При этом РОМИП проводится в более тяжелых условиях – проблемами являются как недостаток культуры лицензирования интеллектуальной собственности для целей некоммерческого использования (в нашем случае для исследований) в РФ, так и практическое отсутствие подготовленных корпусов русскоязычных текстов во многих областях знания (медицина, переводы, юриспруденция и т.д.). Для большинства коллекций устроителям РОМИП приходилось практически решать и организационные задачи по поиску и составлению коллекций, и юридические задачи по их легальному лицензированию.

Целями РОМИП являются:

- создание и развитие информационных ресурсов, обеспечивающих исследования в области информационного поиска (информационно-поисковых систем, экспертных систем, баз данных, других дисциплин);
- проведение независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией;
- формирование среды для исследования феномена информационного поиска на актуальных для российского пользователя задачах;
- формирование требований к оформлению текстовых коллекций для тестирования;
- формирование «правил игры» - этических норм представления и использования результатов.

Результаты работы семинара в целом и каждой из дорожек публично доступны, как в виде трудов семинара, так и используемых корпусов и наборов заданий, а также построенных таблиц релевантности и созданных инструментов.

Большинство этих материалов публикуется на сайте семинара (http://romip.narod.ru), а доступ к корпусам РОМИП можно получить после обращения в оргкомитет и подписания необходимых соглашений с правообладателями.

Успешное завершение первого семинара 2003 года [4] создало благоприятную атмосферу для его развития в 2004 году. В этом году увеличилось число рассматриваемых задач и участвующих систем, появилась новая коллекция нормативных документов и новая обучающая выборка для задачи классификации Веб сайтов.

2. Коллекции

В настоящее время участники РОМИП могут выполнять исследования со следующими коллекциями полнотекстовых документов: Веб-коллекция Narod.ru (предоставлена ООО Яндекс, около 730 тысяч документов), Веб-коллекция DMOZ (предоставлена ООО Рамблер Интернет Холдинг, созданная на основе русскоязычной части каталога dmoz.org с целью получения обучающего множества для задачи

классификации Веб-сайтов, 300 тысяч документов), коллекция нормативных документов законодательства РФ (предоставлена ИК «Кодекс», 60 тысяч документов).

Для предотвращения несанкционированного использования данных участники подписывали специальное соглашение.

2.1. Веб-коллекция Narod.ru (предоставлена ООО Яндекс)

Эта та же коллекция, что использовалась в РОМИП'2003 [4]. Она содержит порядка 22000 случайно выбранных сайтов (около 3%) из домена narod.ru по состоянию на март 2003 года. Всего в коллекции порядка 728000 отдельных страниц (формат HTML, кодировка Window-1251). Каждый сайт представлен полным набором своих страниц, так как коллекция собиралась обходом директорий непосредственно по жесткому диску.

2.2. Веб-коллекция DMOZ (предоставлена ООО Рамблер Интернет Холдинг)

В 2004 году задание для участников дорожки было откорректировано так, чтобы учесть проблемы с обучающей выборкой. На этот раз вместо рубрикатора Narod.ru и соответствующей части корпуса Narod.ru использовался рубрикатор русской части каталога DMOZ (Words/Russian/*). Обучающая выборка состояла из 308967 страниц, автоматически скачанных с 2073 web-сайтов, зарегистрированных в DMOZ. В качестве набора для автоматической классификации использовался корпус Narod.ru.

Обучающая выборка была сформирована следующим образом: из дампа каталога был построен список стартовых страниц всех ресурсов рубрики World/Russian и ее подрубрик. Далее этот список был подан роботу поисковой машины Рамблер, который скачал все перечисленные в списке сайты. Всего было сделано 3 прохода: на первом были скачаны исходные страницы, а на втором и третьем - страницы, на которые были обнаружены ссылки на предыдущем проходе. На каждом проходе скачивалось не более 500 страниц с хоста. Общий объем скачанных роботом страниц – 28 Гб. После скачивания в наборе были найдены и удалены сайты, на страницах которых было указано, что авторы запрещают копирование и распространение публикуемой информации. Затем рубрики, где суммарное количество страниц на всех приписанных к рубрике сайтах было больше 3000, были случайным образом профильтрованы. Преимущество при фильтрации получали средние по размеру сайты.

Необходимо отметить, что в задании на Веб-классификацию прилагался исходный дамп рубрикатора DMOZ, в котором содержалась дополнительная информация. Во-первых, в дампе содержалось исходное, многоуровневое дерево рубрик (в задании все сайты были «стянуты» к двухуровневому рубрикатору). Во вторых, многие сайты были снабжены краткими аннотациями. В третьих – в дампе были указаны перекрестные ссылки между рубриками таксономии DMOZ. Учет этой информации потенциально может помочь построить более точные классификационные признаки в процессе обучения.

2.3. Коллекция нормативных документов (предоставлена ИК «Кодекс»)

В 2004 году на семинар фирмой «Кодекс» была представлена коллекция нормативно-правовых документов. Она содержит основные правовые документы законодательства России, изданные федеральными органами власти, на состояние начала 2004 года. Всего передано 60015 документов, что составляло около 1,6 Гб текста в формате HTML. В коллекции были документы с датой принятия в период с 1918 по 2003 год, большая часть документов относилась ко второй половине 90-х годов. Все документы были представлены последними версиями, старые редакции не включались.

Из текстов документов были удалены все комментарии и другая вспомогательная информация. Это было сделано для того, чтобы, с одной стороны, сделать коллекцию максимально «нейтральной», не привязанной к данному производителю, с другой стороны, на эти материалы накладывались существенные ограничения по авторским правам.

3. Дорожки

Программа РОМИП'2004 состояла из 5 дорожек, каждая из которых была посвящена отдельной задаче.

3.1. Поиск по Веб-коллекции (web adhoc)

Эта дорожка являлась повторением аналогичной дорожки РОМИП'2003. Поиск производился по той же коллекции Narod.ru, но число заданий было увеличено до 24250, и для каждого из них система могла вернуть до 100 результатов.

Задания отбирались из журналов запросов поисковых систем Рамблер и Яндекс. Большое число заданий предотвращало возможность ручной настройки системы под конкретные запросы.

3.2. Поиск по колекции нормативных документов (legal adhoc)

Правила этой дорожки очень похожи на правила дорожки поиска по Веб коллекции. Однако, в этом случае поиск производился по коллекции нормативных документов и число заданий было порядка 13000.

Благодаря содействию компаний Кодекс и Парк.Ру задания отбирались из журналов поисковых систем, специализирующихся на поиске по нормативным документам. Оценка проводилась по отобранному 41 запросу.

3.3. Тематическая классификация Веб-сайтов (web classification)

Эта задача уже рассматривалась в РОМИП'2003, но многие участники жаловались на низкое качество использовавшегося обучающего множества. Поэтому в 2004 году было решено построить новое обучающее множество на основе dmoz.org.

Итоговая таксономия содержала 247 категорий и для каждой категории было не менее 5 обучающих примеров.

Задание состояло в классификации всех сайтов коллекции narod.ru. Для каждого сайта система могла вернуть от 0 до 5 категорий к которым он по ее мнению относится.

3.4. Тематическая классификация нормативных документов (legal classification)

В рамках этой дорожки классификации подвергались все документы из коллекции нормативных документов.

В качестве рубрик использовалось подмножество рубрик разработанного «Кодексом» классификатора правовых документов. Для РОМИП использовался второй уровень иерархии классификатора, содержащий 163 рубрики. В качестве обучающего множества были представлены данные о классификации 6294 документов (не менее 5 документов для каждой из категорий). Все интересующиеся участники могли получить полный классификатор, чтобы иметь представление о взаимном положении в нем рубрик.

Для каждого документа система могла выбрать до 5 категорий из 163 рассматриваемых.

3.5. Поиск фактов по Веб-коллекции (QA)

Данная дорожка посвящена задачам поиска описаний фактов в текстах. К подобным задачам можно отнести вопросно-ответный поиск, автоматический подбор информации к досье.

В 2004 году рассматривалась задача поиска в Веб-коллекции описаний биографических фактов связанных с конкретной персоной. По заданному описанию персоны, включающему полное имя, псевдонимы и краткую справку о роде деятельности, системам-участникам предлагалось найти фрагменты текста, содержащие биографическую информацию о персоне. Кроме того, найденные описания фактов могли быть отнесены системами к одному из заданных типов событий.

В качестве Веб-коллекции документов была использована коллекция Narod.ru. Перечень целевых персон и их описаний был построен на основе информации, опубликованной на сайте «Кроссворд-Кафе» (http://dilet.narod.ru/). Всего получилось чуть более 5000 описаний персон. Для оценки было отобрано около 100 описаний различного вида (с длинными / короткими текстовыми справками, с псевдонимами / без псевдонимов и т.п.).

Асессорам при проверке ответов систем была дана следующая инструкция:

«Цель поиска - составить досье/биографическую справку на заданного человека, то есть найти _все_ события связанные с ним. Полностью релевантный ответ - это ответ, содержащий описание и время возникновения события, а также ссылку на заданного человека (отсутствие части этой информации делает ответ частично релевантным). Если выделенный фрагмент текста не является идеальным (то есть можно выбрать его лучше), то необходимо также уточнить идеальные границы.»

4. Результаты

Всего было получено 11 заявок на участие в РОМИП'2004, но только 9 систем, подавших заявки, дошли до финиша. Отметим, что 8 из подавших заявки коллективов ранее подавали заявки для участия в РОМИП'2003. Было получено 34 варианта ответов от 9 участвующих коллективов (в 2003 году 14 вариантов от 7 команд).

Целью процесса оценки результатов являлось построение таблиц релевантности, которые содержат информацию о правильных и неправильных ответах для каждого задания, а также вычисление итоговых официальных метрик, характеризующих качество работы систем. В пятизначной логике получено десятки тысяч оценок соответствия документов заданиям РОМИП.

Участники продемонстрировали достаточно высокие показатели (Рис.1, Рис.2):

- для дорожки web adhoc в среднем 5-7 релевантных документов среди первых 10, для дорожки legal adhoc 7-8 релевантных в первых 10 документах;
- при этом некоторые относительно низко расположенные кривые результатов объясняются тем, что таким образом участники целенаправленно проверяют влияние тех или иных параметров;
- для дорожки Веб-классификации, даже при оценке на основе сильных требований к релевантности на 19 рубриках из 38 показатель F_1 был больше 0.5. В 2003 году ни одному из участников не удалось достигнуть показателя $F_1 > 0.5$.

Подготовлен сборник статей [5] участников по результатам выполнения заданий РОМИП и анализа результатов. Изложен ряд различных подходов к решению поставленных задач, приведены полученные результаты и анализ этих результатов.

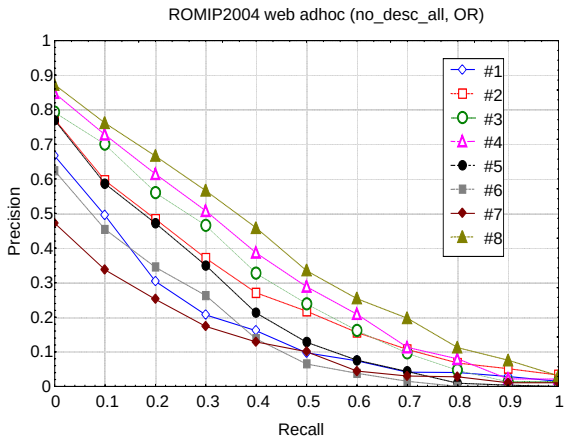

Рис.1. Результаты участников в задаче поиска по Web-коллекции

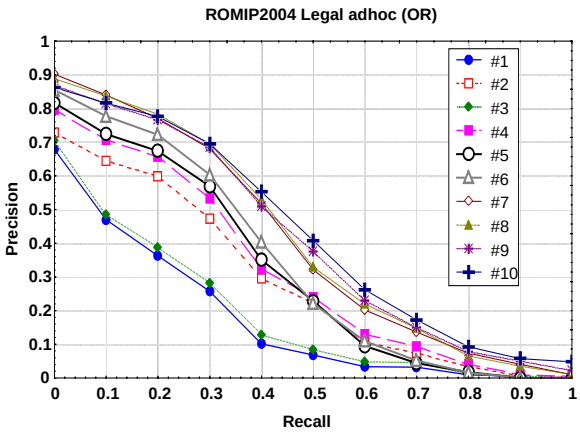

Рис.2. Результаты участников в задаче поиска по Legal-коллекции

Особое место занимает доклад организаторов (И.Некрестьянов, И.Кураленок, СПбГУ), в котором описываются принципы организации семинара, суммируются его итоги, а также приводятся результаты некоторых экспериментов с методологией оценки проведенных в рамках РОМИП'2004 ([5], С.7-26.).

В работе М.В.Губина (ИК «Кодекс») особое внимание уделено заданиям по коллекции нормативных документов, описывается процедура формирования коллекции, делается вывод о важности учета информации о взаимном расположении слов в документе ([5], С.28-38).

И.А.Барков и А.И.Барков, в своей статье приводят результаты выполнения заданий РОМИП для широкоизвестной свободно доступной ИПС mnogosearch, которую они разрабатывают ([5], С.39-42).

Разработчики компании «Гарант-Парк-Интернет» (В.В.Плешко, А.Е.Ермаков, В.П.Голенков) участвовали в выполнении всех 5 дорожек РОМИП в 2004 году, в своей работе провели исследование о влиянии учета информации о влиянии использования русского морфологического анализатора и информации о взаимном расположении слов на качество поиска ([5], С.43-61). По дорожке поиска фактов авторы делают следующие выводы. Задание было выполнено только одной системой из трех заявленных. Поэтому оценивалась только точность ответов системы (метод «общего котла», позволяющий путем объединения ответов разных систем оценить полноту, в данном случае неприменим). Несмотря на достаточно подробную инструкцию, оценки с сильным и слабым требованиями к релевантности ответов отличались в три раза, что наводит на мысль, о различном понимании термина «биографическая справка» асессорами.

Коллектив УИС РОССИЯ (НИВЦ МГУ и АНО Центр информационных исследований) – М.С.Агеев, Б.В.Добров, Н.В.Лукашевич, А.В.Сидоров – сосредоточили свои исследования в рамках РОМИП на получении т.н. «базовой линии» для обеспечения сравнения новых методов в задачах поиска и классификации с описанными в литературе ([5], С.62-89).

В.В.Рыбинкин (компания «БИТ») представил систему рубрикации текстов «Синдбад» ([5], С.90-99).

В работе И.Сегаловича, М.Маслова (ООО Яндекс) подробно описываются подходы к фильтрации и ранжированию результатов поиска, реализованные в известной российской поисковой системе ([5], С.100-109).

В работах аспирантов СПбГУ Н.Осиповой ([5], С.110-118) и М.Кондратьева ([5], С.119-132) исследуются разные алгоритмы классификации правовых и Веб-документов.

Статья А.В.Антонова, М.В.Казачука, В.С.Мешкова (Галактика-Зум) описывает подход к рубрикации Веб-документов на основе построения «портрета» рубрики ([5], С.133-141).

В приложениях к [5] подробно приведены технические детали РОМИП2004.

5. Выводы

Такая форма постоянно действующего семинара, направленного на организацию независимой оценки большого числа различных подходов для разных задач информационного поиска, каким является РОМИП, является новой в РФ.

Следует отметить, что по многочисленным свидетельствам, РОМИП чрезвычайно полезен как для участников, которые имеют возможность проверить свои алгоритмы и технологии, так и для широких слоев исследователей (прежде всего студентов и аспирантов) – как с точки зрения доступа к информации о передовых методах обработки информации (в РОМИП ведущие в области информационного поиска российские

коллективы, в том числе коммерческие компании, раскрывают широкой научной общественности существенные элементы своих подходов), так и с точки зрения использования результатов РОМИП для разработки новых методов и алгоритмов.

Основными результатами РОМИП2004 являются:

- расширение количества участников;
- увеличение количества коллекций документов, доступных для исследований по информационному поиску;
- увеличение объема выполненных исследований по сравнению с РОМИП2003;
- новые матрицы релевантности, которые могут быть использованы для разработки и отладки новых методов поиска информации;
- отчеты участников по результатам выполненных исследований, сравнительный анализ разных подходов для решения поставленных задач.

Насколько известно коллективу авторов РОМИП в настоящее время подготовлено к защите не менее четырех кандидатских диссертаций, использующих материалы РОМИП. Ожидается помещение на сайте РОМИП материалов данных работ.

Со времени проведения итогового семинара о желании участвовать в РОМИП заявили несколько новых организаций, а также представители нескольких исследовательских групп. Кроме того, уже участвующие в РОМИП команды планируют расширить спектр исследований.

Помимо совершенствования методов и алгоритмов решения задач поиска и классификации для новых (обновленных) заданий РОМИП2005, планируется начать несколько новых дорожек – организация новостей, сверхточный поиск персоналий, оценка модальности текстов и т.п.

Благодарности

Настоящее исследование частично поддержано за счет гранта Российского фонда фундаментальных исследований № 04-07-90280.

Список литературы:

1) Voorhees E., Overview of TREC 2001 // NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) - pp. 1-15.
2) Evaluation of Cross-Language Information Retrieval Systems - Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001.Revised papers. - Lecture Notes in Computer Science 2406 // C.Peters, M. Braschler, J.Gonzalo, M.Kluck (Eds.) - Springer 2002.
3) Kando N., Kuriyama K., Yoshioka M., Overview of Japanese and English Imformation Retrieval Tasks (JEIR) at the Second NTCIR Workshop // NTCIR Workshop 2. Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization May 2000 - March 2001. - National Institute of Informatics, Tokyo, Japan – 2001.
4) Труды РОМИП`2003 / Под ред. И.С.Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2003. - 132 с. (http://romip.narod.ru/romip2003/index.html)
5) Труды второго российского семинара РОМИП`2004 / Под ред. И.С.Некрестьянова - Пущино: НИИ Химии СПбГУ, 2004. - 214 с. (http://romip.narod.ru/romip2004/index.html)