ИНИЦИАТИВНЫЙ ПРОЕКТ РОССИЙСКОГО СЕМИНАРА ПО ОЦЕНКЕ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА (РОМИП)

П.И. Браславский, ИМаш УрО РАН; pb@imach.uran.ru,
М.В. Губин, «Кодекс»; max@kodeks.net,
Б.В. Добров, УИС РОССИЯ; dobroff@mail.cir.ru,
В.Ю. Добрынин, И.Е. Кураленок, И.С. Некрестьянов, Е.Ю. Павлова,
СПбГУ; {vdobr, ik}@oasis.apmath.spbu.ru, nis@acm.org, katya@meta.math.spbu.ru
И.В. Сегалович, ООО «Яндекс»; iseg@yandex-team.ru,
http://romip.narod.ru

Российский семинар по оценке методов информационного поиска (РОМИП) направлен на проведение независимой оценки альтернативных методов решения различных задач информационного поиска. В статье описываются основные принципы проведения семинара, включая не только методологические аспекты оценки, но также и организационные и административные принципы.

1. Введение

В последние годы был достигнут значительный прогресс как в теории информационного поиска, так и в создании промышленных информационно-поисковых систем.

Непрерывная эволюция информационного пространства и применение методов поиска в новых контекстах определяет актуальность дальнейших исследований в области теории информационного поиска.

Существует большое число задач, относимых к области информационного поиска:

1. • «традиционный» поиск по коллекциям с фиксированным наборов жанров (правовые системы, архивы СМИ и т.п.);
2. • наиболее востребованный ныне поиск по неконтролируемым коллекциям (Интернет);
3. • тематический или, наоборот, сверхточный поиск, вопросно-ответные системы;
4. • многоязычный поиск;
5. • учет требований разных категорий пользователей - «домохозяек», ученых, чиновников, …;
6. • и т.д.

Системы, реализующие информационный поиск, представляют сложные программные комплексы, решающие помимо собственно поиска массу «сопутствующих» задач – те или иные процедуры автоматического индексирования, аннотирования, классификации/рубрицирования текстов, сложные алгоритмы оценки потребности пользователей, разработка «интеллектуальных» интерфейсов. Для решения многих задач используются развитые лингвистические методы.

Большое количество недостаточно формализованных параметров, например, «релевантность» и «удовлетворенность пользователя», затрудняет даже для профессионалов оценку достоинств и недостатков применения разных методов при решении реальных задач. Для сближения позиций различных исследователей в области информационного поиска уже сложилось несколько форумов, имеющих по сути международный характер – например, американский TREC [1], европейский CLEF [2], японский NTCIR [3].

Однако, следует признать, что современные задачи информационного поиска в коллекциях на русском языке, равно как и других языков народов России, СНГ, не находятся в центре внимания международного научного сообщества. В частности, в TREC пару лет обсуждалась возможность устроить «дорожку» по многоязычному поиску с участием русского языка, но предпочтение было отдано испанскому, китайскому и арабскому. В CLEF2003 потребовался лишь архив русскоязычной газеты федерального уровня за 1994-1995гг.

Круг задач информационного поиска, важных для пользователей Российской Федерации значительно более широк. Для русского языка известны много развитых методов лингвистического анализа текстов, применение которых в разных задачах информационного поиска было бы очень интересно исследовать.

В данной статье описываются шаги, предпринятые группой исследователей из различных организаций, которые объединились осенью 2002 года в открытую для присоединения инициативу - Российский семинар по оценке методов информационного поиска (РОМИП).

Целями инициативы являются:

7. • проведение независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией;
8. • способствовать формированию среды для исследования феномена информационного поиска на актуальных для российского пользователя задачах;
9. • сформировать требования на оформление текстовых коллекций для тестирования (здесь есть совпадение интересов с созданием лингвистически размеченных текстовых корпусов [4]);
10. • сформировать «правила игры» - определенные этические нормы на уровень представления результатов, их использования.

2. Международные форумы по оценке методов информационного поиска

Существует несколько принципов, на которых строится работа таких форумов:

- некоммерческий характер участия в форуме, что подразумевает специально оговариваемые ограничения на использование в рекламе результатов сравнения с другими участниками. Участники находятся в неравных условиях, во-первых, разные участвующие программы «заточены» на разные задачи, а сравниваются по какой-то одной, во-вторых, оцениваются часто не программные комплексы, а методы, подходы, применимость к той или иной задаче;
- определенные требования на описание применяемых для решения задачи методов и подходов (конечно, не раскрывая коммерческих секретов), что способствует взаимному развитию всех участников.

Наиболее известным форумом по информационному поиску является TREC [1] - совместный проект NIST (National Institute of Standards and Technology, USA) и DARPA (Defense Advanced Research Projects Agency, USA), который стартовал в 1992 году.

Целью этой ежегодной конференции является обоснованная оценка эффективности подходов к решению разных задач поиска на основе больших объемов данных и продуманных методологий ее проведения.

TREC состоит из изменяющегося год от года набора секций, каждая из которых посвящена отдельной задаче текстового поиска. Участники конференции сами решают, в каких секциях им участвовать. Например, в 2000 году TREC состояла из 7 секций, и в ней приняло участие 69 организаций. Отметим, что участие в TREC подразумевает, что работа по проведению экспериментов проводится самостоятельно участниками, а вот предварительная подготовка данных и оценка результатов производится централизовано.

Деятельность TREC оказывает существенное влияние на разработку методологии оценки систем текстового поиска [5, 6]. Кроме того, созданные тестовые наборы данных и возможность объективного сравнения различных подходов к задачам поиска значительно стимулировали развитие новых методов поиска. Например, за время существования TREC показатели качества поиска систем участников улучшились в среднем в три раза [7].

Конференция TREC оказала большое влияние на процедуру проведения «родственных» конференций:

11. • CLEF (Cross-Language Evaluation Forum) [2] – европейский форум по многоязычному поиску на европейских языках, в последние годы увеличивается интерес к языкам стран Восточной Европы, а также к русскому языку;

12. • NTCIR [3] - японский семинар с интернациональными участниками по многоязычному поиску, в основном для японского/китайского и английского языков;

13. • SUMMAC (TIPSTER Text Summarization Evaluation Conference) [8] - конференция по оценке качества автоматического аннотирования (1998);

14. • MUC (Message Understanding Conference) [9] – серия конференций, направленных в основном на определении в текстах объектов, соответствующих заданным шаблонам (персоналии, организации и т.п.);

15. • TDT (Topic Detection and Tracking) [10] - проект по обнаружению новых тем в потоке новостей и отслеживанию их развития во времени, особое внимание уделяется системам обрабатывающим речь;

16. • DUC (Document Understanding Conference) [11] – конференция по вопросам автоматического аннотирования, особенно группы связанных документов.

3. Семинар РОМИП

В рамках инициативы по проведению семинара РОМИП/RIRES (Российский семинар по оценке методов информационного поиска / Russian Information Retrieval Evaluation Seminar) предлагается использовать циклический подход. В рамках годового цикла из множества реализуемых проектов по созданию тестовых наборов выбираются один или несколько наборов, которые наиболее интересны участникам. Эти отобранные проекты реализуются, а по завершении этапа с учетом накопленного опыта и текущих приоритетов участников выбираются новые проекты.

Структурно семинар представляет из себя набор «дорожек» - секций, посвященных конкретным проектам (с фиксированной задачей и правилами оценки).

Важнейшим принципом РОМИП является совместное с участниками определение задач для оценки и формирование правил проведения оценки. Оргкомитет лишь координирует проведение секций.

3.1. Структура годового цикла

На подготовительном этапе определяется список участников, уточняется список рассматриваемых задач и методология создания тестовых коллекций и оценки. Оговариваются форматы и способы обмена данными, официальные метрики для оценки. Фиксируется график проведения.

Все участники получают псевдонимы, которые будут использоваться для анонимной оценки и публикации результатов.

Оргкомитет формирует тестовые наборы данных, заданий и распространяет их участникам. В зависимости от происхождения данных может требоваться оформление соглашения о нераспространении и ограничении возможностей использования набора участником.

Участник самостоятельно и на своем оборудовании выполняет поисковые задания.

Оргкомитет организует проведение оценки (с использованием независимых экспертов) полученных ответов. Конкретная методология оценки зависит от рассматриваемой задачи и определяется на подготовительном этапе. Информация о всех оценках будет доступна всем участникам, но эта информация будет использовать псевдонимы для ссылок на участников.

Предполагается, что участники самостоятельно анализируют полученные результаты и подготовят статью, описывающую (общие) принципы их подхода и наблюдаемые результаты. При этом не обязательно раскрывать свое инкогнито и все детали реализации (это зависит от доброй воли участника) - достаточно в общих чертах описать какие известные методы использовались и что отличает их подход от других. Предоставление более подробной информации о системах, результатах и проблемах приветствуется.

Подготовленные статьи будут представлены на очном семинаре, в трудах которого они будут опубликованы.

3.2. Участники

Для того чтобы участвовать в семинаре участник должен подать заявку к рассмотрению оргкомитетом. Пока не решен вопрос с внешним финансированием, участники платят вступительный взнос (компенсирующий начальные затраты на создание, распространение наборов данных, проведение оценки), а также подписывает необходимые соглашения (лицензии).

Семинар открыт для присоединения новых участников. К участию приглашаются все заинтересованные лица - как создатели поисковых систем, так и исследователи, занимающиеся проблемами информационного поиска.

Участник свободен в определении набора дорожек, в которых он хочет участвовать, и может напрямую влиять на правила проведения этих дорожек во время их формирования. Приветствуется также предложение новых вариантов дорожек на общее обсуждение.

3.3. Принципы оценки

Процедура оценки безусловно различается для различных задач информационного поиска и формируется для конкретных дорожек, но можно выделить ряд общих основополагающих соображений:

17. • **Равноправие систем.** Процедура оценки должна по возможности гарантировать равноправие систем при оценке результатов;

18. • **Анонимность источника результата.** При проведении оценки должна соблюдаться анонимность источника результата - то есть, те, кто оценивают результат не должны знать какая система выдала этот результат;

19. • **Использование апробированных подходов.** Предпочтительным является использование апробированных методологий оценки, поскольку это повышает уверенность в получении надежных результатов.

4. Программа РОМИП на 2003 год

Первый год проведения семинара является самым сложным, так как приходится принимать много взаимосвязанных решений, не нарушающих интересы участников.

Затраты на распространение тестовых корпусов и оценку результатов в 2003 году будут совместно компенсироваться (в виде прямых финансовых вкладов или предоставлении ресурсов для проведения оценки) участниками.

4.1. Базовый набор данных

В качестве набора данных для дорожек в 2003 году используется коллекция Веб страниц из домена `narod.ru`. Поскольку общий объем страниц в домене `narod.ru` значительно превышал предполагаемый размер набора, то коллекция была сформирована на основе случайной выборки сайтов из домена. Коллекция содержит порядка 600000 HTML страниц с 22000 сайтов общим объемом более 7 Гб.

Выбор обусловлен не только высокой актуальностью задач поиска в контексте Веб, но также и легальностью доступа к этому набору данных для участников семинара. Рассматривавшиеся альтернативные варианты не Веб-коллекций были отклонены, в основном, по причине невозможности обеспечить легальный доступ за столь ограниченное время.

Для того, чтобы набор данных был максимально приближен к реальному было решено не подвергать содержимое страниц никакой модификации (подавляющее большинство страниц на русском языке в кодировке cp1251).

4.2. Секция по поиску в Веб-ресурсах

Одной из задач 2003 года является классическая задача поиска по запросу (ad-hoc track) по Веб коллекции с оценкой методом «общей котла» (pooling).

Для того, чтобы задания хорошо соответствовали набору данных набор заданий формировался на основе лога запросов к реальной поисковой системе (в данном случае Яндекс).

Для анализа результатов необходимо рассматривать результаты усредненные по группе запросов. Известно, что стабильность результатов требует, чтобы размер группы не был менее 25 запросов (рекомендуемый размер - 50 запросов) [12]. С другой стороны усреднение подразумевает, что запросы имеют приблизительно одинаковые характеристики (тип, сложность, расплывчатость, и т.п.). К сожалению, эти характеристики плохо формализуемы и автоматический выбор запросов на основе этих характеристик невозможен.

Ограниченность доступных ресурсов на проведение оценки накладывает жесткие ограничения сверху на число запросов, которые будут оценены. Однако, это не ограничивает число запросов, которые могут быть выполнены системами.

Набор запросов для выполнения системами состоит из 10000 запросов. Такой подход дает нам возможность не только бороться с возможной фальсификацией результатов, но также и отложить выбор оцениваемых 50 на более поздний срок.

Расширенное множество из 10000 запросов выбирается полностью автоматическим способом. Из протокольной записи запросов Яндекса, начиная с некоторой точки берутся все запросы, удовлетворяющие критериям отбора:

20. • русскоязычные;

21. • без явных грамматических ошибок.

4.3. Процедура оценки

Для оценки предлагается подход «общего котла» (pooling), который используется в TREC. «Общий котел» - это объединенное множество первых N_q документов из выдачи каждой из систем для данного запроса q.

При проведении оценки методом общего котла необходимо, чтобы все оценки релевантности для одного и того же запроса делались исходя из одного и того же понимания информационной потребности (а иначе собранные оценки будут несогласованны).

Очевидно, что короткие запросы, используемые в этой дорожке, зачастую могут трактоваться несколькими разными способами. Для того, чтобы обойти эту проблему в РОМИП используется следующий подход. Небольшое множество экспертов, отбирающих запросы для оценки, для каждого отобранного задания создают расширенную версию задания, которая содержит более детальное описание искомой информации как это понимает эксперт (тем самым уточняется одна из возможных информационных потребностей, выраженная этим запросом).

Именно эта, «расширенная», версия и используется в дальнейшем для сбора оценок релевантности. Для того, чтобы различать разные роли в РОМИП используется термин «эксперт» для лиц, фиксирующих информационную потребность, и «оценщик» для лиц, реально производящих оценку руководствуясь расширенным заданием.

Входная информация для оценки:

22. • набор тестовых заданий;

23. • выдачи систем (упорядоченный набор документов) для каждого из (всех) тестовых заданий;

24. • доступные ресурсы (оценщики и эксперты)

Процедура оценки состоит из нескольких этапов.

Фиксируются параметры оценки. Определяется общее количество оценок N соответствия пар "документ-запрос", на сбор которых достаточно ресурсов. Исходя из N определяется глубина N_T так чтобы общий размер всех котлов глубиной N_T был приблизительно равен N.

Выбирается подмножество (по предварительному плану из 50-ти) тестовых заданий для проведения оценки. Выбор производится экспертами на основе просмотра списка тестовых заданий без использования информации о содержании выдач систем. Для каждого отобранного задания эксперт создает его расширенную версию, которая уточняет трактовку запроса. Для каждого из запросов формируются "котлы" - объединенное множество некоторого количества первых документов из выдачи каждой из систем для данного запроса.

Для проверки оценщику предоставляется документ и расширенное описание задания. Оценщик не будет обладать информацией какими системами и на какой позиции был возвращен данный документ. Документы предоставляются оценщику по одному в случайном порядке (выбор не связанном с порядком выдач или системами). Все оценщики для оценки будут использовать один и тот же интерфейс.

При принятии решения все оценщики будет руководствоваться общей постановкой задачи: «Документ считается релевантным, если, встретив этот документ в процессе поиска информации по данному вопросу (описание расширенного задания), вы сочли бы этот документ достойным дальнейшего прочтения».

В дополнение в вариантам ответа «да» и «нет» на вопрос о релевантности документа оценщик может также ответить «невозможно оценить» в случае если рассматриваемая страница в силу каких-либо причин не понятна (не та кодировка или язык, страница не отображается и т.п.)

Предполагается, что для проверки одного документа оценщику в среднем будет необходимо порядка 1 часа на 100 документов [13].

Вычисление окончательных итоговых оценок систем производится анонимно на основе набора стандартных метрик (точность, полнота, т.п.). Инструмент для вычисления оценок и об оценках пар «документ-запрос» будут доступны всем участникам.

4.4. Тематическая классификация Веб-сайтов

В дополнение к задаче поиска обсуждается дорожка тематической классификации Интернет-сайтов.

Задан список категорий, обучающая выборка и множество сайтов (не документов!). Надо присвоить каждому из сайтов коллекции категорию из этого списка с учетом обучающей выборки.

Множество классов формируется на основе каталога narod.ru (http://narod.yandex.ru/rubrics/).

Процедура оценки организована следующим образом:

25. • выбирается (случайным образом) несколько (заранее неизвестных участникам) категорий;

26. • все сайты, которым хотя бы одна из систем присвоила одну из этих категорий проверяются на соответствие этим категориям. (ответ эксперта бинарен – «да» или «нет»).

Такой подход позволяет оценивать не только точность классификации, но и позволяет аппроксимировать полноту. Количество проверяемых категорий (2-3-4-5-10) определяется исходя из объема доступных ресурсов и размера полученных «котлов» для каждой из категорий.

5. Текущий статус и планы

В феврале 2003 года первый семинар РОМИП вступил в фазу регистрации участников. Предполагается, что годовой цикл будет завершен к октябрю и очная встреча будет совмещена с Российской конференцией по электронным библиотекам (RCDL'2003), которая пройдет в Санкт-Петербурге 29-31 октября.

5.1. Новые «дорожки»

То, из каких дорожек будет состоять следующий семинар определяется исходя из интереса участников и возможностей по организации дорожек. Более формально процедура выбора состоит из следующих шагов:

Формируется множество «возможных реализуемых» дорожек. «Возможная» дорожка - это любая дорожка, подходящая под тематику семинара. Множество возможных дорожек открыто и каждый заинтересованный участник может предлагать свои варианты на общее обсуждение. Для того чтобы дорожка получила статус «реализуемой» необходимо иметь полное описание, а также обоснование доступности необходимых ресурсов (данных, оценке экспертного времени, т.п.)

По каждой дорожке производится открытое голосование. Целью голосования является определить заинтересованность каждого из участников в каждой из возможных дорожек (можно заявляться на участие в нескольких дорожках). Выбираются наиболее популярные дорожки.

Описание дорожки включает в себя ответы на следующие вопросы:

27. • для оценки методов решения какой задачи дорожка предназначена;

28. • какой набор данных предполагается использовать? (с указанием характеристик - объема, легальности, разнородности, и т.п.);

29. • какие будут задания? Сколько? Как они будут формироваться?

30. • в каком виде предполагается получать ответы от систем?

31. • как будет организована процедура оценки результатов? Сколько ручного труда необходимо и каковы предполагаемые затраты на проведение оценки?

32. • какие меры могут быть использованы для оценки

33. • что мотивирует «осмысленность» получаемых цифр и основанных на них выводов о превосходстве тех или иных методов? Например:

34. o стабильность результатов относительно количества заданий;
35. o стабильность относительно процедуры оценки (порядка оценки или других факторов связанных с экспертами);
36. o защищенность от фальсификации результатов участниками.

Литература

1. 1. Voorhees E., Overview of TREC 2001 // NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) - pp. 1-15.
2. 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. (http://clef.iei.pi.cnr.it:2002/)
3. 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. 4. БОКР (Большой Корпус русского языка) (http://bokrcorpora.narod.ru/)
5. 5. Sparck Jones K., Reflections on TREC // Information Processing and Management, 31(3):291-314, 1995.
6. 6. Voorhees E., Variations in relevance judgments and the measurementt of retrieval effectiveness // Proc. of the SIGIR'98 – pp. 315-323.
7. 7. Harman D., What we have learned, and not learned, from TREC // Proc. of the BCS IRSG'2000 - pp 2-20.
8. 8. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/ovview-kando2.pdf)
9. 9. The TIPSTER SUMMAC Text Summarization Evaluation. Final Report // MITRE Technical Report - MTR 98W0000138 – 1998.
10. 10. (http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/final_rpt.html)
11. 11. MUC-7. Message Understanding Conference Proceedings. (http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/muc_7_toc.html)
12. 12. TDT Phase 2 (http://www.ldc.upenn.edu/Projects/TDT2/)
13. 13. DUC 2002. Workshop on Text Summarization, Philadelphia, Pennsylvania, USA // Eds.: U. Hahn, D. Harman – 2002. (http://www-nlpir.nist.gov/projects/duc/pubs.html)
14. 14. Zobel J., How reliable are large-scale information retrieval experiments? // Proc. of the SIGIR'98 - pp. 308-315.
15. 15. Cormack G.V., Palmer C.R., Clarke C.L.A., Efficient construction of large test collections // Proc. of the SIGIR'98 - pp. 282-289.