Система интеллектуального поиска и анализа информации «Exactus» на РОМИП-2010

© Завьялова О.С., Киселёв А.А., Осипов Г.С.,

Смирнов И.В., Тихомиров И.А.

Учреждение Российской академии наук Институт системного анализа РАН

© Соченков И.В.

Российский университет дружбы народов / Учреждение Российской академии наук Институт системного анализа РАН

tih@isa.ru

Аннотация

В статье представлены результаты участия проекта EXACTUS в семинаре РОМИП-2010 по дорожкам поиска и классификации web-страниц. Проведён анализ оценок качества дорожек поиска и классификации.

Введение

С 2005 года система семантического поиска и анализа текстовой информации EXACTUS является неизменным участником семинара РОМИП. В 2010 году система EXACTUS участвовала дорожках

1) web-поиска (по коллекциям BY.WEB и KM.RU),
2) классификации web-страниц;
3) вопросно-ответного поиска.

В 2010 году были проведены эксперименты с алгоритмом ранжирования результатов поиска с целью получения ответа на вопрос о вкладе учёта семантической составляющей в улучшение качества ранжирования результатов поиска. Кроме того, были предложены новые формулы определения информационной

значимости слов естественного языка (ЕЯ) в текстовых документах и модифицирована система обучения поисковых алгоритмов.

Значительные изменения претерпел и метод классификации web-документов: разработан и реализован способ учёта составных (двухсловных) терминов ЕЯ. Целью участия в дорожке классификации web-страниц являлась комплексная проверка модернизированного метода на гипертекстовых коллекциях, предоставленных РОМИП.

1. Анализ результатов участия системы EXACTUS в дорожках web-поиска РОМИП-2010

1.1. Дорожка поиска по коллекции BY.WEB

Анализ результатов участия системы EXACTUS в дорожке поиска по коллекции BY.WEB выполнен на основе оценок *AND-relevant-plus*. На рисунке 1 представлены результаты оценок прогонов участников по дорожке поиска BY.WEB.

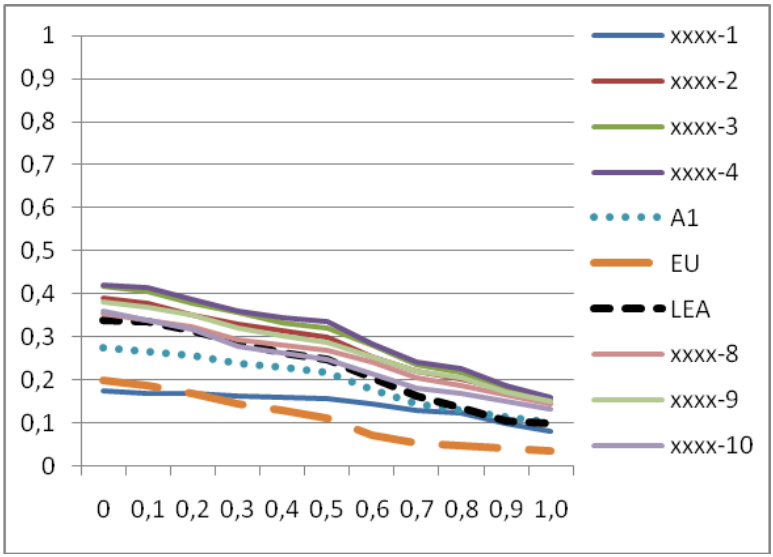

Рисунок 1. Графики TREC для коллекции BY.WEB

В 2010 году коллективом разработчиков Exactus проверены три модификации статистического подхода к

определению значимости слов в текстовых документах. Все три модификации проверялись в составе метода ранжирования, учитывающего семантическую информацию ЕЯ [1], [2]. Приведём описания этих модификаций:

1. LEA – модифицированная версия метода определения информационной значимости слов в гипертекстовых документах:

$$U(w,\tau) = \log_{C(\tau)+1}\left(c(w,\tau)+1\right), \qquad (1),$$

где величина $U(w,\tau)$ зависит от лексемы ЕЯ w и текста τ, представляющего собой совокупность вхождений лексем. $C(\tau)$ – есть общее количество вхождений лексем в текст τ, а $c(w,\tau)$ – число вхождений лексемы w в текст τ.

Для учёта гипертекстовой разметки применяется изменённый вариант формулы (1), где $C(\tau)$ – суммарный вес всех вхождений лексем в текст, $c(w,\tau)$ – суммарный вес всех вхождений лексемы w в текст τ. Вес отдельного вхождения определяется окружающей гипертекстовой разметкой.

2. E1 – метод, основанный на формуле определения информационной значимости слов в гипертекстовых документах:

$$E(w,\tau) = \gamma \cdot TF(w,\tau) \cdot \log_2\left(\frac{1}{\gamma \cdot TF(w,\tau)}\right), \qquad (2),$$

где γ –параметр, а $TF(w,\tau)$ –частота встречаемости лексемы w в текст τ [1].

3. A1 – метод на основе формулы

$$A1(w,\tau) = \log_2\left(1+\sqrt[8]{TF(w,\tau)}\right), \qquad (3).$$

Алгоритм ранжирования, в котором был использован этот метод, показал наилучшие результаты по большинству параметров в дорожках web-поиска на РОМИП 2008 [2]. Для

чистоты эксперимента в 2010 году был сдан тот же прогон алгоритма A1, что и в 2008 году.

Все три метода учитывают гипертекстовую разметку документов: на веса вхождений лексем, рассчитанные по формулам (1)–(3), влияют теги HTML в зависимости от их важности. Например, заголовкам соответсвют большие веса, чем основному тексту. Веса тегов подбирались на основе эмпирических соображений и не подвергались оптимизации.

Указанные методы зависят от небольшого числа параметров (см. [1]), оптимизация значений которых проводилась на основе таблиц релевантности по коллекции BY.WEB 2007–2009 годов (and–relevant-minus).

При обработке поисковых запросов в 2010 году не использовались механизмы коррекции опечаток, транслитерация запросов, методы ссылочного ранжирования, тематические каталоги или иная информацию, не относящуюся к текстовому содержанию web-страниц и коллекции BY.WEB. Основное внимание было сосредоточено исключительно на статистических [1] и лингвистических [2], [4], [5] составляющих метода ранжирования результатов поиска.

Для полноты сопоставления трёх алгоритмов ранжирования приведём сравнение графиков TREC, полученные на основании таблиц релевантности *and-relevant-plus* (рисунок 1) и *or-relevant-minus* (рисунок 2).

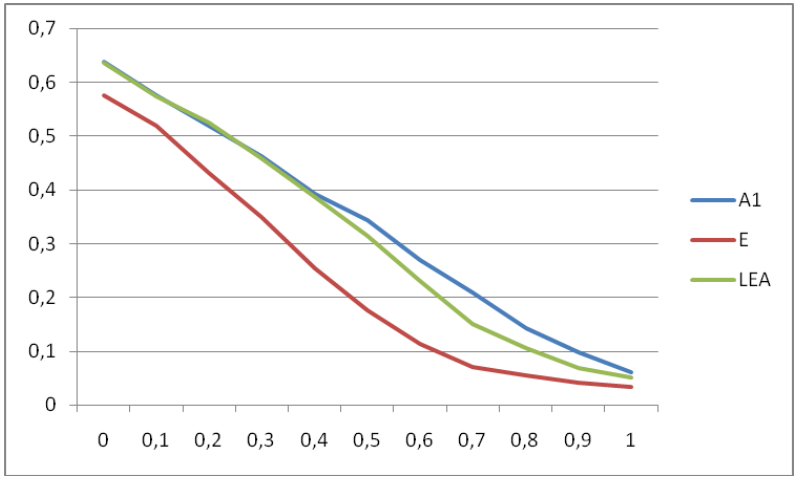

Рисунок 2. Графики TREC прогонов EXACTUS для коллекции BY.WEB (or-relevant-minus)

Из графиков следует, что алгоритм LEA выдаёт результаты, в большей степени удовлетворяющие информационною потребность обоих асессоров, нежели A1 (рисунок 1, соответствующий сильным требованиям к релевантности). При этом оба алгоритма практически идентичны, когда речь идёт о слабых требованиях к релевантности.

В 2010 году были проанализированы оценки качества ранжирования результатов поиска на двух классах запросов: повторно оцениваемых (для которых имелись таблицы релевантности за предыдущие годы) и впервые оцениваемых, не имеющих оценок за предыдущие годы. На рисунке 3 представлено сопоставление показателей качества ранжирования на множествах повторных и новых запросов. Повторным запросам на рисунке 3 соответствует префикс «OLD-» (41 оценённый запрос), оцененным впервые – «NEW-» (456 оценённых запросов).

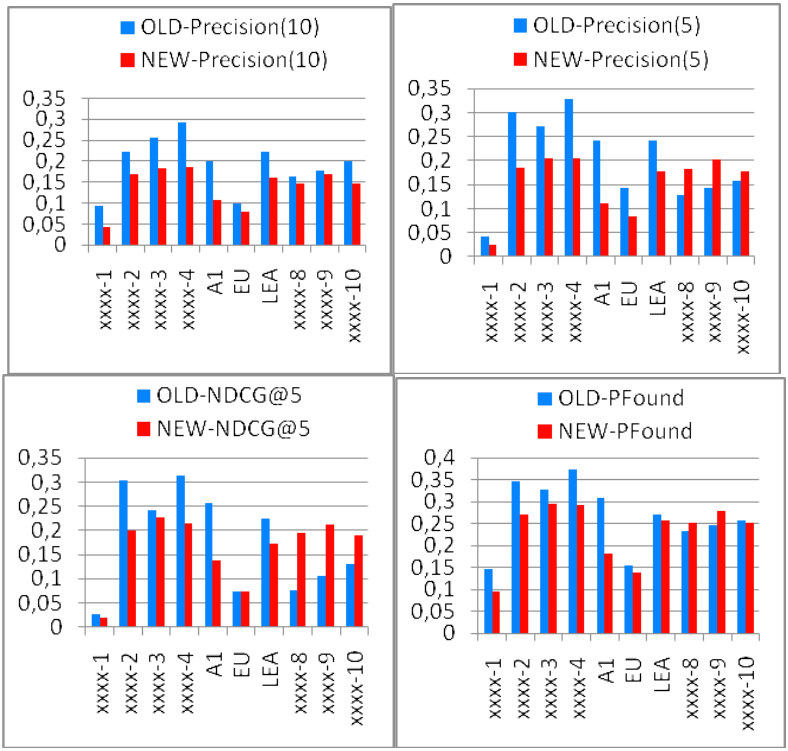

Рисунок 3. Сравнение показателей качества ранжирования на множествах ранее оцененных и новых запросов (дорожка BY.WEB)

Приведённое сравнение позволяет дать ответ на вопрос о стабильности результатов поиска в зависимости от метода ранжирования и способов его обучения. Из рисунка видно, что все три метода EXACTUS показывают на новых запросах более низкие результаты, нежели на тех запросах, для которых имелись таблицы релевантности. Причём такая тенденция имеет место для прогонов xxxx1–xxxx4. Для прогонов xxxx8–xxxx10 имеет место обратная тенденция. По-всей видимости, причиной этого является использование каких-то внешних факторов, влияющих на алгоритм ранжирования, и не зависящих от коллекции BY.WEB.

Следует заметить, что в прогоны системы EXACTUS ранее оцененные web-страницы не подмешивались искусственным образом, а также для них не было создано никаких искусственных признаков, позволяющих повысить их позицию в поисковой выдаче по ранее оцененному запросу.

Интерес представляет тот факт, что наряду с прогонами 2010 года прогон A1 (выполненный в 2008 году!) оказался значительно лучше на повторно оцененных запросах, нежели на новых. Это свидетельствует о нестабильности оценок на относительно небольшом классе повторно оцененных запросов, что не позволяет говорить о репрезентативности подобного сравнения двух классов запросов.

Приведём сопоставление оценок за предыдущие годы для прогонов EXACTUS. На рисунке 4 представлены графики TREC для прогонов по коллекции BY.WEB в 2008—2010 г.г (and-оценка, relevant-minus).

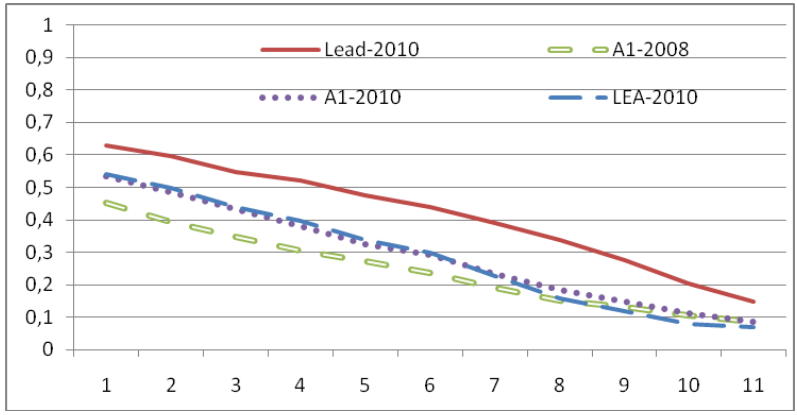

Рисунок 4. Графики TREC для прогонов по коллекции BY.WEB в 2008—2010 г.г (and-оценка, relevant-minus).

Обращает на себя внимание разрыв между прогоном-лидером 2008 года (A1-2008) и прогоном-лидером 2010 (Leader-2010). Отметим, что тот же самый прогон A1 (выполненный в 2008 году), по оценкам 2010 года (A1-2010) показывает результаты лучше, нежели по оценкам 2008 года (A1-2008): разница в левой и средней части графика составляет 0.1, что соответствует 20% абсолютной величины. Это ставит вопрос о сопоставимости оценок асессоров в разные годы. Перечисленные факты свидетельствуют о несравнимости оценок за разные годы, которые зависят от множества оцениваемых запросов и от выдачи участников, попадающей в пул. По этой причине прямое сопоставление абсолютных величин оценок, полученных в разные годы, лишено практического смысла.

1.2. Дорожка поиска по коллекции KM.RU

В дорожке web-поиска по коллекции KM.RU участвовали 2 прогона системы EXACTUS: метод LEA, в котором отключена семантическая составляющая алгоритма («No sem») и полноценный метод, учитывающий семантику запросов ЕЯ, аналогичный тому, что принимал участие в дорожке BY.WEB («Sem»).

В целом, процедура участия не отличалась от аналогичной для BY.WEB. Оптимизация параметров алгоритмов производилась на таблицах релевантности 2007-2009 г.г. для коллекции KM.RU. Для определения информационной значимости слов ЕЯ в текстах использовалась формула (1).

При анализе результатов участия системы EXACTUS в дорожке поиска по коллекции KM.RU рассматривались оценки AND-*relevant-plus*. Такой выбор был продиктован желанием выяснить, насколько система способно удовлетворять потребности нескольких пользователей, предъявляющих строгие требования к релевантности. На рисунке 5 представлены графики TREC для коллекции KM.RU.

На рисунке 6 представлены показатели качества ранжирования результатов поиска по дорожке KM.RU. Из рисунка видно, что результаты оценок всех участников расположены очень «плотно». Разница между «лучшим» и «худшим» по метрикам в среднем менее 0.06. Все участники сопоставимы по точности 10 и полноте. Метод ранжирования с учётом семантики чуть лучше аналога без учёта семантики по метрике PFound и по точности на уровне 10, но уступает по метрикам DCG и по полноте. При сопоставимой (и почти совпадающей для всех участников!) точности, и точности на уровне 10, семантический метод обладает меньшей полнотой в сравнении с другими участниками.

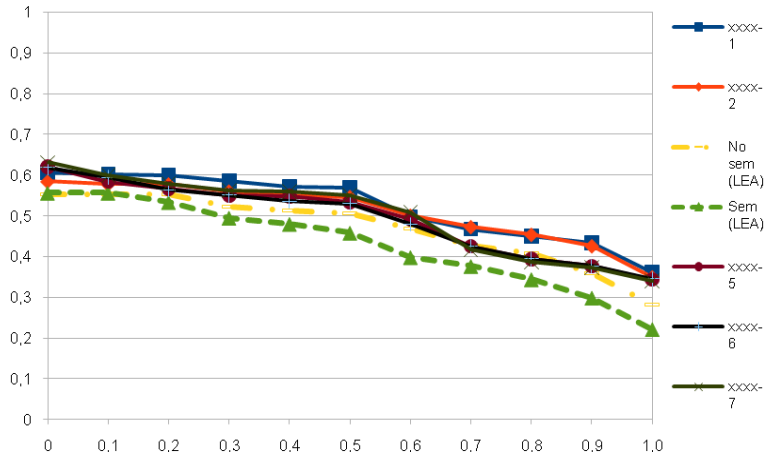

Рисунок 5. Графики TREC для коллекции KM.RU

Подобная плотность результатов объясняется следующим образом. Большинство участников используют схожие признаки и критерии ранжирования, поэтому их результаты достаточно схожи. Оптимизируя свои алгоритмы на таблицах релевантности, полученных в предыдущие годы, участники «сближают» свои методы ранжирования. При этом таблицы релевантности далеко не полны (в силу использования метода «общего котла» и неполноты выдачи участников).

При введении нового критерия ранжирования, оказывающего значительное влияние на состав и порядок документов в выдаче, затруднительна оптимизация параметров алгоритма ранжирования на имеющихся таблицах релевантности. Таблицы релевантности, полученные при оценке документов в «общем котле» зачастую не полны. Оптимизация метода ранжирования на имеющихся таблицах релевантности приводит к тому, что влияние нового критерия нивелируется.

Высказанное предположение подтверждается экспериментально:

1) Из рисунка 6 видно, что результаты всех участников достаточно близки, различия между ними минимальны. Та же самая картина наблюдается для оценок на глубине пула 50, в особенности для метрик точности и полноты.

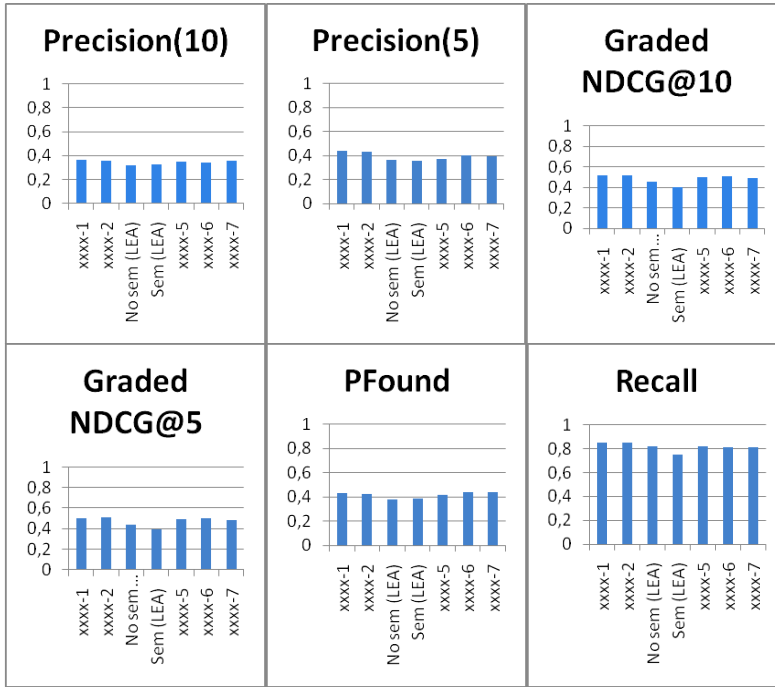

Рисунок 6. Показатели качества ранжирования результатов поиска по дорожке KM.RU

2) Метод ранжирования, учитывающий семантическую информацию, был оптимизирован достаточно слабо: все профили параметров были достаточно «плохими» по оценкам качества поиска на существующих таблицах релевантности. Фактически, был выбран профиль параметров, который можно охарактеризовать, как «наилучший из худших».
3) 30% оцененных в 2010 году запросов содержат семантическую информацию. Однако это не помешало получить результаты, сопоставимые по качеству с результатами прогона метода, не использующего семантическую информацию – см. рисунок 6.
4) Метод ранжирования, не использующий семантическую информацию запроса, показывал высокое качество поиска на таблицах релевантности во время обучения, однако на итоговых оценках он почти совпадает с семантическим методом.

Следует заметить, что метод «общего котла» влияет на итоговые оценки полноты и точности семантического метода.

Однако это влияние не затрагивает метрики на уровне K документов и PFound. Для оценок, полученных на уровне глубины пула, влияние метода «общего котла» также незначительно.

На рисунке 7 представлены диаграммы, сопоставляющие параметры качества поиска для несемантических запросов (префикс “NO-SEM-”) и запросов, содержащих семантическую информацию (префикс “SEM-”).

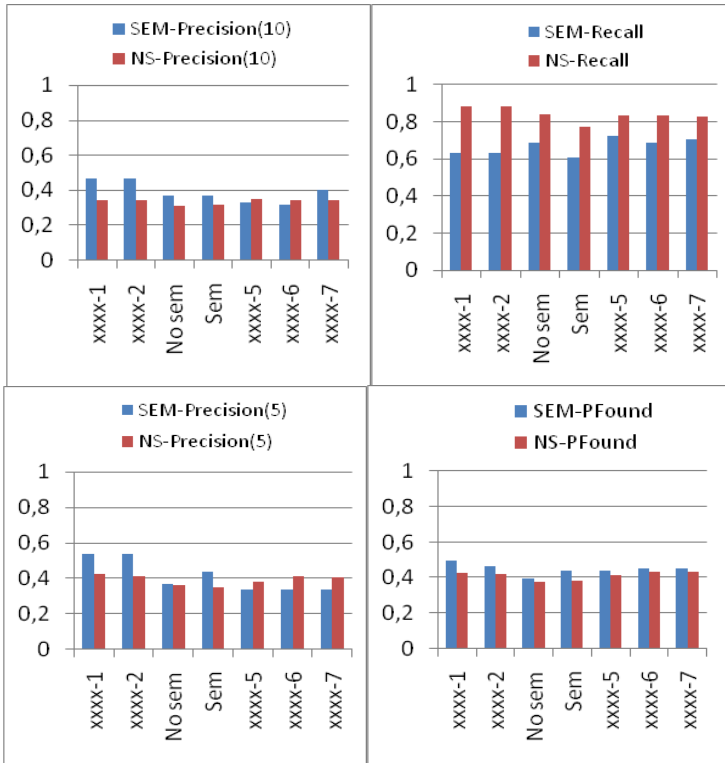

Рисунок 7. Сравнение показателей качества ранжирования на множествах семантических и несемантических запросов (дорожка KM.RU)

Как и ожидалось, метод, использующий семантическую информацию запроса, показывает лучшие результаты на семантических запросах, нежели метод, семантическую информацию не учитывающий (метрики PFound, точность на уровне 5). Так как количество документов, релевантных семантическим

запросам, как правило, меньше числа документов, релевантным несемантическим запросам (см., например, [1]), и исчисляется несколькими документами, то разница между методами становится менее заметной на метрике точности на уровне 10. Это же предположение подтверждается сопоставлением графиков полноты по обоим классам запросов.

2. Метод автоматической классификации Web-страниц и его результаты на РОМИП-2010

В ходе РОМИП-2010 были испытаны 2 модификации метода автоматической классификации гипертекстовых документов на основе характеристики тематической значимости (ХТЗ) [1], [3].

Результатам системы EXACTUS на рисунках 8 и 9 соответствуют прогоны, помеченные “AW”, “2W”. При анализе результатов мы опирались на and–relevant-minus оценку.

Оба прогона системы EXACTUS по дорожке тематической классификации web-страниц учитывали двухсловные термины.

Для выделения двухсловных терминов применялся модифицированный синтаксический анализатор АОТ (http://aot.ru). В построенных синтаксических деревьях предложений по шаблонам выделялись словосочетания следующего вида:

1) существительное + существительное;
2) прилагательное + существительное;

причастие + существительное.

С целью уменьшения количества выделенных терминов на этапе обучения периодически производилось редуцирование тех терминов, которые встречаются только в одном документе на каждые 500 тыс. Аналогичная процедура производилась при обучении для каждой тематической категории: отбрасывались термины, встретившиеся в менее, чем 7% документов в части обучающей выборки, относящейся к категории.

При классификации документов метод AW рассматривал в качестве терминов как отдельные слова, так и двухсловные сочетания. Метод 2W осуществлял классификацию, опираясь исключительно на двухсловные термины.

Для определения информационной значимости терминов ЕЯ в тексте применялась формула (1).

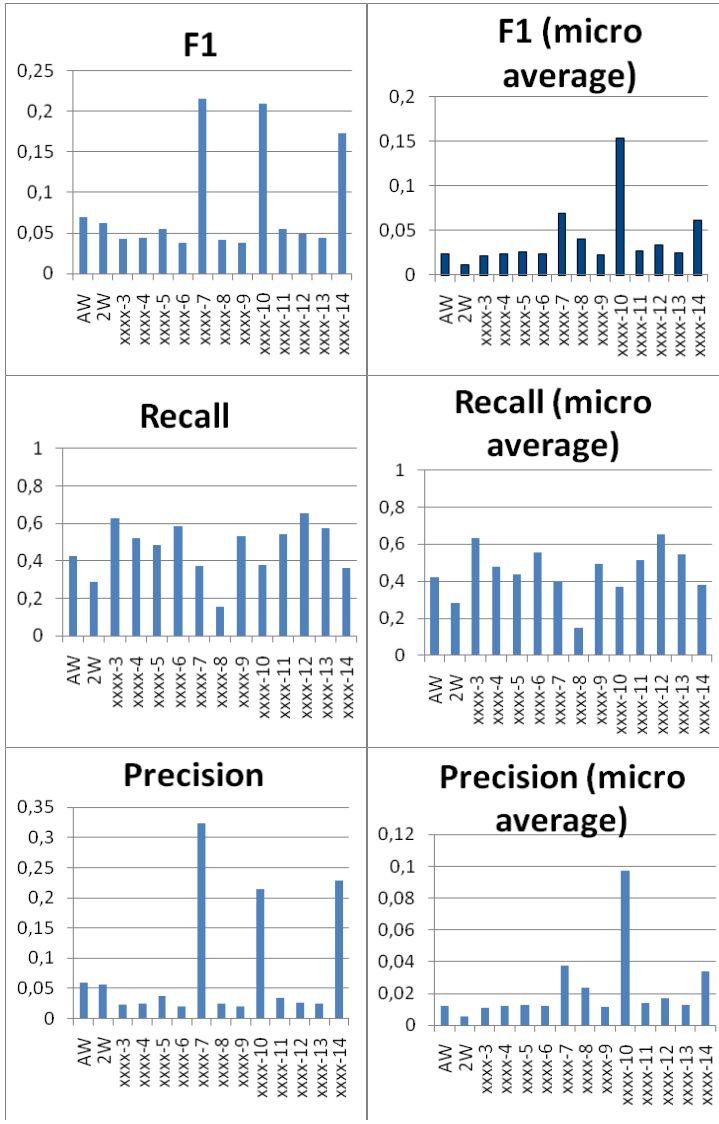

Рисунок 8. Сравнение оценок качества классификации web-страниц (and оценка по всем web-страницам)

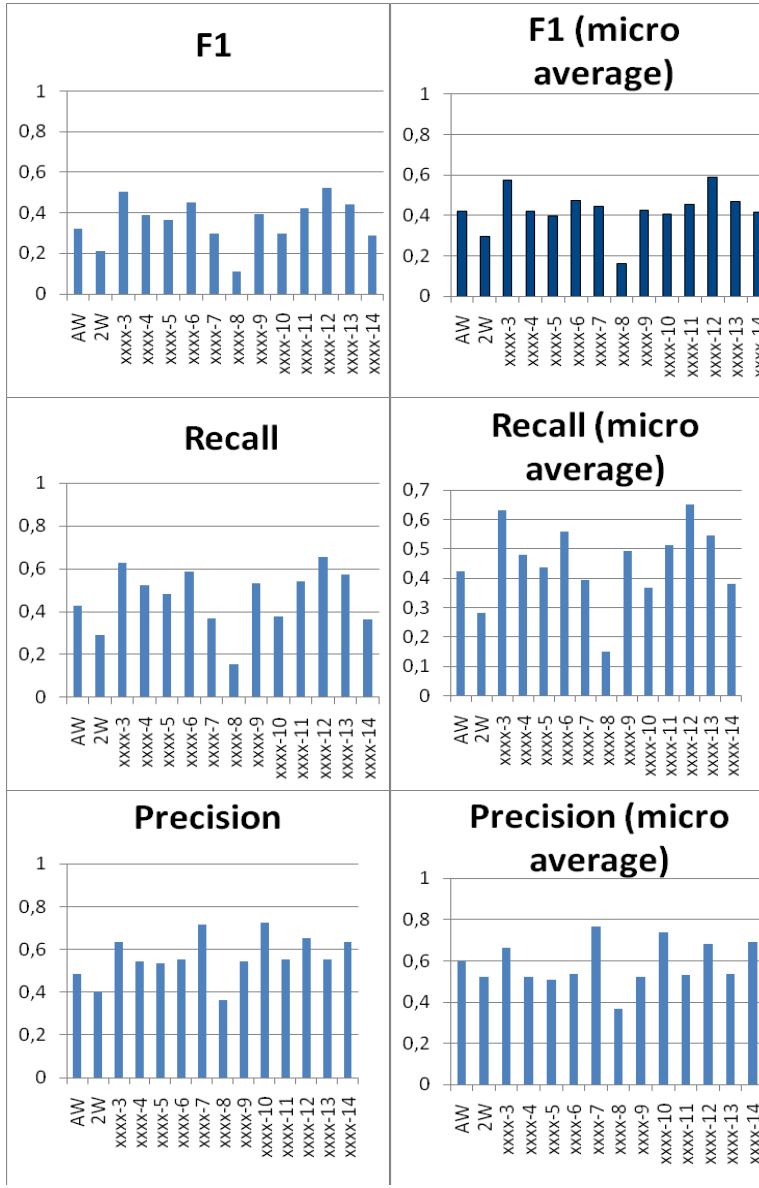

Рисунок 9. Сравнение оценок качества классификации web-страниц (and оценка по web-страницам, имеющим оценки)

Анализ приведённых результатов оценки показывает, что для дорожки классификации РОМИП больше подходит метод AW. Метод 2W уступает AW по полноте.

Сопоставление микро- и макроусреднений оценок говорит о том, что в число оцениваемых категорий попадали как тематически хорошо очерченные категории, с репрезентативной обучающей выборкой, так и категории более общие, для которых обучающая выборка обладает меньшей репрезентативностью. Это подтверждает анализ обучающего множества для данной дорожки.

Интерес представляет тот факт, что явные лидеры по полным оценкам (прогоны xxxx-7, xxxx-10, xxxx-14 на рисунке 8) значительно опускаются на оценках, полученных с учётом только оценённых web-страниц (рисунок 9). Прогоны EXACTUS сопоставимы по результатам полных оценок с большинством участников. По метрикам, рассчитанным только на документах, имеющих оценки, прогоны EXACTUS сопоставимы по качеству с большинством участников, незначительно уступая лидерам (прогоны xxxx-3, xxxx-12 на рисунке 9).

В перспективе планируется испытать другие механизмы оптимизации представленных методов классификации, сделав акцент на отдельную метрику по выбранной оценке с сохранением приемлемого качества по остальным метрикам и оценкам. Методы будут также модифицированы для проверки следующего предположения: если классификатор относит документ более, чем к N классов, то на практике такой документ тематически неоднороден, и де факто не относится ни к одной из рубрик.

3. Результаты участия системы Exactus в дорожке вопросно-ответного поиска РОМИП-2010

В 2010 году в программу семинара РОМИП вернулась дорожка вопросно-ответного поиска, состоявшаяся впервые с 2007 года. Отличительной особенностью этой дорожки являются оцениваемые запросы, сформулированные в виде вопросительного предложения на естественном языке. Это позволяет определять семантическую информацию в запросе и сопоставлять её с информацией, содержащейся в текстах документов коллекции, в соответствии с моделью реляционно-ситуационного поиска [4],[5].

Система Exactus была представлена в дорожке вопросно-ответного поиска одним прогоном: «Exactus-LEA» - алгоритм поиска, учитывающий семантическую информацию запроса. Это та же версия алгоритма с настроечными параметрами, что принимала

участие в дорожке поиска по Web-коллекции BY.WEB (см. п.1.1). Выделение ответа на вопрос и формирование контекста производилось с помощью простейшего алгоритма аннотирования: выдавалось предложение, содержащее, по мнению системы, ответ на заданный вопрос.

Результаты оценок прогонов участников представлены на рисунке 10.

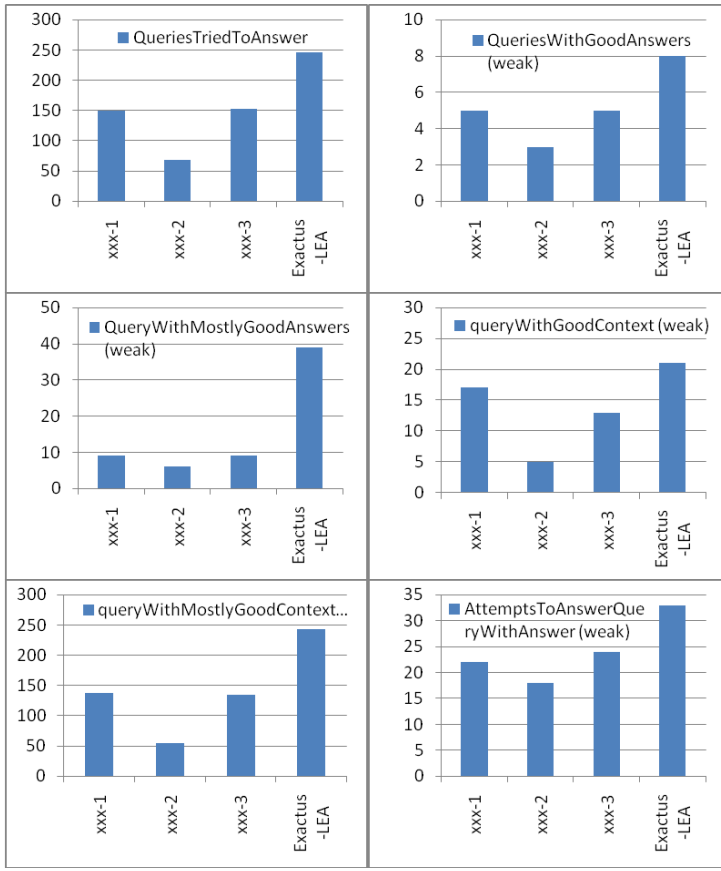

Рисунок 10. Оценки прогонов участников дорожки вопросно-ответного поиска РОМИП-2010

Прогон Exactus-LEA показал лучшие результаты по всем метрикам. Невысокие по абсолютной величине результаты объясняются с одной стороны, несоответствием вопросов

содержанию коллекции BY.WEB. С другой стороны имеют место неточности в работе самого алгоритма. Планируемый детальный анализ оценок асессоров для результатов EXACTUS позволит исправить погрешности в определении границ предложений, снятии омонимии, построении семантической структуры предложения. Впоследствии это позволит повысить точность и полноту вопросно-ответного поиска в системе EXACTUS.

Участие в дорожке вопросно-ответного поиска РОМИП-2010 подтвердило перспективность развития моделей реляционно-ситуационного представления текстов, а также их применимость в прикладных вопросно-ответных поисковых системах, работающих с большими объёмами ЕЯ информации в Web.

Заключение

Подводя итоги участия системы EXACTUS в семинаре РОМИП-2010, хотелось бы отметить следующее.

Введённые организаторами РОМИП метрики DCG, PFound и их вариации позволяют гораздо глубже оценивать качество информационного поиска.

Проведённые эксперименты позволяют сделать вывод, что на основе имеющихся таблиц релевантности поиск новых факторов ранжирования затруднен: обучение алгоритмов поиска приводит к тому, что все участники формируют очень схожие выдачи. Это означает, что алгоритмы «обучаются» на выдачу результатов друг друга за прошлые годы. Настроенные таким образом алгоритмы переносят скрытые закономерности и на новые запросы. Однако, получаемые оценки качества на новых запросах, как правило, ниже аналогичных на старых запросах.

Анализ результатов участия в дорожке тематической классификации web-страниц оставил открытым вопрос, какой из оценок качества классификации (полной или только по оценённым документам) следует отдать предпочтение, если мы хотим проверить качество решения задачи классификации (в классической формулировке).

По результатам участия в РОМИП-2010 целесообразными представляются следующие предложения по проведению будущих семинаров РОМИП:

1. Необходимо создание новой коллекции документов (3-10 млн. web-страниц), представляющей собой фрагмент сети Интернет. При этом коллекция должна обладать развитой ссылочной структурой, актуальностью данных, и иными свойствами,

характерными для сети Интернет. Это должно приблизить дорожку к классической задаче Web-поиска, что позволит учитывать и находить новые факторы, влияющие на качество ранжирования. Это также позволит оценить алгоритмы участников на новых данных (в настоящее время наблюдается сильное сближение и завязывание в плотную косичку результатов участников на TREC-графиках). Отсутствие таблиц релевантности для новой коллекции позволит проверить устойчивость методов поиска, обученных на таблицах релевантности для других коллекций. Смену или актуализацию коллекции по нашему мнению нужно проводить не реже чем раз в три цикла РОМИП.

2. Считаем необходимым актуализировать множество запросов, предлагаемое участникам. В настоящее время ряд оценённых запросов слабо соотносится с тематикой материалов, представленных в коллекциях KM.RU и особенно BY.WEB, что ведет к общему занижению результатов по абсолютной шкале, а также вызывает затруднения у асессоров при оценке документов.
3. Для минимизации потерь информации при согласовании оценок асессоров до уровней relevant-minus, relevant-plus предлагается открыть таблицы релевантности, содержащие исходные мнения асессоров по шкале not-relevant, relevant-minus, relevant-plus, vital. Это позволит участникам оптимизировать свои методы ранжирования, опираясь на метрики Graded NDCG, PFound и др.
4. Для дорожки тематической классификации web-страниц предлагается использовать схему оценки со случайной выборкой N результатов, возвращённых системой для каждой оцениваемой категории. Здесь N – глубина пула. Это позволит адекватно оценить те методы, которые не ранжируют обработанные документы внутри класса (что ближе к задаче классификации, а не поиска).

Литература

[1] Смирнов И.В., Соченков И.В., Тихомиров И. А. Система интеллектуального поиска и анализа информации Exactus на Ромип'2009 //Труды российского семинара по оценке методов информационного поиска РОМИП'2009. Санкт-Петербург: НУ ЦСИ, 2009. - С. 41-52.

[2] Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И. А. Результаты и перспективы поискового алгоритма Exactus. // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. Санкт-Петербург: НУ ЦСИ, 2008, с. 66-76.
[3] Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке. // Вестник НГУ, Информационные технологии, т. 6, Вып. 2, Новосибирск, 2008, с. 94-100.
[4] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Olga Zavjalova. Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. - P. 17-2 - 17-5.
[5] 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.

The system of information search and analysis EXACTUS on ROMIP-2010

© Alexander A. Kiselyov, Gennady S. Osipov, Ivan V. Smirnov, Ilya A. Tikhomirov, Olga S. Zavjalova

Institute for Systems Analysis of Russian Academy of Sciences

Ilya V. Sochenkov

Peoples' Friendship University of Russia / Institute for Systems Analysis of Russian Academy of Sciences

tih@isa.ru

Abstract

The paper describes results of the system EXACTUS on ROMIP-2010 in ad hoc search and classification tracks. The results of the presented approach in search and classification tasks are discussed.

Olga. Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. - P. 17-2