«Отправная точка» для дорожки по поиску в РОМИП (предварительный анализ)

© М.С. Агеев[1, 2, 3], Б.В. Добров[1, 3], Н.В.Лукашевич[1, 3],
А.В. Сидоров[3], С.В. Штернов[1, 3]

[1] Научно-исследовательский вычислительный центр
МГУ им. М.В.Ломоносова
[2] Механико-математический факультет
МГУ им. М.В.Ломоносова
[3] АНО Центр информационных исследований
{ageev, dobroff, louk, alexeys}@mail.cir.ru,
sergs2001@mailru.com

Аннотация

В статье описываются подходы, использованные коллективом разработчиков поисковой машины Университетской информационной системы (УИС РОССИЯ, http://www.cir.ru), для выполнения задания по поиску в рамках «поисковой» дорожки РОМИП2003. Основной целью являлось получение «basic line» (отправной точки) для дорожки по поиску, если использовать классическую схему поиска TF*IDF. Кроме того было выполнено еще два экспериментальных прогона. В работе описываются технические детали проведенной работы, встреченные проблемы, а также соображения по развитию РОМИП.

1. Введение

Существует большое количество разнообразных задач информационного поиска. В рамках РОМИП в 2003 году мы участвовали в задаче поиска по части текстовой коллекции сайта www.narod.ru.

В данной статье мы опишем задачи, которые ставили перед собой, участвуя в РОМИП2003, а также использованные нами методы решения.

В разделе 2 приводятся технические детали применяемой технологии обработки потока документов. В третьем разделе описываются выполненные нами разные прогоны для решения предложенной задачи. В следующих двух разделах обсуждаются проблемы проведения РОМИП в 2003 году, рассматриваются предложения по развитию РОМИП. Последний раздел содержит некоторые выводы и рекомендации.

Мы постарались привести технические подробности, возможно даже, с излишней полнотой, что кажется нам важным на таком этапе РОМИП, когда происходит установление «внутреннего стандарта» на правила оформления результатов участниками. В дальнейшем возможно сокращение оформления за счет ссылок на первое подробное описание.

2. Технические сведения о применяемых технологиях

Приведем технические детали о примененных нами в РОМИП2003 технологиях.

2.1 Решаемая задача

Для анализа было предложено более 700 тысяч документов (согласно нашим данным 727995) сайта www.narod.ru.

Оргкомитет представил список из более чем 15 тысяч запросов (15511 штук), для каждого из которых необходимо было выполнить поиск и выдать не более 100 документов.

2.2 Обработка документов

Тексты подвергались стандартной обработке, которая поддерживается в проекте УИС РОССИЯ:

- деление потока сообщений на архивы по 1000 документов, с одновременным выделением необходимой метаинформации (см. п.2.3);
- обработка текстов по архивам в стандартной последовательности Автоматической лингвистической обработки текстов (АЛОТ) (см. п.2.4);
- загрузка результатов в базу данных УИС РОССИЯ, функционирующую под управлением Oracle 9i (см. п.2.5).

Далее выполнялись прогоны (см. раздел 3) и формировались результаты в соответствии с заданным форматом.

2.3 Формальные параметры

Документы РОМИП были сконвертированы из файлов поставки в архивы по 1000 HTML документов, с одновременным выделением метаинформации - сайт, адрес, описание, ключевые слова, которые можно использовать в качестве поисковых атрибутов, либо элементов для оперативного анализа. Но для выполнения задания РОМИП2003 дополнительная информация не использовалась.

2.4 АЛОТ

АЛОТ – Автоматизированная лингвистическая обработка текстов – включает следующие этапы:

- графематический анализ и очистка документа (разбиение текста на элементы, очистка HTML тэгов, деление на предложения и абзацы);
- морфологический анализ (лемматизация русскоязычных и англоязычных словоформ текста по словарям, порождение гипотез для неизвестных слов);
- тематический анализ текста (определение в тексте терминов Общественно-политического тезауруса, а также оценка их значимости для содержания текста, различного рода аннотирование, рубрицирование).

В настоящее время АЛОТ ориентирован на обработку документов жанра деловой прозы из коллекции УИС РОССИЯ [1] – правовые акты, материалы СМИ, научные материалы. Определенную роль для применения АЛОТ играют специальные фильтры, производящие предварительную очистку текстов разных коллекций, в частности автоматическая конвертация английских букв в русских словах, отсечение «шапок», подписей и т.п.

Текстовая коллекция РОМИП2003 значительно более широкая по жанрам, значительно более сорная.

В результате для исследований в рамках РОМИП использовались только индексы по леммам и словоформам.

2.5 Загрузка в базу данных

Так как предварительной очистки текстов, в частности, конвертации в единую кодировку не производилось, то из-за ошибок процедур обработки (которые, к сожалению, были обнаружены тогда, когда их невозможно было исправить) архивов по тысяче документов, около 9% документов коллекции РОМИП не были загружены в базу данных.

Остальные документы (663077 штук) загружены базу данных и могут быть доступны, в том числе через Интернет, стандартными средствами УИС РОССИЯ. У нас получилось 4476647 уникальных словоформ и 6156962 лемм, и те и другие приведены к верхнему регистру. При этом количество пар лемма-документ 127,1 млн., количество пар словоформа-документ 129,7 млн. То есть в среднем 190 лемм на документ, при том что в коллекциях УИС РОССИЯ в среднем 380 лемм на документ.

До проведения очного семинара загруженная коллекция не будет изменяться, затем она будет полностью перегружена с исправлением замеченных ошибок.

3. Описание результатов

Существует огромное число моделей организации поиска документов. В работающих информационных системах, когда решаются реальные задачи управления информацией, часто используются гибридные методы, параметры классических моделей настраиваются на коллекцию для достижения оптимального результата.

Основным направлением исследований в рамках поисковой машины УИС РОССИЯ является организация тематического поиска, что достигается специальной подготовкой обрабатываемых документов, тематическим индексированием, возможностью использовать тематический поиск по тезаурусу, рубрикатору, в том числе с использованием интерактивной «очистки» результатов запроса [1].

Предложенная в рамках поисковой дорожки РОМИП2003 задача поиска в Web-коллекции является новой для авторов настоящей работы. Поэтому не ставилось задачи достичь наилучших результатов. Интересно было получить некоторую отправную точку, в качестве которой удобно было рассматривать результаты применения какой-нибудь из классических моделей поиска.

Для авторов работы легче всего было использовать реализованную в УИС РОССИЯ модель TF*IDF (см. п.3.1). Кроме того было выполнено два дополнительных экспериментальных прогона для проверки влияния на качество результата – учет использования точных словоформ поискового запроса (см. п.3.2) и расстояния между словами (см. п.3.3).

3.1 Прогон 1. «Отправная точка»

Запросы представленные оргкомитетом РОМИП обрабатывались поисковой машиной УИС РОССИЯ ровно «как есть», не производя фильтрации служебных символов, некоторые из которых («/», «-») имеют в УИС РОССИЯ особый смысл.

Каждый запрос представлялся в виде последовательность словоформ, соединяемых условием «AND», для каждой словоформы вырабатывались леммы, соединяемые условием «OR».

Для получения базовой оценки использовалась модель TF*IDF в формулировке INQUERY [2].

Точнее - вес каждой леммы документа оценивался следующим образом.

$$TFIDF_D(l) = \beta + (1-\beta)\cdot tf_D(l)\cdot idf_D(l) \quad ,$$

где “term frequency” – учет частотности леммы в документе:

$$tf_D(l) = \frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5\cdot \frac{dl_D}{avg\_dl}}$$

$freq_D(l)$ - частотность леммы l в документе, dl_D – мера длины документа, avg_dl – средняя длина документа, $\beta = 0.4$.

“Inverse term frequency” - фактически форма штрафования часто используемых в коллекции слов:

$$idf(l) = \frac{\log\left(\frac{|c|+0.5}{df(l)}\right)}{\log(|c|+1)} \quad ,$$

где $|c|$ - количество документов в коллекции, $df(l)$ - количество документов, где встретилось лемма l.

Аналогичные оценки были построены для словоформ документов.

Известно [3], что можно применять различные модификации данной формулы - все дают примерно одинаковый результат.

Каждый запрос

$$Q = w_1 \quad w_2 \quad w_3 \ldots \quad w_m$$

представлялся в виде формулы

$$L(Q)= L(w_1) \,\&\, L(w_2) \,\&\, L(w_3) \,\&\, \ldots \,\&\, L(w_m) ,$$

где $$L(w) = l_1(w) \; OR \; l_2(w) \; OR \; \ldots OR \; l_q(w) ,$$

$l_k(*)$ - леммы морфологического разбора слова.

Тогда оценка релевантности документа D для запроса Q , что несколько отличается от применяемого в [2]:

$$V_D(Q)=\frac{\sum_{i=1}^{N}\sum_k\left(\theta_{ik}\cdot TFIDF_D(l_{ik}(w_i))\right)}{\sum_{i=1}^{N}\sum_k|\theta_{ik}|} ,$$

где $\theta_{ik} = \theta_i = 1.0$ - “вес” леммы в запросе равен весу, устанавливаемому для соответствующего слова запроса.

Сначала происходит разбор запроса и построение дерева запроса в виде XML структуры (для учета задаваемых пользователем или моделями обработки логических и т.п. операций над элементами запроса). Для некоторых элементов запроса происходит их переопределение в дереве запроса, после исполнения, например, процедуры морфологии, когда появляются разные леммы у введенного слова запроса.

Затем полученное XML-представление запроса преобразуется в SQL-запросы ORACLE: один запрос для поиска релевантных документов и один запрос для вычисления оценки релевантности найденных документов. Полученные запросы исполняются и результаты сохраняются в таблицах ORACLE.

Рассмотрим для примера обработку запроса rb7701:

мыло состав

Дерево разбора запроса выглядит следующим образом:

```
<?xml version="1.0"?>
  <!--to_morfo: "мыло состав"-->
  <AndExpression Creator="QueryMorfo">
    <OrExpression Creator="QueryMorfo">
      <AttributedItem Name="Lemma" Value="МЫЛО"
       ItemID="30070" Flag="+" Словоформа="МЫЛО" />
      <AttributedItem Name="Lemma" Value="МЫТЬ"
       ItemID="3060" Flag="+" Словоформа="МЫЛО" />
    </OrExpression>
    <AttributedItem Name="Lemma" Value="СОСТАВ"
     ItemID="883" Flag="+" Словоформа="СОСТАВ" />
  </AndExpression>
```

Соответствующее SQL выражение:

```
SELECT /*+FIRST_ROWS*/ doc_id
FROM  good_docs
WHERE doc_id
      IN ((((SELECT doc_id FROM   doc_lem_indexs
             WHERE lem_id=30070 AND class_id=182)
             UNION
             (SELECT doc_id FROM  doc_lem_indexs
              WHERE lem_id=3060 AND class_id=182)
            )INTERSECT
             SELECT doc_id FROM  doc_lem_indexs
             WHERE lem_id=883 AND class_id=182))) ,
```

условие *class_id = 182* показывает, что поиск идет по коллекции РОМИП (в УИС РОССИЯ можно искать одновременно по нескольким коллекциям).

В результате формируется временная таблица doc_list_543436, затем производится ранжирование результата:

```
SELECT SUM(rnk*weight)
       /SUM(DECODE(SIGN(weight),1, weight, 0))rank,
       MIN(rwd) rwd
FROM  (SELECT /*+ORDERED */ L.doc_id, x.rank rnk,
                            1.0 weight, L.rowid rwd
       FROM doc_list_543436 l, doc_lem_indexs x
       WHERE x.doc_id = L.doc_id AND class_id=182
         AND x.lem_id = 30070
       UNION ALL
       SELECT /*+ORDERED */ L.doc_id, x.rank rnk,
                            1.0 weight, L.rowid rwd
       FROM doc_list_543436 L, doc_lem_indexs X
       WHERE X.doc_id = L.doc_id AND X.class_id=182
       AND   X.lem_id = 3060
       UNION ALL
       SELECT /*+ORDERED */ L.doc_id, X.rank rnk,
                            1.0 weight, L.rowid rwd
       FROM doc_list_543436 L, doc_lem_indexs X
       WHERE X.doc_id = L.doc_id AND X.class_id=182
       AND   X.lem_id = 883 )
  GROUP BY doc_id
```

3.2 Прогон 2. Влияние точных словоформ запроса

Второй прогон был предназначен для исследования влияния учета точных словоформ запроса.

Во втором прогоне исполнялся запрос по словам с учетом морфологического словоизменения.

Например, запрос РОМИП rb7701 (*мыло состав*) преобразуется в запрос:

(мыло состав)
"мыло состав":OnlyRank(20)

здесь символы кавычек означают, что рассматриваются точные словоформы слов запроса.

Запрос вычисляется также как и в п.3.1 – по леммам, но при ранжировании для лемм как и ранее $\theta_{ik} = \theta_i = 1.0$, для словоформ же $\theta_i = 20.0$.

Кроме того, при разборе строки запроса производилась фильтрация специальных символов, наличие которых приводило к неправильной интерпретации запроса (из-за различий в формате языка запросов в Яндекс и УИС РОССИЯ).

3.3 Прогон 3. Влияние расстояния между словами

Третий прогон был предназначен для исследования влияния на качество информационного поиска учета расстояния между словами запроса в индексируемом документе.

В настоящее время в УИС РОССИЯ не поддерживается индекс по близко расположенным словам. Определенной альтернативой является индекс по терминам-словосочетаниям тезауруса (последовательность слов, непосредственно следующих друг за другом).

Поэтому для моделирования такого индекса была применена технология, аналогичная используемой в УИС РОССИЯ для рубрикации текстов [4]. Был образован псевдотезаурус из первых лемм морфологического разбора слов всех запросов.

Вес формулы вычислялся по правилу:

$$V(Q) = \frac{\sum_{j=1}^{m} \psi(w_j) + \sum_{j<k} S(w_j, w_k)}{m + C_m^2},$$

здесь $\psi(w_j)$ – вес слова в документе, из-за отсутствия связи между «понятиями» в псевдотезаурусе – фактически мера частотности в

документе, $S(w_j, w_k) = \min\{1.0; \frac{\sum s(l_{jq} \in L(w_j), l_{kz} \in L(w_k))}{\max s(v \in W_D, u \in W_D)}\}$ - сила «текстовой связи», равна единице если слова в тексте часто встречаются «рядом» и близка к нулю, если редко.

При рубрикации [3] текстовая связь между элементами формулы устанавливается, если между ними находится не более трех терминов (ориентируясь на «волшебное число семь», характеризующее размер кратковременной человеческой памяти при восприятии текстовой информации).

В данном случае, расстояние между элементами формулы измерялось в словах, входящих в состав хотя бы одного запроса. Эта достаточно грубая модель учета связи между словами была применена из-за наличия готового программного обеспечения.

Документы в выдаче упорядочивались по убыванию вычисленного веса запроса для документа, при значении веса меньше порога документ не включался.

Следует отметить:

1) прогон 3 не является результатом работы собственно поисковой машины, так как поисковый индекс не создавался заранее (не зная запросов), наоборот, он представляет собой специализированную систему ответа именно на предложенные 15000 запросов. Вместе с тем соответствующий индекс может быть создан и на этапе обработки текстов.
2) решение о запуске прогона 3 было принято непосредственно перед конечным сроком подачи результатов, поэтому индекс был построен только примерно по 40% коллекции. Так как цель прогона 3 – проверка влияния на точность, то тем не менее было принято решение о представлении результатов.

3.4 Сравнение результатов разных прогонов по общей полноте-точности результатов

Как представлено на Рис.1 прогон 1, который планировался в качестве базового, предназначенного отправной точкой для улучшения за счет учета дополнительных параметров, показывает в методике подсчета WEAK (хотя бы один из экспертов назвал документ релевантным) лучшие результаты.

Прогон 3, демонстрируя сравнимые с прогоном 1 показатели в начале списка, далее резко падает. Это может быть объяснено как дополнительным требованием текстовой близости слов запроса и

отбрасыванием документов с весом ниже заданного, так, к сожалению, и тем, что результаты прогона 3 были получены не для всех документов.

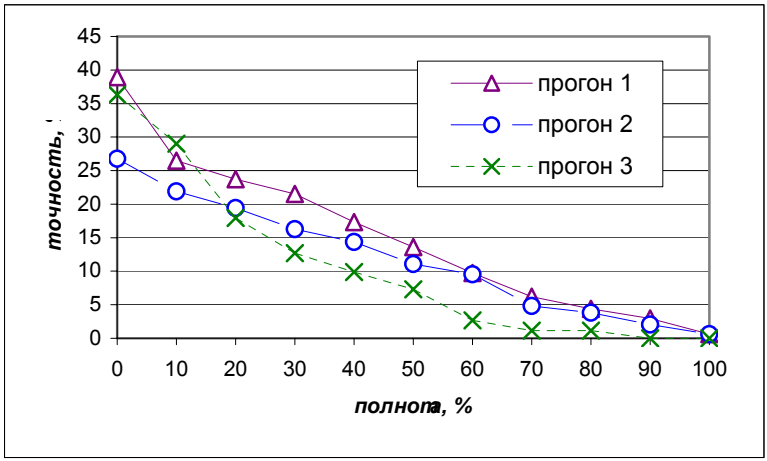

Рис.1. Сравнение результатов разных прогонов по метрике TREC для методики оценки WEAK

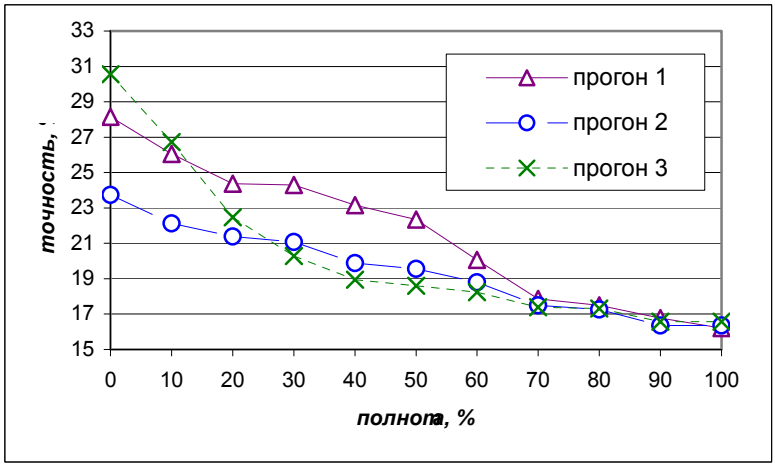

Рис.2. Сравнение результатов разных прогонов по метрике TREC для методики оценки STRONG

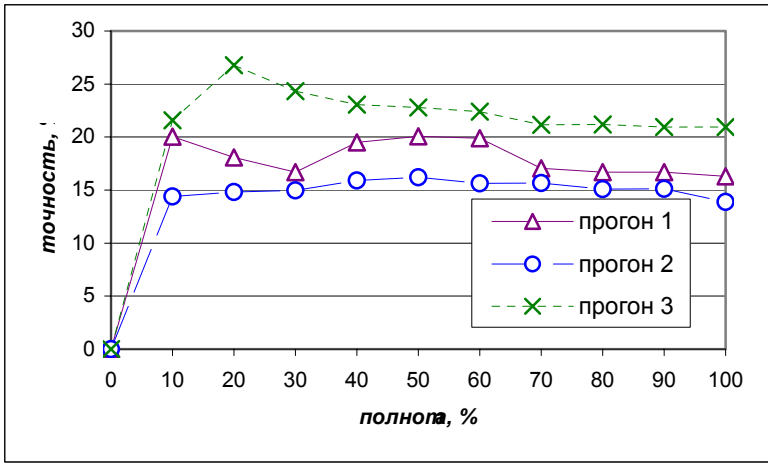

Рис.3. Сравнение результатов разных прогонов
по метрике RIRES для методики оценки WEAK

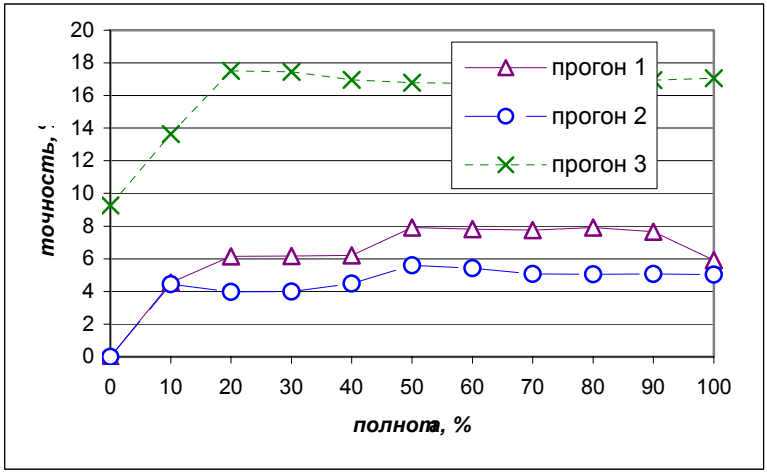

Рис.4. Сравнение результатов разных прогонов
по метрике RIRES для методики оценки STRONG

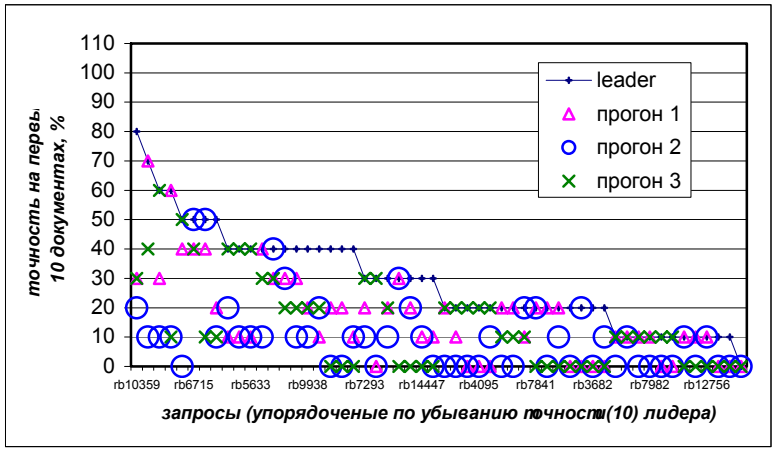

Рис.5. Сравнение точности с лучшим результатом в 10 первых документах разных прогонов для методики оценки WEAK

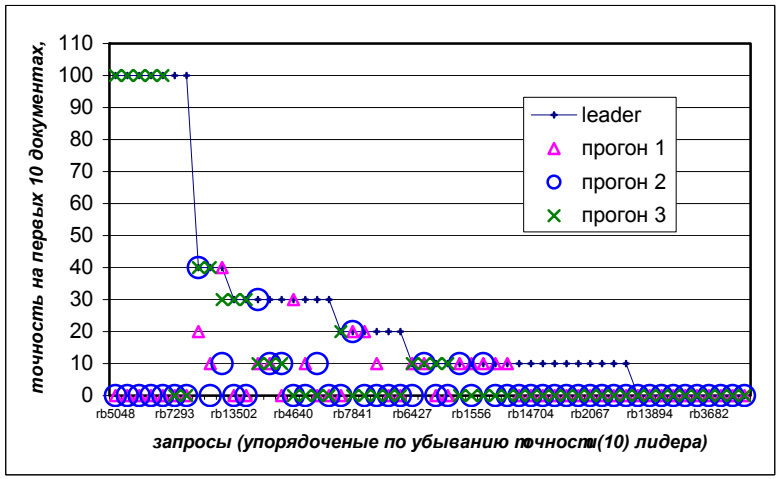

Рис.6. Сравнение точности с лучшим результатом в 10 первых документах разных прогонов для методики оценки STRONG

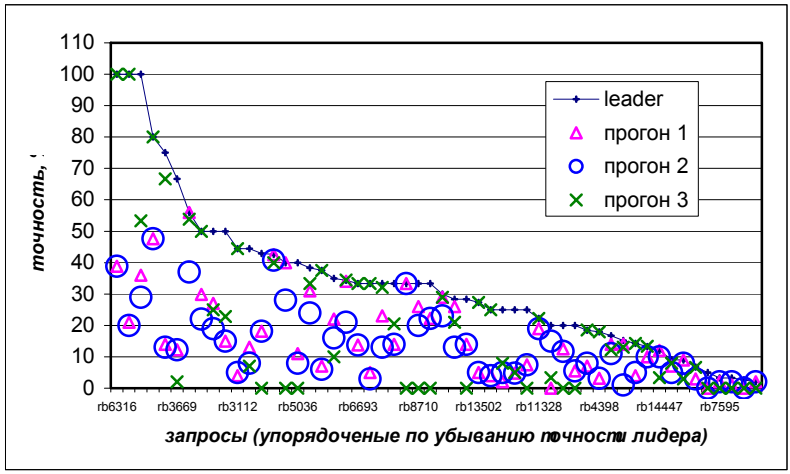

Рис.7. Сравнение точности с лучшим результатом по всем документам разных прогонов для методики оценки WEAK

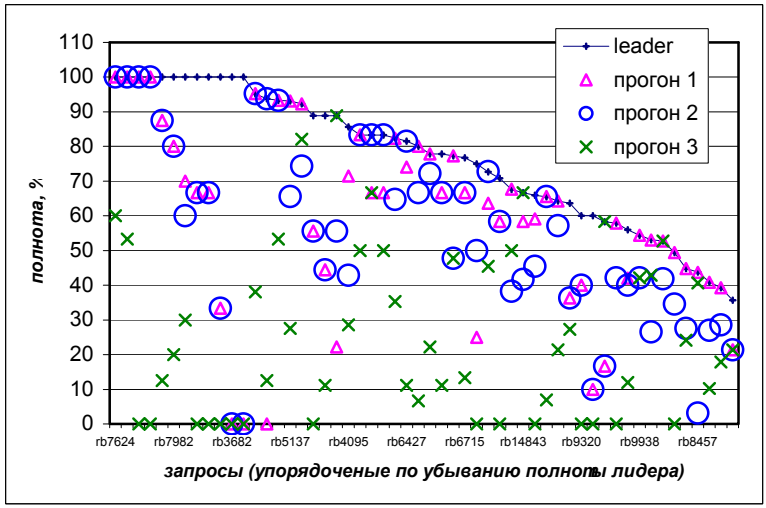

Рис.8. Сравнение полноты с лучшим результатом по всем документам разных прогонов для методики оценки WEAK

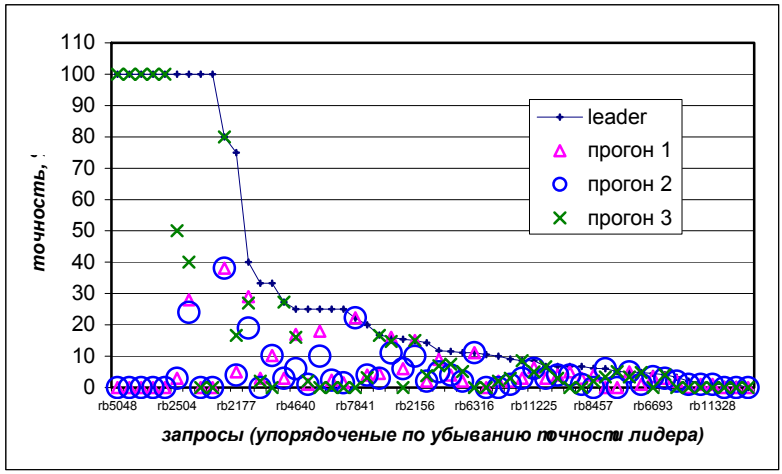

Рис.9. Сравнение точности с лучшим результатом по всем документам разных прогонов для методики оценки STRONG

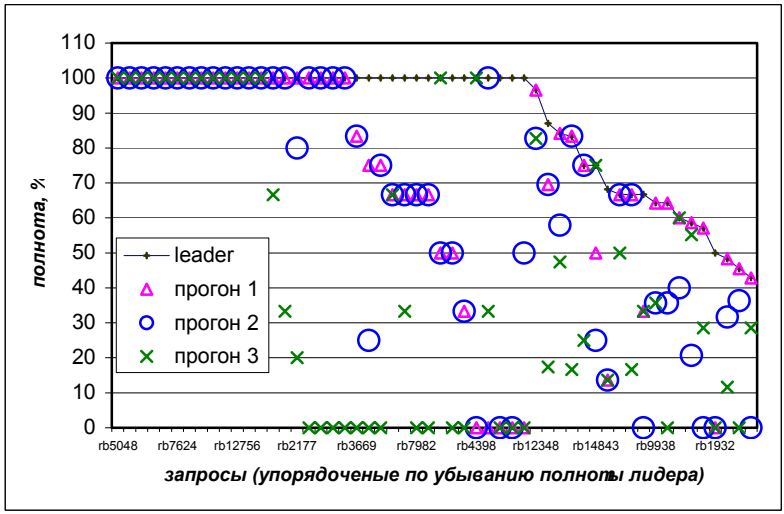

Рис.10. Сравнение полноты с лучшим результатом по всем документам разных прогонов для методики оценки STRONG

Несколько неожиданным для нас явилось то, что результаты прогона 2 (с более высоким ранжированием вхождения точных словоформ запроса) оказались хуже результатов прогона 1. Объяснение этого требует дополнительного изучения.

Результаты сравнения прогонов для методики сравнения STRONG (совпадение мнений обоих экспертов) отличаются от WEAK. Лучшие результаты (см. Рис.2) демонстрирует прогон 3, а первый и второй прогоны показывают примерно одинаковые результаты.

На Рис.3 и Рис.4 приведены те же результаты в метрике RIRES, которая отличается от метрики TREC [5] в следующих вопросах:

- если система X не достигает заданного уровня полноты a, то точность системы X на уровне полноты a считается равной точности системы на множестве результатов (в TREC считается 0.0).

  Рассмотрим пример. Пусть имеется 2 релевантных документа. Система А выдает всего один релевантный документ. Система Б выдает всего три документа: первый и третий релевантны, второй - нет. Тогда по метрике RIRES система А имеет преимущество перед Б. В метрике TREC - наоборот, система Б будет иметь преимущество перед А на уровнях полноты >50%;
- не производится процедура интерполяции результатов, применяемая в TREC. Интерполяция TREC обеспечивает монотонное убывание PR-графика и позволяет более аккуратно учитывать ситуацию, когда указанное значение полноты достигается не точно. Кроме того, интерполяция TREC позволяет получить полезное значение в нуле: максимальная точность на начальном отрезке результатов.

3.5 Сравнение результатов разных прогонов для точности результатов в первых 10 документах выдачи

На рисунках 5 и 6 представлено сравнение результатов разных прогонов с наилучшими, показанными участниками РОМИП (кривая с прямыми крестиками и обозначением «leader») по точности в первых документах выдачи, что является весьма важным при поиске в Интернет.

По оси абцисс нанесены номера запросов (отличающиеся для графиков), упорядоченных по убыванию точности лидера.

Как можно видеть, результаты прогона 3 (с учетом связи между словами) демонстрируют лучшую точность, при этом по методике

STRONG прогоны 1 и 2 показывают плохие результаты в первых десяти документах.

3.6 Сравнение полноты/точности результатов разных прогонов с лучшим достигнутым

На Рис.7 и Рис.8 приведены сравнение результатов точности и полноты по отдельным запросам с наилучшим достигнутым участниками РОМИП2003 уровнем по методике WEAK, на Рис.9 м Рис.10 – по методике STRONG.

Как нетрудно видеть, для методики WEAK в целом наблюдается разброс результатов по отдельным вопросам, при этом полнота результатов прогона 3 в среднем хуже прогонов 1 и 2 (Рис.6).

Рис.9 и Рис.10 демонстрируют, что при подсчете результатов по методике STRONG по достаточно многим запросам наблюдается 100% покрытие, прогоны 1 и 2 показывают иногда плохую точность при том, что прогон 3 достигает лучшего результата.

3.7 Предварительное сравнение с результатами других конференций

Следует отметить, что в аналогичных зарубежных конференциях участники публикуют результаты с более высокими показателями.

Однако, обычно там рассматриваются более длинные запросы. Кроме того, требует специального изучения возможность и разрешение использования «обратной связи» при анализе результатов. Только в РОМИП2003 была реализована процедура, практически полностью исключающая ручную подгонку результатов (невозможно вручную «подчистить» результаты для 15 тысяч запросов).

3.8 Особые случаи

Вопрос о том, как надо обрабатывать ситуацию, когда количество релевантных документов запроса равно 0, является спорным. В текущей реализации метрики TREC считается, что точность системы по запросу для которого нет релевантных документов равна 100%.

С одной стороны, логично считать правильным поведением системы, когда на запрос без релевантных документов система так и говорит: "ничего не найдено", и неправильным, если система в таком случае что-то выдает. В таком случае имеет смысл считать, что точность системы равна 100% если система ничего не выдала, и 0%, если система выдала какие-то результаты. Но, с другой стороны,

есть проблема - правильно ли штрафовать точность по запросу на целых 100% в случае, если система выдала всего один лишний документ? Ведь если есть релевантные документы, то наличие небольшого количества лишних документов в результатах влияет на метрики не значительно.

Имеется следующая особенность метрики TREC, реализованной в текущей версии программы оценки: если количество релевантных документов для данного запроса равно 0, то точность полагается равной 100% при любом уровне полноты.

Такая ситуация реализуется при оценке методом STRONG для 8 запросов. По всей видимости, в TREC такой ситуации не возникает и в описании метрик TREC данный случай не рассматривается.

Из-за такой реализации оценки возникает забавный эффект: для метрики STRONG все системы показывают точность выше 0.14 для любого уровня полноты.

Кроме того резко повышается влияние ошибок асессоров на оценки систем.

По метрике WEAK для всех запросов есть релевантные документы, а для STRONG - восемь "пустых" запросов. Это можно интерпретировать как то, что часть документов являются спорными, "релевантными наполовину".

Если асессоры ошибаются, то "пороговая" оценка точности начинает ранжировать системы с точностью до наоборот: если некоторая система выдала релевантный документ, то ее точность будет равна нулю, а если, наоборот, не выдала ничего - то ее точность 100%.

Такая ситуация вполне реальна. Например, для запроса rb5048 "история рентгенологии" имеется документ, который выдала наша система: http://nmgazette.narod.ru/Our_Sanitation.htm . Полагаем, что данный документ скорее релевантен, чем нет. Тем не менее, по метрике STRONG для данного запроса нет релевантных документов (а по WEAK - есть).

Текущая реализация оценки посчитала, что точность системы на запросе rb5048 равна 100%.

3.9 Предварительные выводы

Сейчас можно делать только предварительные выводы – слишком маленькая статистика сравнения результатов накоплена, слишком высока вероятность ошибки из-за неправильного учета случайных факторов.

Тем не менее, любопытным является сильная зависимость результатов в зависимости от методики оценки WEAK/STRONG.

Если для методики STRONG (особенно для первых 10 документов) хорошие результаты показываются при учете связи между словами, то для WEAK поиск без учета связи тоже оценивается неплохо. При этом прогоны 1 и 2 демонстрируют значительно большую полноту поиска по сравнением с прогоном 3 (что важно при сложном поиске, например уточнении запроса по результатам первого поиска).

Предварительный анализ результатов позволяет сформулировать формальный вывод – простое объединение результатов прогонов 1 и 3, позволяет получить и достаточно высокую точность в начале выдачи и необходимую полноту результатов в целом.

4. РОМИП в 2003

Авторы могут оценивать организацию РОМИП только в той дорожке, в которой принимали участие – по поиску в Web-коллекции.

В 2003 году была проведена большая работа:

- сформирована приличного размера коллекция;
- группа участников выполнила задания;
- создан комплекс программного обеспечения для оценки релевантности, сравнения результатов;
- в основном сформирован пакет согласованных между участниками документов.

К недостаткам работы РОМИП в 2003 году следует отнести многократные переносы сроков, ошибки. Хочется надеяться, что с этими проблемами удастся справиться в следующих циклах.

Представляется, что при надлежащем уровне планирования можно было бы распределить создание необходимого программного обеспечения между всеми (частью) участниками, чтобы снизить нагрузку на коллектив из Санкт-Петербурга.

4.1 Коллекция

Для дорожки поиска РОМИП2003 использовалась часть текстовой коллекции мегасайта www.narod.ru. Такой выбор коллекции представляется очень удачным, так как коллекция представляет полное жанровое разнообразие российского Интернет.

Вместе с тем, отсутствие замкнутости коллекции, делает проблематичным использование столь важного в Интернет-поиске фактора как взаимные ссылки между страницами.

4.2 Запросы и оценка релевантности

Методика выбора реальных запросов пользователей из лог-файла поисковой системы Яндекс является наилучшей. Однако необходимо привести дополнительные технические данные, так как включение в лог-файл некоторого количества запросов во внутреннем формате поисковой системы (которые не могли быть заданы пользователями) вызывает определенные вопросы.

Требует большего описания процедура отбора запросов для оценивания.

Желательно, чтобы запросы для оценки отбирались сериями «со смыслом» - для проверки коротких/длинных запросов, похожих запросов и т.п.

Большое количество вопросов вызывает «расширенная интерпретация» формулировок запросов. Мы проставили свои оценки согласия/несогласия с расширенной интерпретацией запроса. Из 47 проанализированных расширения мы согласны только с 22, с 12 полностью несогласны, с 13 несогласны частично. Основные причины несогласия – расширенное описание запроса содержит только часть из возможных интерпретаций, отвергая остальные. В частности, систематически отвергались документы жанра коммерческих предложений.

Запросы «как выбирать гитару», «кто такие рэперы» подразумевают необходимость применения специальных лексических фильтров. Расширение запроса «Лермонтов в 1841 году» вносит «сверхзнание» - гибель Лермонтова, что может влиять на поведение оценщиков.

При этом, учитывая большое расхождение между оценщиками даже при наличии расширенного толкования запроса, возможно, следует отказаться от расширенного толкования, как вносящего дополнительный фактор неопределенности.

Желательно сделать свободно доступным (не только среди участников РОМИП) программное обеспечение для получения сопоставимых результатов.

4.3 Процедура сравнения результатов

Следует обратить внимание, чтобы все формулы, по которым проводятся оценки были опубликованы и с необходимой степенью точности задокументированы.

Желательно опубликовать данные о работе экспертов – время на размышление, совпадение с другими экспертами и т.п.

4.4 Интерпретация результатов и открытость информации об участниках

Один из самых сложных вопросов РОМИП (как и других подобных форумов) – предотвращение нездоровых спекуляций на основе сравнения полученных результатов.

Участники РОМИП представляют собой сложившиеся коллективы, каждый из которых занимает свою нишу на «рынке» информационных систем. Это значит, что каждый коллектив имеет набор технологий эффективно решающих те или иные задачи информатизации РФ. Причем эти задачи разные у разных коллективов, что не противоречит возможности конкуренции в тех или иных смежных подзадачах.

Рассматривая текущую (равно как и будущие) дорожку РОМИП, можно констатировать, что участники решают не свою стандартную задачу, но «задачу, где могут быть применены их методы».

Нам представляется, что лучший путь интерпретации результатов – трактовать их как соревнование определенных моделей, вообще говоря, совершенно не связанных с «родными» системами.

Это позволит открыто обсуждать и обмениваться результатами, приведет к улучшению показателей всех участников.

5. Развитие РОМИП

Нам представляется, что проведенная работа поставила много интересных вопросов, которые необходимо далее исследовать.

При этом надо максимально сохранить то, что было наработано за 2003 год:

- текстовые коллекции;
- коллекции запросов с их интерпретацией;
- коллекции результатов участников;
- обслуживающее программное обеспечение;
- согласованные методические материалы.

5.1 Текстовые коллекции, задания

Желательно расширять набор коллекций, как расширяя Web-коллекцию, так и добавляя коллекции специальных жанров. Возможно, стоит выделить в рамках Web-коллекции подколлекции специальных жанров – новостные сайты, сообщения форумов и т.п.

Любопытные задачи можно поставить, если распространять коллекции частями, сначала, скажем, каждый второй документ - для настройки, затем – остальные.

Можно было бы предложить некоторое количество новых заданий:

- повторить старое задание с целью оценить максимально возможный результат, при этом стараться расширить количество оценок релевантности документов;
- формирование тематических подборок в Web (за ограниченное время);
- формирование тематических каталогов в Web-коллекции;
- информационные системы «играющие в Кубок Яндекса», либо помогающие играть пользователю;
- оценка релевантности при поиске по коллекциям специальных жанров – есть ли отличия от Web-коллекции – по материалам СМИ, правовым документам;
- поиск по сложным запросам, здесь же поиск документов, похожих на заданный;
- задание на аналитическую обработку результатов запроса, в частности на деление на тематическую и коммерческую составляющие;
- «вопросно-ответные» задания;
- и т.д.

5.2 Популяризация РОМИП и новые участники

Следует иметь в виду, что постоянно появляется масса идей на тему «какой должен быть информационный поиск», лучшим способом популяризации РОМИП, на наш взгляд, является создание среды для упрощения проверки таких идей на коллекциях РОМИП. Это, с одной стороны, приведет к быстрому отсеиванию несерьзных предложений, с другой стороны, в случае хороших идей – к их скорейшему распространению.

Нам кажется, что участники РОМИП должны содействовать упрощению процедуры доступа к результатам и программному обеспечению, создаваемому в рамках семинара.

5.3 Сотрудничество РОМИП с «родственными» конференциями

Новые возможности для развития РОМИП (как сообщества участников) появляются при расширении сотрудничества с аналогичными исследовательскими форумами по информационному поиску, функционирующими зарубежом:

- обмен текстовыми коллекциями;
- сравнение с новыми участниками

Наиболее близким по текущим задачам РОМИП является Cross-Language Evaluation Forum (CLEF) [6], в котором есть устойчивый интерес к анализу российских текстовых коллекций. С 2003 года CLEF получил права на использование в целях оценки архива газеты «Известия» за 1995 год.

При этом российские участники могут попробовать свои силы на иноязычных текстовых коллекциях.

6. Выводы

Текущие полученные результаты РОМИП следует оценивать как предварительные. Необходимо «повторить» задания РОМИП2003, чтобы отработать процедуру, избавиться от замеченных технических ошибок (это не исключает добавления новых заданий).

Нам представляется, что РОМИП как сообщество заинтересованных исследовательских коллективов находится в самом начале чрезвычайно интересного пути. Необходимо приложить максимально возможное количество усилий для развития и расширения проекта. Наиболее важным представляется решение вопросов повышения устойчивости проекта, для чего необходимо решить вопросы с финансированием и с расширением круга участников. Авторам текущей работы наиболее важным кажется вовлечение более широких слоев исследователей в мероприятия РОМИП.

Литература

[1] Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н., Технологические аспекты организации доступа к разнородным информационным ресурсам в университетсткой информационной системе РОССИЯ // Электронные библиотеки – 2002 – Том.5 – Выпуск 2 http://www.elbib.ru/journal/2002/200202/ADZLSJU/ADZLSJU.html

[2] Callan J.P., Croft W.B. and Harding S.M., The INQUERY Retrieval System // A.M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3rd International Conference on Database and Expert Systems Applications. - Springer Verlag, New York. - 1992. - pp.78-93. http://citeseer.nj.nec.com/26307.html

[3] Salton G, Buckley C., Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management. 1988. 24, 5. - pp.513-523.

[4] Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. 7-12 октября 2002, Коломна – М.: Физматлит – Т.1 – С.178-186.

[5] The Eleventh Text Retrieval Conference (TREC 2002) Appendix A: Common Evaluation Measures // NIST Special Publication: SP 500-251 http://trec.nist.gov/pubs/trec11/appendices/MEASURES.ps.gz

[6] 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

Basic Line for RIRES2003 Information Retrieval Web-Track

M.S.Ageev[1, 2, 3], B.V.Dobrov[1, 3], N.V.Loukachevitch[1, 3], A.V.Sidorov[3], S.V.Shternov[1, 3]

[1] Research Computing Center of Moscow State University

[2] Faculty of mechanics and mathematicsof Moscow State University

[3] NCO Center for Information Research

{ageev, dobroff, louk, alexeys}@mail.cir.ru, sergs2001@mailru.com

This article describes approaches used by team of UIS RUSSIA (University Information System of Russian inter-University Social Science Information and Analytical consortium, http://www.cir.ru/eng/) search engine for RIRES2003 (Russian Information Retrieval Evaluation Seminar) ad hoc track. Our main task was to obtain a "basic line" for ad hoc track using standard TF*IDF algorithm. We also made two experimental runs for ad hoc track. We describe a technical details and a problems of the implementation. We also outline our view on the future of RIRES.