УИС РОССИЯ в РОМИП 2007: поиск и классификация

© М.С. Агеев[1, 3], Б.В. Добров[1, 3], П.В. Красильников[2]
Н.В. Лукашевич[1, 3], А.М. Павлов[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@mail.cir.ru, dobroff@mail.cir.ru,
p.krasilnikov@gmail.com, louk@mail.cir.ru,
apavlov@ttk.ru, alexeys@mail.cir.ru, sergey@shternov.ru

Аннотация

В статье описываются подходы, использованные коллективом разработчиков Университетской информационной системы РОССИЯ (УИС РОССИЯ, http://www.cir.ru), для выполнения заданий РОМИП 2007 по поиску в Веб коллекции, поиску в коллекции правовых документов и классификации Веб-страниц и Веб-сайтов.

1. Введение

В цикле РОМИП 2007 мы принимали участие в дорожках по поиску в Веб коллекции, поиску в коллекции правовых документов, классификации Веб-страниц и Веб-сайтов.

Для дорожек ad hoc поиска документов по запросу описания проведенных экспериментов и выводы описаны в разделе 2, для дорожек классификации веб-страниц и веб-сайтов – в разделе 3.

2. Дорожки ad hoc поиска документов по запросу

Наш коллектив принимал участие в дорожках поиска РОМИП в 2003, 2004 и 2005 году [1, 2]. Исследования, проведенные в рамках РОМИП'2007, являются продолжением исследований 2003-2005 годов. Развитие алгоритмов и технологий велось по двум направлениям.

Во-первых, для ранжирования документов применялся комплекс различных факторов, расширявшийся с каждым годом. Использовались следующие факторы ранжирования документов:

- классическая векторная модель ранжирования TF*IDF (2003-2005, 2007);
- учет близости слов запроса в найденном тексте – размер «минимального окна» в тексте, содержащего все слова запроса (2005, 2007);
- «поиск по кворуму» - поиск документов, содержащих не все слова запроса (2005, 2007);
- близость по парам слов (2007).

Для выбора оптимальной формулы ранжирования для каждого из факторов, а также подбор коэффициентов для объединения данных факторов использовались наборы данных и результаты оценок РОМИП 2004-2006. Методика подбора параметров оптимальной формулы ранжирования и результаты описаны в статье [3].

Во-вторых, в этом году мы использовали новую технологическую платформу индексирования и поиска документов, основанную на стандартном представлении индекса в виде инвертированных списков.

Мы приняли участие в двух дорожках ad hoc поиска документов по запросу:

- дорожке поиска по коллекции нормативных документов;
- дорожке поиска по коллекции Веб-страниц «белорусский интернет» (BY.WEB).

Использовался один и тот же алгоритм поиска для обеих дорожек.

Изложение данного раздела построено следующим образом: в разделе 2.1 мы опишем использованный алгоритм ранжирования, в разделах 2.2 и 2.3 – параметры дорожек поиска и полученные результаты, в разделе 2.4 – анализ результатов и выводы.

2.1 Алгоритм ранжирования документов

Подробное описание алгоритма ранжирования документов приведено в [3].

Для ранжирования использовались 3 фактора: TF*IDF, вес по парам слов, вес по минимальному окну.

1) Сумма TF*IDF-весов слов запроса:

$$\mathrm{w}_{\mathrm{tfidf}}(d,Q)=\sum_{t\in Q}(0.4+0.6\cdot tf(d,t)\cdot idf(t))$$

$$tf(d,t)=\frac{freq(d,t)}{freq(d,t)+0.5+1.5\cdot\frac{docLen(d)}{380}}$$

$$idf(t)=1-0.16\cdot\log_{10}(dc(t))$$

где

- $freq(d,t)$ - частота встречаемости слова t в документе d
- $docLen(d)$ – длина документа в различных леммах
- $dc(t)$ – количество документов коллекции, содержащих лемму t

2) Вес по парам слов:

Для каждой пары слов (t, s), встречающихся в запросе на расстоянии не более 5 и в документе на расстоянии не более 3, вычислялся вес пары $\mathrm{p}(d,t,s)$.

$$\mathrm{p}(d,t,s)=\begin{cases}idf(t)+idf(s), \text{ если } |t-s|_d\le 3\\ 0 \text{ иначе}\end{cases}$$

Вес документа по парам слов вычислялся как нормированная сумма весов всех пар слов, входящих в запрос:

$$\mathrm{w}_{\mathrm{pair}}(d,Q)=\frac{\sum_{t,s\in Q,|t-s|_Q\le 5}\mathrm{p}(d,t,s)}{\sum_{t,s\in Q,|t-s|_Q\le 5}(idf(t)+idf(s))},$$

3) Вес по минимальному окну

$$w_{\min-window}(d,Q)=\frac{1}{\ln(mv(d,Q)-|Q|+4)}$$

mv(d, Q) – размер минимального «окна», содержащего все слова запроса *Q*, |*Q*| - длина запроса.

4) Всё вместе: вес соответствия документа запросу вычислялся как сумма трех факторов с коэффициентами:

$$W(d,Q) = 0.9 \cdot W_{\text{tfidf}}(d,Q) + 0.1 \cdot W_{\text{pair}}(d,Q) + 0.3 \cdot W_{\text{min-window}}(d,Q)$$

2.2 Дорожка поиска по Веб-коллекции BY.web

Коллекция документов состояла из 1 525 585 текстов, полученных выборкой из страниц сайтов домена .by из индекса поисковой системы yandex.ru.

Оргкомитет представил список из 19627 запросов, полученных смешением запросов из логов поисковых систем yandex.ru и km.ru, а также оцененных запросов РОМИП 2003-2006 годов. Для каждого из запросов необходимо было выполнить поиск и выдать не более 100 документов.

На Рис. 1-4 представлены графики метрик оценки качества поиска, полученных с использованием «сильной» и «слабой» шкалы оценки релевантности.

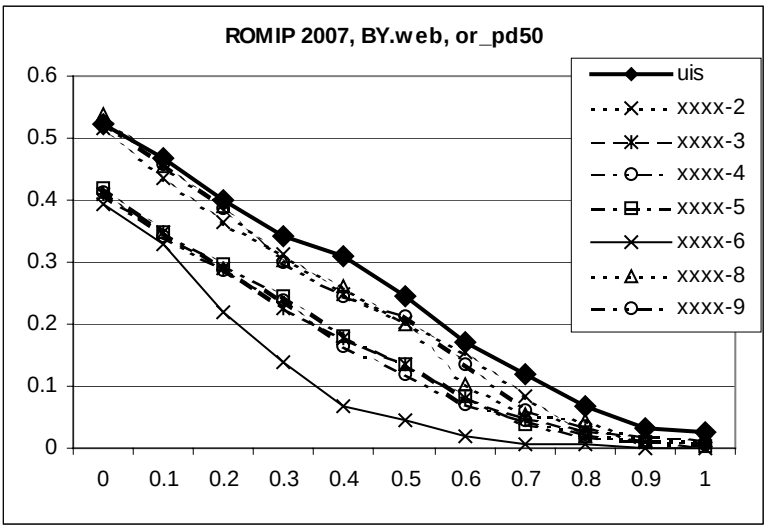


Рис. 1 11-точечный график полноты/точности для дорожки поиска по BY.WEB, таблица релевантности «OR»

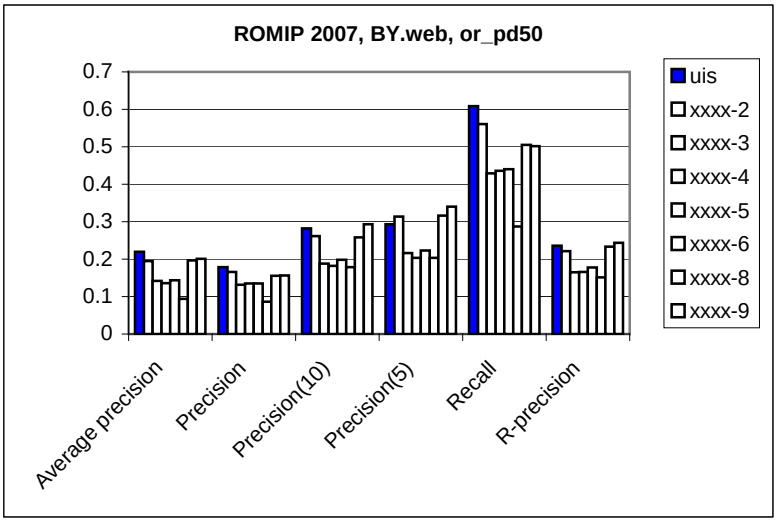


Рис. 2 Различные метрики для дорожки поиска по BY.WEB, таблица релевантности «OR»

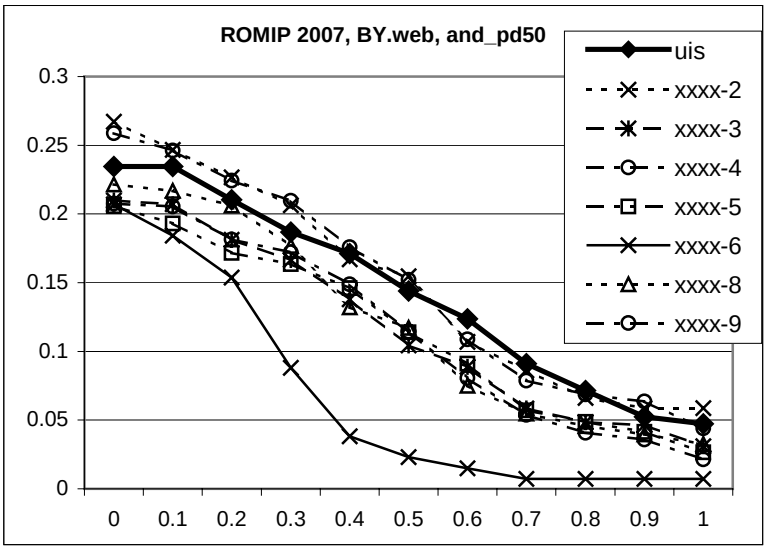


Рис. 3 11-точечный график полноты/точности для дорожки поиска по BY.WEB, таблица релевантности «AND»

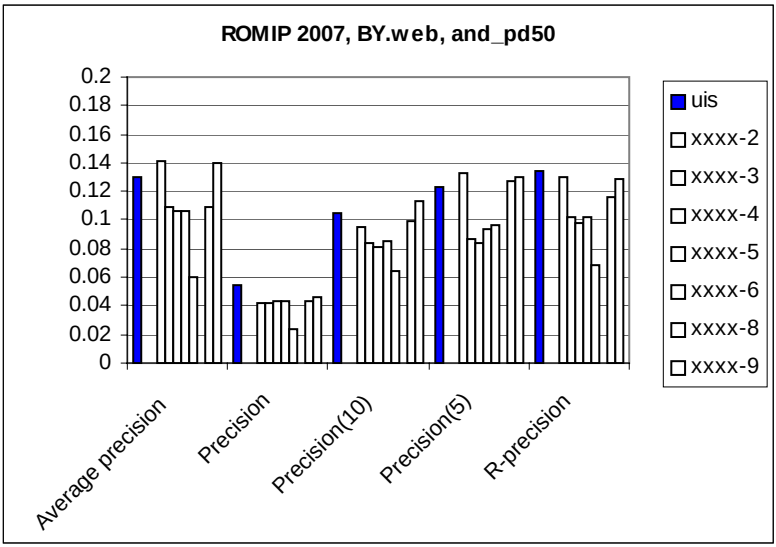


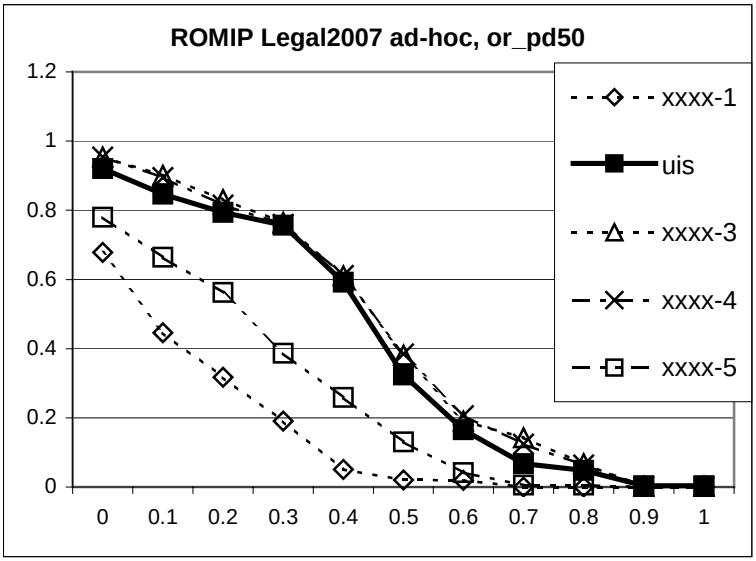
Рис. 4 Различные метрики для дорожки поиска по BY.WEB, таблица релевантности «AND»

2.3 Дорожка поиска по коллекции нормативных документов

Перед участниками поставлена задача классического ad hoc поиска документов по запросу пользователя. Коллекция документов состояла из 348410 текстов нормативно-правовых актов РФ, предоставленных компанией Кодекс.

Оргкомитет представил список из 14797 запросов, выделенных из лога запросов к правовому разделу портала www.kodeks.ru. Для каждого из запросов необходимо было выполнить поиск и выдать не более 100 документов.

На Рис. 5-8 представлены графики метрик оценки качества поиска, полученных с использованием «сильной» и «слабой» шкалы оценки релевантности.



Р

ис. 5 11-точечный график полноты/точности для дорожки поиска по нормативной коллекции, таблица релевантности «OR»

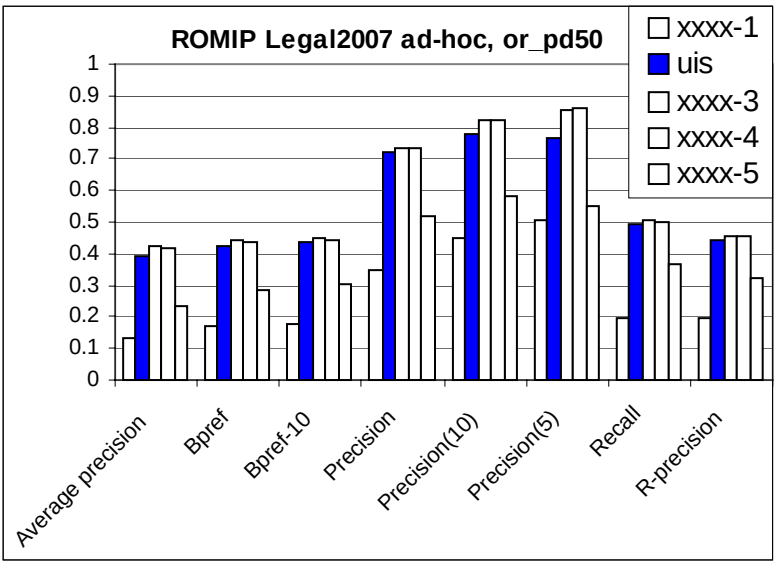


Рис. 6 Различные метрики для дорожки поиска по нормативной коллекции, таблица релевантности «OR»

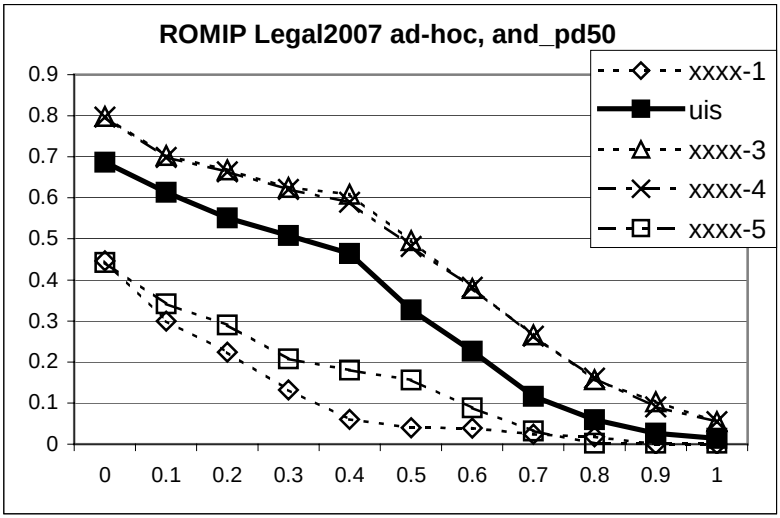


Рис. 7 11-точечный график полноты/точности для дорожки поиска по нормативной коллекции, таблица релевантности «AND»

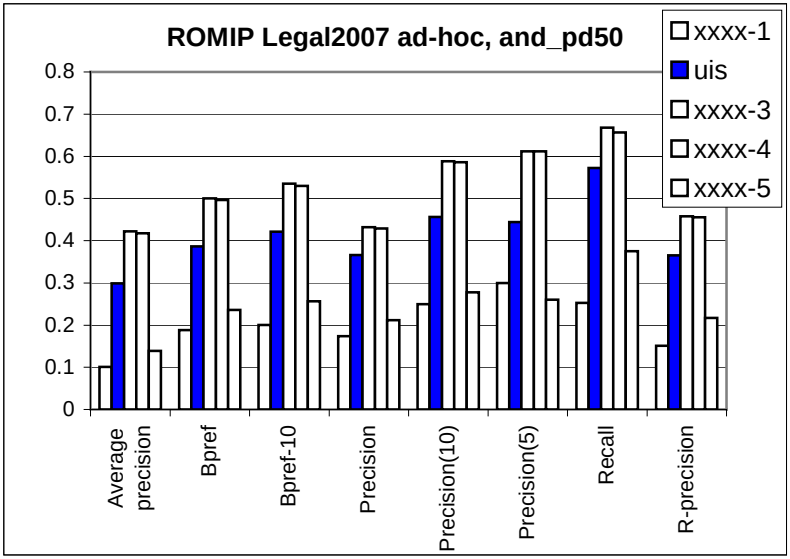


Рис. 8 Различные метрики для дорожки поиска по нормативной коллекции, таблица релевантности «AND»

2.4. Дорожка поиска: анализ результатов и выводы

Анализ полученных графиков позволяет сделать следующие выводы:

1) Данный алгоритм показал лучший результат в дорожке поиска по коллекции BY.WEB по метрике average precision, при «слабых» требованиях к релевантности, и близкий к лучшему результат в дорожке поиска нормативных документов по той же метрике.
2) Показатели метрик P(10), P(5) несколько хуже (в сравнении с другими алгоритмами, представленными в РОМИП'2007)
3) Результаты сильно зависят от таблицы релевантности – при «сильных» требованиях к релевантности результаты значительно хуже.

Отметим, что при подборе оптимальных коэффициентов для алгоритма использовались те же (AveragePrecision, OR_pd50) метрики, которые показали лучший результат. Можно предположить, что, если подбирать оптимальные параметры на основе других метрик, то можно улучшить результаты по соответствующим метрикам.

Дальнейший анализ результатов основан не на средних показателях метрик по всем запросам, а на анализе результатов по отдельным запросам и сравнении результатов с «лидером» - лучшим результатом по каждому запросу алгоритмов, представленных на РОМИП'2007.

Рассмотрим детально несколько запросов, по которым происходит наибольшее отставание от «лидера». Анализировались следующие запросы:

arl39274 «ликвидация учреждения»
arl44427 «федеральный закон о рекламе»
arl37798 «добросовестный приобретатель»
arl39432 «международные договоры»
arl37624 «гражданский кодекс»
arl33576 «о статусе судей»
arl37010 «банкротство индивидуального предпринимателя»

По каждому запросу было просмотрено 5-10 случайных сильно релевантных документов не попавших в первые 100 выдачи.

- Было обнаружено, что ~3% документов не было загружено в базу, что могло незначительно, но негативно сказаться на результатах.

- Можно сделать предположение, что качество поиска было бы выше, если бы система придавала больший вес тем документам, у которых слова запроса содержатся в заголовках документа:
  - По запросу «международные договоры» могли бы найти документы с заголовками:
    1. Релевантный документ d126877, «О порядке заключения, исполнения и денонсации международных договоров СССР (не применяется. Следует руководствоваться Федеральным законом "О международных договорах от 15 июля 1995 года N 101-ФЗ)»;
    2. Релевантный документ d32333, «О Федеральном законе «О внесении изменения и дополнений в Федеральный закон "О международных договорах Российской Федерации» (с изменениями на 5 декабря 1996 года)»;
  - По запросу «гражданский кодекс» могли бы найти документы с заголовками:
    1. Релевантный документ d28500, «Гражданский кодекс Российской Федерации (часть первая статьи 1- 453) (с комментарием) (с изменениями на 10 января 2006 года)», однако сам документ очень большой ~600kb;
    2. Релевантный документ d243819, «О внесении дополнения в Федеральный закон «О введении в действие части третьей Гражданского кодекса Российской Федерации»;
    3. Релевантный документ d317834, «Гражданский кодекс Российской Федерации (часть первая статьи 1- 453) (с комментарием) (с изменениями на 27 июля 2006 года) (редакция, действующая с 1 сентября 2006 года)».
- Возможно, следует учитывать структуру текста по предложениям, то есть искать все слова (или пары слов) запроса в одном предложении и за такие вхождения давать документу больший вес:
  - Например, по запросу «федеральный закон о рекламе» релевантной была любая ссылка вроде: «Федеральным законом от 18.07.95 N 108-ФЗ «О рекламе»;
  - По запросу «ликвидация учреждения», релевантной фразой является: «О порядке создания, реорганизации и ликвидации предприятий, объединений, организаций и учреждений».

3. Классификация Веб-страниц по классификатору dmoz: машинное обучение vs. работа экпертов

В проблеме автоматического рубрицирования документов существует определенная дискуссия между методами машинного обучения (когда по имеющемуся обучающему множеству автоматически строится решающее правило) и, так называемым, «инженерным подходом» (когда решающее правило формируется экспертами). В рамках настоящей работы – решении задачи по автоматической классификации документов коллекции BY.web по 247 рубрикам Веб коллекции DMOZ – у нас была возможность сравнить два этих подхода.

3.1. Тематический анализ текста на основе Общественно-политического тезауруса

Авторы работы ранее разработали метод автоматического построения тематического представления содержания текста на основе Общественно-политического информационно-поискового тезауруса для автоматического концептуального индексирования [4, 5] – лингвистического ресурса, содержащего 35 тысяч понятий, 90 тысяч текстовых входов, в среднем каждое понятие связано с учетом иерархии отношений с 30 другими.

Приведем основные сведения о тематическом представлении текста, которые нам понадобятся при дальнейшем изложении результатов участия в РОМИП 2007.

Тематическое представление текста (Рис.9) строится в результате последовательности следующих шагов:

1) в тексте определяются текстовые вхождения терминов тезауруса;
2) используя иерархию тезаурусных связей производится разрешение многозначности – определяются понятия тезауруса, упоминавшиеся в тексте;
3) далее сеть понятий текста кластеризуется на «тематические линии» вокруг, так называемых, «центров тематических линий» - наиболее значимых понятий;
4) среди всех тематических линий выделяются «главные тематические линии» - все имеющие сильные текстовые связи между собой (элементы разных тематических линий находятся близко друг от друга), «локальные тематические линии» - связанные с «главными», все остальные тематические линии – рассыпаются.

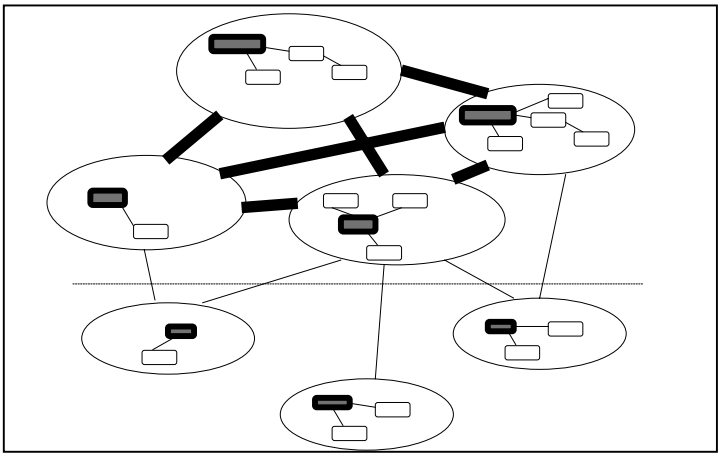


Рис.9 Структурная схема тематического представления текста. Выше линии схематически изображены главные тематические линии, ниже линии – локальные тематические линии

Место понятия тезауруса в тематическом представлении определяет оценку значимости $\omega(d; D)$ понятия относительно содержания текста – наибольший вес получают центры главных тематических линий, наименьший – понятия не входящие в главные или локальные тематические линии, а также многозначные.

Окончательно вес понятия сглаживается учетом частотности понятия в документе ($\alpha = 0.7$):

$$\theta(d) = \alpha \cdot \omega(d; D) + (1 - \alpha) \cdot \frac{freq(d; D)}{\max\limits_{c} freq(c; D) + 2}$$

3.2. Использование методов машинного обучения

Для каждой рубрики в Веб коллекции DMOZ представлен набор сайтов. В качестве множества положительных примеров для обучения мы использовали множество веб-страниц данных сайтов, в качестве множества отрицательных примеров – остальные документы.

Мы рассматривали разрабатываемый нами метод машинного обучения ПФА [6], который строит логические формулы вида

$$R = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right]$$

где d_{ijk} – может быть либо леммой, либо понятием Общественно-политического тезауруса, либо понятием тезауруса с учетом иерархии.

В силу некоторых ограничений для обеспечения быстродействия при работе над заданиями РОМИП мы ограничивали одной тысячей количество положительных примеров и пятью тысячами количество отрицательных. При этом три тысячи отрицательных примеров набиралось случайным образом из положительных примеров «соседних веток» иерархии рубрикатора.

Метод ПФА удобен тем (он для этой цели и разрабатывается), что позволяет экспертам оценить качество получаемого решающего правила, и даже при необходимости подправить решающее правило.

Наши эксперименты для обучающей коллекции показали, что применяемый метод склонен к переобучению.

Рассмотрим, например, рубрику №135 «Спорт -- Боевые искусства».

Отметим, что метод ПФА получает достаточно простую формулу (с показателями качества рубрицирования на обучающем множестве полнота = 0.52, точность = 0.88, F-мера = 0.82) описания смысла документов, отнесенных к данной рубрике:

```
[Тип = T | Имя = БОЕВЫЕ ИСКУССТВА ]
```

здесь и далее запись «Тип = T» означает понятие тезауруса с расширением по иерархии, «Тип = C» означает понятие тезауруса без расширения по иерархии, «Тип = L» означает лемму.

Однако в конечном счете алгоритм предпочитает более сложную формулу с лучшими показателями на обучающем множестве (полнота = = 0.82, точность = 0.98, F-мера = 0.96):

```
    (   [Тип = L | Имя = КАРАТЭ ])
OR (    {   [Тип = C | Имя = ХОККЕЙНЫЙ КЛУБ ]
         OR [Тип = T | Имя = ОХРАННОЕ ПРЕДПРИЯТИЕ ]}
    AND
         [Тип = T | Имя = БЕДСТВИЕ ])
OR (    {   [Тип = C | Имя = КУЛЬТУРА ]
         OR [Тип = C | Имя = СЕВЕРО-ЗАПАДНАЯ ЧАСТЬ ]}
    AND
         [Тип = C | Имя = ОДЕЖДА ]
    AND
         [Тип = T | Имя = ВЕРОВАТЬ ])
OR (    {   [Тип = C | Имя = МЕДИЦИНСКОЕ УЧРЕЖДЕНИЕ ]
         OR [Тип = C | Имя = КРЫЛАТСКОЕ ]}
   AND
         [Тип = T | Имя = ВОСТОЧНЫЕ ЕДИНОБОРСТВА ])
```

```
OR ( [Тип = C | Имя = МАСЛЕНИЦА ])
OR ( [Тип = L | Имя = ДЗЭНИН ])
OR ( [Тип = C | Имя = САМООБОРОНА ]
   AND [Тип = в дереве | Имя = ИСТОРИЧЕСКИЕ НАУКИ ])
```

Причины эффекта переобучения становятся понятны, если более внимательно посмотреть на список сайтов обучающего множества. Среди сайтов типа «www.karate.ru», «aikido.kuban.net», «capoeira.narod.ru» и т.п., встречаются также:

- tornado.spb.ru – не только «спортивный клуб таэквон-до», но и хоккейный клуб, а также охранные услуги и системы сигнализации;
- kryltkd.narod.ru – (Спортивный Клуб "Олимп" в Крылатском и Кунцево) не только «тхэквондо, кикбоксинг, самооборона», но и «развитие гибкости, ОФП».

Поэтому чем точнее мы будем стараться приблизить обучающее множество, тем больше возникает опасность построить решающее правило на основе «случайных» с точки зрения формулировки рубрики свойств.

В результате экспертного изучения качества получаемых формул нами был сделан вывод о трудности реализации метода машинного обучения без специальных мер по очистке обучающего множества.

Поэтому для решения задачи классификации Веб-страниц мы стали применять другой метод.

3.3. Экспертное описание смысла рубрик

Недостатком при использовании «инженерного подхода» для решения задачи автоматической рубрикации считается высокая трудоемкость описания решающего правила рубрики.

Авторы работы используют «инженерный подход» в течение достаточно длительного времени (с 1996 года [7]) и за это время разработали технологию, снижающую трудоемкость работы.

Одной из мотиваций авторов настоящей работы было получить оценки трудоемкости для построения описания рубрикатора.

Всего для решения задачи описания рубрикатора было затрачено 8 человеко-часов двух экспертов (2 эксперта по 4 часа).

Основная идея состоит в существенном использовании иерархии тезаурусных связей. То есть смысл рубрики описывается достаточно короткой формулой над понятиями тезауруса, а затем производится автоматическое расширение построенной формулы.

Подробнее [8] – смысл рубрики описывается как дизъюнкция:

$$R = \bigcup_i D_i$$

Приведем необходимые численные данные. Всего было описано 234 рубрики из 247. Для 234 рубрик описано 265 дизъюнктов.

Каждый дизъюнкт представляется конъюнкцией (всего 334 конъюнкта):

$$D_i = \bigcap_j K_{ij}$$

Каждый конъюнкт представляется в виде совокупности «положительных» и «отрицательных» «опорных концептов», которые задают в регулируемом порядке применения правила добавления и удаления множеств подчиненных по иерархии концептов (здесь *f(.)* – схематическое обозначения правил учета иерархии):

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn})$$

Всего было использовано 899 опорных концептов.

$$R = \bigcup_i D_i = \bigcup_i \left[\bigcap_j K_{ij} \right] = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right]$$

Далее, расширяя по иерархии тезауруса, получаем полное представления рубрики (где уже для всех рубрик задействовано 40161 концептов - естественно, с учетом повторения – и 107897 текстовых входов).

Отметим, что некоторые отношения в тезаурусе снабжены пометкой «аспекта», что при автоматическом расширении ведет к простановке флага «необходимости подтверждения» - рубрика не будет выводиться для текста при наличии только «неподтвержденных» понятий, при наличии же подтверждения – подтвержденные понятия учитываются в полной мере.

Вес рубрики определяется как максимум весов дизъюнктов. Вес дизъюнкта определяется как взвешенная сумма весов конъюнктов и текстовых связей между конъюнктами:

$$\theta(D_i) = \frac{\sum_{j=1}^{m} \theta(K_{ij}) + \sum_{j<k} S(K_{ij}, K_{ik})}{m + C_m^2}$$

где вес конъюнкта – максимум из весов приписанных понятий (с учетом наличия подтверждения):

$$\theta(K_{ij}) \quad = \quad \min\left\{1.0; \max\left(\theta(d_{ijk}),\ \chi \cdot \theta(p_{ijm})\right)\right\}$$

вес текстовой связи между конъюнктами:

$$S(K_{ij}, K_{ik}) = \min\{1.0; \left(\sum s(c_{ijq} \in K_{ij}, d_{ikw} \in K_{ij})\right) \Big/ \max s(c \in D, d \in D)\}$$

При описании рубрик классификатора эксперты в основном ориентировались на формулировку рубрики. В единичных случаях эксперты заходили на сайт dmoz.org для уточнения объема рубрики.

Общественно-политический тезаурус покрывает практически все предметные области, отражаемые в деловой прозе – нормативных актах, СМИ федерального уровня. Поэтому для решения задачи описания рубрик потребовалось ввести в тезаурус дополнительно только восемь понятий, для описаний специфических экстремальных видов спорта.

Рассмотрим пример описания экспертами для рубрики №135 «Спорт -- Боевые искусства».

Опорное булевское выражение состоит из одного понятия *БОЕВЫЕ ИСКУССТВА (E)* с меткой «E» полного расширения по тезаурусу.

В состав расширенного булевского выражения входят помимо исходного следующие понятия: *АЙКИДО, ДЖИУ-ДЖИТСУ, ДЗЮДО, КАРАТЭ, САМБО, ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ*.

Понятия тезауруса, соответствующие людям (*ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ)* входят в рубрику с пометкой подтверждения, поскольку появление соответствующих слов в тексте еще не означает, что текст посвящен боевым искусствам.

Рассмотрим пример более сложного описания на примере рубрики №43 «домашний ремонт»:

```
          (     РЕМОНТ(N)
           OR   КАПИТАЛЬНЫЙ РЕМОНТ(N)
           OR   ТЕКУЩИЙ РЕМОНТ(N)
           OR   РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ(N) )
AND
          (     ЖИЛОЕ ЗДАНИЕ(L)
           OR   ЖИЛОЕ ПОМЕЩЕНИЕ(L)
           OR   КВАРТИРА(L) )
```

здесь пометка «L» означает, что предусматривается только расширение по отношениям типа «ВЫШЕ-НИЖЕ», пометка «N» означает отсутствие расширения.

3.4. Результаты применения «инженерного подхода» по дорожке классификации Веб-страниц

На Рис.10, Рис.11 приведены результаты при слабом и сильном согласии между оценщиками для дорожки классификации Веб-страниц коллекции ROMIP.BY.

Как справедливо отмечают организаторы, не вполне корректно сравнивать результаты нашего прогона «thescateg» с результатами других систем – из-за разницы в представлении результатов – мы представили сортированный список с контролем, чтобы на каждый сайт приходилось не более пяти документов, в то время как другие участники представляли несортированные данные по пять документов на сайт.

Тем не менее, достижение показателей качества рубрицирования при нестрогом согласии между экспертами:

- полнота = 81.7%;
- точность = 68.2%;
- F-мера = 72.9%;

следует признать весьма успешным для 8 часов трудозатрат экспертов.

Отметим, что полученные результаты позволяют произвести улучшение путем анализа ошибок и внесения соответствующих изменений в описание рубрик.

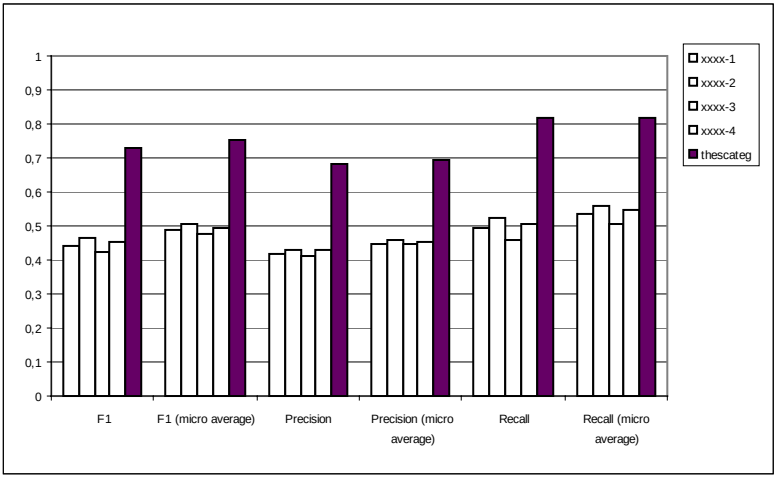


Рис.10 РОМИП2007: классификация веб-страниц [ог]

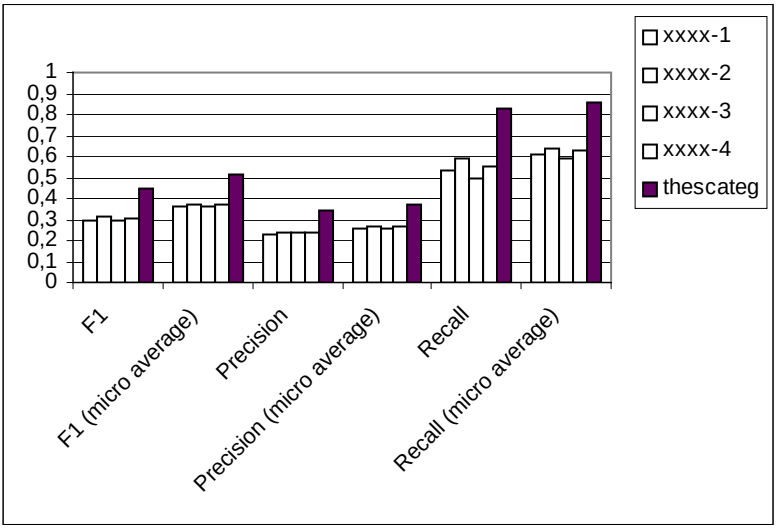


Рис.11 РОМИП2007: классификация веб-страниц [and]

3.5. Выводы по дорожке классификации Веб-страниц

Полученные результаты позволяют сделать следующие выводы:

- Существуют задачи классификации текстов, когда нет достаточно качественной обучающей коллекции:
  - Нет достаточного множества обучающих примеров или ручная классификация проведена недостаточно последовательно;
  - В таких условиях применения методов машинного обучения очень проблематично;
- При машинном обучении системы извлекают некоторые знания о языке и мире, которые можно условно подразделить на:
  - Общие знания о языке и мире, необходимые для работы различных приложений в разнообразном круге предметных областей, и
  - «Текущие знания», характерные именно для текущей задачи, текущей коллекции, данного типа пользователей и т.п.;
  - Значимую часть знаний о современной жизни общества и современном языке деловой прозы нам удалось упорядочить в рамках понятийных структур Общественно-политического тезауруса;

- В текущем эксперименте у нас не было возможности сделать предварительный прогон, оценить и исправить ошибки и неточности описания рубрик:
  - В обычной практике проводится несколько итераций, консультаций с экспертами;
  - Имеются средства анализа расхождения между системой и экспертами, расхождения также описываются через понятия тезауруса;
  - Поэтому имеются определенные возможности улучшения полученных результатов на основе тезаурусных знаний.

4. Классификация Веб-сайтов по классификатору dmoz

Мы впервые приняли участие в дорожке по классификации Веб-сайтов. Для расчета тематической оценки была предложена формула, которая казалась авторам достаточно логичной:

$$Rank\ R_i(S_j) = Avg\ R_i(S_j) \bullet \frac{cnt80\ R_i(S_j)+1}{cnt\ R_i(S_j)+1} \bullet$$

$$\bullet \min\left\{\frac{cnt\ R_i(S_j)}{N_S}, 1.0\right\} \bullet \min\left\{\frac{cnt\ R_i(S_j)}{cnt\ S_j \bullet {}^{PS}\!/_{100}}, 1.0\right\}$$

здесь: $Avg\ R_i(S_j)$ – средний вес страниц, отнесенных к рубрике; $cnt\ R_i(S_j)$– кол-во страниц сайта, отнесенных к рубрике; $cnt80\ R_i(S_j)$ – кол-во страниц сайта с высоким весом по тематическому представлению, отнесенных к рубрике; $cnt\ S_j$ – общее кол-во страниц сайта; PS – минимальный процент страниц, посвященных рубрике (=40%); NS – мин. кол-во страниц, посвященных рубрике (=200).

То есть преимущество должны получать сайты, где:
- больше средний тематический вес страниц;
- больше удельный вес высоко оцененных страниц;
- больше чем 200 страниц отнесено к рубрике;
- больше чем 40% страниц отнесено к рубрике.

Результаты по классификации Веб-сайтов представлены на Рис.12 и Рис.13.

Как нетрудно видеть, по сравнению, с классификацией Веб-страниц – полнота рубрикации в целом не упала, в то время как точность уменьшилась.

По-видимому, для повышения качества классификации сайтов стоит устанавливать более жесткие критерии на долю страниц сайта, отнесенных к рубрике.

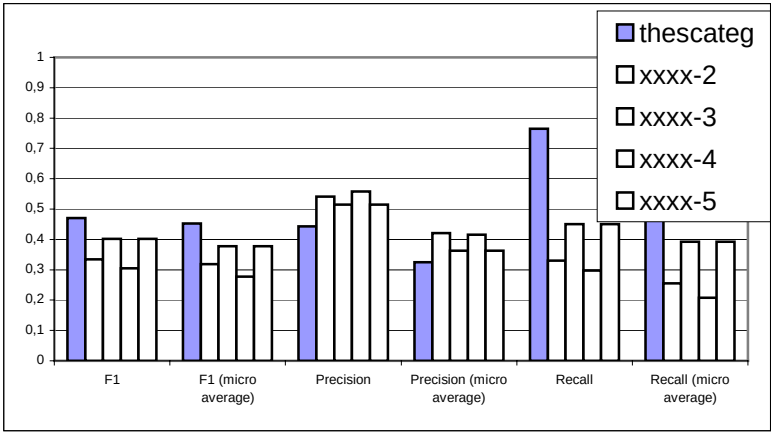


Рис.12 РОМИП2007: классификация Веб-сайтов [or]

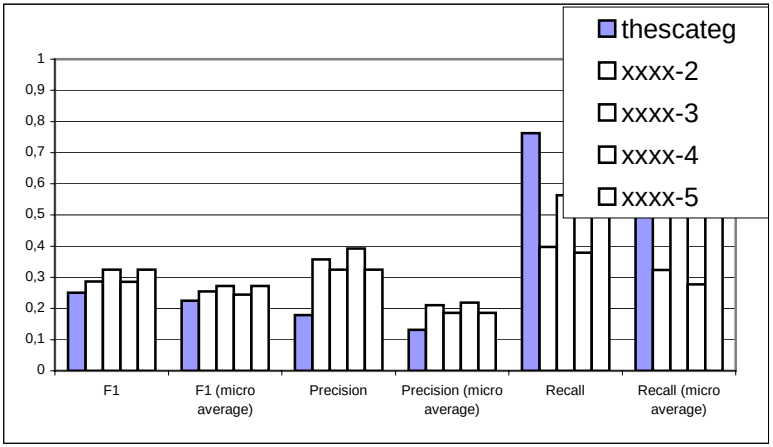


Рис.13 РОМИП2007: классификация Веб-сайтов [and]

Заключение

Наш коллектив использует возможность участия в РОМИП как эффективный способ уточнения параметров применяемых нами методов.

Опыт показывает, что для получения высоких результатов в дорожках РОМИП является полезным более полный учет специфики решаемых задач, настройка на особенности коллекций.

Литература

[1] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Российский семинар по оценке методов информационного поиска. Труды второго российского семинара РОМИП'2004 (Пущино, 01.10.2004) – СПб: НИИ Химии СПбГУ. – 2004. – С.62-89.

[2] Агеев М.С., Добров Б.В., Оптимизация параметров алгоритма поиска на основе анализа оценок экспертов // Российский семинар по Оценке Методов Информационного Поиска. Труды третьего российского семинара РОМИП'2005 (Ярославль, 6 октября 2005г.) – СПб.: НИИ Химии СПбГУ, 2005. – С.78-88

[3] Красильников П.В., Воспроизведение лучших результатов ad hoc поиска семинара РОМИП // Интернет-математика 2007: Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. ун-та, 2007. - С.91-97.

[4] Лукашевич Н.В., Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.

[5] Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни – М.: Наука – 2002. – Т.2 - С.338-346.

[6] Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: Дис. канд. физ-мат. наук: 05.13.11 / Московский гос. унив. - Москва, 2005. (http://www.cir.ru/docs/ips/publications/2005_diss_ageev.pdf)

[7] Лукашевич Н.В., Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. - 1996. - N 10. - С.22-30.

[8] 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.

UIS RUSSIA at ROMIP 2007: ad hoc search and text categorization

Mikhail S. Ageev, Boris V. Dobrov, Pavel V. Krasilnikov, Natalia V. Loukachevitch, Andrey M. Pavlov, Alexey V. Sidorov, Sergey V. Shternov

In the paper we describe methods used by the 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 ROMIP 2007 (Russian Information Retrieval Evaluation Seminar) tracks. We participated in the ad hoc track on a web collection, the ad hoc track on a legal documents collection, and text categorization of web-pages and web-sites. In ad hoc tracks we used an retrieval model based on ranking of several factors: tf*idf weights, proximity of query words in a text and quorum search. In text categorization tracks we employed the knowledge-based categorization technique, which we traditionally used for text categorization since 1996 and which utilises knowledge described in Sociopolitical thesaurus of UIS RUSSIA.