RCO на РОМИП 2004

© Плешко В.В., Ермаков А.Е., Голенков В.П.

ООО «Гарант-Парк-Интернет»
rco@metric.ru

Аннотация

Настоящая работа является отчетом об экспериментах, проведенных в рамках инициативы РОМИП. В результате выполнения дорожек по поиску web-страниц и поиску правовых документов были получены численные оценки влияния учета словоформ и словосочетаний на показатели полноты и точности. Получены предварительные результаты для дорожек по классификации web-сайтов и классификации правовых документов. Приведено описание экспериментов по поиску биографических фактов, связанных с заданными персонами.

Введение

В 2003 году прошел первый Российский семинар по Оценке Методов Информационного Поиска. Основными целями семинара являются создание русскоязычных текстовых корпусов для получения количественных оценок работы информационно-поисковых систем (ИПС), а также создание сообщества и консолидация усилий разрозненных групп разработчиков и исследователей в области построения ИПС. Среди зарубежных инициатив наиболее близкими к РОМИП являются американская конференция TREC и европейский форум CLEF.

Результаты работы РОМИП'2003 в целом можно считать положительными. Во-первых, был приобретен уникальный опыт создания корпусов, а также разработки специализированного программного обеспечения для оценки качества ответов ИПС. Во-вторых, было создано сообщество разработчиков и исследователей, поддержавших инициативу РОМИП. Тем не менее, в силу ряда факторов, недостаточное качество полученных в процессе первого годового цикла РОМИП корпусов не позволило придать им

«официальный» статус. Именно повышению качества оценок и исходных данных (обучающих выборок) было посвящено подавляющее большинство сообщений на web-форуме, посвященном деятельности семинара.

Второй семинар собрал большее число участников и предметом его изучения стало большее количество дорожек (заданий). Помимо дорожек прошлого года по поиску web-страниц и классификации web-сайтов, в 2004 году участникам были предложены задания по классификации и поиску правовых документов, а также задание по поиску биографических фактов, связанных с персонами. Осознавая важность инициативы РОМИП для всей отрасли, наш коллектив принял участие во всех перечисленных дорожках.

1. Документальный поиск

Предложенное системам-участникам задание заключалось в выполнении большого числа поисковых запросов, которые были представлены в виде последовательностей ключевых слов, описывающих информационную потребность пользователя, против коллекции документов. Ответом на запрос был список документов, упорядоченных по убыванию соответствия запросу. Для оценки был случайно отобран ряд запросов, для каждого из которых в рассмотрение принимались первые 100 документов, выданных системой.

Было организовано две дорожки по документальному поиску. В дорожке web adhoc в качестве коллекции было использовано подмножество сайтов домена narod.ru размером около 6.3 Gb (более 720000 web-страниц), а в качестве поисковых запросов использовались фрагменты журналов поисковых машин Яндекс и Рамблер (всего более 24000 запросов). Во второй дорожке, legal adhoc, в качестве коллекции было взято подмножество документов из справочно-правовой системы «Кодекс» размером около 1.6 Gb (более 67000 документов), в качестве запросов использованы фрагменты журналов справочно-правовых систем «Кодекс» и «Гарант-WWW» (около 13000 запросов).

По каждой дорожке система могла предоставить несколько вариантов выполнения заданий – прогонов.

1.1 Описание системы

Основной целью авторов при выполнении заданий по документальному поиску было количественно оценить, как влияют факторы учета словоформ и порядка следования слов в запросе на

основные показатели качества поиска. На прошлом семинаре был проведен эксперимент по учету указанных факторов [1], однако окончательных выводов сделано не было.

Для указанной цели был создан прототип поисковой системы, использующей классический tf*idf метод вычисления степени соответствия документа запросу.

По умолчанию, система производила поиск документов, используя оператор «NEAR» (все слова запроса должны встретиться в тексте документа в пределах 100 слов в любом порядке), без учета словоформ русского языка. Кроме того, не использовался список стоп-слов.

При этом можно было независимо задать режим учета словоформ всех слов запроса, а также режим ранжирования документов, содержащих все слова запроса в виде одного словосочетания, выше остальных документов (для краткости назовем его «режимом учета словосочетаний»).

Среди особенностей реализации можно выделить обработку глагольных форм. Все словоформы глагола можно разбить на 6 групп: прямые основные, прямые причастные в действительном залоге, прямые причастные в страдательном залоге, а также три группы, соответствующие возвратным формам. При синтезе всех словоформ глаголов по словоформе, встретившейся в запросе, отбирались только формы, входящие с ней в одну группу.

1.2 Описание экспериментов

Для дорожки legal adhoc было выполнено 4 прогона:

1. базовый, настройки по умолчанию,
2. включен режим учета словоформ,
3. включен режим учета словосочетаний,
4. одновременно включены режимы учета словоформ и словосочетаний.

В силу внешних причин было принято решение выполнить только 2 прогона для дорожки web adhoc:

1. базовый, настройки по умолчанию,
2. включен режим учета словоформ.

1.3 Результаты

При оценке основные показатели были получены 4-мя способами для legal adhoc и 8-ю способами для web adhoc. В таблице 1 приведены результаты для способов оценки legal-or-pd50 (хотя бы один из оценщиков счел документ соответствующим запросу;

показатели считались только для первых 50 документов из выдачи системы) и web-or-withdesc-pd50-2004 (хотя бы один из оценщиков счел документ, соответствующим запросу; оценщикам были предоставлены расширенные описания запросов; показатели считались только для первых 50 документов из выдачи системы; показатели рассчитаны только для запросов, отобранных для цикла 2004 года).

Таблица 1. Результаты прогонов для дорожек web adhoc и legal adhoc.

Run	Recall	Precision(5)	Average precision	Precision(10)	R-precision	Precision
web 1	0.2269	0.4417	0.1753	0.3438	0.2236	0.3955
web 2	0.4299	0.5125	0.2825	0.4354	0.3427	0.3777
legal 1	0.2026	0.5461	0.1650	0.4753	0.1942	0.5300
legal 2	0.3729	0.5843	0.2718	0.5618	0.3367	0.5659
legal 3	0.2202	0.5663	0.1809	0.4910	0.2110	0.5526
legal 4	0.3910	0.6584	0.3086	0.6225	0.3611	0.5994

На рисунках 1 и 2 приведены 11-точечные графики точность-полнота для дорожек web adhoc и legal adhoc соответственно.

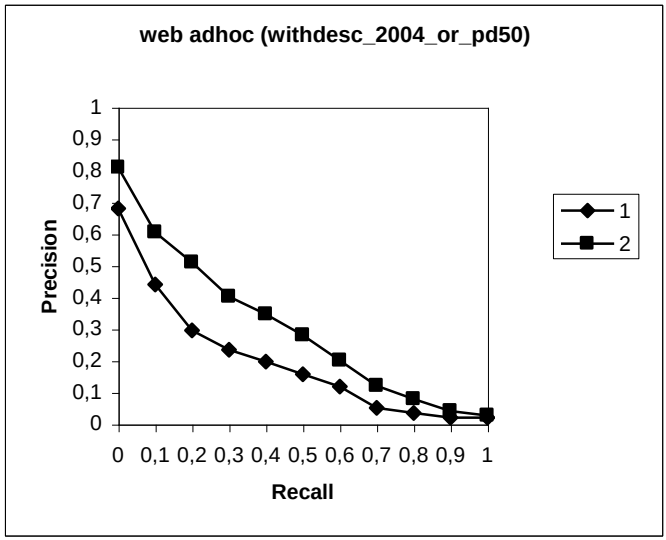

Рисунок 1. 11-точечные графики для дорожки web adhoc

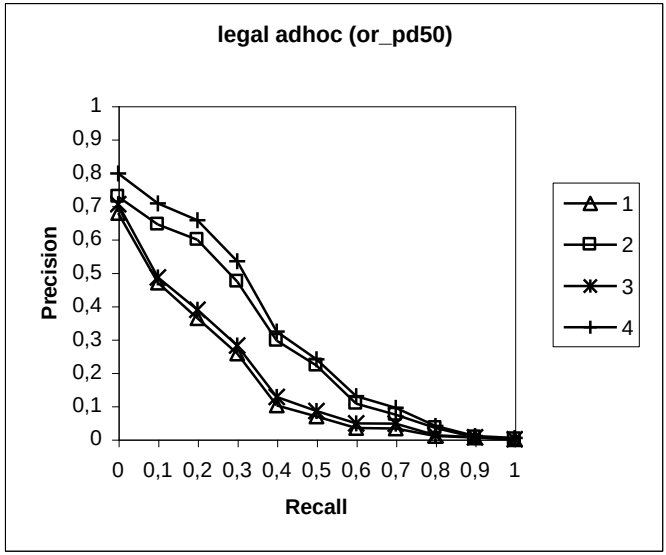

Рисунок 2. 11-точечные графики для дорожки legal adhoc

1.4 Анализ влияния учета словоформ и словосочетаний

Оценим зависимость показателей Recall, Precision, Precision(5), Precision(10) методом линейной регрессии. Рассмотрим два прогона из одной дорожки. Для каждого из показателей составим список пар значений на этих прогонах для всех способов оценки. Каждую пару значений можно представить как точку на плоскости. Затем найдем прямую, проходящую через начало координат, имеющую наименьшей сумму квадратов уклонений вдоль оси y по всей совокупности точек. Коэффициент уклона этой прямой примем за оценку вклада фактора в значение показателя.

Итак, будем для каждого из упомянутых показателей рассчитывать зависимости на следующих парах прогонов:

– legal-2 vs legal-1 (учет словоформ – правовые документы);
– legal-4 vs legal-3 (учет словоформ при учете словосочетаний – правовые документы);
– web-2 vs web-1 (учет словоформ – web-страницы);
– legal-3 vs legal-1 (учет словосочетаний – правовые документы);
– legal-4 vs legal-2 (учет словосочетаний при учете словоформ – правовые документы);

– legal-4 vs legal-1 (учет словоформ и словосочетаний – правовые документы).

Для иллюстрации подхода в таблице 2 приведена зависимость legal-2 vs legal-1 для показателя Recall, а на рисунке 3 графически показано решение задачи линейной регрессии.

Таблица 2. Зависимость legal-2 vs legal-1 для показателя Recall.

Тип оценки \ Прогон	legal-1	legal-2
and	0.3163	0.6027
and_pd50	0.2871	0.4960
or	0.2324	0.4857
or_pd50	0.2026	0.3729

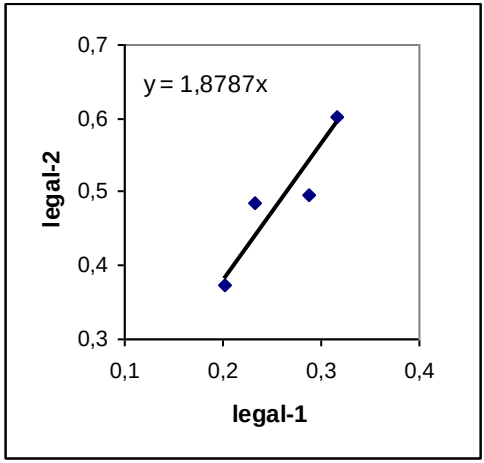

Рисунок 3. Иллюстрация решения задачи линейной регрессии для зависимости legal-2 vs legal-1 для показателя Recall.

Результаты расчетов для всех показателей и всех видов зависимости приведены в таблице 3. В первом столбце таблицы приведены названия зависимостей, в остальных столбцах – оценки линейных коэффициентов регрессии для каждого из показателей. В последних двух строках даны значения произведений линейных коэффициентов соответствующих зависимостей.

Таблица 3. Оценка влияния учета словоформ и словосочетаний на показатели Recall, Precision, Precision(5), Precision(10).

Зависимость \ Показатель	Recall	Precision	Precision(5)	Precision(10)
legal-2 vs legal-1	1.8787	1.0053	1.0573	1.1843
legal-4 vs legal-3	1.8627	1.0220	1.1787	1.2816
web-2 vs web-1	1.6821	0.9561	1.1664	1.2606
legal-3 vs legal-1	1.0610	1.0457	1.0416	1.0378
legal-4 vs legal-2	1.0508	1.0629	1.1596	1.1230
legal-4 vs legal-1	1.9767	1.0690	1.2275	1.3298
legal-4 vs legal-2 * legal-2 vs legal-1	1.9741	1.0685	1.2260	1.3300
legal-4 vs legal-3 * legal-3 vs legal-1	1.9763	1.0687	1.2277	1.3300

По данным, приведенным в таблице 3, можно сделать следующие наблюдения:

- учет словоформ повышает полноту на 70-90% (независимо от того, учитывать или нет при поиске словосочетания);
- при учете словоформ общая точность не меняется либо падает на 5% (находится больше документов, не все из них релевантные);
- точность на первых 5 документах при учете словоформ повышается на 6-17%, а на первых 10 документах – на 18-28% (за счет роста частот терминов при учете словоформ ранжирование по релевантности получается более точным);
- учет словосочетаний без учета словоформ «равномерно» повышает плотность релевантных документов в ответе системы на 4% (прирост всех показателей точности одинаков);
- учет словосочетаний при учете словоформ повышает общую точность на 6%, в то время как точность в начале выдачи повышается на 12-16%;
- из последних трех строк таблицы следует, что влияние факторов учета словоформ и учета словосочетаний является независимым друг от друга и прирост значений показателей точности равен сумме вкладов обоих факторов.

Полученные результаты весьма далеки от того, чтобы считаться исчерпывающими, однако полученные количественные оценки могут служить ориентиром для дальнейших исследований.

1.5 Выводы и пожелания

Большое разнообразие оценок, с одной стороны, дало возможность получить устойчивую оценку факторов, но, с другой стороны, затрудняет представление результатов. На взгляд авторов, на следующем семинаре нужно остановиться на единой оценке.

В первую очередь, это касается способа сведения оценок от ассессоров. Оценок для каждой пары документ-запрос должно быть нечетное число, чтобы результирующая оценка была получена путем «голосования». Кроме того, тяжело однозначно оценить использование в оценке вторых 50 документов из ответа системы, которые не входят в общий пул запросов. Учет неоцененных документов как нерелевантных, а случайно попавших в пул как релевантных, вносит дополнительную погрешность в оценку ответа системы.

2. Классификация правовых документов

Участникам был предложен перечень рубрик первого уровня классификатора справочно-правовой системы «Кодекс» (173 рубрики), в качестве обучающей выборки было отобрано подмножество документов (всего 13771), входящих в рубрики данного классификатора. Задача состояла в отнесении оставшейся части коллекции документов к рубрикам классификатора. Каждому документу допускалось присвоить от нуля до пяти рубрик.

2.1 Описание метода

В отличие от экспериментов на РОМИП'2003 [2], где была использована стратегия обучения на положительных примерах, в этом году был опробован метод, использующий также отрицательные примеры при обучении. Опишем метод для случая одного класса.

Пусть дано множество документов D, разделенное на два непересекающихся подмножества Tr и Ts, называемых обучающей и тестовой выборками.

Пусть также дано соответствие между документами и некоторым классом c в виде $\Phi : D \to \{0,1\}$, устанавливающее для каждого документа значение 1, в случае принадлежности документа классу, и 0 – в противном случае.

Требуется построить, используя только информацию из обучающей выборки, функцию $\hat{\Phi} : D \to \{0,1\}$, аппроксимирующую Φ, чтобы число ошибок на тестовой выборке было наименьшим:

$$Error = \sum_{Ts} \left| \Phi - \hat{\Phi} \right| \to \min . \quad (1)$$

Пусть T – множество терминов, каким-либо образом выделенных из документов класса c. Тогда документ можно представить в виде терминологического вектора

$$d_j = (d_{j1}, \ldots, d_{j|T|})^T, \quad (2)$$

где d_{jk} – вес термина t_k в документе d_j.

Описания каждого из классов представим в виде векторов той же размерности, что и вектора документов:

$$c = (c_1, \ldots, c_{|T|})^T, \quad (3)$$

c_k - вес термина t_k в описании класса c.

Степень соответствия между классом c и документом d_j определим как скалярное произведение между их векторными представлениями:

$$CSV(c, d_j) = c \cdot d_j = \sum_k c_k d_{jk} . \quad (4)$$

Будем принимать решение о принадлежности документа классу, если степень соответствия достигнет заданного порога τ. Таким образом,

$$\hat{\Phi}(c, d_j) = \begin{cases} 1, & CSV(c, d_j) \geq \tau, \\ 0, & CSV(c, d_j) < \tau. \end{cases} \quad (5)$$

Искомое значение порога τ^* определяется из отношения правдоподобия, которое минимизирует суммарную ошибку:

$$\frac{\#\{\Phi(c,d) = 1 \,\&\, CSV(c,d) = \tau^*\}}{\#\{\Phi(c,d) = 0 \,\&\, CSV(c,d) = \tau^*\}} = 1, \quad (6)$$

где $\#\{\Phi(c,d) = 1 \,\&\, CSV(c,d) = \tau^*\}$ - количество документов из обучающей выборки, *из* класса c, степень соответствия которых данному классу равна τ^*,

$\#\{\Phi(c,d) = 0 \,\&\, CSV(c,d) = \tau^*\}$ - количество документов из обучающей выборки, *вне* класса c, степень соответствия которых данному классу равна τ^*.

2.2 Описание эксперимента

Набор терминов T, входящих в документы из класса c_i, получался в результате процедуры анализа каждого документа из класса c, в

результате которой выделялись слова и словосочетания документа, имеющие высокий коммуникативный ранг для автора документа [3].

Кроме того, полученный набор терминов подвергается фильтрации по следующим двум условиям:

$$\frac{\#\{\Phi(c,d)=1 \,\&\, t_k \in d\}}{\#\{\Phi(c,d)=1\}} \geq \alpha, \tag{7}$$

$$\frac{\#\{\Phi(c,d)=1 \,\&\, t_k \in d\}}{\#\{t_k \in d\}} \geq \beta. \tag{8}$$

где $\#\{\Phi(c,d)=1 \,\&\, t_k \in d\}$ - число документов из класса c, содержащих термин t_k, $\#\{t_k \in d\}$ - общее число документов, содержащих термин t_k. При вычислениях значения порогов α и β были взяты равными 0.2%.

Вес термина в документе считался следующим образом:

$$d_{jk} = \begin{cases} 1, & t_k \in d_j, \\ 0, & t_k \notin d_j. \end{cases} \tag{9}$$

Вес термина в описании класса вычислялся по следующей формуле:

$$c_k = \left[\frac{\#\{\Phi(c,d)=1 \,\&\, t_k \in d\}}{\#\{t_k \in d\}} \right]^2. \tag{10}$$

На практике чаще используют «классическую» инверсную документную частоту $\log(|Tr| / \#\{t_k \in d_j\})$ или ее квадрат. Однако если инверсная частота отражает разделительную способность термина на всем множестве, то предлагаемый коэффициент отражает способность термина отделять от других документы из заданного класса.

В среднем было получено 100 терминов на рубрику. Среди отобранных терминов оказалось 83% словосочетаний.

2.3 Результаты

Было произведено 4 способа оценки. Первые три способа - для небольшого числа случайно отобранных рубрик:

- документ соответствует рубрике, если все оценщики с этим согласны;
- документ соответствует рубрике, если с этим согласен хотя бы один оценщик;

– документ соответствует рубрике, если он отнесен к ней в классификаторе «Кодекс».

Четвертый, способ оценки совпадал с третьим по сути, но для оценки были отобраны рубрики, которые содержали не менее 40 документов из обучающей выборки.

Результаты прогона приведены в таблице 4.

Таблица 4. Результаты классификации правовых документов.

показатель \ способ	1	2	3	4
F1 (macro average)	0.1027	0.1186	0.1558	0.3355
Recall	0.0627	0.0774	0.0765	0.2327
Precision (macro average)	0.1508	0.4484	0.3333	0.5069
F1	0.0608	0.1107	0.1039	0.2866
Recall (macro average)	0.0779	0.0684	0.1017	0.2507
Precision	0.1727	0.4513	0.3013	0.4216
Accuracy	0.9913	0.9780	0.9888	0.9719
Error	0.0087	0.0220	0.0111	0.0281

Кроме того, при оценке ассессорами-непрофессионалами в качестве одной из систем-участников выступал сам классификатор «Кодекс», названный «идеальной» системой. В таблице 5 приведены результаты «идеальной» системы.

Неожиданно низкий результат «идеальной» системы, а также большой разброс в результатах оценки точности для способов оценки 1 и 2 говорит о том, что мнение оценщиков-непрофессионалов в области права сильно разошлись как между собой, так и с мнением юристов. Результаты, показанные другими системами, близки к приведенным.

Таблица 5. Результаты «идеальной» системы.

показатель \ способ	1	2
F1 (macro average)	0.2329	0.2259
Recall	0.2137	0.1533
Precision (macro average)	0.1852	0.3390
F1	0.1470	0.1763
Recall (macro average)	0.3135	0.1694
Precision	0.1290	0.2430
Accuracy	0.9928	0.9808
Error	0.0072	0.0192

Гораздо большие вопросы вызвал низкий результат при третьем способе оценки. Причина крылась в том, что распределение документов в рубриках на обучающую и тестовую выборки было несбалансированным – ряд рубрик содержал по 1-2 документа в качестве данных для обучения.

Четвертый последний способ оценки явился наиболее адекватным. Среди результатов систем есть прогоны с показателями полноты свыше 40%, а точности – свыше 60%.

2.4 Выводы и пожелания

Эксперимент с оценкой непрофессионалами нормативно-правовых документов следует признать неудачным.

Качество коллекции оставляет хорошее впечатление. Задача классификации нормативно-правовых документов является сложной и представляет прекрасный полигон для развития методов автоматической классификации документов.

В следующем году интересно было бы решить задачу в постановке, когда каждому документу должна быть присвоена хотя бы одна рубрика. Ведь все документы справочно-правовой системы классифицированы.

Также интересно было бы использовать гипертекстовые связи. Например, приложения к документам чаще всего тематически нейтральны, но относятся к тому же классу.

3. Классификация web-сайтов

Участникам было предложено подмножество интернет-каталога dmoz.org размером около 7 Gb (более 300000 web-страниц, более 2000 web-сайтов), используя которое в качестве обучающей выборки, требовалось соотнести с рубриками каталога dmoz.org (247 рубрик) web-сайты коллекции narod.ru (более 23500 web-сайтов).

3.1 Описание метода

На РОМИП’2003 [2] авторами был использован так называемый метод суперстраниц, когда все страницы сайта конкатенировались в одну, после чего задача классификации web-сайтов сводилась к задаче классификации больших документов. Оказалось, что такой подход имеет явные недостатки, а именно:

– текст навигационных элементов получал неоправданно большой вес, так они повторялись на всех страницах сайта;

– словосочетания, потенциально являющиеся хорошими классификационными признаками, несмотря на высокий коммуникативный ранг, получали заниженный вес из-за малых относительных частот в тексте.

В этом году мы принимали решение о принадлежности web-сайта заданному классу на основе принадлежности этому классу каждой из его страниц поотдельности.

Таким образом, задача классификации сайтов решалась в два этапа. На первом этапе была решена задача классификации web-страниц как описано в п.п. 3.1. На втором этапе планировалось решить собственно задачу классификации, но уже в пространстве размерности 2. В качестве классификационных признаков были взяты:

– общее число страниц сайта (обозначим через S),
– число страниц сайта, относящихся к заданной рубрике (обозначим через s_i).

Планировалось провести классификацию страниц обучающей выборки dmoz.org, в полученном двумерном пространстве классификационных признаков провести разведочный анализ данных и затем выбрать метод классификации. Предполагалось, что сайты можно будет разбить на группы по числу страниц S и далее вычислить пороги, используя отношение правдоподобия для распределения величины s_i / S, доли страниц сайта, относящихся к рубрике. К сожаленью, времени на расчеты и проработку метода не хватило.

3.2 Описание эксперимента

Было выполнено 2 прогона. Первый этап классификации страниц был одинаковым для обоих прогонов. В качестве классификационных признаков для каждой рубрики было отобрано около 600 терминов. При этом 45% терминов оказались словосочетаниями.

При выполнении первого прогона решающее правило было выбрано в виде двух условий:

$$s_i \geq 5, \tag{11}$$

$$s_i / S \geq 0.2, \tag{12}$$

то есть для отнесения сайта к рубрике на нем должно найтись не менее 5 страниц, отнесенных к данной рубрике, и доля этих страниц должна составлять не менее 20% от контента сайта.

Таким образом, вне рассмотрения остались, в основном, сайты размером менее 5 страниц, что составляет около 43% сайтов тестовой выборки.

Кроме того, при выполнении первого прогона сайт мог быть отнесен только к одной рубрике, для которой доля страниц сайта максимальна. Данный прогон был мотивирован свойством обучающей выборки – каждому сайту соответствует не более одной рубрики.

Второй прогон заключался в отнесении сайта к первым 5-ти рубрикам с наибольшей долей контента. При этом в качестве решающего правила было использовано только условие (12).

3.3 Результаты

В таблице 6 приведены результаты обоих прогонов для способа оценки or (хотя бы один ассессор согласился с решением системы).

Большие различия между оценками точности в микро и макроусреднении говорят о наличии небольшого числа «провальных» рубрик. Анализ причин планируется провести позже.

Также предметом дополнительного исследования является невысокая полнота результатов.

Таблица 6. Результат классификации web-сайтов (оценка or).

показатель \ прогон	1	2
F1 (macro average)	0.1662	0.3083
Recall	0.0762	0.1692
Precision (macro average)	0.8876	0.7207
F1	0.1228	0.2265
Recall (macro average)	0.0917	0.1962
Precision	0.4382	0.4678
Accuracy	0.9913	0.9916
Error	0.0087	0.0084

3.4 Выводы и пожелания

Судя по неплохим результатам, показанным отдельными системами, качество коллекции dmoz.org выглядит достаточным для создания корпуса. Требуется только исправить ряд технических недостатков, обнаруженных при распространении коллекции среди участников (в частности, некорректно закодированные entity в url web-страниц, а также сайты, не принадлежащие ни одной рубрике).

Интересной новой дорожкой на следующий годовой цикл выглядит задача классификации web-страниц с использованием в качестве обучающей выборки каталога web-сайтов. На наш взгляд с прикладной точки зрения эта задача не менее актуальна, чем задача классификации сайтов.

4. Поиск биографических фактов

Изначально данная задача называлась “фактографическим поиском” и появление соответствующей дорожки на РОМИП было стимулировано нами в связи с недавним появлением технологии поиска описаний фактов заданного типа, кратко описанной нами в [4], а также успехов, достигнутых в распознавании в тексте обозначений персон и организаций. Однако, после длительных споров с другими участникам по поводу постановки данной задачи, она, на наш взгляд, корректно сформулирована не была, что и подтвердили результаты экспериментов, описанные далее.

В итоге участникам был предложен список персон (более 5000) с краткими описаниями, для которых в коллекции narod.ru требовалось найти фрагменты текста, описывающие нечто, относящееся к биографии персоны.

В качестве ответа системы для каждой персоны принимался набор четверок <идентификатор документа, смещение фрагмента в документе, длина фрагмента>.

4.1 Описание системы

Поиск описаний фактов проводился в 2 этапа. На первом этапе производился предварительный отбор документов для анализа путем традиционного контекстного поиска для каждой персоны, в результате чего были отобраны документы, содержащие имена и фамилии персоны. Поиск производился с учетом всех словоформ имен и фамилий, которые были автоматически построены при помощи нашего морфологического анализатора даже для неизвестных персон.

На втором этапе был использован разработанный нами алгоритм выделения и отождествления всех обозначений персон в тексте, включая полные, краткие и косвенные наименования (по части полного ФИО, должности). Данная процедура является вычислительно емкой, так как требует проведения серьезного анализа текста документа, в том числе синтаксического анализа и снятия омонимии. Именно поэтому для предварительного отбора

документов, которые могут содержать упоминания о целевых персонах, использовался первый этап.

Поскольку точная постановка задачи поиска биографических фактов отсутствовала, в качестве ответа система выдавала просто перечень всех предложений, которые содержали упоминания о целевых персонах. В итоге результат решения задачи следовало бы оценивать на точность и полноту поиска упоминаний о персонах. Тем не менее, и в такой вырожденной постановке оценка качества решения задачи и появление соответствующей дорожки на РОМИП имели смысл ввиду востребованности задачи (прежде всего, аналитическими агентствами) и фактического отсутствия систем, производящих поиск текста по персонам и организациям. В частности, других участников в этой дорожке, кроме нас, в этом году на РОМИП'е не оказалось.

4.2 Результаты

Предварительный анализ полученных результатов показал, что в результатах встречается большое количество тривиальных упоминаний (заголовки, текст ссылок) персон, поэтому из ответа были удалены все фрагменты, состоящие менее чем из четырех слов.

При проведении оценки ассессоры проверяли наличие описания биографического факта в пассаже и соответствие его персоне.

Как видно из таблицы 5, в которой приведены результаты работы системы, она содержит только один традиционный показатель – точность. Причина этого - невозможность применения метода общего котла, когда ответы всех систем объединяются, так как дорожку выполнила только одна система.

Таблица 5. Результаты фактографического поиска.

показатель \ способ	or	and
Precision	0.6909	0.2391
Precision(macro average)	0.6941	0.2475

В ходе оценок выяснилось еще одно досадное обстоятельство. Смещения фрагментов текста, выданные нашей системой отличались от смещений, воспроизводимых инструментом оценки. Причина различий - использование авторами собственной утилиты для распаковки архивов коллекции, которая некорректно сохраняла на диск символы перевода каретки. Разработчики инструмента оценки путем его доработок постарались минимизировать погрешность оценки точности, за что им особая благодарность.

На коррекцию смещений и длин пассажей из ответа системы, чтобы оценить, насколько сильно они отличаются от предложенных ассессорами, у авторов, к сожалению, не хватило времени.

Трехкратное различие между значениями показателей при различных способах подсчета свидетельствует о том, что ассессоры по-разному понимали постановку задачи (“биографичность факта”) и критерии оценки. Не исключен и более банальный фактор – влияние упомянутой выше ошибки в результатах системы.

4.3 Выводы и пожелания

Первый эксперимент прошел не очень удачно. Главная тому причина – неопределенная постановка задачи плюс отказ остальных участников от решения этой задачи на последнем этапе, когда постановку уже нельзя было подогнать хотя бы под единственного участника. Следствие этого – невозможность оценить характеристики системы, а также получить размеченный корпус, полезный хотя бы кому-либо из участников для дальнейшего развития своей системы.

Тем не менее, мы довольны вообще появлением этой дорожки на РОМИП и в следующем году продолжим работу над постановкой задачи поиска описаний фактов в надежде на появление участников, заинтересованных в этой задаче. Пока же нами лишь брошен вызов...

По мнению авторов, формулировка задачи в перспективе должна быть ближе к вопросно-ответной системе, т.е. система должна возвращать точные ответы на точно заданные вопросы, например, дату рождения, место учебы, занимаемые должности, а текст, в котором эти сведения можно найти, должен служить только для проверки корректности ответа системы. Предлагаемая нами схема решения данной задачи описана в работе [4].

Заключение

Основным достижением второго годичного цикла РОМИП является, по мнению авторов, создание хороших коллекций для оценки качества решения задач поиска и классификации. Пока сообщество РОМИП не пришло к единому мнению по поводу способа оценки результатов, но чувствуется, что «истина где-то рядом».

Среди возможных задач на следующий годичный цикл РОМИП хотелось бы предложить следующие:

– Добиться единого способа вычисления оценок по всем дорожкам;

- Опубликовать корпуса, вызывающие доверие у большинства членов сообщества;
- Организовать дорожку, исходными данными для которой были бы материалы СМИ;
- Сформулировать постановку задачи фактографического поиска и совместно с другими участниками получить реальные и полезные результаты по соответствующей дорожке.

Также хотелось бы благодарить Игоря Некрестьянова и Игоря Кураленка, уже второй год несущих на себе тяжелую ношу по подготовке семинара, Максима Губина, без усилий которого не состоялись бы дорожки по коллекции правовых документов, Владислава Шабанова, без которого не было бы столь качественной обучающей выборки для классификации web-сайтов, а также компанию Парк.Ру, любезно предоставившую журнал запросов справочно-правовой системы «Гарант-WWW».

Литература

[1] *Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Штернов С.В.* "Отправная точка" для дорожки по поиску в РОМИП (предварительный анализ) // Труды первого российского семинара по оценке методов информационного поиска. *Под ред. И.С. Некрестьянова* - Санкт-Петербург: НИИ Химии СпбГУ – 2003 - с. 87-109.
http://romip.narod.ru/romip2003/8_romip_uisrussia.pdf

[2] *Плешко В.В., Ермаков А.Е., Митюнин В.А.* RCO на РОМИП 2003: отчет об участии в семинаре по оценке методов информационного поиска // Труды первого российского семинара по оценке методов информационного поиска. *Под ред. И.С. Некрестьянова* - Санкт-Петербург: НИИ Химии СПбГУ – 2003 - с. 42-51.
http://romip.narod.ru/romip2003/3_RCO_ROMIP2003.pdf

[3] *Ермаков А.Е.* Значимость элементов текста в свете теории синтаксической парадигмы // Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. - Москва: МГУ - 2004. http://www.rco.ru/article.asp?ob_no=484

[4] *Киселев С.Л., Ермаков А.Е., Плешко В.В.* Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва: Наука – 2004. http://www.rco.ru/article.asp?ob_no=629

RCO at RIRES 2004

Pleshko V.V., Ermakov A.E., Golenkov V.P.

This article presents report on experiments in IR that were driven as a part of RIRES initiative. In web adhoc and legal adhoc tasks, quantitative evaluation on how much stemming improves recall and precision was performed. Preliminary results on web site classification and legal document classification were obtained. Also some consideration was taken to qa task.