RCO на РОМИП 2005

Плешко В.В., Ермаков А.Е.,
Голенков В.П., Поляков П.Ю.
rco@metric.ru

Аннотация

Настоящая работа является отчетом об экспериментах, проведенных в рамках цикла семинара РОМИП 2005 года. Проведены исследования различных факторов, влияющих на качество алгоритмов тематической классификации. Также получены предварительные результаты по контекстно-зависимому аннотированию, выявлению наименований персон и организаций, поиску описаний фактов.

1. Введение

РОМИП растет! Третий годовой цикл РОМИП порадовал и выросшим числом дорожек и количеством участников. Помимо дорожек документального поиска и тематической классификации, уже ставших традиционными, каждый год появляются экспериментальные дорожки. В этом году их было три: контекстно-зависимое аннотирование, фактографический поиск, новостная агрегация.

В настоящей статье приводится отчет об экспериментах авторов по выполнению заданий по тематической классификации нормативно-правовых документов, Веб-страниц, Веб-сайтов, а также заданий по контекстно-зависимому аннотированию и фактографическому поиску.

2. Классификация нормативно-правовых документов

2.1 Постановка задачи

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 не более 5 рубрик. Также допускалось не присваивать документу ни одной рубрики.

2.2 Общий подход

Исследования проводились в рамках векторной модели представления документов. Во всех прогонах строились классификаторы вида

$$d*c > h \quad (1)$$

где d – вектор документа, c – вектор профиля категории, h – пороговое значение, которое должно превысить скалярное произведение упомянутых векторов для отнесения документа к категории.

Исследовались три параметра построения классификаторов:

1. Способ выбора терминов для профиля категории (классификационных признаков);
2. Способ построения весов терминов и порога (алгоритм обучения);
3. Способ вычисления весов терминов в документе.

2.3 Отбор терминов

Отбор терминов проводился из набора положительных примеров для каждой из категорий при помощи алгоритмов синтактико-семантического анализа [1] с различными настройками (см. Табл. 1). Во всех случаях число терминов, выделяемых из одного документа, не превышало 500.

Таблица 1. Настройки алгоритма выделения терминов.

	Однословные термины	Многословные термины
число слов термина	1	5

минимальный вес термина	5	10
категории терминов, не включавшиеся в обработку	названия организаций, географические наименования, даты	названия организаций, географические наименования, имена прилагательные, даты

Для исследования было отобрано три способа выделения терминов:

1. Однословные термины, условное обозначение – «L» (lemmas);
2. Многословные термины – «T» (themes);
3. Теоретико-множественное объединение двух предыдущих способов – «TL» (themes+lemmas).

В силу особенностей реализации алгоритмов, если выделенный термин являлся неделимой сущностью (например, «пакет акций»), то он попадал в перечень однословных терминов. К сожаленью, эта особенность была замечена достаточно поздно, чтобы повторить расчеты. Таким образом, отобранные леммы содержали некоторое количество словосочетаний (в среднем 1-2%).

2.4 Алгоритм обучения

Опробовались три алгоритма обучения:

1. Алгоритм, основанный на вычислении весов терминов по априорной информации (документных частотах терминов и принадлежности документов рубрикам) – «SIMPLE»;
2. Метод опорных векторов [2] с линейным ядром – «SVM»;
3. Метод опорных векторов с линейным ядром и измененным алгоритмом выбора порогового значения - «OURSVM».

Метод «SIMPLE» является модификацией алгоритма, использованного авторами в экспериментах цикла РОМИП 2004 года [3]. Отличие текущей реализации заключается в использовании другого целевого показателя при вычислении порога (в 2004 году – Accuracy, сумма ошибок I и II рода, в текущей реализации – F1).

К особенности реализации выбора порога следует отнести следующее. В силу конечности обучающей выборки, выбор порога всегда имеет произвол с точки зрения показателя F1. Если показатель F1 максимален при значении порога h, то он максимален и на некотором интервале [h, h1). В экспериментах порог «прижимался» к левому либо правому краю интервала в пропорции c/(b + c), где b – число ошибок I рода (пропуск релевантного

документа); c – число ошибок II рода (отнесение к категории нерелевантного документа).

Для апробации метода «SVM» была взята его реализация SVMLight [4]. Как уже было отмечено в [5], метод чувствителен к отношению весов ошибок I и II рода (параметр j). В проводимых экспериментах было взято значение j = 10.

Отличие метода «OURSVM» от «SVM» заключается в алгоритме вычисления порога. Веса терминов вычислялись по методу «SVM», а порог, как в методе «SIMPLE» - путем максимизации показателя F1.

В результате расчета профиля рубрик каждым из опробованных методов получался нормированный вектор весов терминов и порог – число из интервала (0,1).

2.5 Взвешивание терминов

Было опробовано три способа взвешивания терминов при построении векторов документов:

1. бинарное (1 – если термин встретился в документе, 0 – в противном случае) – «B» (binary);
2. пропорционально частоте термина – «F» (frequency);
3. пропорционально числу слов, составляющих термин – «L» (length).

При расчете весов терминов в документе создавался вектор, ненулевыми элементами которого служили перечисленные характеристики для всех найденных в нем терминов из множества отобранных на этапе синтактико-семантического анализа. Затем вектор документа нормировался.

2.6 Описание прогонов

Всего было сделано 24 прогона. Обозначим для удобства прогоны тройками вида

Term-Algorithm-Weight,

где Term, Algorithm и Weight – условные обозначения перечисленных выше параметров классификатора.

Полный план эксперимента составлял 27 прогонов (3*3*3), однако прогоны TL-OUR_SVM-* не были выполнены, так как после выполнения прогонов {L,T}-OURSVM-* метод «OURSVM» был признан неудачным.

Кроме того, из-за указанных выше особенностей работы алгоритма синтактико-семантического анализа, прогоны– L-*-L и L-*-B давали различные, в пределах 1-2%, результаты (при

корректной реализации взвешивание однословных терминов по длине должно было совпасть с бинарным взвешиванием).

2.7 Результаты на матрицах релевантности 2004 года

Все прогоны предварительно проверялись на полной матрице релевантности ideal40 (40 категорий, содержащие не менее 10 положительных примеров в обучающей выборке), использованной на прошлогоднем семинаре РОМИП-2004.

В таблице 2 приведены усредненные и наилучшие достигнутые по каждому параметру классификатора показатели F1 и F1 (macro).

Таблица 2. Усредненные показатели на матрице 2004 года.

	F1		F1 (macro)		Лучш. метод F1 / F1 (macro)
	avg	max	avg	max	
Отбор терминов					
L	0.3866	0.4777	0.4528	0.5086	L-SVM-F L-SVM-L
T	0.3966	0.4674	0.4737	0.5386	T-SVM-B T-SVM-B
TL	0.4305	0.4818	0.4990	0.5513	TL-SVM-B TL-SVM-B
Метод обучения					
OURSVM	0.3751	0.4465	0.4541	0.5074	L-OURSVM-F L-OURSVM-F
SIMPLE	0.3529	0.4150	0.4325	0.4956	TL-SIMPLE-F TL-SIMPLE-F
SVM	0.4673	0.4818	0.5239	0.5513	TL-SVM-B TL-SVM-B
Взвешивание терминов					
B	0.3918	0.4818	0.4658	0.5513	TL-SVM-B TL-SVM-B
F	0.4250	0.4777	0.4884	0.5137	L-SVM-F T-SVM-F
L	0.3871	0.4680	0.4623	0.5469	TL-SVM-L TL-SVM-L

Предварительные выводы, которые можно сделать на основании таблицы 2, достаточно очевидны:

1. В среднем лучшие результаты показывают

- способ отбора терминов с использованием как слов, так и словосочетаний,
- методы обучения, основанные на выборе весов терминов путем оптимизации целевой функции,
- учет частоты при расчете веса термина внутри документа.

2. Комбинация наиболее удачных в среднем параметров не дает лучший результат (лучший прогон – TL-SVM-B, а не TL-SVM-F).
3. Неожиданно критичным оказался способ выбора порога – метод «OURSVM» показал результаты, близкие к «SIMPLE».

2.8 Результаты оценки в 2005 году

Для оценок в 2005 году были по тому же принципу, что и в 2004 году, отобраны 40 категорий из ранее не оцененных. Причем средние значения и среднеквадратичные уклонения числа положительных примеров в обоих наборах очень близки (среднее ~ 80, среднеквадратичное уклонение ~ 11).

Так как оценка проводилась автоматически (не требовала работы асессоров), для оценки были представлены все прогоны.

Соотношения полученных результатов по отношению к усредненным показателям не изменились. Однако задание 2005 года оказалось немного труднее – показатели F1 и F1 (macro) получились ниже на 1-4%.

Перечень результатов прогонов в сравнении с прогонами других участников приведен в таблице 3.

Таблица 3. Результаты оценки legal-classification

Прогон	F1 (macro)	F1	Прогон	F1 (macro)	F1
1	0.1511	0.1953	L-SIMPLE-L	0.3927	0.2245
2	0.1126	0.1618	L-SIMPLE-F	0.4539	0.2857
3	0.2757	0.2039	L-SIMPLE-B	0.4057	0.2286
4	0.2884	0.2181	L-SVM-L	0.5024	0.3830
5	0.3963	0.1906	L-SVM-F	0.5106	0.4092
6	0.5841	0.3956	L-SVM-B	0.5023	0.3844
7	0.5916	0.4318	T-OURSVM-L	0.4025	0.2525
TL-SIMPLE-L	0.4425	0.2890	T-OURSVM-F	0.4687	0.3245
TL-SIMPLE-F	0.4973	0.3393	T-OURSVM-B	0.4221	0.2735
TL-SIMPLE-B	0.4496	0.2982	T-SIMPLE-L	0.4436	0.2891

TL-SVM-L	0.5373	0.3836	T-SIMPLE-F	0.4781	0.3129
TL-SVM-F	0.5132	0.3990	T-SIMPLE-B	0.4520	0.2971
TL-SVM-B	0.5385	0.3975	T-SVM-L	0.5254	0.3685
L-OURSVM-L	0.4581	0.3006	T-SVM-F	0.4977	0.3781
L-OURSVM-F	0.5103	0.3862	T-SVM-B	0.5238	0.3768
L-OURSVM-B	0.4592	0.2961	32	0.1853	0.0758

Лучший результат по показателю F1 (macro) среди представленных авторами методов показал TL-SVM-L (0.5373), среди систем – 0.5916. Лучший результат по показателю F1 из представленных авторами – L-SVM-F (0.4092), среди систем – 0.4318, командный результат (усреднение лучших результатов среди систем на каждой категории) – 0.4691.

Полученные результаты говорят о том, что возможности для исследований на коллекции legal далеко не исчерпаны. Разница между лучшим прогоном систем-участников и командным результатом составляет почти 10%.

Кроме того, существуют сложные для всех систем категории (см. рис. 1). В частности, для всех систем представляли трудности (F1 < 0.2) следующие категории:

- «Учет денежных средств» (Id=9002010),
- «Экономический механизм охраны окружающей среды и природопользования» (Id=901716394),
- «Международное сотрудничество в социальной и культурной сферах» (Id=901716583).

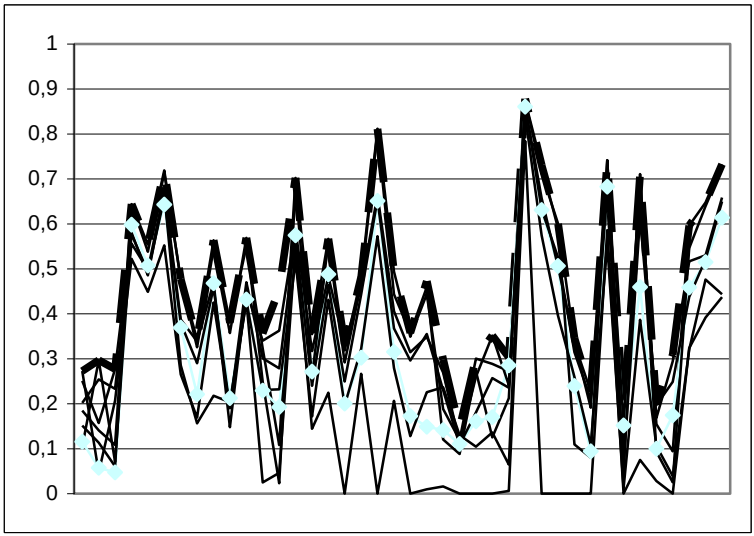

Рис.1. Командный результат по показателю F1 на категориях 2005 года коллекции legal

3. Тематическая классификация Веб-страниц

3.1 Постановка задачи

Участникам было предложено подмножество интернет-каталога dmoz.org (300000 страниц), используя которое в качестве обучающей выборки, требовалось соотнести с категориями каталога dmoz.org (247 категорий) страницы коллекции narod.ru (728000 страниц).

Среди особенностей задачи следует отметить «зашумленность» обучающей выборки. Если сайт относился к категории, то все его страницы относились к той же категории.

3.2 Общий подход

План эксперимента по тематической классификации Веб-страниц изначально совпадал планом по классификации нормативно-правовых документов. Однако в процессе выполнения в план были внесены следующие коррективы.

Наборы терминов, полученные методами «L» и «T», для ускорения вычислений подверглись фильтрации аналогично [3]:

- доля страниц, содержащих термин, среди положительных примеров должна превысить заданное значение (полнота);
- доля положительных примеров среди страниц, содержащих термин, должна превысить заданное значение (точность).

Также из-за недостатка времени не делались прогоны с выбором терминов «TL».

Кроме того, алгоритм обучения «OURSVM» не использовался, так как дал плохой результат на коллекции legal.

Таким образом, было проведено 12 прогонов:
{L, T}-{SIMPLE, SVM}-{B, F, L}.

3.3 Результаты на матрицах релевантности 2004 года

Для предварительной оценки результатов на основе неполной матрицы релевантности для сайтов, полученной на семинаре в цикле 2004 года, была создана неполная приближенная матрица релевантности для страниц (точнее, 2 ее варианта, для оценок and и or). Страница считалась релевантной для категории, если сайт был отнесен к категории.

Оценки на построенной матрице проводились по методу judjedonly – страницы, отсутствующие в матрице исключались из рассмотрения. В таблице 4 приведен результат усредненной оценки для параметров классификаторов по методу and.

Таблица 4. Усредненная оценка на матрице 2004 года (Веб-страницы)

	F1		F1(macro)		Лучш. метод F1 / F1 (macro)
	avg	max	avg	max	
Отбор терминов					
L	0.1650	0.1878	0.2014	0.2361	L-SIMPLE-F L-SVM-F
T	0.1579	0.1820	0.1930	0.2170	T-SVM-F T-SVM-F
Метод обучения					
SIMPLE	0.1687	0.1877	0.1896	0.2358	L-SIMPLE-F L-SIMPLE-F
SVM	0.1543	0.1860	0.2047	0.2361	L-SVM-F L-SVM-F
Взвешивание терминов					
B	0.1462	0.1677	0.1825	0.2148	T-SIMPLE-B T-SVM-F

F	0.1810	0.1877	0.2204	0.2361	L-SIMPLE-F L-SIMPLE-F
L	0.1572	0.1766	0.1886	0.2316	L-SVM-L L-SVM-L

Среди неожиданных (и ошибочных!) результатов следует отметить, что метод «SIMPLE» ведет себя, не хуже «SVM».

3.4 Результаты оценки в 2005 году

Для оценок асессорами были отобраны прогоны L-SIMPLE-F, L-SVM-F, T-SVM-F. Оценивалось только подмножество пула, собранного из ответов всех систем. Таким образом были получены четыре оценки (как для полной матрицы, judjedonly * and, or). Результат оценки and, judjedonly приведен в таблице 5.

Основной вывод, который можно сделать из таблицы 5 – метод «SVM» все-таки работает лучше и на Веб-страницах. Следует также отметить, что результаты, полученные методом оценки для полной матрицы, выглядят гораздо скромнее (командный результат по показателю F1 равен 0.1619).

Таблица 5. Результаты оценки webpage-classification по методу and, judjedonly

Прогон	F1 (macro)	F1
1	0.2489	0.2334
2	0.2657	0.2346
3	0.0534	0.0484
4	0.2934	0.2855
L-SIMPLE-F	0.2367	0.1904
L-SVM-F	0.2855	0.2510
T-SVM-F	0.3203	0.2581
8	0.1093	0.1051

5. Тематическая классификация Веб-сайтов

5.1 Постановка задачи

Участникам было предложено подмножество интернет-каталога dmoz.org (2100 Веб-сайтов), используя которое в качестве обучающей выборки, требовалось соотнести с категориями каталога

dmoz.org (247 категорий) Веб-сайты коллекции narod.ru (более 22000 Веб-сайтов).

5.2 Общий подход

Для дорожки по классификации сайтов было реализовано два метода:

1. Метод суперстраниц – «S» (superpage);
2. Метод гистограмм – «H» (histogram).

В методе суперстраниц сайт представлялся одним вектором (аналогично вектору документа). В состав вектора сайта включались все термины, содержащиеся на страницах сайта. Отбор терминов для профиля рубрики осуществлялся таким же методом, что и при классификации Веб-страниц. Приэтом фильтрации списка терминов не проводилось, метод отбора терминов «TL» не использовался. Обучение классификатора осуществлялось методами «SIMPLE» и «SVM».

Метод гистограмм использовал результаты классификации страниц. В качестве классификационного признака для отнесения сайта к категории использовалась доля релевантных категории страниц сайта. При обучении использовались результаты классификации страниц обучающей выборки. Выбор порога осуществлялся путем максимизации показателя F1 (аналогично методам «SIMPLE» и «OURSVM»).

Была также сделана попытка использовать метод опорных векторов с гауссовым ядром на двумерном пространстве классификационных признаков:

- Доля релевантных страниц,
- Общее число страниц сайта.

Однако уже первые полученные результаты выявили неприменимость такого подхода в данном случае. Размер обучающей выборки оказался слишком мал по сравнению с «емкостью» использованного ядра.

В итоге было сделано 24 прогона:
{S, H}-{L, T}-{SIMPLE, SVM}-{B, F, L}

5.3 Результаты на матрицах релевантности 2004 года

Для предварительных оценок прогонов была использована неполная матрица релевантности (варианты and, or) 2004 года. В таблице 6 приведены оценки для метода and, judjedonly.

Таблица 6. Усредненная оценка на матрице 2004 года (Веб-сайты)

	F1		F1 (macro)		Лучш. метод F1 / F1 (macro)
	avg	max	avg	max	
Метод представления сайта					
H	0.2924	0.3348	0.3834	0.4469	H-L-SVM-B H-L-SVM-B
S	0.2157	0.3291	0.2823	0.4340	S-T-SIMPLE-B S-L-SVM-F
Отбор терминов					
L	0.2649	0.3348	0.3435	0.4469	H-L-SVM-B H-L-SVM-B
T	0.2437	0.3291	0.3221	0.4197	S-T-SIMPLE-B S-T-SIMPLE-B
Метода обучения					
SIMPLE	0.2657	0.3291	0.3467	0.4197	S-T-SIMPLE-B S-T-SIMPLE-B
SVM	0.2424	0.3348	0.3189	0.4469	H-L-SVM-B H-L-SVM-B
Взвешивание терминов					
B	0.2367	0.3348	0.3103	0.4469	H-L-SVM-B H-L-SVM-B
F	0.2971	0.3291	0.3939	0.4340	S-T-SIMPLE-B S-L-SVM-B
L	0.2284	0.3134	0.2942	0.4032	H-L-SIMPLE-L H-L-SIMPLE-L

На основе таблицы 6 можно сделать два (неверных!) вывода:

1. Метод гистограмм в среднем гораздо лучше метода суперстраниц;
2. Метод «SIMPLE» ведет себя на хуже «SVM».

5.4 Результаты оценки в 2005 году

Для оценки были представлены методы H-L-SIMPLE-B, H-L-SVM-B, S-L-SVM-F, S-T-SIMPLE-F. Результаты оценок представлены в таблицах 7 и 8.

Таблица 7. Результаты оценки website-classification по методу and

Прогон	F1 (macro)	F1
1	0.2203	0.2027
2	0.1181	0.1192

3	0.1761	0.2133
H-L-SIMPLE-B	0.1291	0.1197
H-L-SVM-B	0.1114	0.1275
S-L-SVM-F	0.2465	0.2382
S-T-SIMPLE-F	0.1506	0.1582
8	0.1820	0.1917

Таблица 8. Результаты оценки website-classification по методу or

Прогон	F1 (macro)	F1
1	0.2383	0.2116
2	0.2782	0.2584
3	0.2323	0.2451
H-L-SIMPLE-B	0.1589	0.1582
H-L-SVM-B	0.1932	0.1880
S-L-SVM-F	0.2161	0.2232
S-T-SIMPLE-F	0.1784	0.1784
8	0.3374	0.3023

Оценки на основе полных матриц (оценивались все ответы систем) внесли коррективы в предварительные оценки на основе неполных матриц релевантности:

1. Метод «SVM» все-таки показывает лучшие результаты, чем «SIMPLE»;
2. Неожиданно плохо показал себя метод гистограмм.

6. Контекстно-зависимое аннотирование

6.1 Постановка задачи

Участникам были предоставлены колекции narod.ru и legal и набор заданий вида "запрос и документ". Система должна была предоставить аннотацию этого документа по этому запросу.

Число символов аннотации без учета разметки не должно было превышать заданное число L (300 символов).

6.2 Описание метода

Исследуемый алгоритм построения аннотаций состоял из двух этапов:

1. Поиск множества вхождений слов запроса в тексте документа (окна), удовлетворяющего заданным критериям

2. Формирование фрагмента(ов) текста, содержащего найденные вхождения слов запроса для выдачи пользователю.

На первом этапе для каждого из слов запроса находился список его упоминаний в документе с точностью до словоформы. Затем, на основе полученных списков находилось окно длины не более L символов, такое, что

- Число упоминаний различных слов запроса в нем максимально
- Суммарное число символов между упоминавшимися в окне словами запроса минимально.

Далее делалась попытка найти два окна, размер каждого из которых не превышает L/2 и суммарное число упоминаний различных слов запроса в обоих окнах превышает соответствующее число в найденном окне длины L.

В результате получалось одно окно длины не более L символов, либо два окна длины не более L/2 символов каждое.

На втором этапе производилось выравнивание полученного окна (окон) на границы предложения с учетом ограничений на длину окна. Если полученное предложение состояло из одного слова, границы предложения расширялись дальше.

К сожаленью, в последний момент перед отправкой заданий был обнаружен ряд дефектов в реализации второго этапа, приводивших к превышению числа символов аннотации. Перед отправкой все «бракованные» аннотации были усечены справа.

6.3 Результаты оценки

Первоначально планировалось провести предварительные оценки на матрицах релевантности 2004 года. Однако времени на осуществление оценок не хватило.

В результате оценки аннотаций асессорами несколькими методами были построены матрицы релевантности для аннотаций. Затем полученные матрицы были сопоставлены с матрицами релевантности документов на тех же запросах. В итоге было рассчитано два основных показателя:

- AnnotationAccuracy - Доля релевантных аннотаций (по мнению экспертов), которым соответствуют реально релевантные документы;
- AnnotationError - Доля нерелевантных аннотаций, которым соответствуют релевантные документы.

В таблице 9 для иллюстрации приведены перечисленные показатели полученные авторами и другими системами участниками.

Таблица 9. Результаты оценки summarization методом vital,and,and

Прогон	AnnotationAccuracy	AnnotationError
1	0.9022	0.4806
2	0.9377	0.4696
4	0.9491	0.4674
4	0.9491	0.4674
my_1	0.8782	0.4800
6	0.8888	0.4765
7	0.9146	0.4812
8	0.9384	0.4813

Как видно, разброс между лучшим и худшим результатами составляет менее 10%, что не типично для других дорожек РОМИП. Тем не менее, задачу нельзя считать легкой – показатель AnnotationError одинаково высок у всех систем.

7. Фактографический поиск по новостной коллекции

7.1 Постановка задачи

Предметом обработки являлась новостная коллекция, предоставленная компанией Яндекс. Коллекция содержит 24 000 сообщений из 16 источников общим объемом 50 Мб.

Участникам предлагалось решить 2 задачи

1. Для каждого сообщения система должна построить список именованных сущностей (классифицировать найденные сущности, выдать список ссылок на упоминания сущностей в тексте).
2. Для найденных сущностей выделить факты следующих типов:
   - Кто работал\работает в данной организации?
   - Где работал\работает данный человек?
   - Кто владеет или владел данной организацией?
   - Какими предприятиями владеет или владела данная организация/персона?

При этом выделение фактов должно было осуществляться для всех упоминаний персон и организаций, содержащих в себе имя собственное, без учета кореферентных обозначений.

В качестве описания факта система должна была выдать тип факта, ссылку на фрагмент текста, два стандартизированных имени фигурантов и ссылки на них в тексте.

7.2 Описание метода

Для извлечения упоминаний об объектах использовалась схема, описанная в [6]. Для поиска заданных типов фактов были применены разработанные авторами универсальные алгоритмы поиска описаний фактов в тексте на основе поиска изоморфизмов в сетях синтактико-семантических отношений [7].

7.3 Результаты оценки

К сожаленью, на момент написания статьи оценки еще не были готовы.

8. Заключение

Цикл 2005 года можно считать первым циклом РОМИП, в котором появилась возможность исследовать методы информационного поиска не вслепую, а основываясь на опыте и результатах прошлых лет. Этот факт мы считаем признаком качественного роста семинара.

Также отрадно, что мы не смогли принять участие во всех дорожках, как в прошлые годы, хотя и жаль, что не хватило сил и времени поучаствовать в новостной дорожке.

В заключение хотелось бы поблагодарить Игоря Некрестьянова и его соратников из Санкт-Петербургского Государственного Университета за самоотверженную работу по организации семинара и проведения оценок.

Литература

[1] *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.

[2] *Burges C.J.C.* A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery – 1998, V.2, No.2 – pp.121-167.
[3] *Плешко В.В., Ермаков А.Е., Голенков В.П.* RCO на РОМИП 2004 // Труды второго российского семинара РОМИП'2004. (Пущино, 1 октября 2004г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2004 - с. 43-61.
[4] *Joachims T.* Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998.
[5] *М.С. Агеев, Б.В. Добров, Н.В.Лукашевич, А.В. Сидоров* Экспериментальные алгоритмы поиска/классификации и сравнение с "basic line" // Труды второго российского семинара РОМИП'2004. (Пущино, 1 октября 2004г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2004 - с. 62-89.
[6] *Ермаков А.Е.* Референиция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. – Москва, Наука, 2005.
[7] *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.

RCO at RIRES 2005

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

This article presents report on experiments in IR that were driven as a part of RIRES seminar. The main research was taken on different factors that affect performance of classification task. Also preliminary results on document summarization and entity and fact extraction tasks were obtained.