

RCO на РОМИП 2005

Плешко В.В., Ермаков А.Е., Голенков В.П., Поляков П.Ю.

ООО «Гарант-Парк-Интернет»

rco@metric.ru

Дорожки

- Классификация правовых документов
- Классификация Веб-страниц
- Классификация Веб-сайтов
- Аннотирование
- Фактографический поиск
- Агрегация новостей (только мысли)

Классификация

Общий подход

- Векторная модель:
 - $d * c > h$
- Предмет исследования:
 - отбор терминов
 - алгоритм обучения
 - значимость терминов в документе
- Предварительная оценка:
 - матрицы РОМИП 2004

Классификация

Отбор терминов

Средство:

- Технология синтактико-семантического анализа
=> информационный портрет документа

Способы:

- ЛЕММЫ (1 слово)
- ТЕМЫ (до 5 слов)
- ЛЕММЫ + ТЕМЫ

Информационный портрет документа

Разрабатывая с 1999 года уникальные лингвистические технологии, компания "Гарант-Парк-Интернет" выпускает серию продуктов для компьютерного анализа текста.

Категория	Элемент содержательного портрета	Значимость (1-100)
<i>Самостоятельный</i>		
Организация	"ГАРАНТ-ПАРК-ИНТЕРНЕТ"	100
Событие	ВЫПУСК ПРОДУКТА	64
Событие	РАЗРАБОТКА ЛИНГВИСТИЧЕСКОЙ ТЕХНОЛОГИИ	40
Событие	КОМПЬЮТЕРНЫЙ АНАЛИЗ ТЕКСТА	24
<i>Производный</i>		
Предмет	ПРОДУКТ	25
Предмет	ЛИНГВИСТИЧЕСКАЯ ТЕХНОЛОГИЯ	24
Событие	АНАЛИЗ ТЕКСТА	16
Предмет	КОМПЬЮТЕРНЫЙ АНАЛИЗ	16
Предмет	ТЕХНОЛОГИЯ	12
Предмет	ТЕКСТ	9

Классификация

Алгоритмы обучения

- SIMPLE
 - РОМИП 2004
 - $F1(h) \rightarrow \max$
- SVM
 - SVMLight, линейное ядро, $j = 10$
- OURSVM
 - SVMLight, линейное ядро, $j = 10$
 - $F1(h) \rightarrow \max$

Классификация Значимость терминов

- Бинарное (0/1, есть/нет)
- Частотное (число вхождений в документ)
- По длине (число слов в термине)

Классификация

План эксперимента

<Term, Algorithm, Weight>

{L,T,TL}-{SIMPLE,SVM, OURSVM}-{B,F,L}

27 прогонов

Классификация

Метод оценки

- Матрица РОМИП 2004
 - В случае неполных матриц - judjedonly
- Показатель качества
 - При одном фиксированном параметре классификатора
 - avg F1, max F1

Классификация правовых док-тов

Матрицы РОМИП 2004

	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

Классификация правовых док-тов

Матрицы РОМИП 2005

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

24 прогона

Классификация правовых док-тов

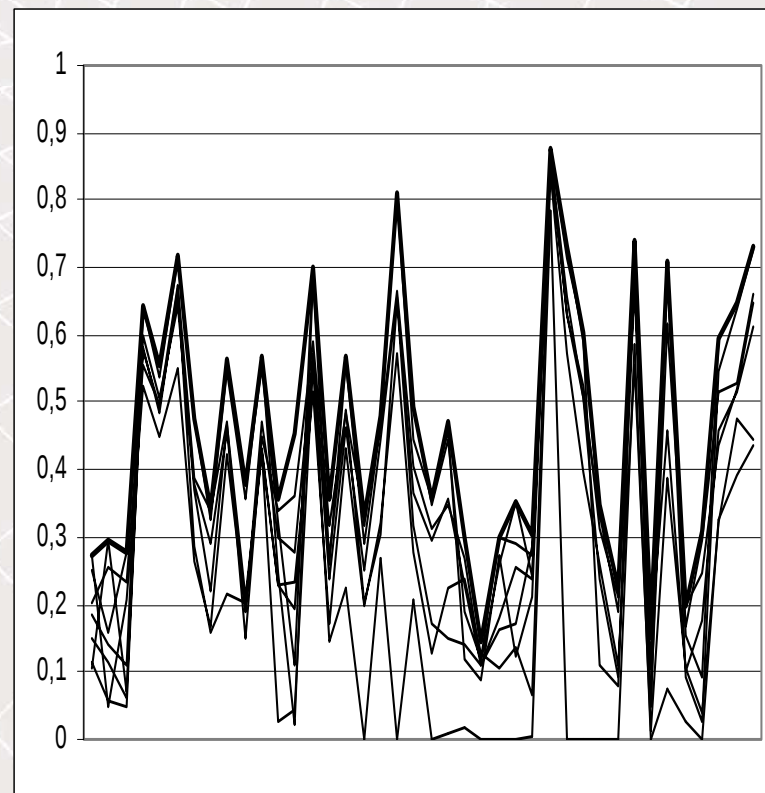
Анализ результатов

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

Классификация правовых док-тов

Командный результат

- Разница F1 между лучшим среди систем (0.432) и командным (0.469) результатами – 10%
- Сложные категории ($F1 < 0.2$)
 - «Учет денежных средств» (Id=9002010),
 - «Экономический механизм охраны окружающей среды и природопользования» (Id=901716394),
 - «Международное сотрудничество в социальной и культурной сферах» (Id=901716583).



Классификация Веб-страниц

Особенности реализации

- Приведение к $sr1251$
- Прогоны
 - без OURSVM, TL (всего 12)
- Фильтрация терминов
 - доля страниц, содержащих термин, среди положительных примеров должна превысить заданное значение (полнота);
 - доля положительных примеров среди страниц, содержащих термин, должна превысить заданное значение (точность).
- Матрицы релевантности
 - Изготовлены на основе матриц для сайтов 2004 года

Классификация Веб-страниц

Матрицы РОМИП 2004

	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

and, judjedonly

Классификация Веб-страниц

Матрицы РОМИП 2005

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

and, judjedonly

Классификация Веб-страниц

Выводы

- Оценка на неполной матрице методом judjedonly
 - SVM ~ SIMPLE, Леммы > Темы
- Задача сложная
 - Командный F1 = 0.1619 (and, как для полной матрицы)
- В среднем лучшие показатели
 - Темы, SVM, Частоты

Классификация Веб-сайтов

Особенности реализации

- Способ представления сайта (на базе классификации Веб-страниц)
 - «суперстраницы» (объединение терминов со всех страниц сайта) = 12 прогонов
 - «гистограммы» (доля страниц сайта, отнесенных к категории) = 12 прогонов
- Неудачный эксперимент
 - Гистограммы, (доля страниц, размер сайта), SVMLight, гауссово ядро

Классификация Веб-сайтов

Матрицы РОМИП 2004

	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
Метод 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

and, judged only

Классификация Веб-сайтов

Матрицы РОМИП 2005

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

and

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

or

Классификация Веб-сайтов

Выводы

- Оценка на неполной матрице методом judjedonly
 - SVM ~ SIMPLE, Гистограммы > Суперстраницы
- Гистограммы – неожиданно плохо
- Разные системы «оптимизированы» под разные методы оценки (and, or)

Аннотирование

Описание метода

1. Выбор окна (окон) – совокупности вхождений терминов запроса в документ, соотв. заданному критерию
2. Формирование результата – выдача фрагмента текста, содержащего найденные вхождения

Аннотирование

Выбор окна

1 фрагмент:

- Длина не более L
- Наибольшее число разных терминов
- Наибольшая плотность

2 фрагмента:

- Оба длиной не более $L/2$
- Суммарно содержат наибольшее число разных терминов запроса

Выбирается вариант, содержащий наибольшее число разных терминов запроса

Аннотирование

Формирование результата

- Фрагмент расширился до границы предложения с ограничением на длину (было реализовано с дефектом)
- Предложения из одного термина
- Сколько же символов в аннотации
 - Фрагмент1 *
* Фрагмент2
 - Html entities - *"e;*

Аннотирование

Результат

- AnnotationAccuracy - доля релевантных аннотаций, которым соответствуют реально релевантные документы;
- AnnotationError - доля нерелевантных аннотаций, которым соответствуют релевантные документы.

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

vital, and, and

Аннотирование Выводы

- Кучный результат (разброс $< 10\%$)
 - не типично для РОМИП
- Задача не из легких
 - AnnotationError ~ 0.50

Фактографический поиск

Постановка

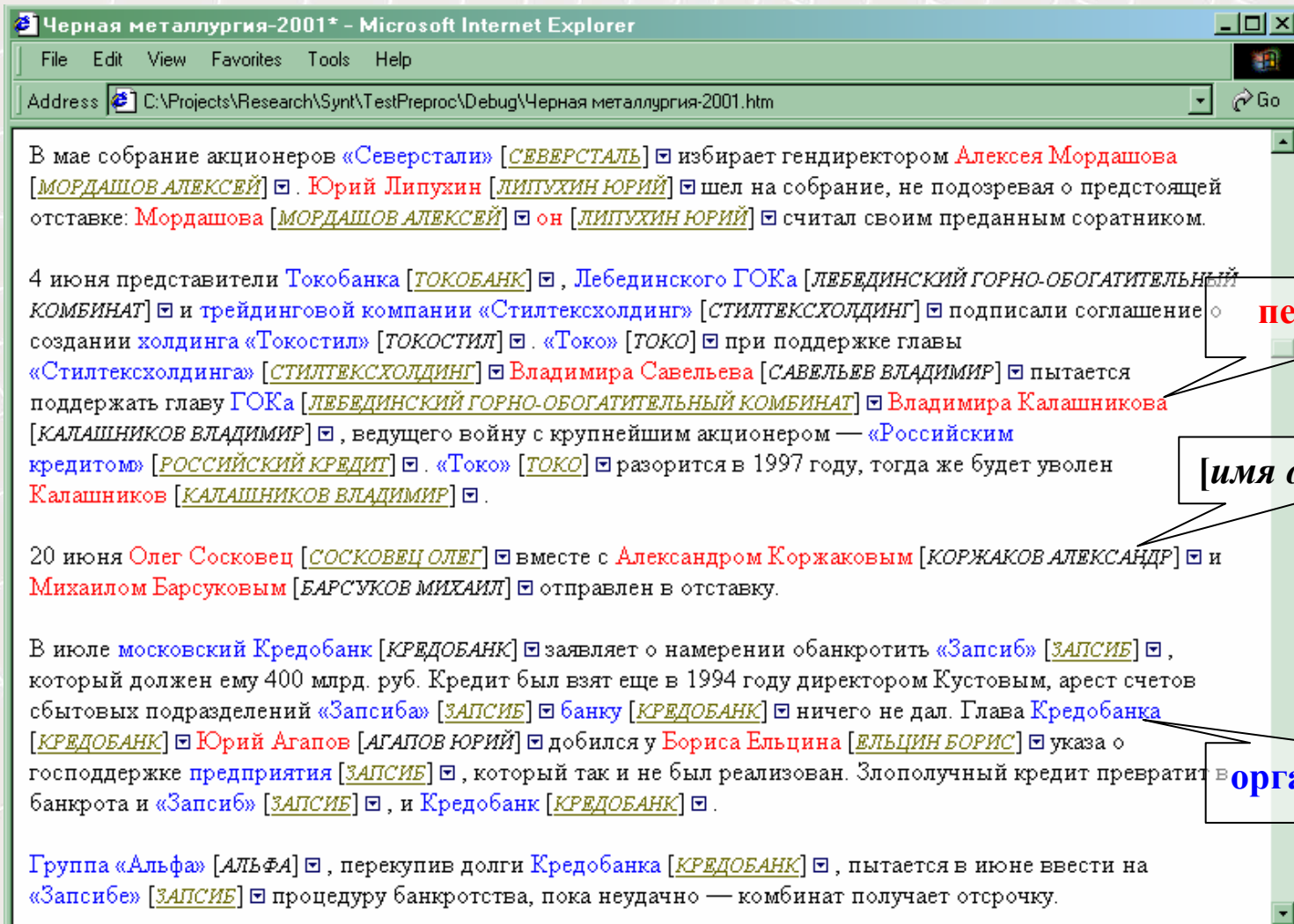
- Коллекция
 - 24000 сообщений, 16 источников, 50Мб
- Найти и классифицировать именованные сущности
 - Персона
 - Организация
- Найти описания фактов и фигурантов
 - Кто работал\работает в данной организации?
 - Где работал\работает данный человек?
 - Кто владеет или владел данной организацией?
 - Какими предприятиями владеет или владела данная организация\персона?

Фактографический поиск

Инструменты

- Алгоритм выявления наименований персон и организаций
- Синтактико-семантический анализ текста
- Поиск изоморфизмов в семантической сети

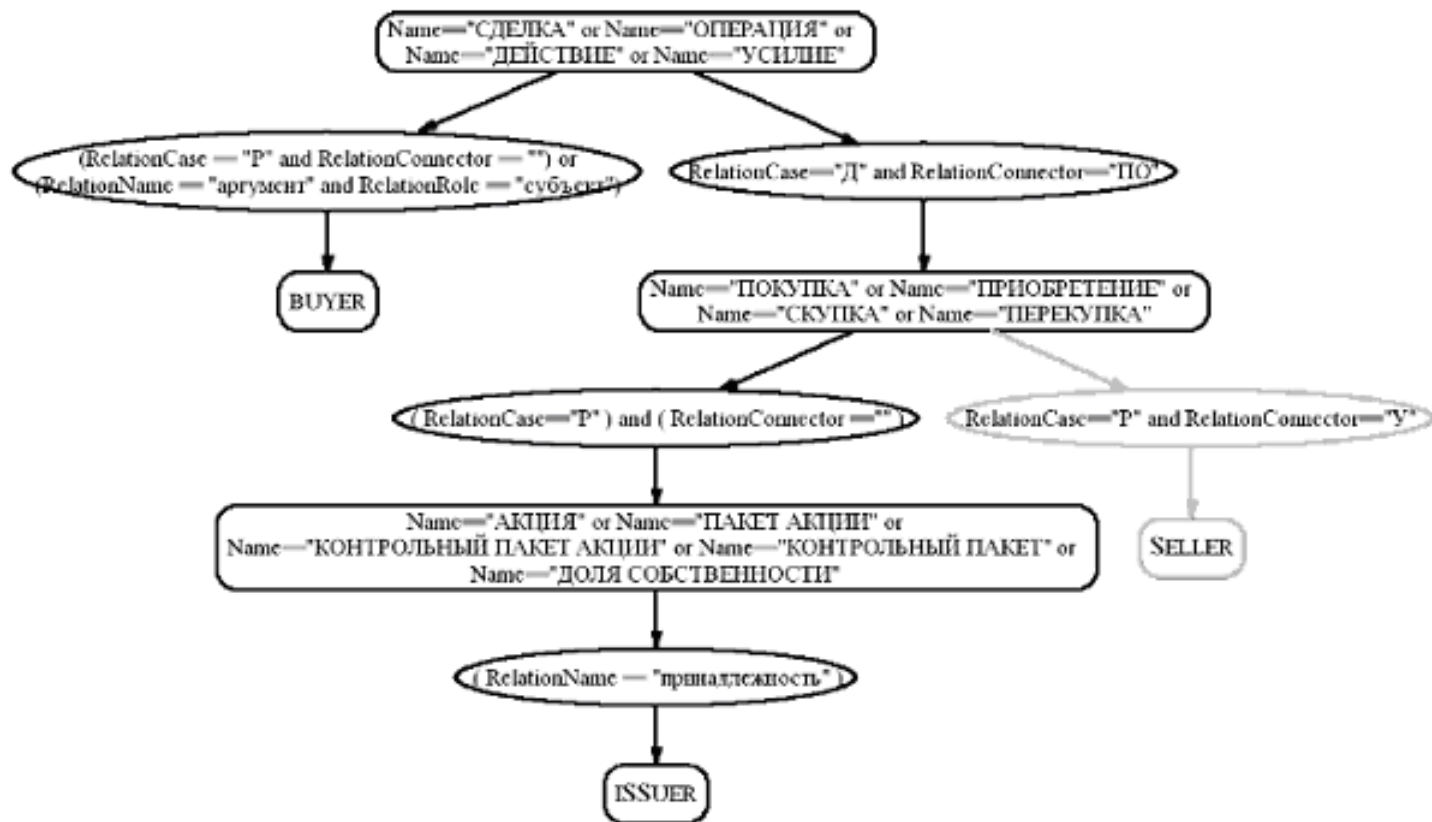
Программа отождествила: он → Липухин Юрий, ГОК → Лебединский горно-обогатительный комбинат, банк → Кредобанк, предприятие → Запсиб.





В ноябре 2003 года Ивановым была совершена сделка по покупке акций Лукойла у Петрова.

Схема ситуации



Покупатель совершает действие по приобретению у продавца акций предприятия.

Поиск описаний фактов

Поиск фактов в тексте - RCO Fact Extractor

Файл Правка Вид Объект Атрибут Анализ Помощь

Объекты

- VIP (90)
 - Владимир Путин (76)
 - Юрий Лужков (14)
- ФПГ (781)
 - МДМ (776)
 - группа МДМ (769)
 - Владеет предпр...
 - Владельцы (16)
 - Договора (23)
 - Покупает (64)
 - Продает (4)
 - Связанные объ...
 - Мельниченко Андре...
 - Юкос (5)
 - компания Юкос (3)
 - Ходорковский Миха...

Атрибуты

- Общие
- Атрибуты персон
- Атрибуты организаций
- Владеет предприятиями
- Владельцы
- Встречи
- Договора
- Поездки
- Покупает акции
- Покупает
- Продает
- Участие в событиях
- Связанные объекты

Объект мо...	Атрибут досье	Значение	Частота	Комментарий
группа МДМ	Покупает	"КОМИСОЦБАНК"	8	
группа МДМ	Покупает	"ПЕТРОВСКИЙ НАРОДНЫЙ Б...	6	
группа МДМ	Покупает	ПАКЕТ "КОНВЕРСБАНКА"	4	
группа МДМ	Покупает	"РЕСО-ГАРАНТИЯ"	4	
группа МДМ	Покупает	РЕГИОНАЛЬНЫЙ БАНК	3	
группа МДМ	Покупает	"ЭКСПОБАНК"	3	
группа МДМ	Покупает	БОЛЬШИНСТВО ЭМИССИИ АК...	2	Покупатель: БАНК
группа МДМ	Покупает	КРУПНАЯ РОССИЙСКАЯ СТРА...	2	
группа МДМ	Покупает	БИЗНЕС	2	
группа МДМ	Покупает	БАНК	2	Покупатель: БАНК
группа МДМ	Покупает	ЭМИССИЯ АКЦИИ	2	Покупатель: "КОНВЕРСБАНК"
группа МДМ	Покупает	"КУЗНЕЦКИЕ ФЕРРОСПЛАВЫ"	2	Продавец: "ФЕРРОТРАНС"
группа МДМ	Покупает	ДОПЭМИССИЯ	2	
группа МДМ	Покупает	ПАКТ АКЦИИ "РЕСО-ГАРАНТИ...	1	
группа МДМ	Покупает	ПАКЕТ "МИНАТОМА"	1	
группа МДМ	Покупает	ЭМИССИЯ КОМПАНИИ	1	

Файл: 17203.htm
([исходный документ с подсветкой>>](#))
Этой группой из пяти страховых компаний владел недавно приобретенный **МДМ-банком Петровский народный банк (ПНБ)**.

Файл: 19501.htm
([исходный документ с подсветкой>>](#))
Эта покупка еще раз доказывает, что финансовый мир очень тесен - предыдущим владельцем Петровской страховой группы был **Петровский народный банк (ПНБ)**, также недавно купленный **МДМ-банком**.

структура
досье

факты по
атрибуту

цитаты из
документов

Фактографический поиск

Результаты

	my_1	2
Precision (macro average)	0.8186	0.9353
Precision	0.9527	0.9825
Precision(owns) (macro average)	0.5974	0.6786
Precision(owns)	0.8615	0.8788
Precision(employs) (macro average)	0.8358	0.9403
Precision(employs)	0.9542	0.9833

and

Фактографический поиск

Выводы

- Система, оптимизированная на конкретную предметную область, показывает лучший результат
 - по крайней мере, в Precision
- Полностью оценки еще не проведены

Агрегация новостей

Мысли

Сообщение → Событие → Сюжет

- Сообщение → Событие
 - Сравнение информационных портретов
 - Сравнение семантических сетей
- Событие → Сюжет
 - Фреймовая структура (2006-...)

Заключение

- Матрицы релевантности прошлых лет
 - Распространять среди участников матрицы и утилиту для подсчета оценок
- Коллекция материалов СМИ
 - Можно ли на ней повторить задания по фактографическому поиску и агрегации новостей в следующем цикле?