RCO на РОМИП 2004

Плешко В.В., Ермаков А.Е., Голенков В.П.
ООО «Гарант-Парк-Интернет»
rco@metric.ru

Дорожки

- Документальный поиск
- Классификация правовых документов
- Классификация web сайтов
- Поиск биографических фактов

Документальный поиск
Мотивация

Получение количественных оценок влияния некоторых факторов на основные показатели качества документального поиска, а именно:

- учет словоформ
- учет словосочетаний

Документальный поиск
Описание системы

Базовые функции

- Ранжирование tf*idf
- NEAR-100
- Без стоп-слов

Опции

- Морфология (см. Диалог'2004, особый учет глагольных форм)
- Альтернативное ранжирование (док-ты со словосочетаниями получают больший ранг)

Документальный поиск
Прогоны

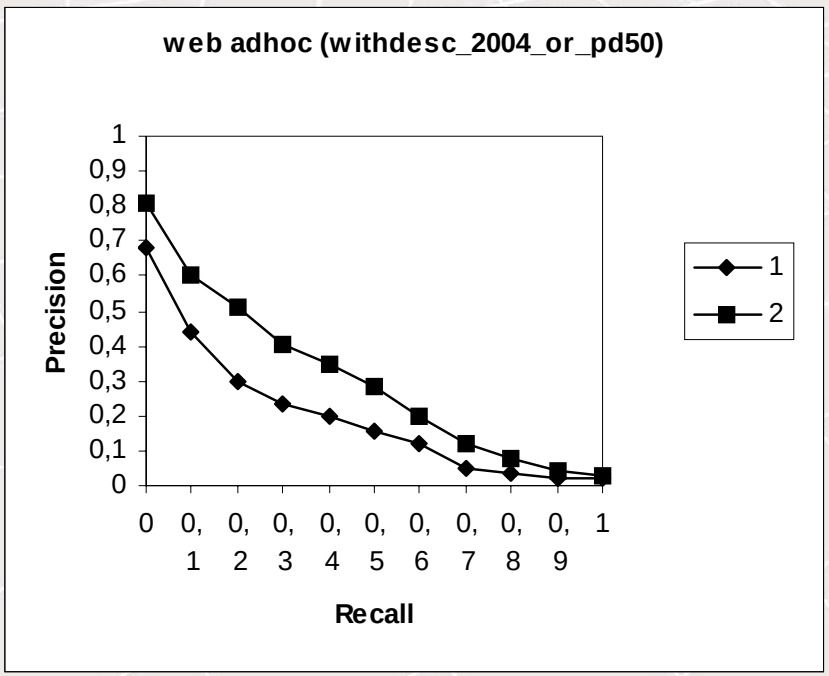
- legal adhoc
  1. базовый
  2. учет словоформ
  3. учет словосочетаний
  4. учет словоформ и словосочетаний
- web adhoc
  1. базовый
  2. учет словоформ

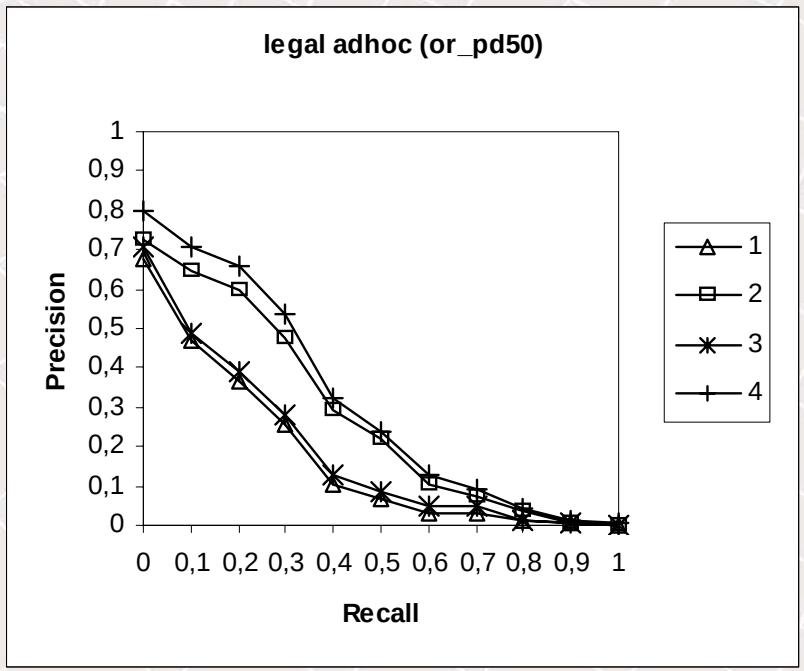
Документальный поиск Результаты (таблица)

Run	Recall	Prec(5)	Avg prec	Prec (10)	R-prec	Prec
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

- web-or-withdesc-pd50-2004
- legal-or-pd50

Документальный поиск Результаты (графики)

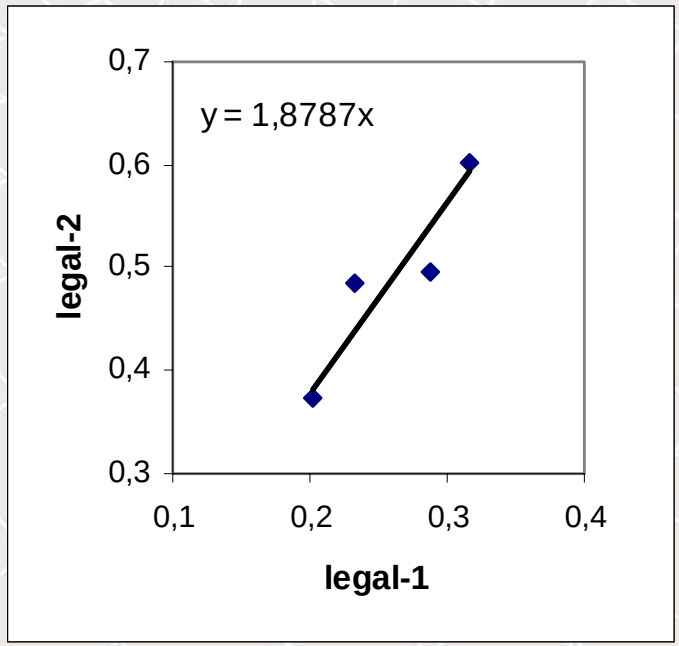




Документальный поиск
Метод оценки факторов

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

•показатель Recall



Документальный поиск
Оценка факторов (таблица)

Зависимость \ Показатель	Recall	Prec	Prec(5)	Prec(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

Документальный поиск Оценка факторов (выводы)

•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.

Документальный поиск Пожелания

- Нечетное число оценок каждой пары <документ,запрос> => релевантность определяется путем «голосования»
- Остальные оценки (and, or) для проверки
- pd50 – хорошая идея
- Описания – открытый вопрос...

Классификация правовых док-тов Мотивация

- Использовать отрицательные примеры наряду с положительными
- Отбор терминов в качестве признаков на основе «коммуникативного ранга»
- Реализация процедуры классификации при помощи полнотекстовых запросов

Классификация правовых док-тов
Постановка задачи

D – множество док-тов
Tr – обучающая выборка
Ts – тестовая выборка
Ф – отношение принадлежности
док-тов заданному классу

$$Error = \sum_{Ts} \left| \Phi - \hat{\Phi} \right| \rightarrow \min$$

Классификация правовых док-тов Описание метода - 1

c – терминологический вектор описания рубрики

d_j – терминологический вектор документа

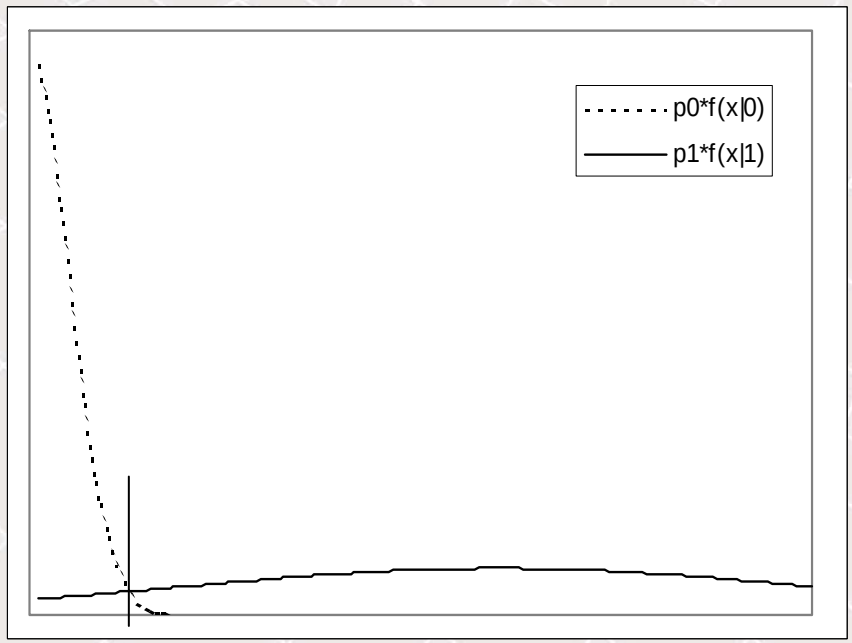
$$CSV(c, d_j) = c \cdot d_j = \sum_k c_k d_{jk}$$ - степень сходства

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

Классификация правовых док-тов Описание метода - 2

Выбор порога – из отношения правдоподобия

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



Классификация правовых док-тов Описание эксперимента - 1

Коммуникативный ранг:

- Фактор тема – рема
- Вес члена предложения (подлежащее, субстантивное сказуемое, дополнение, глагольное сказуемое, обстоятельства (места, времени, причины, цели), определения и прочие обстоятельства (образа действия, меры, степени).
- Число слов в составе группы
- Число слов в составе наиболее полной группы, содержащей термин

- Вес термина в док-те – сумма его КР по всему тексту док-та
- Отбирались термины с весом не менее 15% от макс. в док-те

Классификация правовых док-тов Описание эксперимента – 2

Отбор терминов

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

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

Веса в терм-векторах

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

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

Классификация правовых док-тов Описание эксперимента – 3

Терминов в описании рубрики – ок 100
Словосочетаний среди терминов – 83%

Терминов в описании документа – ок 40

Классификация правовых док-тов Результаты

показатель \ способ	and	or	ideal	ideal40
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

Классификация правовых док-тов Пожелания

- Видимо, должны оценивать профессионалы
- Постановка с обязательным отнесением хотя бы к одному классу
- Учет гипертекста (ссылки на приложения, на исходные редакции док-тов)

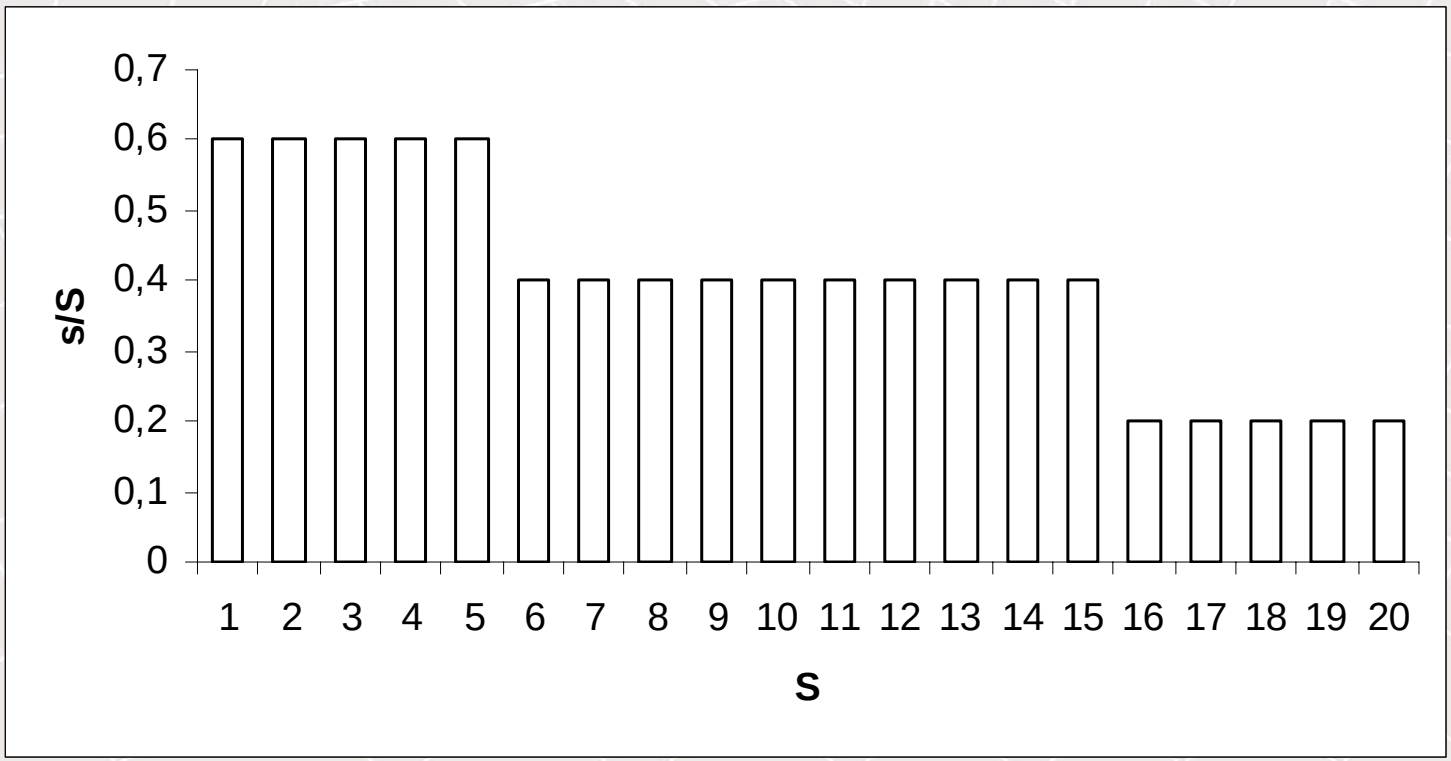
Классификация web-сайтов Мотивация

- Отказ от представления сайта в виде «суперстраницы», а также от описания сайта набором терминов
- Классификация страниц поотдельности (обучающая выборка будет содержать шум – служебные страницы)
- Вывод о принадлежности сайта классу по принадлежности его страниц

Классификация web-сайтов Описание метода - 1

1. Классификация страниц
2. Классификация сайтов в пространстве небольшой размерности (по результатам шага 1)
   - *S* – общее число страниц сайта
   - *s* – число страниц сайта, отнесенных к заданному классу

Классификация web-сайтов
Описание метода - 2



Классификация web-сайтов
Описание эксперимента – 1

Терминов в описании рубрики – ок 600
Словосочетаний среди терминов – 45%

Терминов в описании документа – ок 50
Уникальных терминов на сайте – ок 1600

Классификация web-сайтов
Описание эксперимента – 2

Прогон 1

$s_i \geq 5$

$s_i / S \geq 0.2$

Только класс с макс. s_i / S

Прогон 2

$s_i / S \geq 0.2$

Первые 5 классов по убыв. s_i / S

Классификация web-сайтов Результаты

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

- оценка or

Классификация web-сайтов
Пожелания

- Дорожка по классификации страниц
  - РОМИП’2005
- Информация о «глубине» страницы
  - число переходов по ссылкам от стартовой

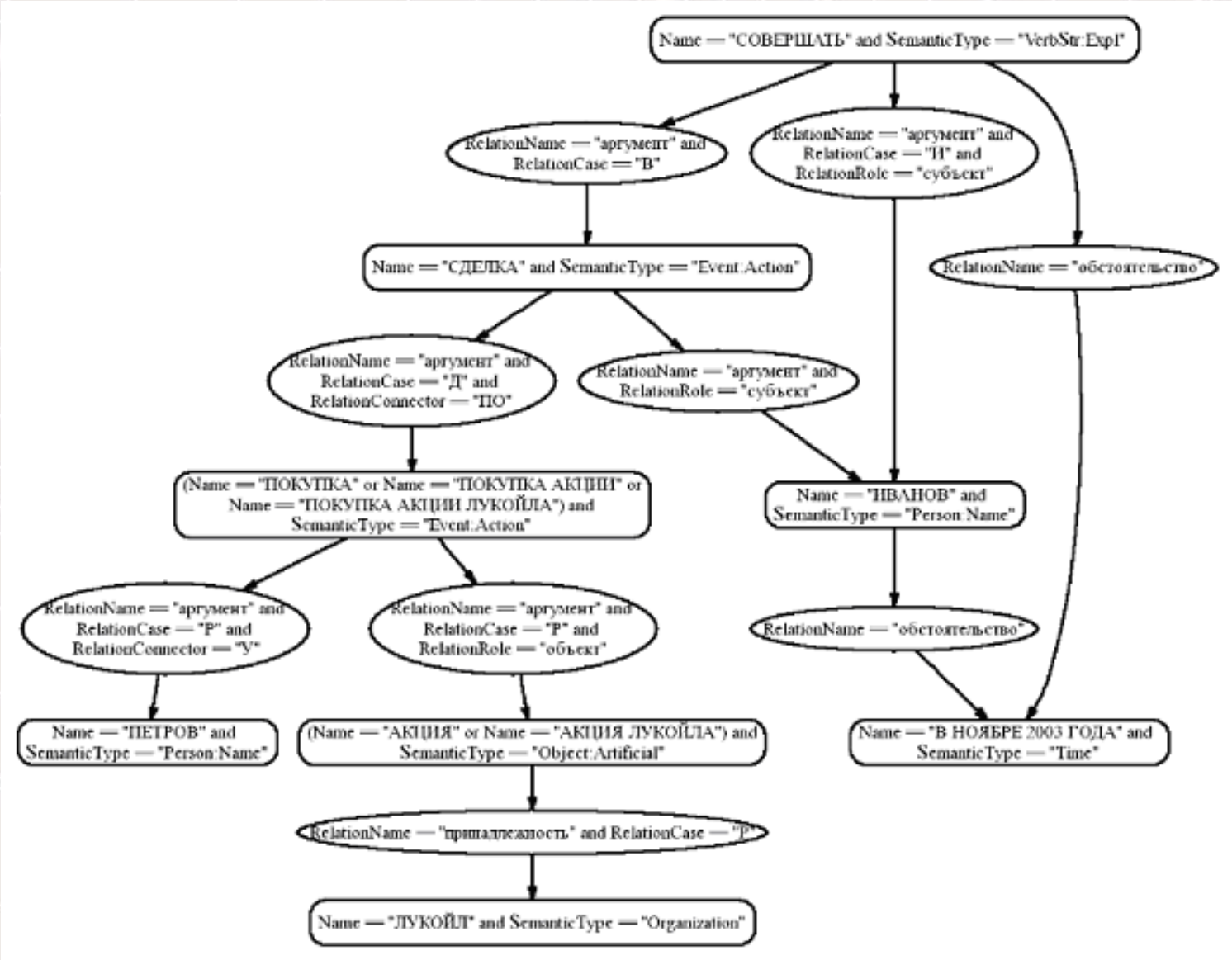
Поиск биографических фактов Постановка задачи

- 5000 записей о персонах
  - ФИО
  - Краткое описание
- Коллекция narod.ru
- Требовалось найти «биографические» факты
  - Фрагмент (ИД док-та, смещение, длина)
- Опционально – картирование найденных фрагментов

Поиск биографических фактов Описание метода

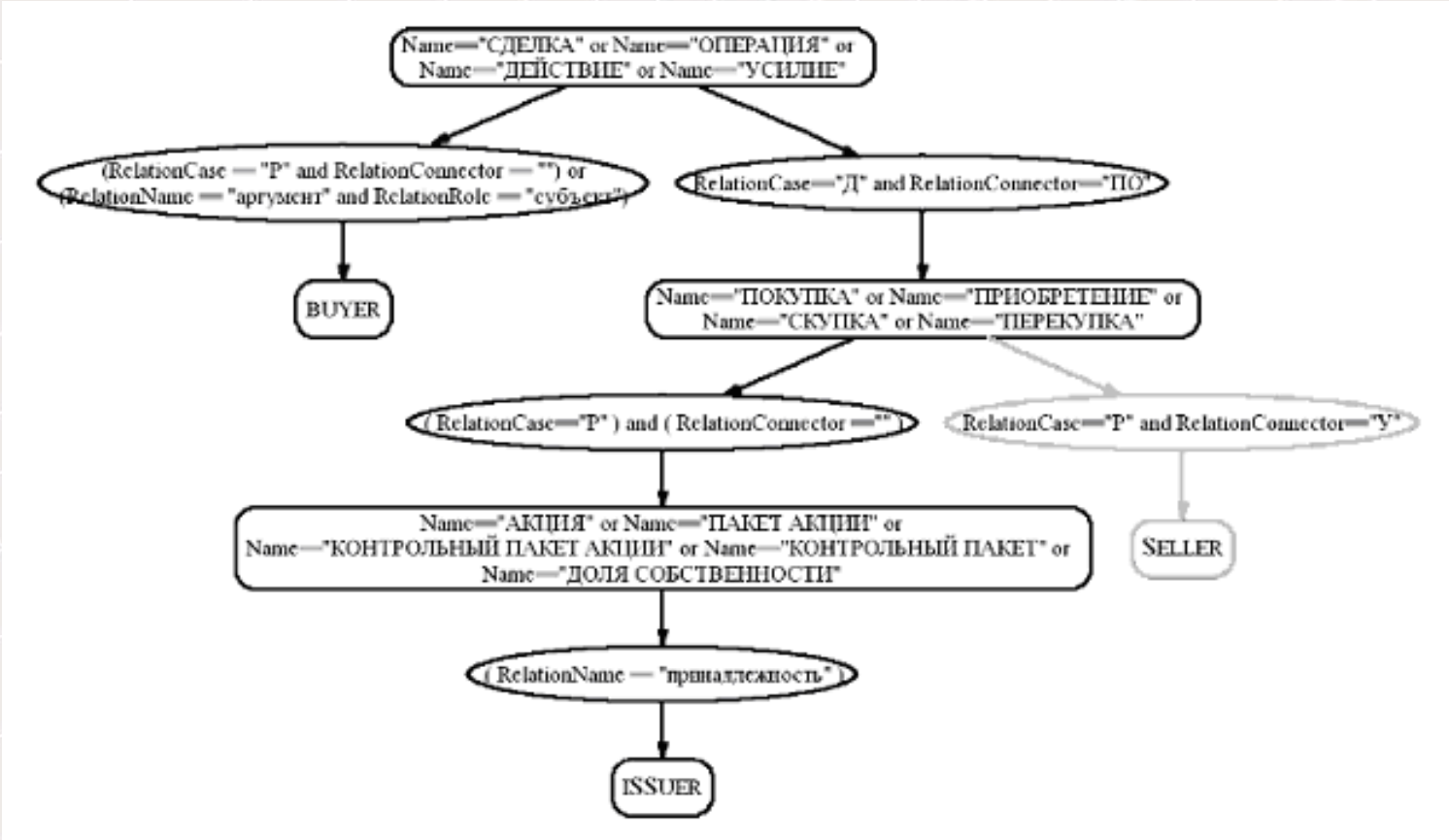
1. Полнотекстовый поиск
   - Необходимое условие (фамилия, имя)
   - Гипотеза о поле персоны
2. Анализ отобранных документов
   - Поиск предложений, содержащих полные, краткие, косвенные наименования персоны

Семантическая сеть предложения



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 биографических фактов Описание эксперимента

- Вырожденная постановка задачи, с точки зрения метода
  - Одна вершина в описании ситуации
  - SemanticType = PersonName
- Неточности при выполнении задания
  - Много отдельно стоящих упоминаний (заголовки, ссылки); Удалили упоминания < 4 слов
  - Погрешность в смещениях (свой распаковщик)

Поиск биографических фактов Результаты

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

- Только одна система (нельзя применить метод общего котла)
- Трехкратное различие между or и and

Поиск биографических фактов Пожелания

- Более четкие задания
  - в идеале, вопросно-ответного типа (дата рожд., должность, место учебы, «кто-где-когда»)
  - поиск упоминаний персон, организаций
- Картирование фактов
  - только если по образцу, с обучающей выборкой (набором фрагментов)

Заключение

- Единый способ оценки
- Публикация корпусов
  - по поиску и классификации результаты заслуживают доверие
- Коллекция материалов СМИ
- Фактографический поиск