Автоматическая рубрикация текстов-2

Рубрицирование в реальных системах

Методы рубрицирования

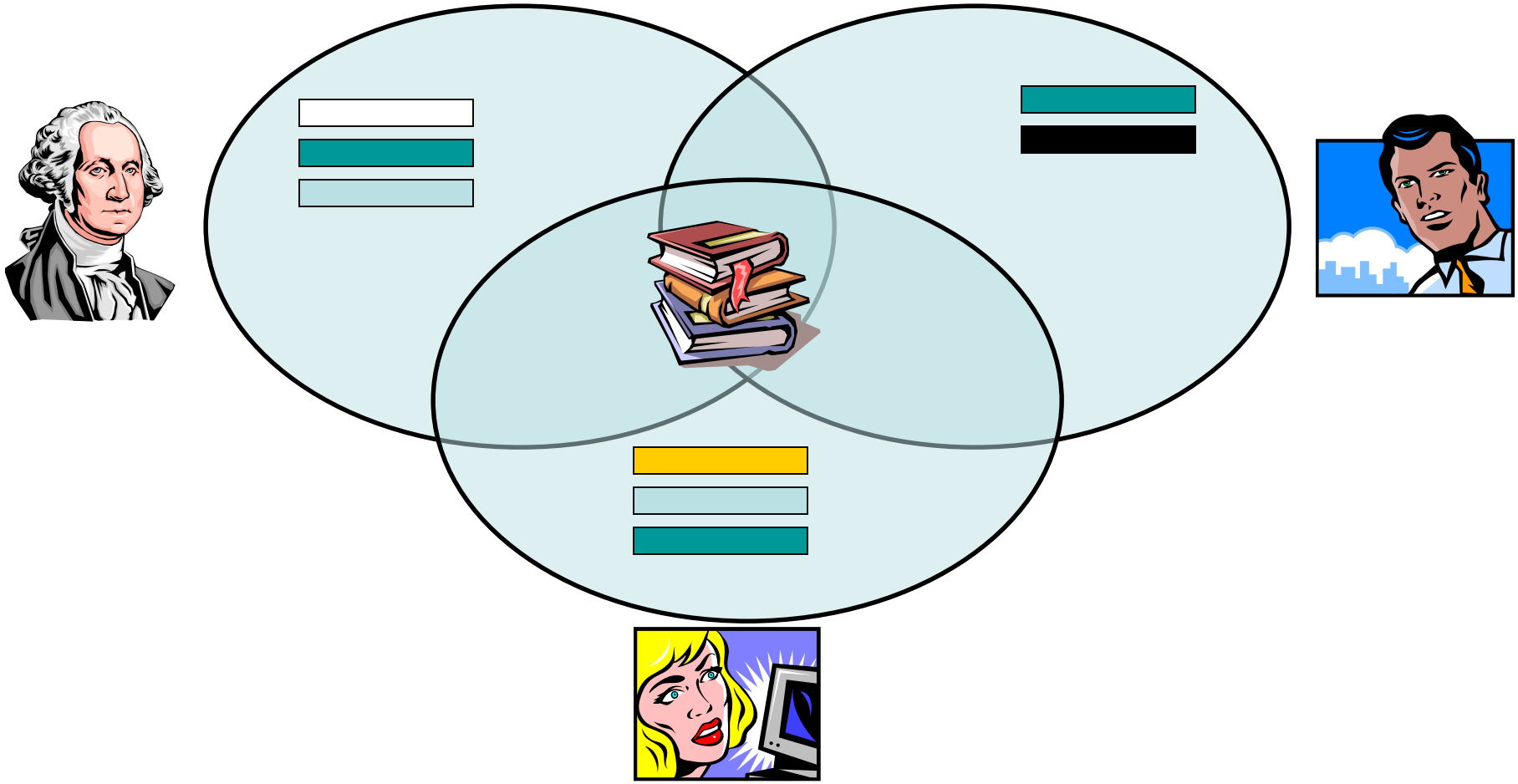
- Ручное рубрицирование
  - Используется для оценки качества автоматического рубрицирования
- Автоматическое рубрицирование
  - Инженерный подход (методы, основанные на знаниях)
  - Машинное обучение

Ручное рубрицирование

- Высокая точность рубрицирования
  - Обычно процент документов, в которых проставлена явно неправильная рубрика, чрезвычайно мал
- Низкая полнота рубрицирования
  - одна-две основных рубрики, характеризующие основное содержание документа, хотя документ может быть отнесен и к ряду других рубрик.
  - В результате получается, что
    - Процент совпадения результатов рубрицирования различных экспертов весьма низкий - 60 %.
    - В результате похожие документы могут получить достаточно разные наборы рубрик
    - Непоследовательность ручного рубрицирования
    - Низкая скорость обработки документов

Субъективизм экспертов

Совпадение при ручной рубрикации между разными экспертами 60%



Мало отличающиеся документы имеют разные наборы рубрик: как обучаться?

Инструмент-1, ФЕДЕРАЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО, Верховный Совет, Постановление № 5436-I от 14.07.1993 (51%)

Об индексации минимального размера пенсий с учетом изменения индекса цен за второй квартал 1993 года

070060110010 Минимальный размер пенсии

Инструмент-1, ФЕДЕРАЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО, Верховный Совет, Постановление № 4810-I от 15.04.1993 (50%)

Об индексации минимального размера пенсий с учетом изменения индекса цен за первый квартал 1993 года

070060110 Исчисление пенсии. Надбавки. Перерасчет пенсий
020030100010 Общие вопросы\Цены и ценообразование

Инструмент-1, ФЕДЕРАЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО, Верховный Совет, Постановление № 4296-I от 15.01.1993 (49%)

Об индексации минимального размера пенсий с учетом изменения индекса цен за четвертый квартал 1992 года

070060110010 Минимальный размер пенсии
020030100010 Общие вопросы\Цены и ценообразование

Субъективность ручного рубрицирования

Федеральные законы 1995-1997 гг. «О повышении минимального размера оплаты труда»

	1(95)116	2(95)159	3(96)40	4(97)6	5(95)43	Итого
010140030010		+	+	+	+	4
060020090040	+	+	+	+	+	5
070010		+				1
070070010070		+		+	+	3
080100010				+		1
080100020010				+		1
080050030020040				+		1
130010040060				+		1
130010040080				+		1
130010040090				+		1
130010070030030	+	+	+	+	+	5
	2	5	3	10	5	(11)
Полнота	0,18	0,45	0,27	0,91	0,45	47,2%
Точность	1,00	1,00	1,00	1,00	1,00	100%

Инструмент-1, ФЕДЕРАЛЬНОЕ ЗАКОНОДАТЕЛЬСТВО, Президент, Указ № 1422 от 18.12.2006

О Боевом знамени воинской части

010140030010 Отмена, изменение и дополнение нормативных правовых актов
010020010010 Государственные герб, гимн, флаг Российской Федерации
010140040045020 Иные положения
150020 Вооруженные Силы Российской Федерации, другие войска, воинские формирования и органы, привлекаемые к выполнению задач в области обороны

ПОЛОЖЕНИЕ
ВОИНСКИЙ
ЧАСТЬ
БОЕВОЙ
ЗНАМЯ

☐ версия для печати; ☐ карточки документов; ☐ текст; ☐ трансп.; ☐ пересчет по отмеченным соседям

обновить

Рубрика	док-в	F (train)	док-рубр dist	threshold	nn_rubr	0	1	2	3	4	5	6	7	8	9	10
						1.00	0.70	0.63	0.63	0.62	0.60	0.57	0.55	0.55	0.45	0.35
010140 Правотворческая деятельность органов государственной власти	32887	88.55	**1.0000**	**0.5249**	10	+	+	+	+	+	+	+	+	+	+	+
150020 Вооруженные Силы Российской Федерации, другие войска, воинские формирования и орга...	673	41.27	**0.7165**	**0.2054**	7	+		+	+	+	+	+		+	+	
010020 Государственные символы Российской Федерации и субъектов Российской Федерации. Стол...	340	48.40	**0.7756**	**0.2094**	8	+		+	+	+	+		+	+	+	+
010090 Президент Российской Федерации	2429	56.53	**0.7611**	**0.3026**	7		+	+	+	+	+	+		+		
020010 Органы исполнительной власти	16741	66.53	**0.7107**	**0.3090**	7		+	+	+		+		+	+		+
010180 Государственные награды. Высшие степени и знаки отличия. Почетные звания. Знаки, зн...	399	52.26	**0.2351**	**0.2003**	2		+		+							
060020 Труд (см. также 200.160.020)	5362	56.93	0.2067	0.3014	2						+	+				

1257600: О Боевом знамени Президентского полка Службы коменданта Московского Кремля Федеральной службы охраны Российской Федерации

Инженерный подход

Проблемы методов, основанных на знаниях

- Содержание рубрики сложнее, чем это выглядит по формулировке
- Лексическая многозначность
- Ложная корреляция
- Нестандартный контекст употребления терминов
- Упоминание терминов вне главной темы
- Неполнота описания рубрики

Ошибки: появление лишних рубрик (1)

Содержание рубрики сложнее, чем это выглядит по формулировке

Например, к рубрике «Выборы» при автоматической рубрикации при обработке материалов СМИ может быть отнесен следующий текст

ГАЗЕТА "КОММЕРСАНТЪ" № 135(3466) ОТ 26.07.2006

Мишель Платини хочет возглавить UEFA

Чемпион мира француз Мишель Платини будет баллотироваться на пост президента Европейского союза футбольных ассоциаций (UEFA). Об этом проинформировали во Французской футбольной федерации (FFF). 51-летний Платини стал пока единственным конкурентом 76-летнего шведа Леннарта Юханссона, который возглавляет UEFA с 1990 года и намерен вновь баллотироваться на эту должность. В настоящее время господин Платини занимает пост вице-президента FFF, входит в исполкомы FIFA и UEFA.

В прошлом месяце немец Франц Беккенбауэр признался, что выдвинул бы свою кандидатуру, если бы Леннарт Юханссон не стал баллотироваться. Выборы пройдут в немецком Дюссельдорфе 25-26 января 2007 года.

Ошибки: появление лишних рубрик (2)

- **Лексическая многозначность - текст может быть отнесен не к той рубрике из-за того, что некоторые слова, сопоставленные рубрике, в конкретном тексте употреблены в таком значении, которое не соответствует данной рубрике.**

  - **МОРСКИЕ СУДА;**
    РЕШЕНИЕ СУДА;
    СТАРИННОЕ ЗДАНИЕ СУДА

  - **ПРОИЗВОДСТВО ТОВАРОВ;**
    ПРОИЗВОДСТВО ПО УГОЛОВНОМУ ДЕЛУ

Ошибки: появление лишних рубрик (3)

- **Нестандартный контекст употребления терминов. Например, следующий текст может быть отнесен к рубрике "Средства массовой информации", по такому же словосочетанию, употребленному в тексте, но по сути текст не является релевантным данной**

Жертвами жары во Франции стали около 40 человек.

26.07.2006 07:19:20, Париж:

Около 40 человек умерли во Франции в результате установившейся в стране в последние две недели жары. Об этом сообщил государственный Институт здоровья Франции. Правительство и средства массовой информации следят за ситуацией и сообщают населению, как следует себя вести в условиях высокой температуры, которая в последние дни колеблется между 35 и 40 градусами по Цельсию, передает (С) Associated Press.
В 2003г. жертвами необычайно жаркой погоды во Франции стали 15 тыс. человек, преимущественно пожилого возраста.

Ошибки: пропуск нужной рубрики

- Правильная рубрика не определена, поскольку в тексте упомянуты слова, не описанные в словаре системы рубрицирования.
- Например, следующий текст может быть не отнесен к рубрике **"Политические партии и движения",** поскольку партии и движения упомянуты посредством их сокращенных названий

Результаты "Родины" и РПЖ на последних региональных выборах

0

ГАЗЕТА "КОММЕРСАНТЪ" № 135(3466) ОТ 26.07.2006

Дата	Регион	Партия	Результат (%)
16.10.05	Белгородская область	"Родина"	6,42
		РПЖ	1,27 (не прошла)
27.11.05	Чеченская республика	"Родина"	2,39 (не прошла)
04.12.05	Москва	"Родина"	Снята судом за нарушение правил агитации
		РПЖ	4,77 (не прошла)
04.12.05	Костромская область	"Родина"	9,06
		РПЖ	4,71
04.12.05	Ивановская область	"Родина"	10,51
11.12.05	Хабаровский край	"Родина"	10,57

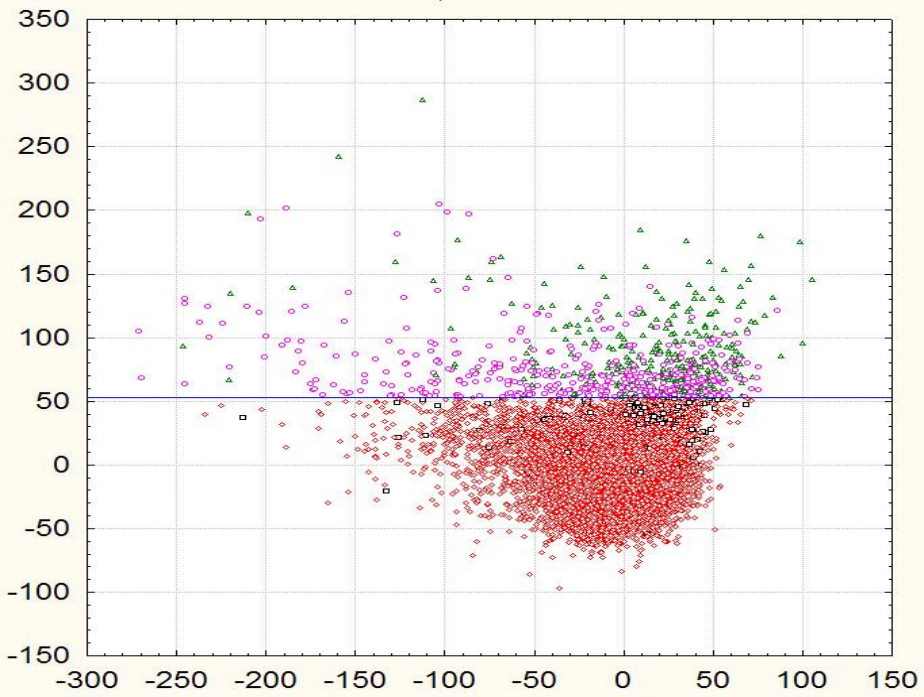
Машинное обучение

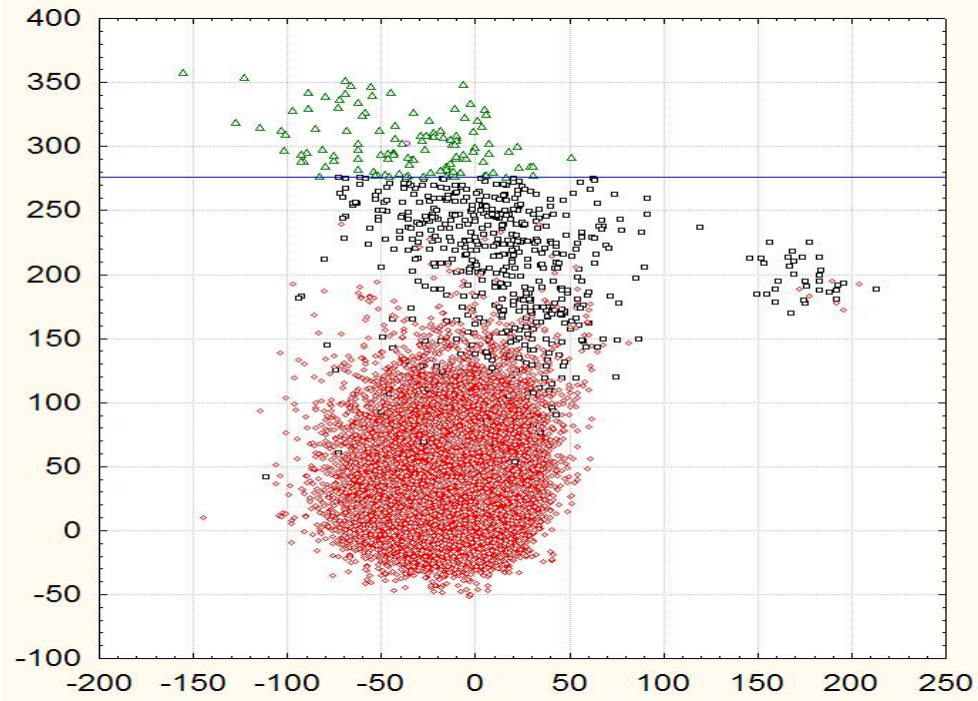
Оценка метода на коллекции документов

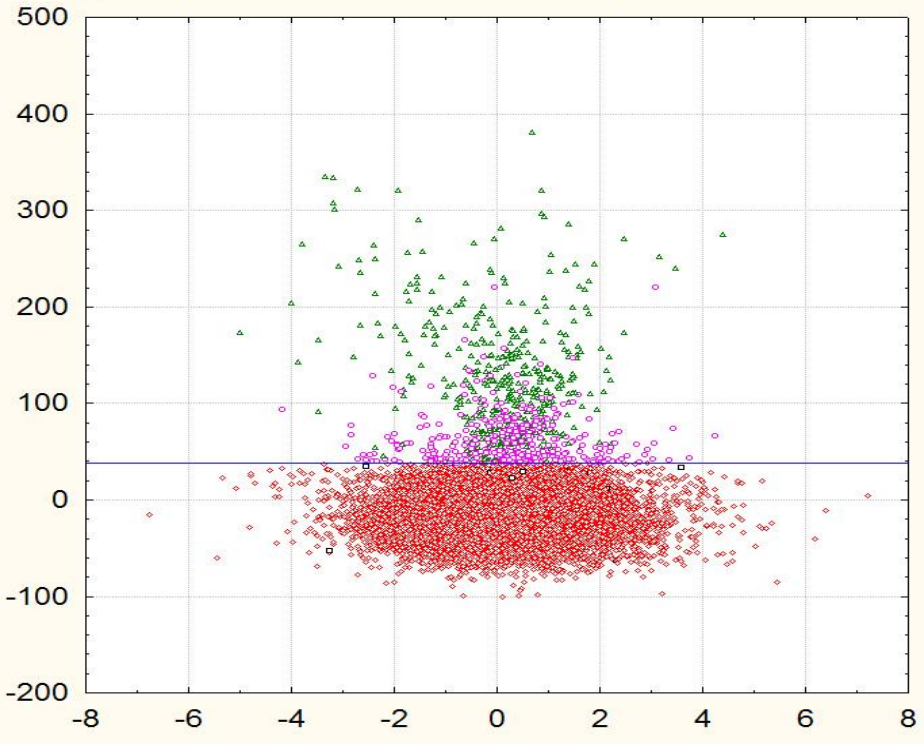
- Оценка метода на коллекции обучения
  - Позволяет оптимизировать параметры алгоритма в процессе обучения
  - Возможно «переобучение» алгоритма

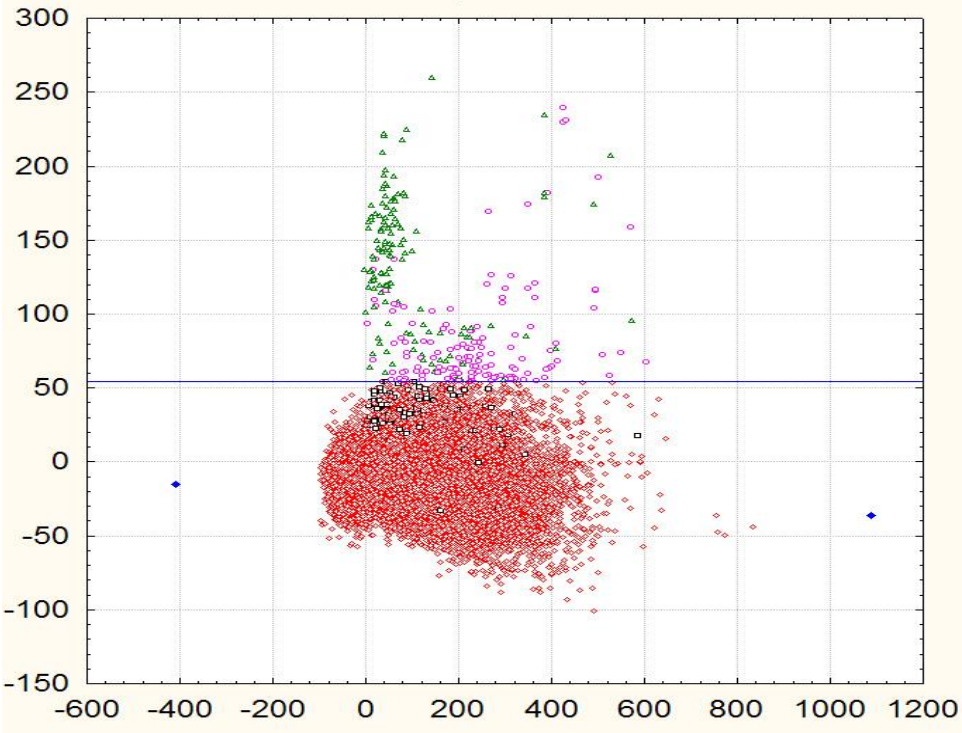
- Оценка на разбиении TRAIN/TEST
  - Коллекция обучения разбивается на множество для обучения и множество для тестирования (например, в пропорции 70%/30%)

- Усреднение оценок для различных разбиений TRAIN/TEST: *cross-validation*


350
300
250
200
150
100
50
0
-50
-100
-150
-300 -250 -200 -150 -100 -50 0 50 100 150


400
350
300
250
200
150
100
50
0
-50
-100
-200 -150 -100 -50 0 50 100 150 200 250


500
400
300
200
100
0
-100
-200
-8 -6 -4 -2 0 2 4 6 8


300
250
200
150
100
50
0
-50
-100
-150
-600 -400 -200 0 200 400 600 800 1000 1200

Сложные задачи автоматической рубрикации текстов: проблемы машинного обучения

- размер рубрикатора больше 300-500 рубрик, обычно со сложной иерархией

- трудно обеспечить достаточную по качеству и количеству обучающую коллекцию, субъективизм ручного индексирования (обучающей коллекции) значительно возрастает

- сложные задачи решаются на основе инженерных подходов или с помощью частичной автоматизации

Множество примеров отсутствует и не может быть создано в короткое время

- Российский социологический архив (www.socialpolicy.ru)
- Данные соцопросов разных организаций
- 350 рубрик, 4 уровня иерархии
- Новый проект => отсутствие примеров

Множество примеров существует, но отсутствовали требования к качеству

- Международное научное сообщество RePec (www.repec.org), SocioNet (www.socionet.ru)
- Архив исследовательских материалов по экономике и социологии
- Рубрикатор: Journal of Economic Literature Classification System (JEL)
- Более 700 рубрик
- Автор сам приписывает рубрики к своей работе

Множество примеров противоречиво и недостаточно для большинства рубрик (очень большие классификаторы)

- Российские правовые документы
- Президентский классификатор (Указ №511 15.03.2000) - 1168 рубрик
- Множество примеров – 10,000 документов классифицированных вручную
- Только для 47 рубрик – более чем 100 док., только для 200 рубрик – более чем 20 док.
- Inconsistency: мало отличающиеся документы имеют разные наборы рубрик

Множество примеров для обучения из другой коллекции

- Примеры: документы федерального уровня
- Проблема: рубрицирование 600,000 региональных документов
- Тот же рубрикатор
- Похожие документы, похожая проблема

НО!!!

- Стандартный метод SVM-light, обученный на федеральных документах не приписывает ни одной рубрики для 50% документов

CS276: Information Retrieval and Web Search

Christopher Manning, Pandu Nayak, and Prabhakar Raghavan

The Real World

P. Jackson and I. Moulinier. 2002. *Natural Language Processing for Online Applications*

- "There is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate intranets, government departments, and Internet publishers"

- "Understanding the data is one of the keys to successful categorization, yet this is an area in which most categorization tool vendors are extremely weak. Many of the 'one size fits all' tools on the market have not been tested on a wide range of content types."

The Real World

- Gee, I'm building a text classifier for real, now!
- What should I do?

- How much training data do you have?
  - None
  - Very little
  - Quite a lot
  - A huge amount and its growing

Manually written rules

- No training data, adequate editorial staff?
- Never forget the hand-written rules solution!
  - If (wheat or grain) and not (whole or bread) then
    - Categorize as grain
- In practice, rules get a lot bigger than this
  - Can also be phrased using tf or tf.idf weights
- With careful crafting (human tuning on development data) performance is high:
  - Construe: 94% recall, 84% precision over 675 categories (Hayes and Weinstein 1990)
- Amount of work required is huge
  - Estimate 2 days per class … plus maintenance

С тезаурусом как у нас намного быстрее!!!

Very little data?

- If you're just doing supervised classification, you should stick to something high bias
  - There are theoretical results that Naïve Bayes should do well in such circumstances (Ng and Jordan 2002 NIPS)
- The interesting theoretical answer is to explore semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …
- The practical answer is to get more labeled data as soon as you can
  - How can you insert yourself into a process where humans will be willing to label data for you??

A reasonable amount of data?

- Perfect!
- We can use all our clever classifiers
- Roll out the SVM!

- But if you are using an SVM/NB etc., you should probably be prepared with the “hybrid” solution where there is a Boolean overlay
  - Or else to use user-interpretable Boolean-like models like decision trees
  - Users like to hack, and management likes to be able to implement quick fixes immediately

A huge amount of data?

- This is great in theory for doing accurate classification…
- But it could easily mean that expensive methods like SVMs (train time) or kNN (test time) are quite impractical

- Naïve Bayes can come back into its own again!
  - Or other advanced methods with linear training/test complexity like regularized logistic regression (though much more expensive to train)

How many categories?

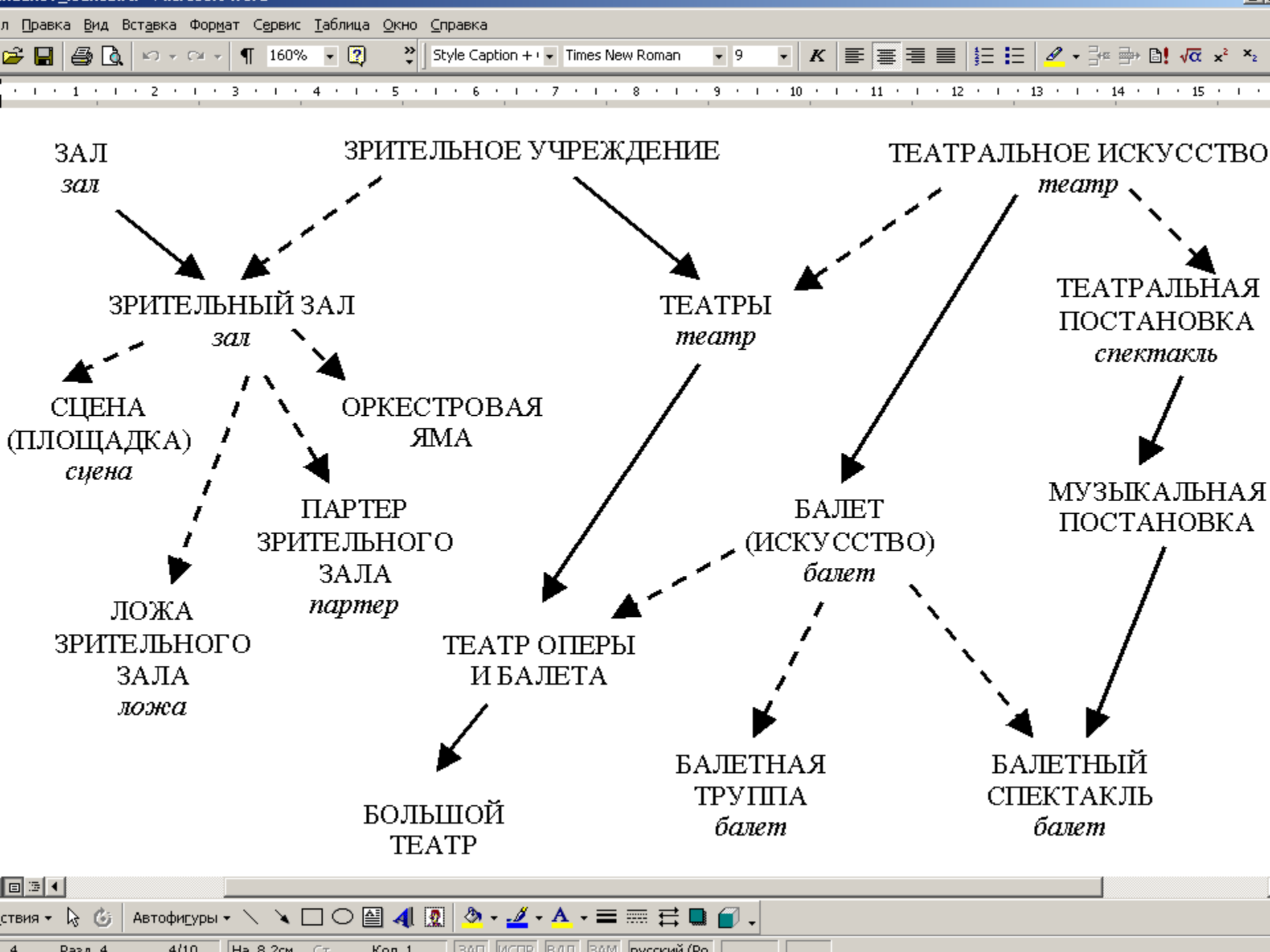
- A few (well separated ones)?
  - Easy!
- A zillion closely related ones?
  - Think: Yahoo! Directory, Library of Congress classification, legal applications
  - Quickly gets difficult!
    - Classifier combination is always a useful technique
      - Voting, bagging, or boosting multiple classifiers
    - Much literature on hierarchical classification
      - (Tie-Yan Liu et al. 2005)
    - May need a hybrid automatic/manual solution

Инженерный подход на основе тезаурусных знаний

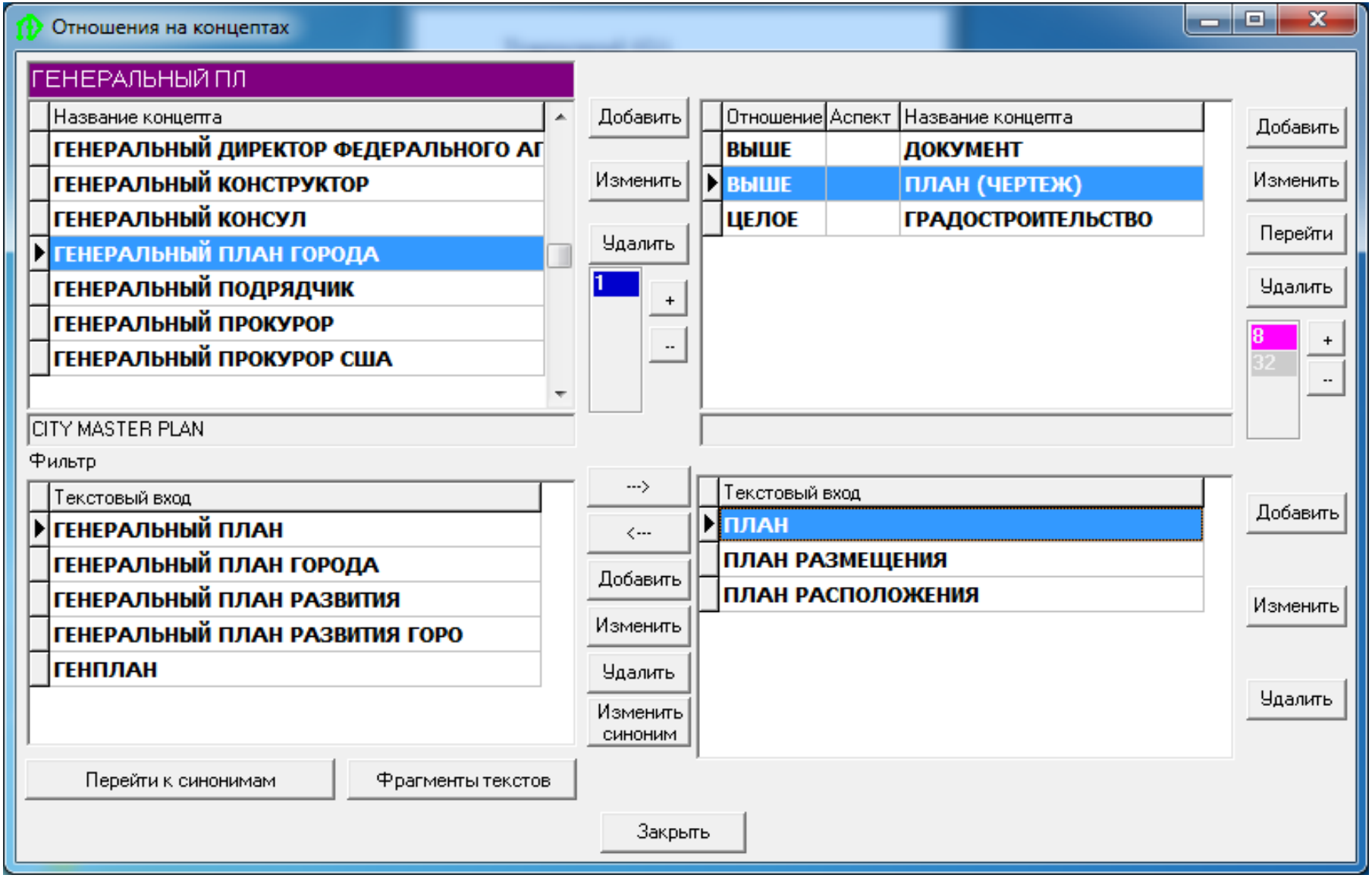
Лингвистическая онтология Тезаурус РуТез

- Понятие:
  - Имя понятия
  - Набор текстовых выражений
  - Отношения между понятиями
- 53 тыс. понятий,
  156 тыс. текстовых выражений,
  210 тыс. отношений (более 2 млн. с иерархией)

- Переведен на английский язык:
  130 тысяч слов и выражений


ЗАЛ
зал
ЗРИТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ТЕАТРАЛЬНОЕ ИСКУССТВО
театр
ЗРИТЕЛЬНЫЙ ЗАЛ
зал
ТЕАТРЫ
театр
ТЕАТРАЛЬНАЯ ПОСТАНОВКА
спектакль
СЦЕНА (ПЛОЩАДКА)
сцена
ОРКЕСТРОВАЯ ЯМА
ПАРТЕР ЗРИТЕЛЬНОГО ЗАЛА
партер
БАЛЕТ (ИСКУССТВО)
балет
МУЗЫКАЛЬНАЯ ПОСТАНОВКА
ЛОЖА ЗРИТЕЛЬНОГО ЗАЛА
ложа
ТЕАТР ОПЕРЫ И БАЛЕТА
БОЛЬШОЙ ТЕАТР
БАЛЕТНАЯ ТРУППА
балет
БАЛЕТНЫЙ СПЕКТАКЛЬ
балет

Пример экрана тезауруса



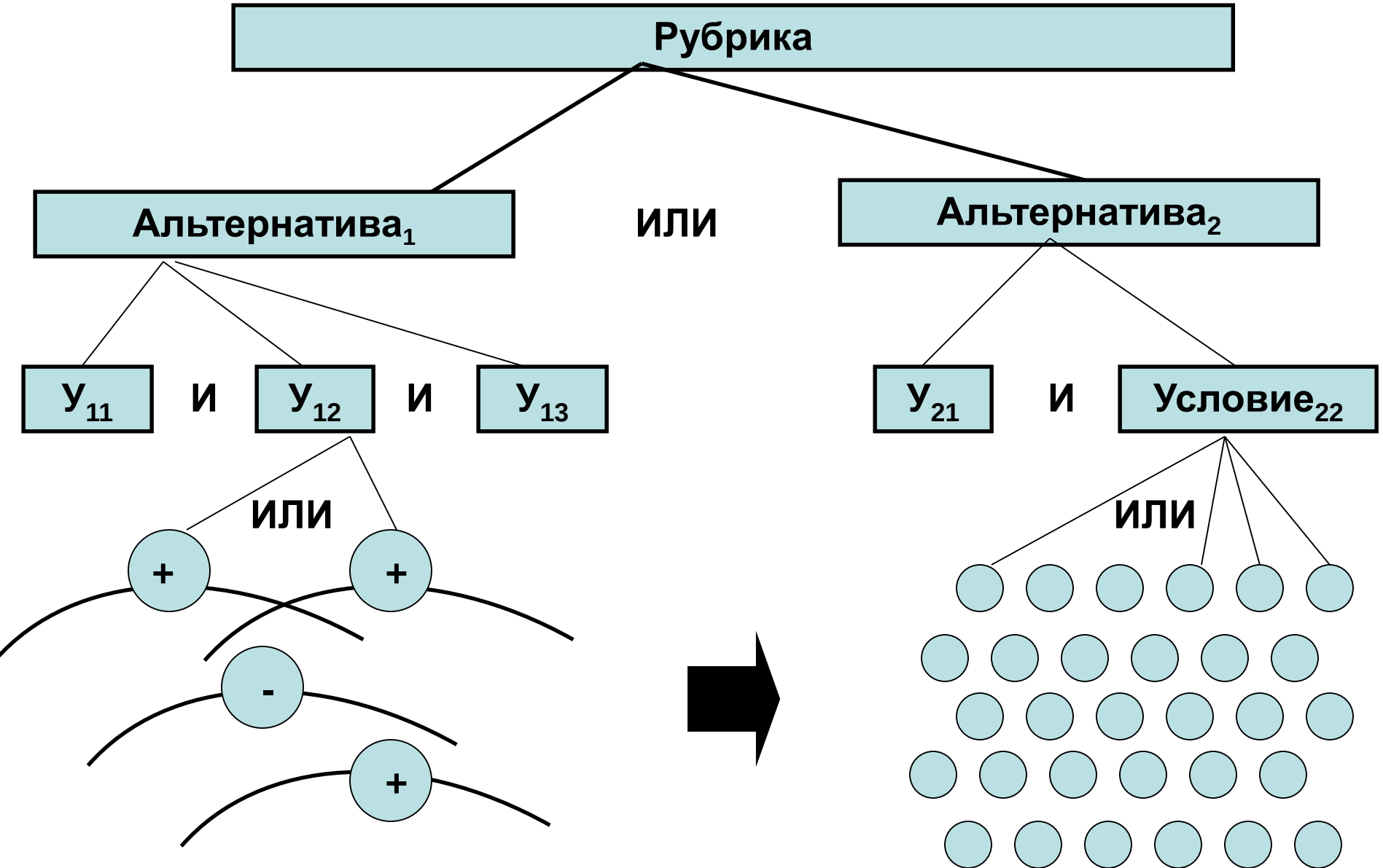
Технология автоматического рубрицирования

- Опора на знания, описанные в тезаурусе
- Представление рубрики в виде булевской формулы для небольшого числа *ОПОРНЫХ* концептов, затем автоматическое расширение с использованием иерархической структуры Тезауруса
- Независимый от конкретного рубрикатора (изменения состава рубрикатора) автоматический тематический анализ текста – выявление в тексте совокупностей близких терминов, выявление терминов, характеризующих основную тему и подтемы документов
- Ранжирование документов с учетом весов

технологии автоматической классификации на основе тезауруса для автоматического индексирования

- По общему тематическому правовому классификатору Центральной избирательной комиссии РФ (450 рубрик, 4 уровня)
- По терминам верхнего уровня тезауруса Исследовательской службы Конгресса США (80 рубрик)
- По правовому рубрикатору Центра информационных исследований (180 рубрик, 3 уровня)
- По Классификатору правовых актов РФ (Указ Президента РФ N511 от 15 марта 2000 г., 1169 рубрик)
- По Классификатору НПП «Гарант» (3200 рубрик)
- Journal of Economic Literature Classification System (JEL), более 700 рубрик

Схема описания рубрики



Представление смысла рубрики опорными понятиями

200.020.020 ВСТРЕЧИ НА ВЫСШЕМ УРОВНЕ

{ встреча на высшем уровне $_{Y}$

OR

{

(переговоры $_{N}$)

(международные переговоры $_{Y}$)

(международные контакты $_{N}$)

(встреча $_{N}$) √

AND

(глава государства $_{L}$)

}

Расширенное представление рубрики понятиями тезауруса

200.020.020 ВСТРЕЧИ НА ВЫСШЕМ УРОВНЕ
{
(встреча на высшем уровне $_{Y}$)
(встреча в верхах, саммит, переговоры на высшем уровне)
OR
{ (переговоры N)
(международные переговоры $_{Y}$)
(межгосударственные переговоры, международный диалог, межправительственные переговоры, переговоры(м), переговоры правительственных делегаций)
(международные контакты $_{N}$)
(встреча $_{N}$) √
AND
(глава государства $_{L}$)
(высшая государственная власть, глава страны, лидер государства, правитель(м), правительница(м), руководитель государства, руководитель страны, президент государства, гарант конституции, ..., монарх, эмир, эмир Кувейта, ..., царь, ...)

}

РОМИП'2007
дорожка классификации web-страниц

- **Рубрикатор: DMOZ,**
  247 рубрик 2го уровня Top/World/Russian/*/*
- **Коллекция обучения «DMOZ»**
  - 300 000 документов с 2100 сайтов
  - Русскоязычные сайты, упоминающиеся в категориях второго уровня, на страницах которых не было явного запрещения копирования содержимого этих сайтов. Для снижения размеров коллекции до разумных пределов для каждого сайта в коллекцию включалось не более 500 страниц, полученных обходом в ширину, начиная со стартовой страницы.
  - Собрано и предоставлено компанией Рамблер в 2004 году.
- **Коллекция тестирования «BY.web»**
  - 1 500 000 документов с 19 000 сайтов
  - построена компанией Яндекс как выборка из страниц домена .by, присутствовавших в индексе поисковой системы Яндекс по состоянию на май 2007 года. С каждого известного сайта из домена .by брались все страницы на глубину 3 ссылки от стартовой.

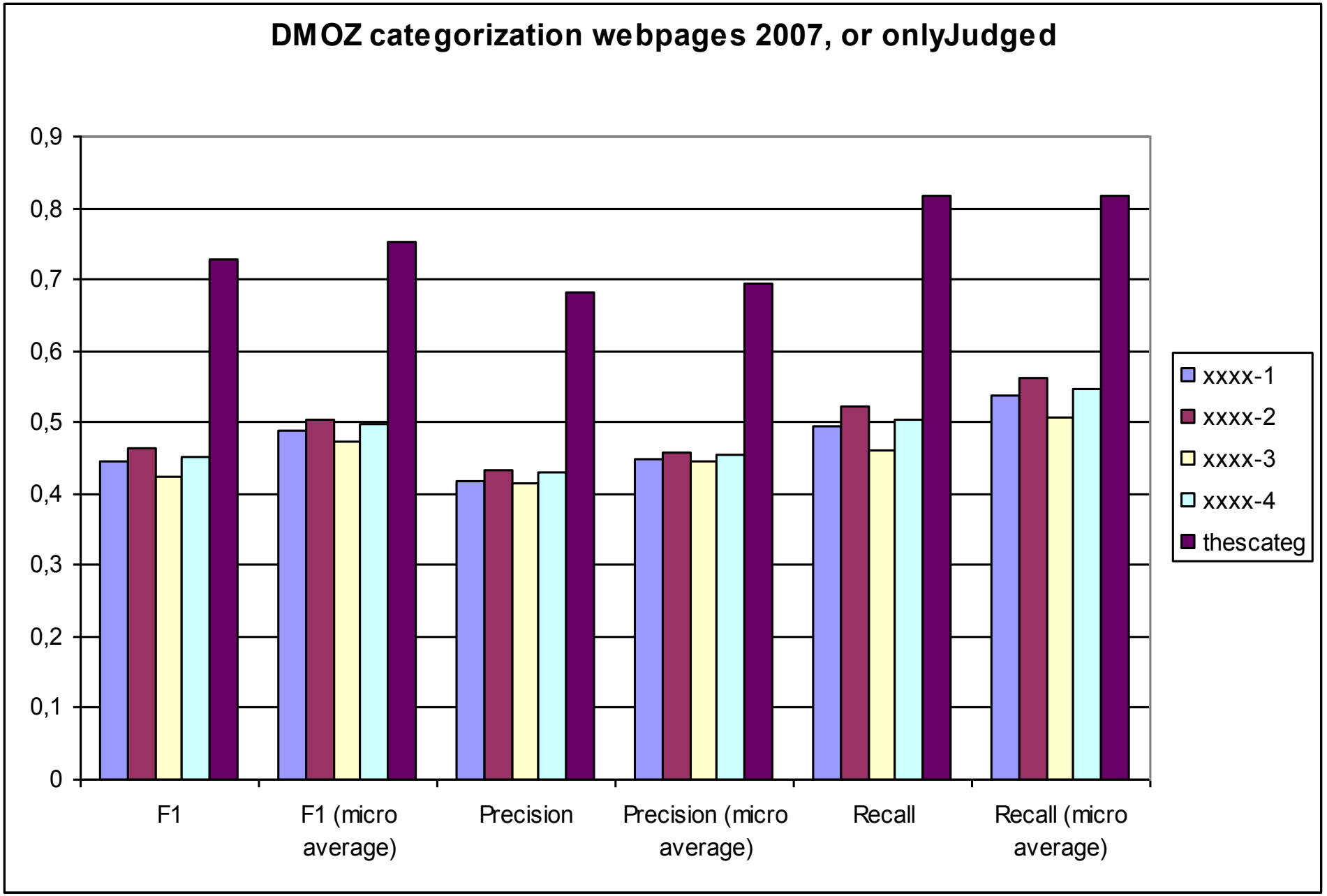
Инженерный подход (8 чел*час): пример простого описания рубрики

- Рубрика 135 «Боевые искусства» (F1-мера [OR] = 0.97, R=0.98, P= 0.96)
- 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

РОМИП2007: классификация веб-страниц



Заключение

- Методы машинного обучения применимы для небольших рубрикаторов (до 50-80 рубрик), когда несложно получить непротиворечивую размеченную коллекцию
- Инженерные методы применимы для рубрикаторов средних размеров, когда имеется возможность детального экспертного описания смысла каждой рубрики
- Для больших рубрикаторов (более 1000-2000 рубрик) из-за сложностей с подготовкой непротиворечивой коллекции для обучения – по-прежнему открытая задача, часто с необходимостью изменения организационной архитектуры

Задание

- Зайти news.yandex.ru
- Набрать 10 разных текстов из рубрики Политика
- И 10 текстов из другой рубрики, например, Спорт (или другая)

- Итого – 20 текстов в двух рубриках
- Система классификации методом Байеса
  - Можно задать текст и система будет классифицировать
  - Нужно показывать веса документа, посчитанные для каждой рубрики

Пример задачи для контрольной

- Система рубрикации должна классифицировать поток документов по двум рубрикам.
- Эксперт отнес к первой рубрике 75 документов, ко второй рубрике – 50 документов.
- Система отнесла:
- - к первой рубрике 100 документов, из них 50 правильно.
- - ко второй рубрике 40 документов, из них 30 правильно.
- Найти макро-характеристики качества классификации (точность, полноту, F-меру) - и микро-характеристики (точность, полноту, F-меру).