МГУ им. М.В.Ломоносова
Научно-исследовательский
вычислительный центр

АНО Центр
информационных
исследований

Проблемы автоматической рубрикации текстов

Лукашевич Н.В.

Louk@mail.cir.ru

План презентации

- **Методы автоматической рубрикации текстов**
- **Проблемы ручной и автоматической рубрикации**
- **Наши исследования**
  - **Лаборатория анализа информационных ресурсов НИВЦ МГУ**
  - **АНО Центр информационных исследований**
  - **УИС РОССИЯ (www.cir.ru)**

Рубрикация текстов

- **Классификация/рубрикация информации - отнесение порции информации к одной или нескольким категориям из конечного множества рубрик.**
- **Применение:**
  - **Навигация по коллекции документов**
  - **Поиск информации**
  - **Замена сложного запроса**
  - **Иерархическое упорядочение знаний предметной области**
    - **Анализ распределения документов по тематике**
  - **Фильтрация потока текстов:**
    - **Тематический сбор новостей**
    - **Персонализированная фильтация потока текстов**
    - **Фильтрация спама**
    - **Тематический сбор информации из интернет**

Примеры рубрикаторов

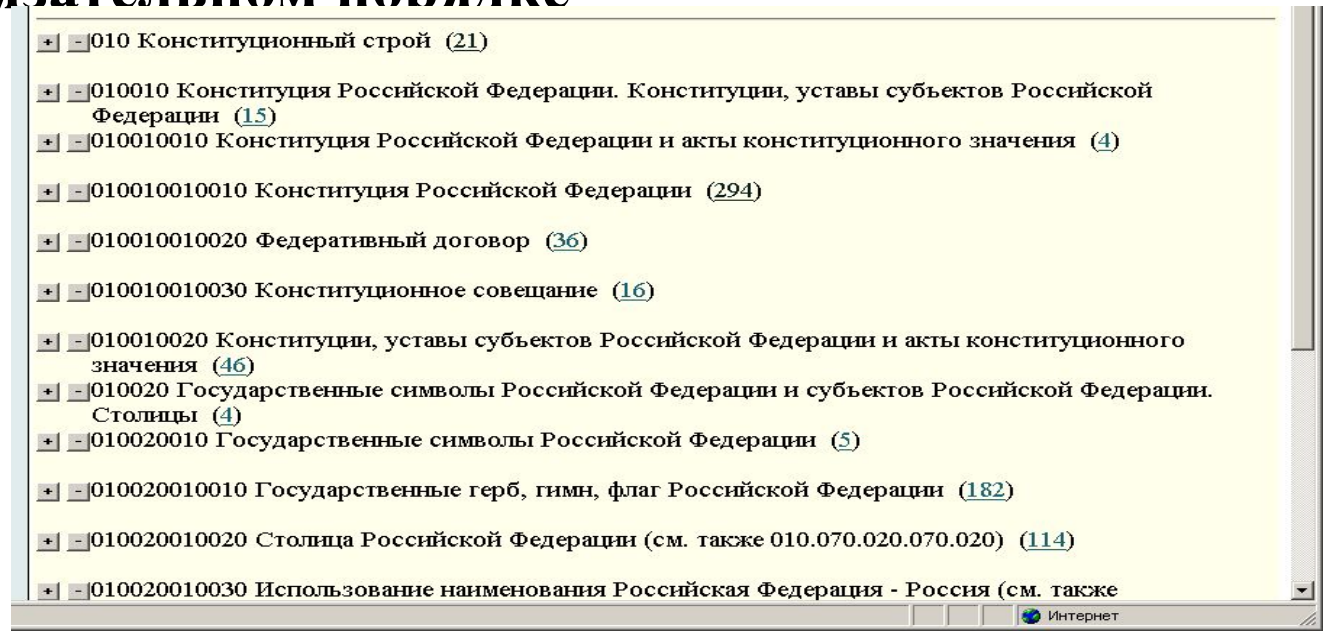
- Каталог Интернет-сайтов: Open Directory Project – dmoz.org
  - 4,830.584 sites. 75.151 editors. over 590.000 cat
  - Сет ками

Каталог Яндекс - Фасетная классификация

- Тематическая
  - Иерархический классификатор, имеет порядка 600 значений и описывает предметную область интернет-ресурса
- Регион
  - 230 географических областей. Определяется географическим расположением представляемого объекта, сферой управления и влияния, потенциальной аудиторией информации или информационным содержанием ресурса
- Жанр
  - художественная литература; научно-техническая литература; научно-популярная литература; нормативные документы; советы; публицистика
- Источник информации
  - Официальный, СМИ, Неформальный, Персональный Анонимный
- Адресат информации
  - Партнеры, Инвесторы, Потребители, Коллеги
- Сектор экономики
  - Государственный, Коммерческий, Некоммерческий

Рубрикатор нормативно-правовых актов

- **Президентский классификатор (Указ №511 15.03.2000)**
- **Иерархия рубрик - 1168 рубрик**
- **Все НПА рубрицируются экспертами в обязательном порядке**



Коллекция и рубрикатор Reuters для автоматического рубрицирования

- **Более 21 тысячи информационных сообщений из области биржевой торговли и слияния предприятий**
- **Массив разделен на две части: документы для обучения, документы для тестирования**
- **Большинство текстов имеют рубрики, проставленные людьми**
- **Основные рубрики: 135 без иерархии**
- **Примеры рубрик: Золото (товар), Свинец (товар), Кофе и др. товары, Торговля**
- **Средняя длина текста - 133 слова**
- **156 публикаций по автоматическому рубрицировнаю на сайте CiteCeer**

Методы рубрицирования текстов

- **Ручное рубрицирование**
- **Полуавтоматическое**
- **Автоматическое**
  - **Инженерный подход (=методы, основанные на знаниях, экспертные методы)**
  - **Методы машинного обучения**

Методы оценки эффективности автоматического рубрицирования

Основа: сравнение результатов автоматического и ручного рубрицирования

$$\textbf{полнота } r(u) = \frac{|u \cap C|}{|C|}, \textbf{ точность } p(u) = \frac{|u \cap C|}{|u|},$$

где C — множество документов, отнесённых к рубрике человеком,
u — множество документов, отнесённых к рубрике автоматически

$$\textbf{F-мера } F_\beta(u) = \frac{1+\beta}{\frac{\beta}{p(u)} + \frac{1}{r(u)}}, \ \beta > 0$$

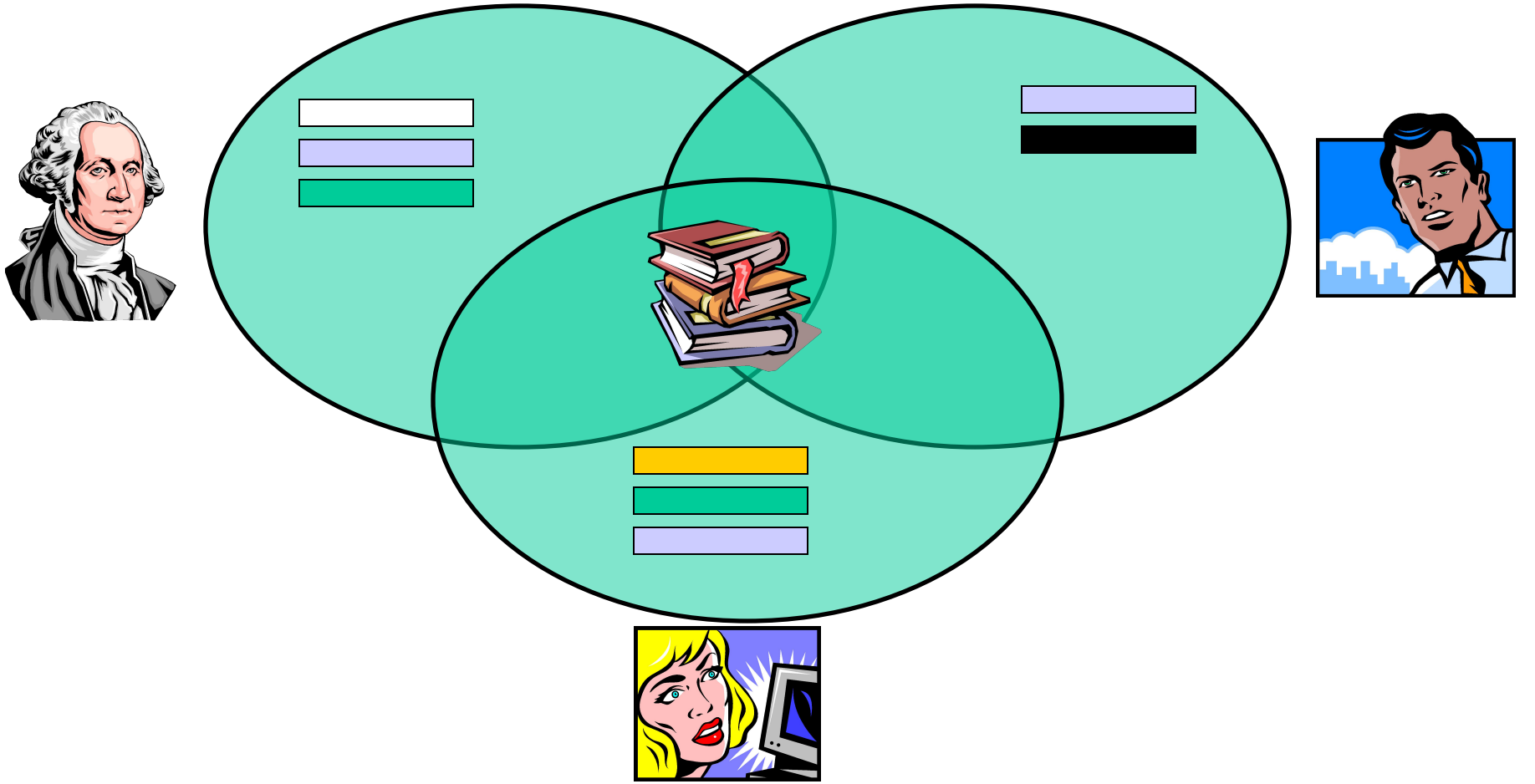
Ручное рубрицирование

- **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%



Инженерный подход

- **Основное предположение: рубрикатор создается осмысленно, содержание рубрики можно выразить ограниченным количеством понятий в виде формулы**
- **Эксперты описывают смысл рубрики в виде булевских выражений, правил продукции**
- **Construe system (Hayes)**
  - **Reuter news story**
  - **674 рубрики: 135 тематических рубрик + география…**
  - **4 человеко-года**
  - **94 % полноты и 84 % точности на 723 текстах**

Reuters: пример описания рубрики

```
if      (wheat & farm) or
        (wheat & commodity) or
        (bushels & export) or
        (wheat & tonnes) or
        (wheat & winter and (¬ soft))
then
        WHEAT
else
   (not WHEAT)
```

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 рубрике:**

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 (Родина и РПЖ), видимо, неизвестных системе рубрицирования.

Результаты "Родины" и РПЖ на последних региональных выборах

0

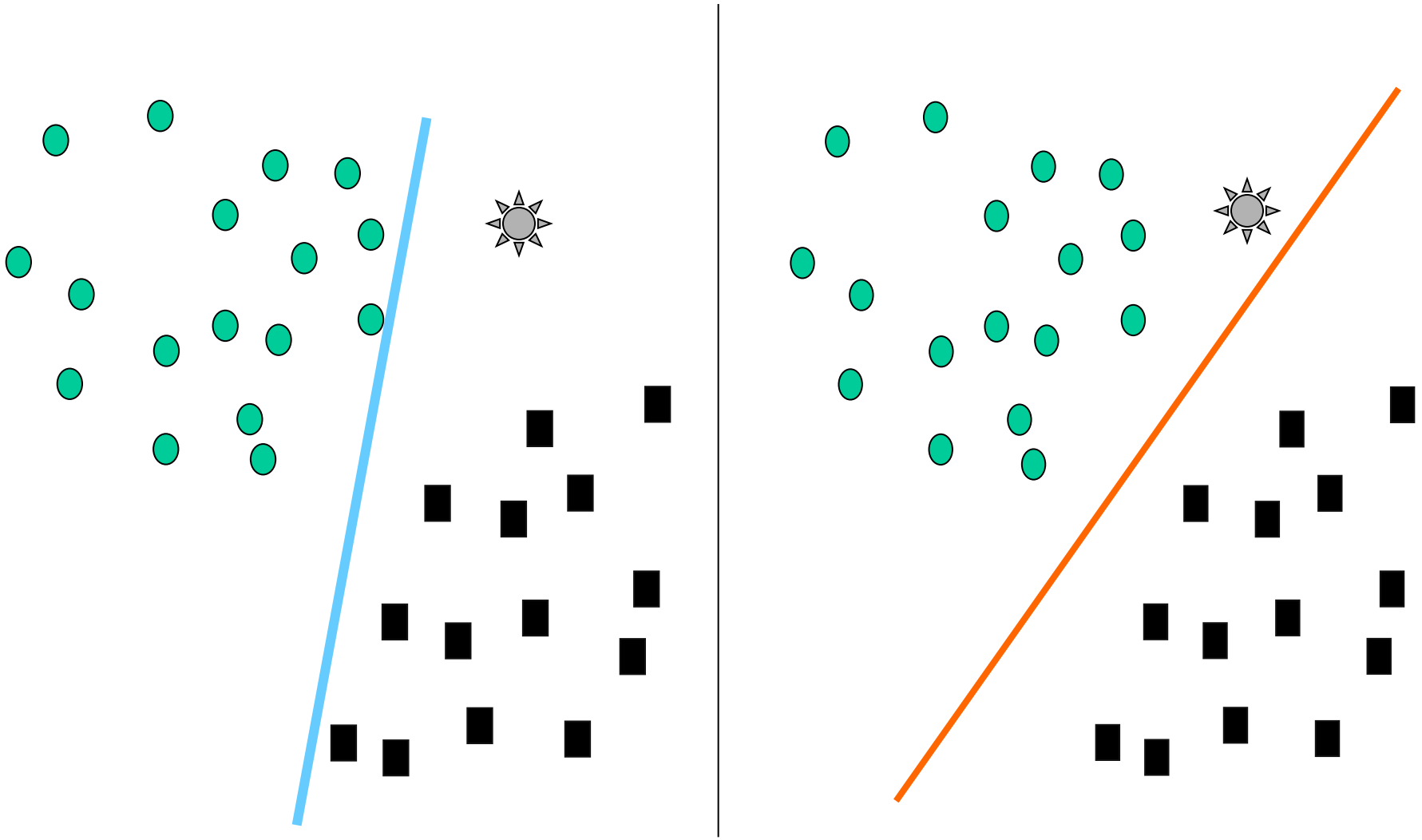
ГАЗЕТА "КОММЕРСАНТЪ" № 135(3466) ОТ 26.07.2006

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

Методы машинного обучения в задачах рубрикации

- Имеется коллекция отрубрицированных людьми текстов.=>
- Для каждой рубрики имеется множество положительных и отрицательных примеров

Положительные и отрицательные примеры: как лучше отделить



Векторная модель: основные этапы

- **Задача: преобразовать множество текстов в векторы пространства R^n**
- **Пословная модель – bag of words**
- **Удаление стоп-слов (предлоги, союзы…), которые заданы списком**
- **Приведение к нормальной морфологической форме (stemming, лемматизация – приведение к словарной форме)**
- **Определение весов слов**
- **Построение вектора слов документа**

Вычисление весов слов

- **Частота встречаемости слова в документе**
- **Количество документов коллекции, содержащих данное слово**
- **Длина документа, средняя длина документов коллекции**
- => **формула TF*IDF**
- **Расположение слова в тексте, заголовках?**

TF*IDF

- **Наиболее общепринятый способ вычисления веса терма: tf • idf**

  tf – частотность терма в документе (term frequency)

  idf – величина, обратная к количеству документов, содержащих терм (inverse document frequency)

$$tf_D(t) = freq_D(t)$$

$$idf(t) = \log(|c|/df(t))$$

Формула tf•idf [Okapi BM25 – cir.ru]

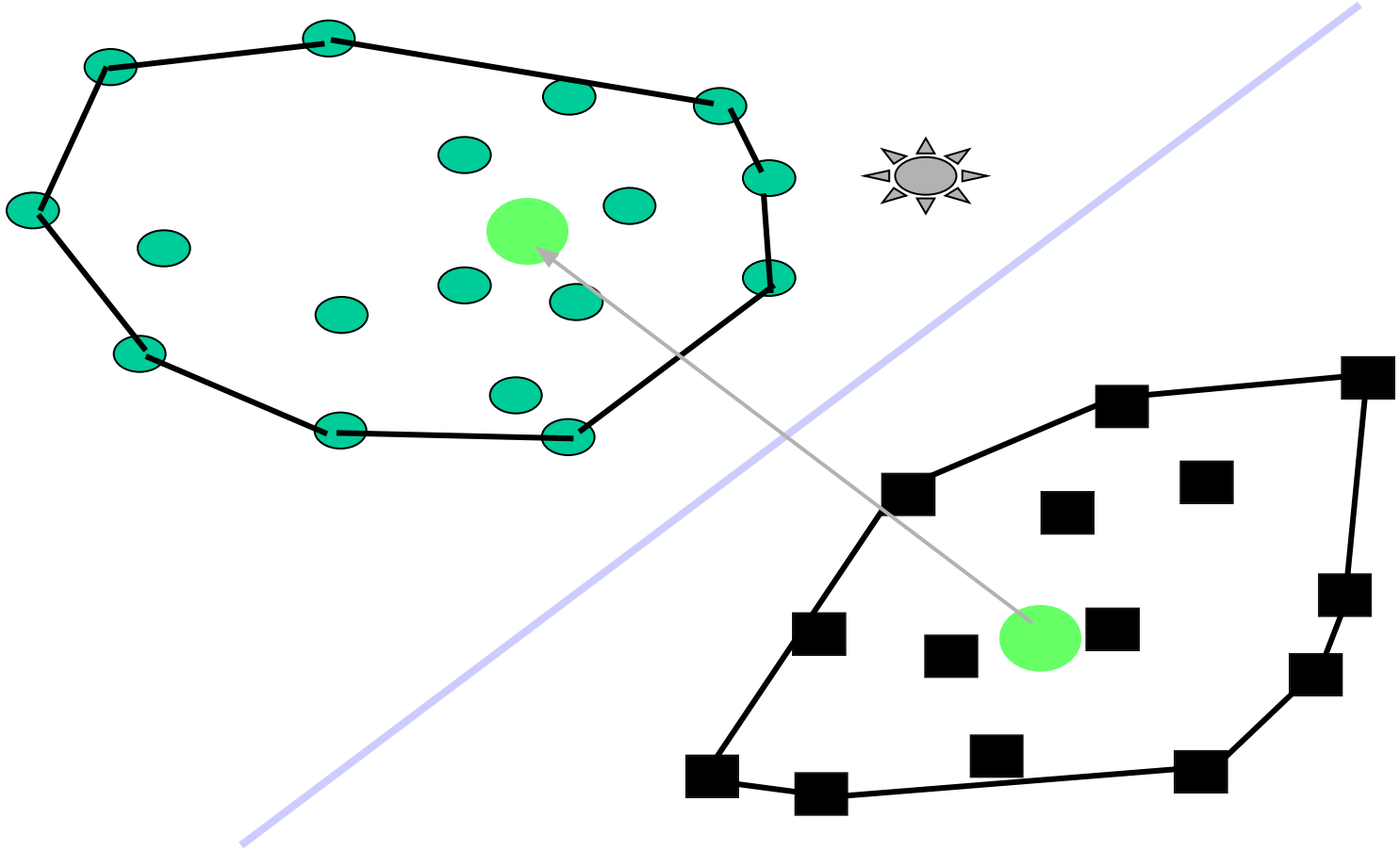
Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA, November 1994.

$$w_D(t) = 0.4 + 0.6 \cdot tf_D(t) \cdot idf(t)$$

$$tf_D(t) = \frac{freq_D(t)}{freq_D(t) + 0.5 + 1.5 * \frac{dl_D}{avg\_dl}}$$

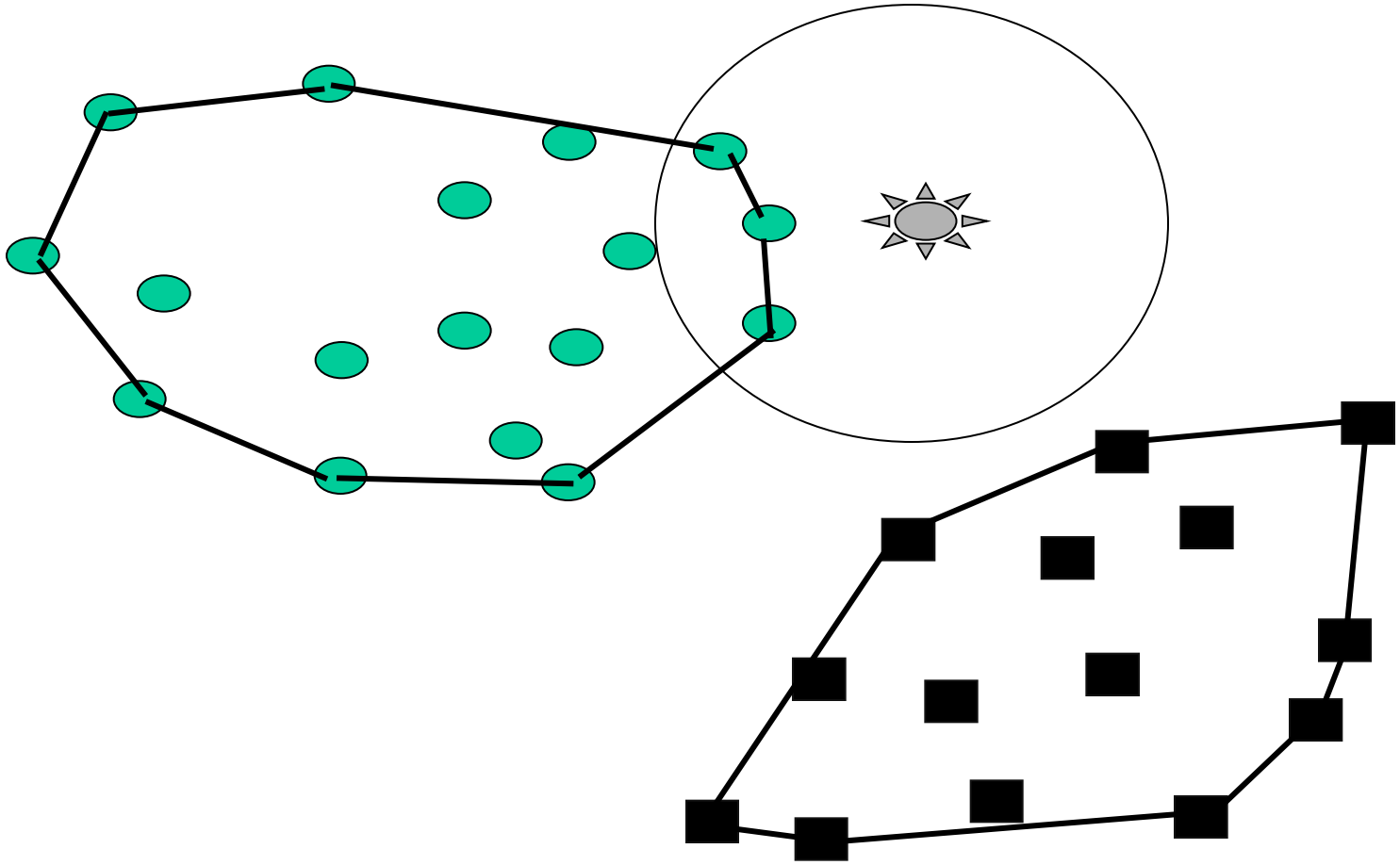
$$idf(t) = \frac{\log\left(\frac{|C| + 0.5}{df(t)}\right)}{\log(|C| + 1)}$$

Отсечение по центрам тяжести

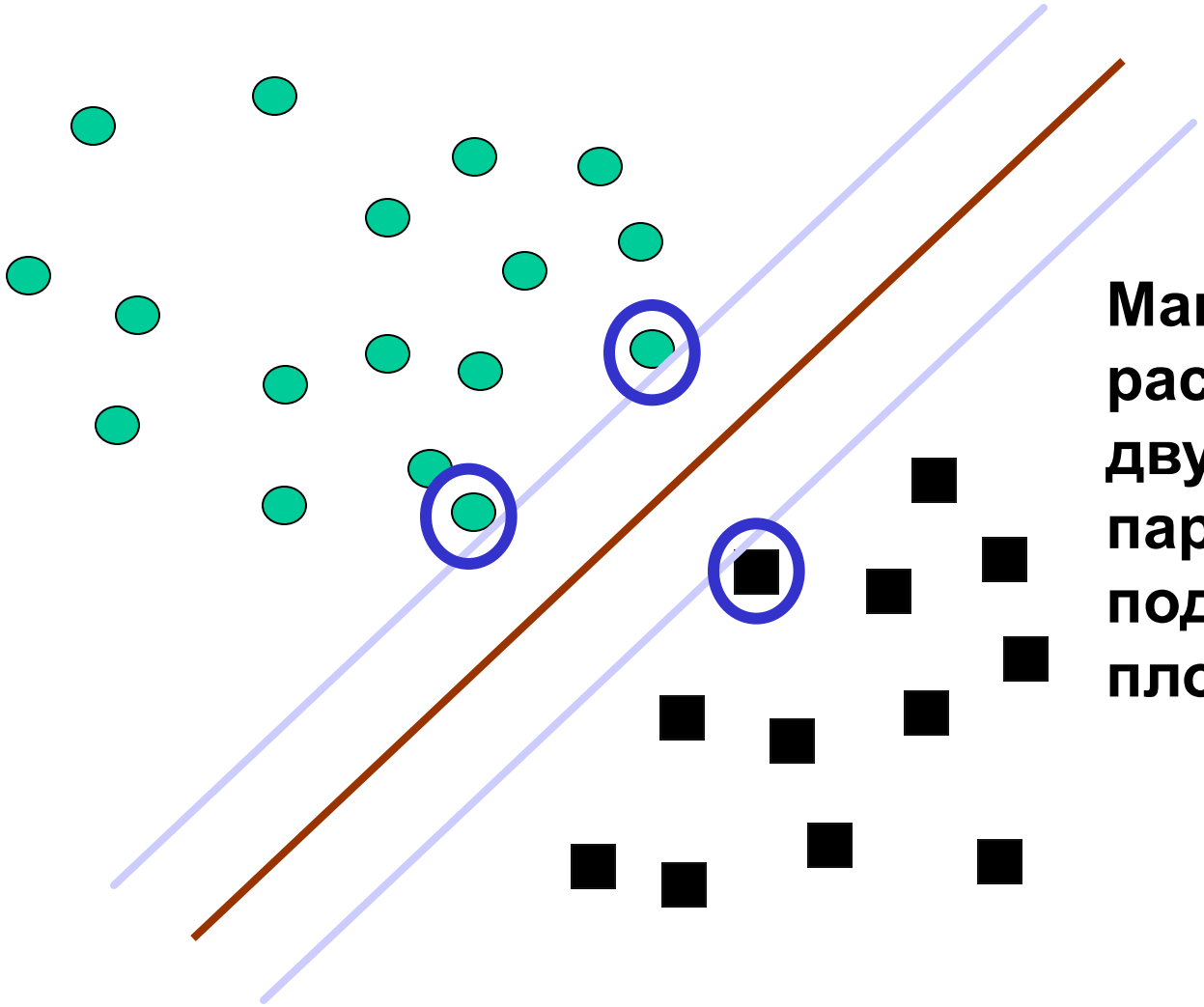


$$\text{Rocchio}_t = w^t_{filter} + 2w^t_{rel} - \frac{1}{2}w^t_{non-rel}$$

Отсечение по ближайшим соседям (kNN)



Оптимальный линейный сепаратор
SVM (Support Vector Machines)



Максимизация расстояния между двумя параллельными поддерживающими плоскостями

$$\frac{2}{\|w\|}$$

Reuters-21578, применяем SVM

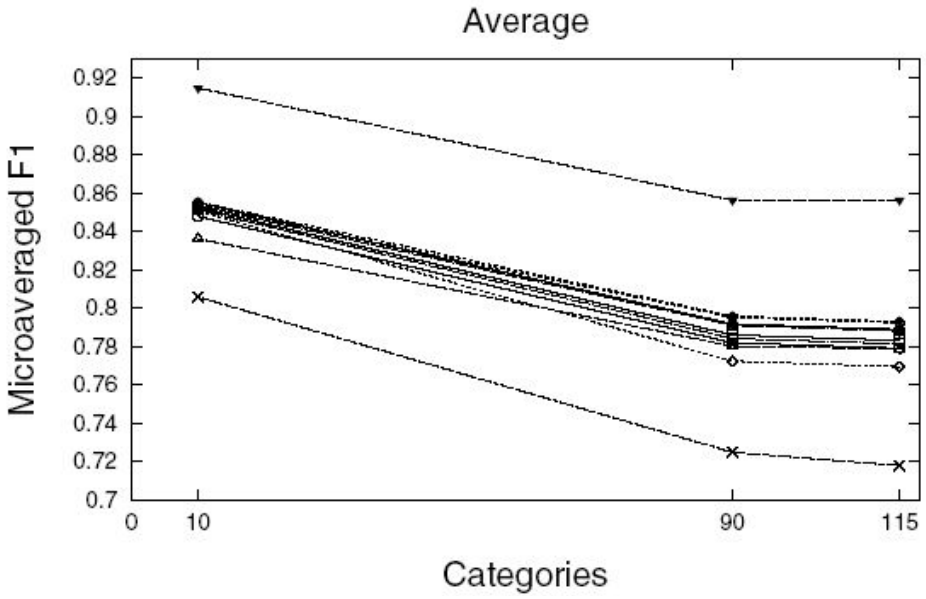
[1] *Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. // Proceedings of ECML-98, 10th European Conference on Machine Learning — 1998.*

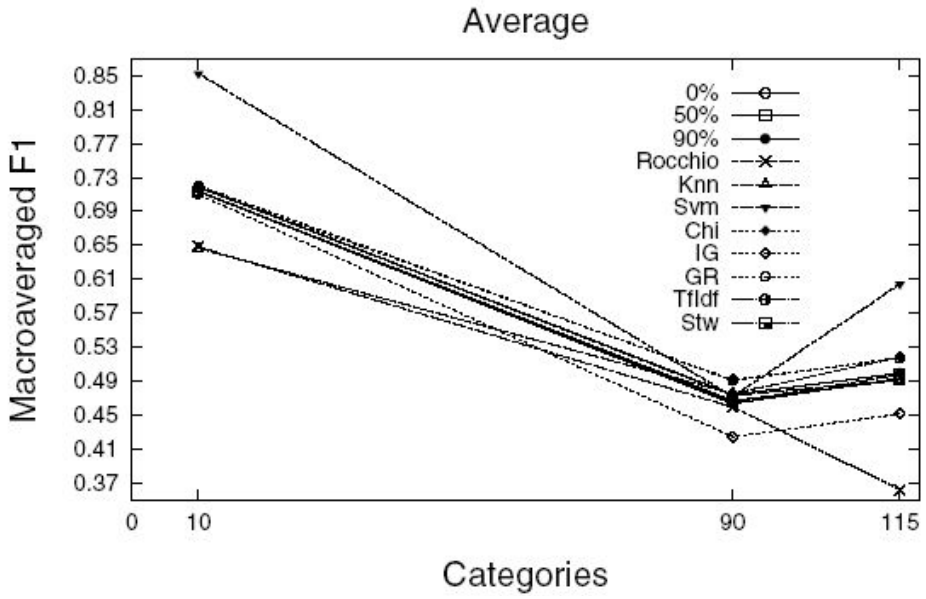
[2] *Dumais S., Platt J, Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization. // In Proc. Int. Conf. on Inform. and Knowledge Manage., 1998.*

NAME	DOC_CNT	Joachims P/R b.p.	Dumais et.al. P/R b.p.	Our SVM	disj formulae
earn	3964	98,20	98,00	97,79	90,70
acq	2369	92,60	93,60	95,69	82,01
...	2108			83,72	56,06
money-fx	717	66,90	74,50	72,83	58,54
grain	582	91,30	94,60	89,00	88,89
crude	578	86,00	88,90	82,82	69,31
trade	486	69,20	75,90	77,45	64,52
interest	478	69,80	77,70	75,57	56,59
ship	286	82,00	85,60	74,55	69,60
wheat	283	83,10	91,80	89,59	89,74
corn	237	86,00	90,30	86,31	90,32
dlr	175			69,81	51,79
money-sup	174			74,01	48,54
oilseed	171			65,96	78,57
sugar	162			88,54	85,37
coffee	139			92,72	91,80
gnp	136			83,57	75,56
veg-oil	124			77,56	70,97
gold	124			64,48	61,54
soybean	111			61,56	74,70
nat-gas	105			61,03	44,44
bop	105			69,13	53,52

An Analysis of the Relative Hardness of Reuters-21578 Subsets

Franca Debole and Fabrizio Sebastiani. //In proc. of LREC-04, 4th int. conf. on Language Resources and Evaluation, pp.971--974, Lisbon, PT, 2004.





- 90 из 135 категорий имеют хотя бы один положительный пример для обучения и для тестирования
- Лучший результат на R(90): в среднем 50% F-меры

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 (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**

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

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**

Два основных подхода к автоматическому рубрицированию

- **Методы, основанные на знаниях («инженерный» подход)**
  - **высокая эффективность**
  - **«прозрачность» получаемых результатов**
  - **трудоемкость описания рубрик**
- **Машинное обучение**
  - **эффективно при наличии качественно размеченной обучающей коллекции**
  - **низкая эффективность при большом числе рубрик**
  - **трудно интерпретируемые результаты («черный ящик»)**

Основные направления исследований по автоматической рубрикации

- **Лаборатория (ЛАИР) НИВЦ МГУ**
- **УИС РОССИЯ (www.cir.ru) - 1 млн. современных российских документов**
- **Инженерный подход – использование знаний Общественно-политического тезауруса**
- **Машинное обучение - автоматическое формирование формул**
- **Смешанные подходы**
- **Современные техники: bagging, boosting**

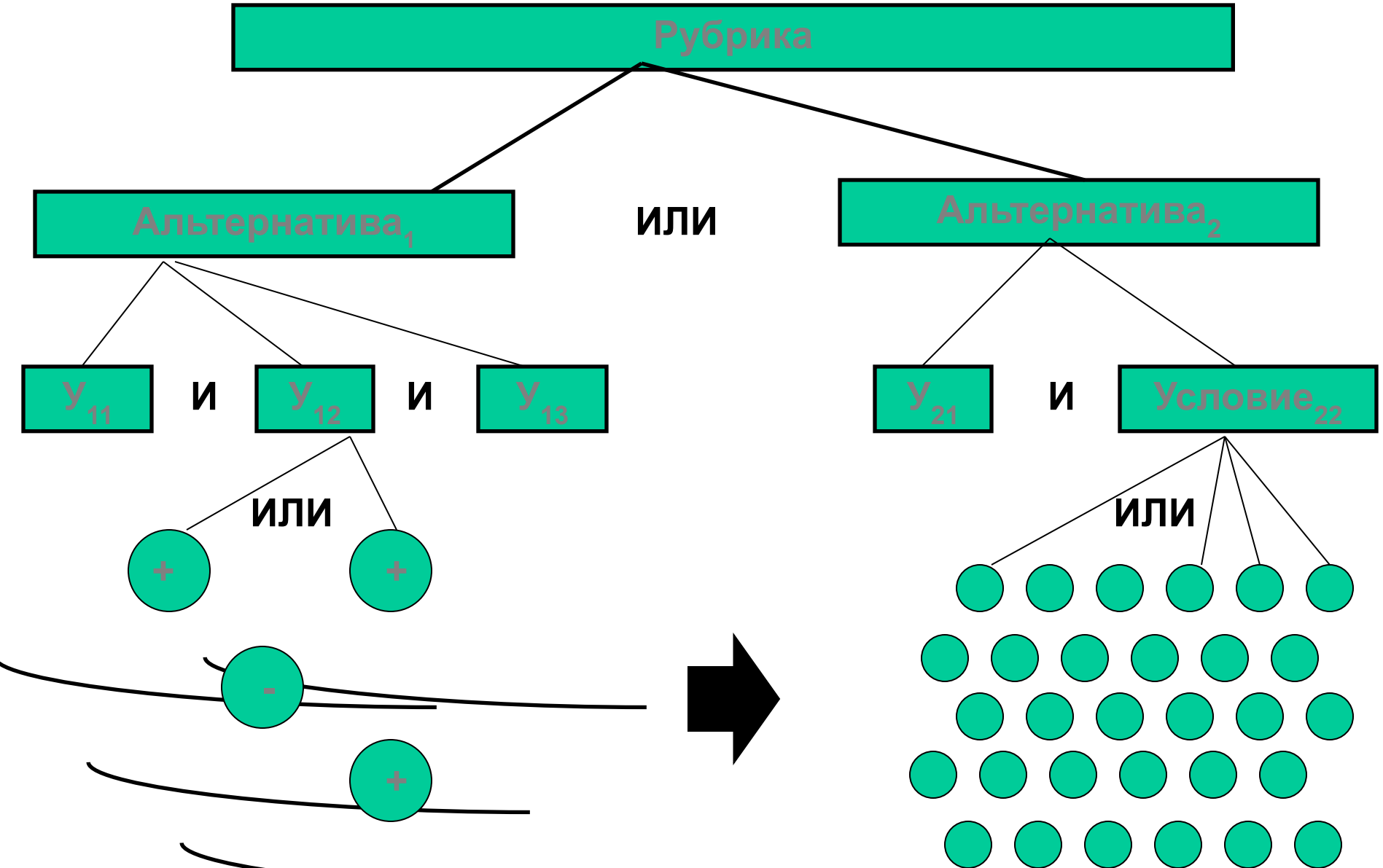
Технологии автоматической классификации на основе УИС РОССИЯ

- 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

Технология автоматического рубрицирования

- Опора на знания, описанные в Общественно-политическом 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

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, ...)

}

Метод машинного обучения, основанный на моделировании логики рубрикатора

- Алгоритм строит формулы вида:

$$U = \bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{i,j}$$

где $t_{i,j}$ — множество документов, содержащих некоторое понятие тезауруса. Конъюнкции, составляющие формулу, имеют длину J_i от 1 до 3.

- Пример (рубрика «Право международных договоров»)

```
/Термин="РАТИФИКАЦИЯ"
  OR (/Термин="ПОСТАНОВИТЬ"
      AND /Термин="СССР"
      AND /Термин="КРЕМЛЬ")
  OR /Термин="КОНСУЛЬСКАЯ КОНВЕНЦИЯ"
```

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.

Машинное обучение: метод ПФА

- **Рубрика 135 «Боевые искусства»**

```
Recall = 0.52  Precision = 0.88    FMeasure = 0.82

    [Тип = в дереве | Имя = БОЕВЫЕ ИСКУССТВА ]

Recall = 0.82  Precision = 0.98    FMeasure = 0.96

     ( [Тип = лемма | Имя = КАРАТЭ ])
OR   ( { [Тип = в тексте | Имя = ХОККЕЙНЫЙ КЛУБ ]
          OR [Тип = в дереве | Имя = ОХРАННОЕ ПРЕДПРИЯТИЕ ]}
     AND
          [Тип = в дереве | Имя = БЕДСТВИЕ ])
OR   ( { [Тип = в тексте | Имя = КУЛЬТУРА ]
          OR [Тип = в тексте | Имя = СЕВЕРО-ЗАПАДНАЯ ЧАСТЬ ]}
     AND
          [Тип = в тексте | Имя = ОДЕЖДА ]
     AND
          [Тип = в дереве | Имя = ВЕРОВАТЬ ])
OR   ( { [Тип = в тексте | Имя = МЕДИЦИНСКОЕ УЧРЕЖДЕНИЕ ]
          OR [Тип = в тексте | Имя = КРЫЛАТСКОЕ ] }
     AND [Тип = в дереве | Имя = ВОСТОЧНЫЕ ЕДИНОБОРСТВА ])
OR   ( [Тип = в тексте | Имя = МАСЛЕНИЦА ])
OR   ( [Тип = лемма | Имя = ДЗЭНИН ])
OR   ( [Тип = в тексте | Имя = САМООБОРОНА ]
     AND [Тип = в дереве | Имя = ИСТОРИЧЕСКИЕ НАУКИ ])
```

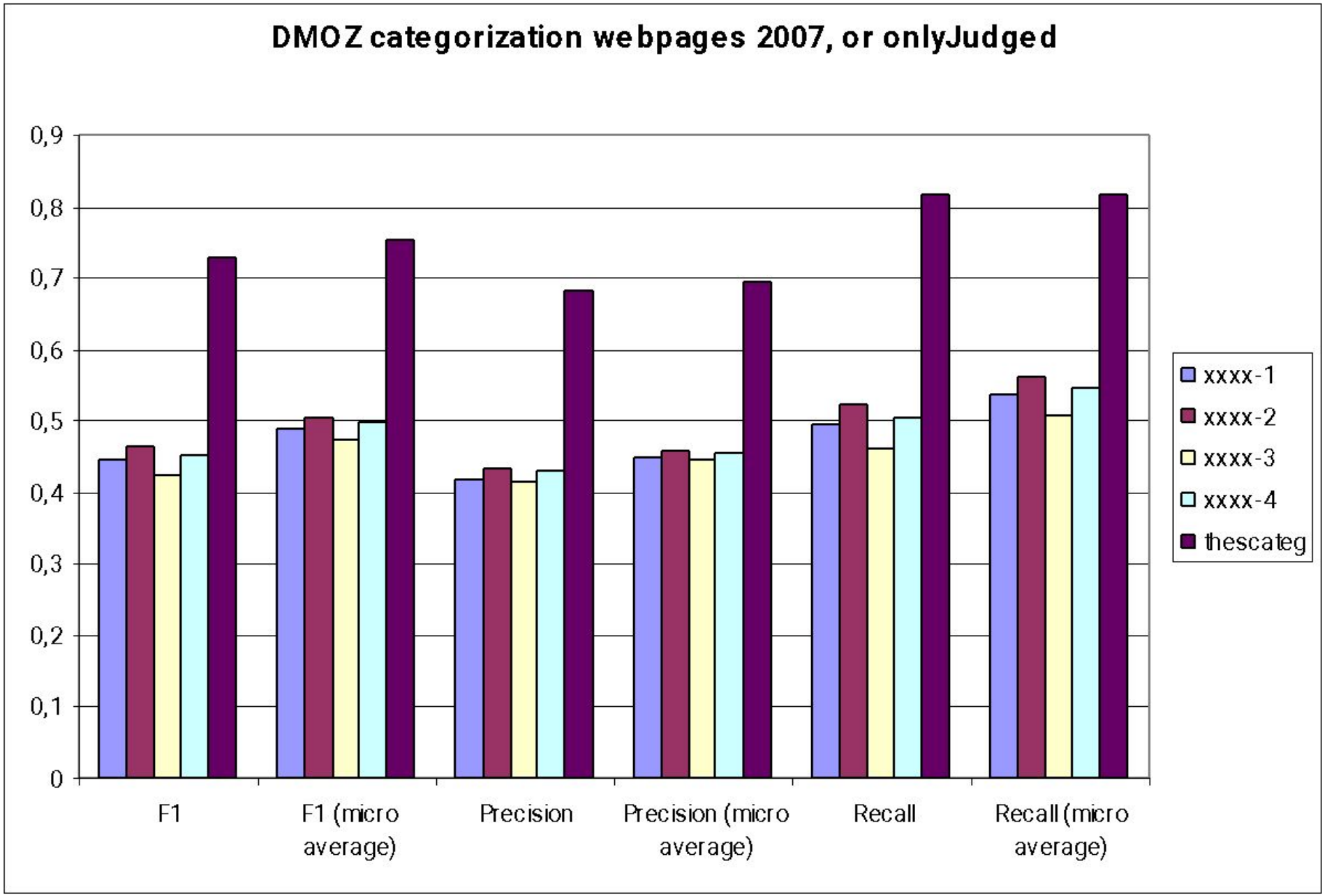
Инженерный подход (8 чел*час): пример простого описания рубрики

- 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

- **Каждый из методов классификации текстов:**
  - **Ручное рубрицирование (РР)**
  - **Машинное обучение (МО)**
  - **Методы, основанные на знаниях (МЗ)**

  имеет ограниченную область применения
- **Улучшить результаты можно при помощи комбинации различных методов**
  - **МО+РР, МЗ+РР (ручная проверка результатов алгоритма)**
  - **МО+МЗ (ПФА, полуавтоматическое описание рубрик)**
  - **МЗ+МО (тематический анализ, поиск расхождений)**
- **Использование базы знаний о связях понятий языка – Тезауруса – позволяет повысить скорость и качество описаний рубрик для автоматических методов рубрицирования**