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

АНО Центр
информационных
исследований

Университетская информационная система
РОССИЯ

УИС РОССИЯ:
поиск и классификация в РОМИП 2007

Агеев М.С., Добров Б.В., Красильников П.В.,
Лукашевич Н.В., Сидоров А.В., Штернов С.В.

Классификация веб-страниц:

машинное обучение vs. работа экспертов

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])

Машинное обучение: метод ПФА (2)

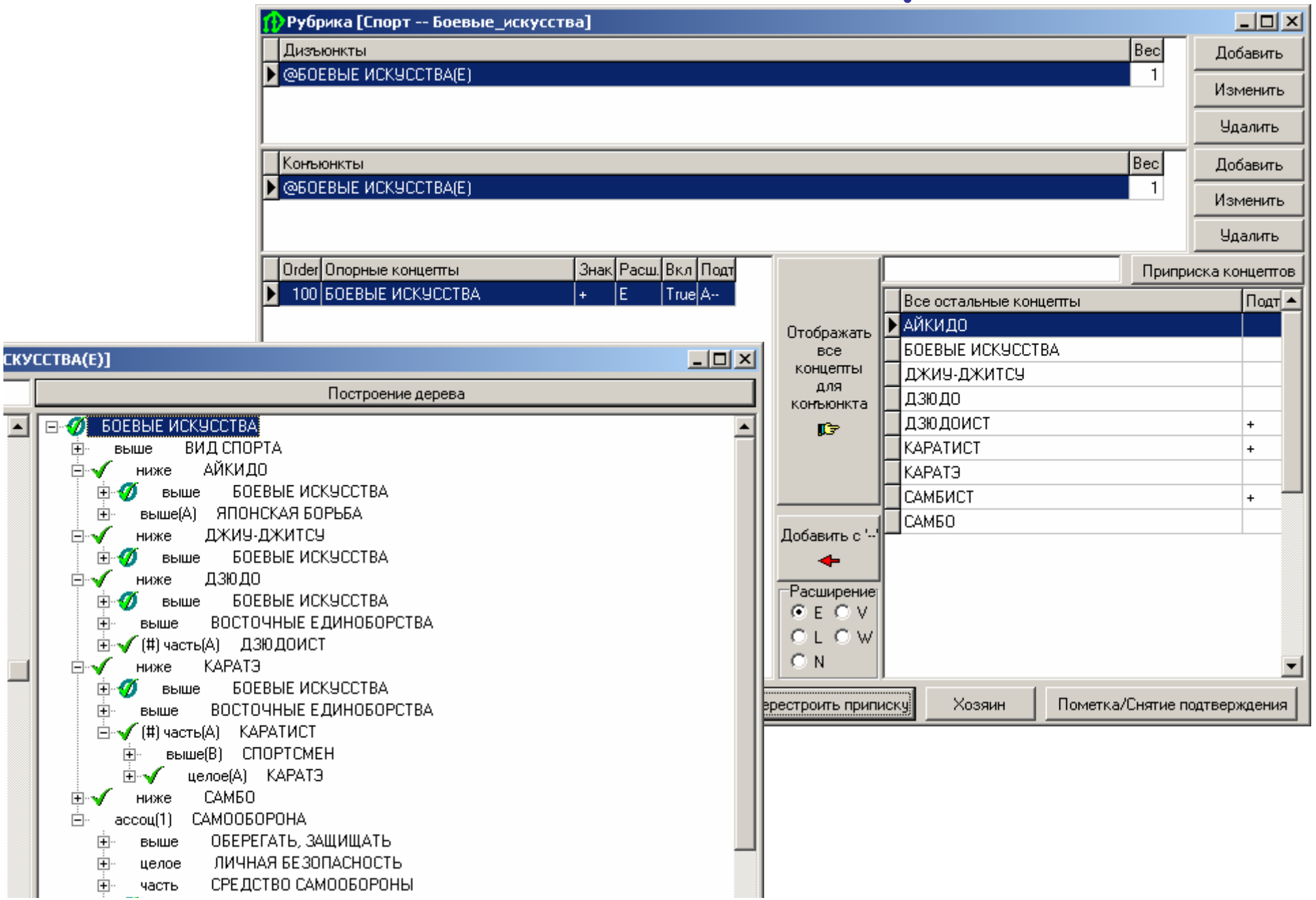
- Рубрика 14 «Страны_и_регионы ->Азия»
- Recall = 0.56 Precision = 0.95 FMeasure = 0.88

```
{ ( [Тип = в дереве | Имя = ДУНГАНЕ ])
OR ([Тип = в тексте | Имя = СРЕДСТВА МАССОВОЙ ИНФОРМАЦИИ ]
        AND [Тип = в дереве | Имя = ПОЛИТИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ ]
        AND [Тип = в дереве | Имя = СРЕДНЯЯ АЗИЯ ] )}
```

- Рубрика 82 «Досуг -> Туризм»
- Recall = 0.92 Precision = 0.99 FMeasure = 0.98

```
{       ( { [Тип = в дереве | Имя = ТУРИСТИЧЕСКИЙ СЕРВИС ]
              OR [Тип = в дереве | Имя = РЕЗЕРВИРОВАТЬ ]}
        AND [Тип = в дереве | Имя = ПАССАЖИРСКИЙ ТРАНСПОРТ ]
        AND [Тип = в дереве | Имя = АВИАЦИОННЫЕ РАБОТЫ ]
        AND [Тип = в дереве | Имя = РАЗРЕШИТЕЛЬНЫЙ ДОКУМЕНТ ])
OR
        ( { [Тип = в дереве | Имя = ГОСТИНИЦА ]
              OR [Тип = в дереве | Имя = РУМЫНИЯ ]}
        AND [Тип = в дереве | Имя = ПУТЕШЕСТВИЕ ]
        AND [Тип = в дереве | Имя = ТУРИСТИЧЕСКИЙ СЕРВИС ])
}
```

Рабочее место эксперта



Представление смысла рубрики понятиями тезауруса (8 чел.-час.)

$$R = \bigvee_i D_i \quad [234] \qquad D_i = \bigwedge_j K_{ij} \quad [265]$$

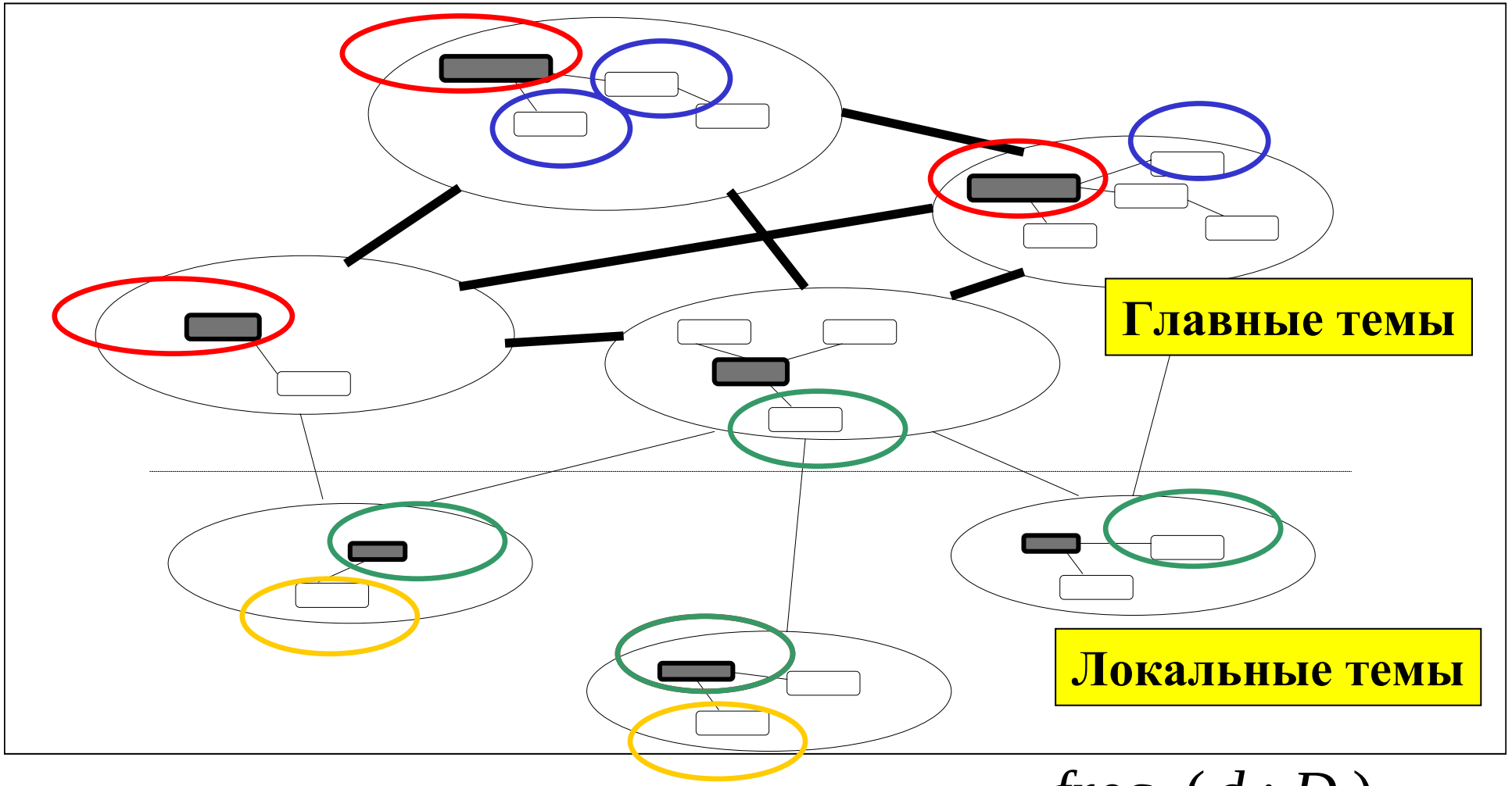
$$K_{ij} = \bigvee_m f_m(c_{ijm}) \setminus \bigvee_n f_n(e_{ijn}) = \bigvee_k d_{ijk}$$

[334] [899]

$$R = \bigvee_i D_i = \bigvee_i \left[\bigwedge_j K_{ij} \right] = \bigvee_i \left[\bigwedge_j \left(\bigvee_k d_{ijk} \right) \right]$$

[40161]

Вес понятия в тексте: тематическое представления



$$\theta(d) = \alpha \cdot \omega(d; D) + (1 - \alpha) \cdot \frac{freq\ (d; D)}{\max\limits_{c} freq\ (c; D)}$$

Расчет веса рубрики: тематическое представление и текстовая связность

$$\theta(D_i) = \frac{\sum_{j=1}^{m} \theta(K_{ij}) + \sum_{j<k} S(K_{ij}, K_{ik})}{m + C_m^2}$$

$$\theta(K_{ij}) = \min\left\{1.0; \max\left(\theta(d_{ijk}), \chi \cdot \theta(p_{ijm})\right)\right\}$$

$$S(K_{ij}, K_{ik}) = \min\{1.0; \left(\sum s(c_{ijq} \in K_{ij}, d_{ikw} \in K_{ij})\right) / \max s(c \in D, d \in D)$$

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 текст посвящен боевым искусствам**

Более сложное описание рубрики

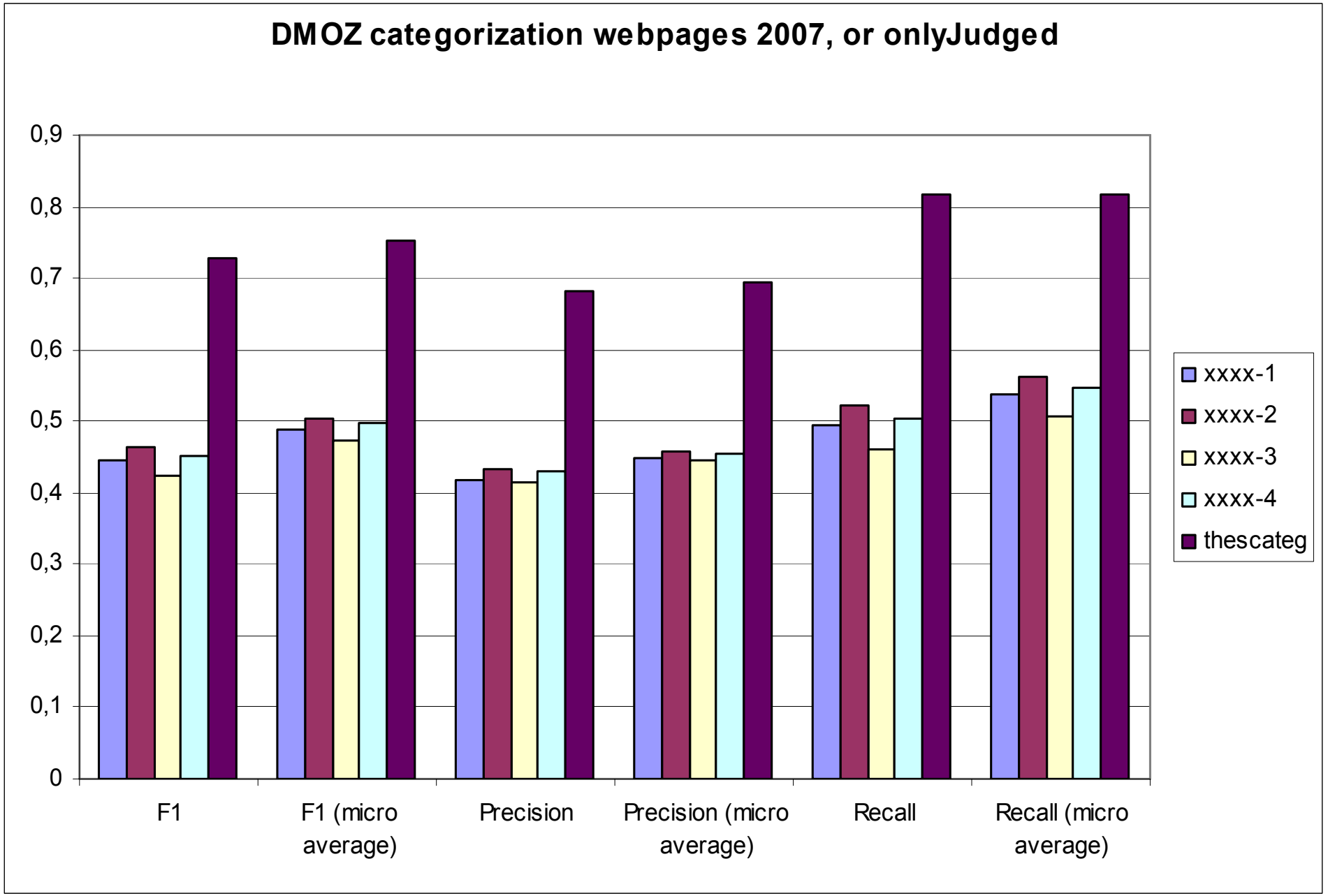
- **Рубрика 43 «Домашний ремонт»**
  F1-мера [OR]= 0.658, R=0.71, P=0.61)

(*РЕМОНТ*
OR *КАПИТАЛЬНЫЙ РЕМОНТ*
OR *ТЕКУЩИЙ РЕМОНТ*
OR *РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ* **)**

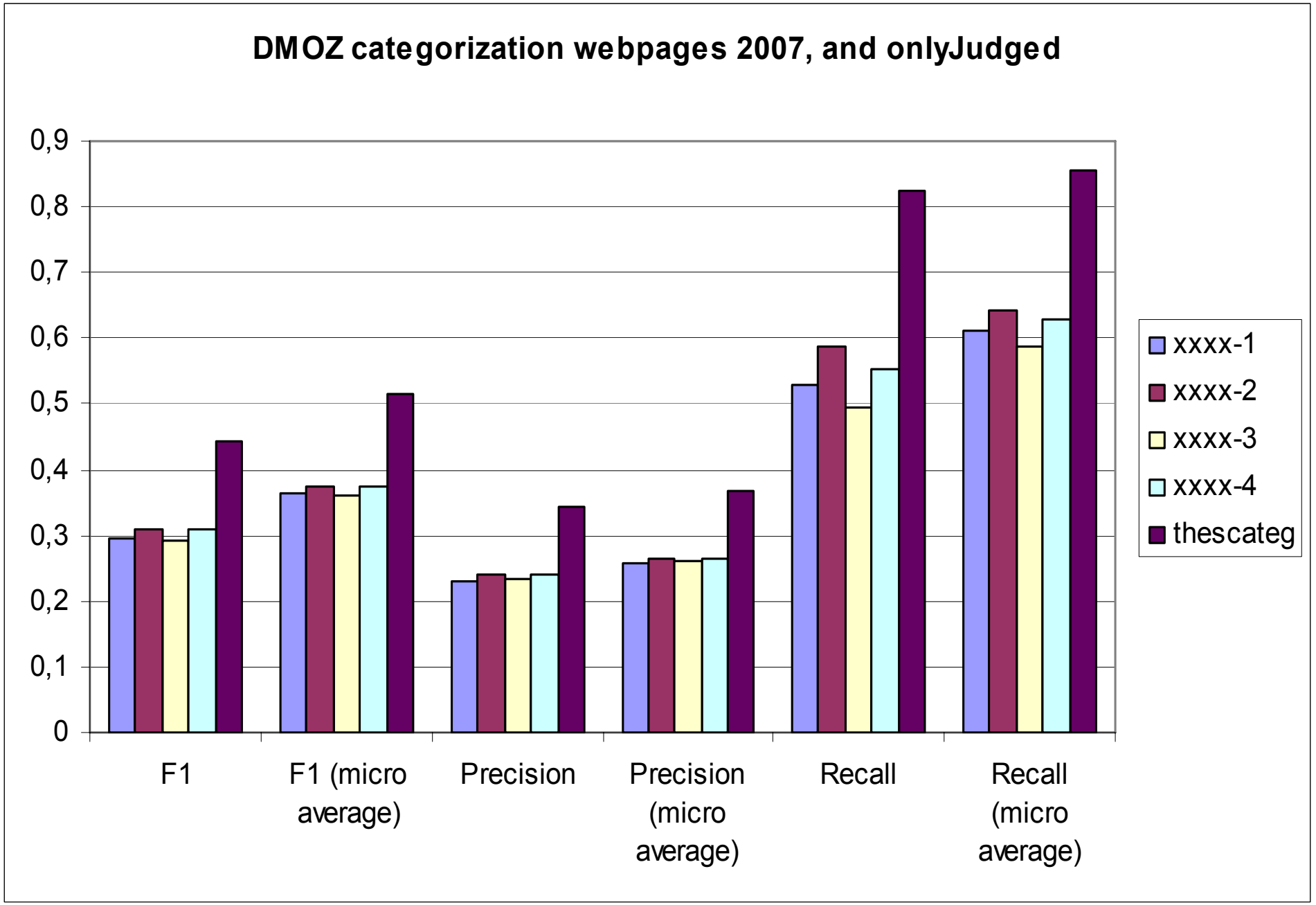
AND

(*ЖИЛОЕ ЗДАНИЕ* $_{L}$
OR *ЖИЛОЕ ПОМЕЩЕНИЕ* $_{L}$
OR *КВАРТИРА* $_{L}$ **)**

РОМИП2007: классификация веб-страниц [or]



РОМИП2007: классификация веб-страниц [and]



Проблемы при описании рубрик

- **Наиболее низкие по качеству результаты были показаны для рубрики 033 «Сад и огород»**
  мера F1 по метрике OR составила всего 0.32

- **Это связано с тем, что в состав булевского выражения рубрики были введены понятия *ДАЧНЫЙ УЧАСТОК* и *ЗАГОРОДНАЯ ДАЧА*, без дополнительных условий конъюнкции.**

  Это описание недостаточно точно, поскольку появление в тексте соответствующих слов еще не гарантирует то, что текст обсуждает проблемы сада и огорода

Выводы
по задаче web-классификации РОМИП 2007

- **Существуют задачи классификации текстов, когда нет достаточно качественной обучающей коллекции:**
  - Нет достаточного множества обучающих примеров или ручная классификация проведена недостаточно последовательно
  - В таких условиях применения методов машинного обучения очень проблематично
- **При машинном обучении системы извлекают некоторые знания о языке и мире, которые можно условно подразделить на:**
  - Общие знания о языке и мире, необходимые для работы различных приложений в разнообразном круге предметных областей, и
  - «Текущие знания», характерные именно для текущей задачи, текущей коллекции, данного типа пользователей и т.п.
  - Значимую часть знаний о современной жизни общества и современном языке деловой прозы нам удалось упорядочить в рамках понятийных структур Общественно-политического тезауруса

Выводы (2)
по задаче web-классификации РОМИП 2007

-
- **В текущем эксперименте у нас не было возможности сделать предварительный прогон, оценить и исправить ошибки и неточности описания рубрик**
  - **В обычной практике проводится несколько итераций, консультаций с экспертами**
  - **Имеются средства анализа расхождения между системой и экспертами, расхождения также описываются через понятия тезауруса**
  - **Поэтому имеются определенные возможности улучшения полученных результатов на основе тезаурусных знаний**

Классификация веб-сайтов

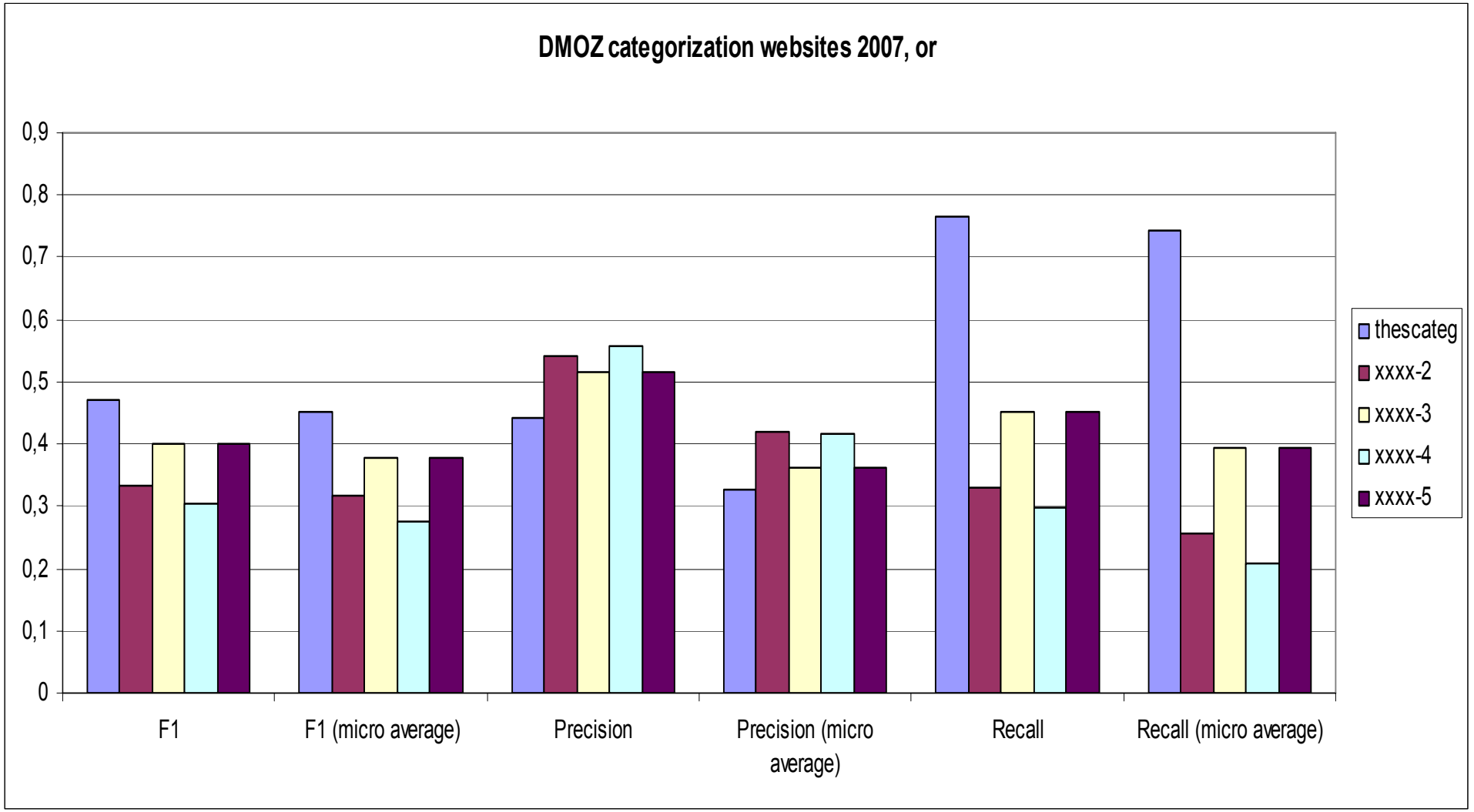
Расчет веса рубрики для сайта (только вес страниц)

$$Rank\ R_i(S_j) = Avg\ R_i(S_j) \bullet \frac{cnt80\ R_i(S_j)+1}{cnt\ R_i(S_j)+1} \bullet$$

$$\bullet \min\left\{\frac{cnt\ R_i(S_j)}{N_S}, 1.0\right\} \bullet \min\left\{\frac{cnt\ R_i(S_j)}{cnt\ S_j \bullet {}^{PS}/_{100}}, 1.0\right\}$$

Avg $R_i(S_j)$	**– средний вес страниц, отнесенных к рубрике**
cnt $R_i(S_j)$	**– кол-во страниц сайта, отнесенных к рубрике**
cnt80 $R_i(S_j)$	**– кол-во страниц сайта с весом 80+, отнесенных к рубрике**
cnt S_j	**– общее кол-во страниц сайта**
PS	**– мин.процент страниц, посвященных рубрике**
NS	**– мин. кол-во страниц, посвященных рубрике**

РОМИП2007: классификация веб-сайтов [or]



РОМИП2007: классификация веб-сайтов [and]