
Яндекс

КС классификатор и дорожка классификации сайтов РОМИП'2010

Маслов М. Ю., Пяллинг А.А.
Яндекс

1. КС классификатор: основные принципы
2. Отбор признаков: перекрестная проверка
3. КС классификатор: настройка на данных РОМИП, результаты РОМИП'2010
4. Свойства метрик дорожки классификации сайтов РОМИП
5. Как улучшить метрики?

Я

КС классификатор: веса терминов

$$W_i^j(L) = \ln\left(\frac{P_{ij}(L)}{\overline{P_{ij}(L)}}\right)$$ - вес W слова i для рубрики j в документе длиной L

$$P_{ij}(L) = 1 - (1 - P_{ij})^L$$ - вероятность слова в документе

$$P_{ij} = \frac{N_{ij}}{N^j}$$ - вероятность слова

Для классификации отбирается

R_j^g положительных

R_j^b отрицательных

признаков

Слово	Вес	
airsoftgun	14.07	R_j^g
paintball	12.11	
x-ball	11.67	
…	0	
…	0	
издательство	-4.45	R_j^b
беременности	-4.46	
кредит	-4.48	
гост	-4.49	

Я

КС классификатор: классификация документов и сайтов

Близость документа к рубрике:

$$F_j = \frac{\sum_i W_{ij} \cdot N_i}{\sum_i N_i}$$

$F_j > R_j^{\min}$ Документ в рубрике N^+

$F_j < R_j^{\max}$ Документ не в рубрике N^-

Сайт относится к теме, если $N^+ > N^-$

R_j^g R_j^b $R_j^{\min}$ $R_j^{\max}$

Последовательно подбирались для максимизации F меры классификатора сайтов

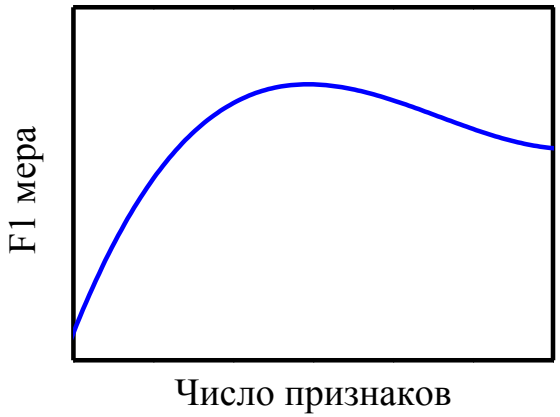
Я

Отбор признаков: проблема шума в web

Качество классификации зависит от числа признаков

Одна из причин - в web много шумящих данных

1. На сайте слово много раз написано с ошибкой
2. Ник пользователя
3. Многократное (тысячи раз) цитирование фразы



Случайные признаки оказываются значимыми для рубрик

Задача – автоматически удалить случайные для темы признаки.

Я

Отбор признаков: правила фильтрации

Сайты делятся на обучающие (train) и проверочные (valid).

Проверяется согласованность статистики слов

Слово **игнорируется,** если

Вероятность встретить слово в рубрике на train и valid сильно отличается

$$\frac{\left|\mathrm{P}_{ij}^{tr} - P_{ij}^{vld}\right|}{\sqrt{P_{ij}\left(1-P_{ij}\right)\left(\frac{1}{N^{j\,vld}} + \frac{1}{N^{j\,tr}}\right)}} > L_P$$

Вес слова для рубрики на train и valid сильно отличается

$$\frac{\left|\mathrm{W}_{ij}^{tr} - \mathrm{W}_{ij}^{vld}\right|}{\min\left(\left|\mathrm{W}_{ij}^{tr}\right|, \left|\mathrm{W}_{ij}^{vld}\right|\right)} > L_W$$

Веса разных знаков

$$\mathrm{W}_{ij}^{tr} \cdot \mathrm{W}_{ij}^{vld} < 0$$

Слово редкое в теме с положительным весом

$$\mathrm{W}_{ij} > 0 \quad \& \quad \left(N_i < N^j \cdot L_N\right)$$

Слова редкое и его нет в рубрике

$$N_i^j < 2 \quad \& \quad \frac{N^j}{N^j} N_i < 2$$

Я

Обучение классификаторов

Обучение проводилось на обучающем, проверочном (ОП) подмножествах.

Тестирование на тестовом (Т)

Сделаны разбиения – random, balanced и full.

1. Random: хосты относятся к ОПТ случайно.
2. Balanced: хосты каждой рубрики последовательно относим к ОПТ.
3. Full: нет теста. Хосты каждой рубрики последовательно относим к ОП.

Разбиение	Самопроверка		Полное множество	
	MicroF1	MacroF1	MicroF1	MacroF1
Random	22.3	12.8	н/д	н/д
Balanced	30	18.5	53.6*	47.8*
Full	н/д	н/д	64.3*	56.3*
Dmoz	62.7	35.3	38.0	32.1

Построен классификатор по данным Dmoz

Я

Таблицы релевантности РОМИП'07-09: тестирование

Сравнение F1 меры (AND)

Результаты Full лучше чем Balanced на 5% за счет увеличения обучающего множества на 30%

Год	Тип	bestRomip	Balanced	Full	Dmoz
2007	Macro	32	35.1	41.2	***53.5***
	Micro	28	37.4	41.8	***59.2***
2008	Macro	***38***	27.5	31.9	20.6
	Micro	38	31.5	***39.9***	21.3
2009	Macro	***39***	33.1	36.8	17.8
	Micro	***51***	39.9	45.2	21.1

Сравнение полноты и точности

Точность гораздо выше полноты **Обучающее множество РОМИП и дооценка сильно отличаются**

		Balanced		Dmoz		full
		Тест	РОМИП 2007_OR	Тест	РОМИП 2007_OR	РОМИП 2007_OR
Micro	Precision	36.3	***87.7***	55.9	***85.5***	***79.2***
	Recall	25.5	16.4	71.3	39.2	24.4
Macro	Precision	21.3	***71.2***	37.8	***81.1***	***89.0***
	Recall	16.3	18.5	33.1	36.9	25.7

Классификатор не оптимальный

Я

Таблицы релевантности РОМИП'07-09: повышение полноты

full3, dmoz3 – увеличили в 3 раза порог R_j^g

dmoz5 – увеличили в 5 раз порог

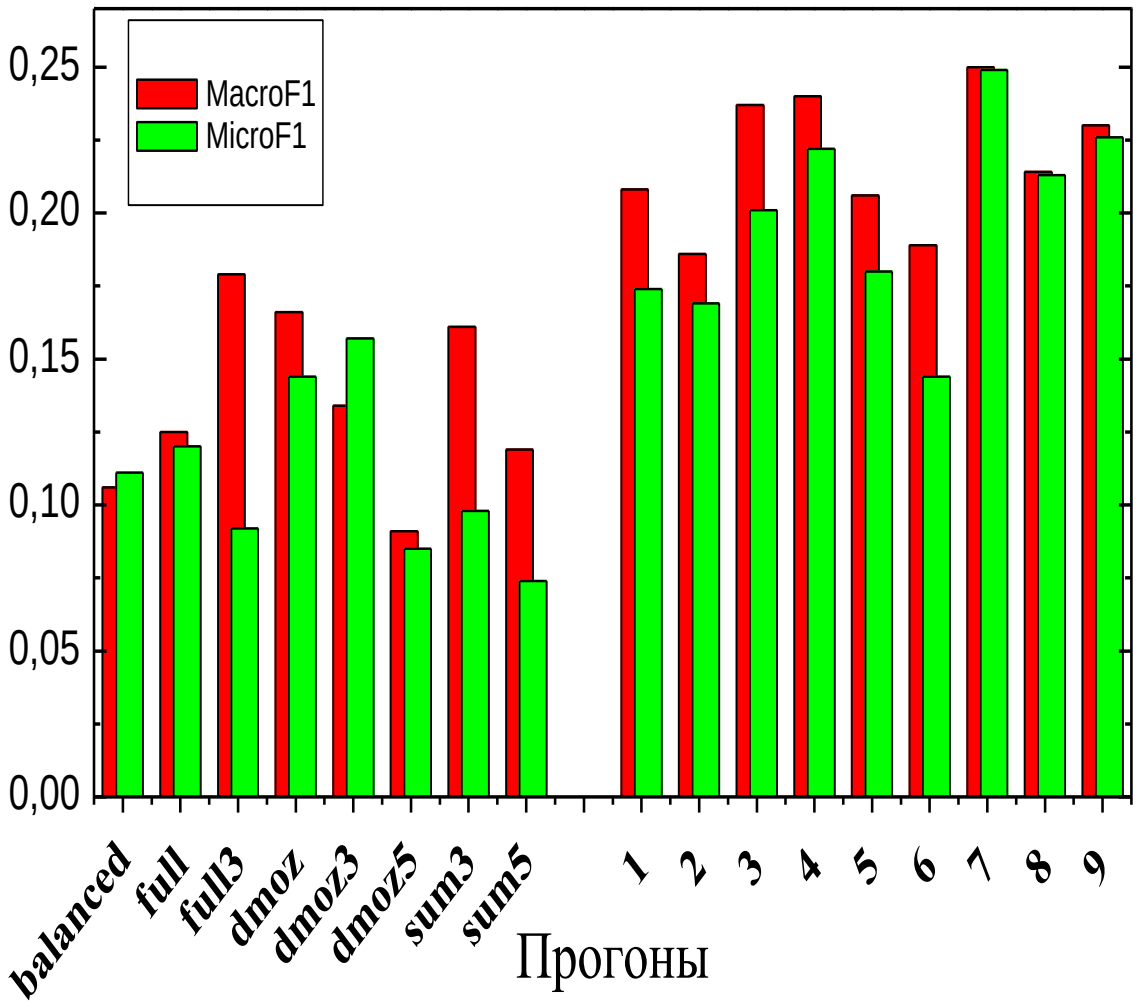
sum3(sum5) – объединили full3 и dmoz3(5)

Сравнение F1 меры (AND)

Год	Тип	Best	full3	dmoz3	sum3	sum5	sum3*	sum5*
2007	Macro	32	66.4	63.1	***72.7***	69.2	73.3	***73.6***
	Micro	28	51.3	***54.5***	51.5	48	51.9	48.8
2008	Macro	38	42.6	28.3	51.1	***52.5***	52.6	***54.2***
	Micro	38	48.8	27.2	***49.9***	48.7	***50.6***	48.8
2009	Macro	39	52	33	58.7	***59***	60.9	***66.4***
	Micro	51	61.6	41.3	***63.5***	63.1	63.3	62.8

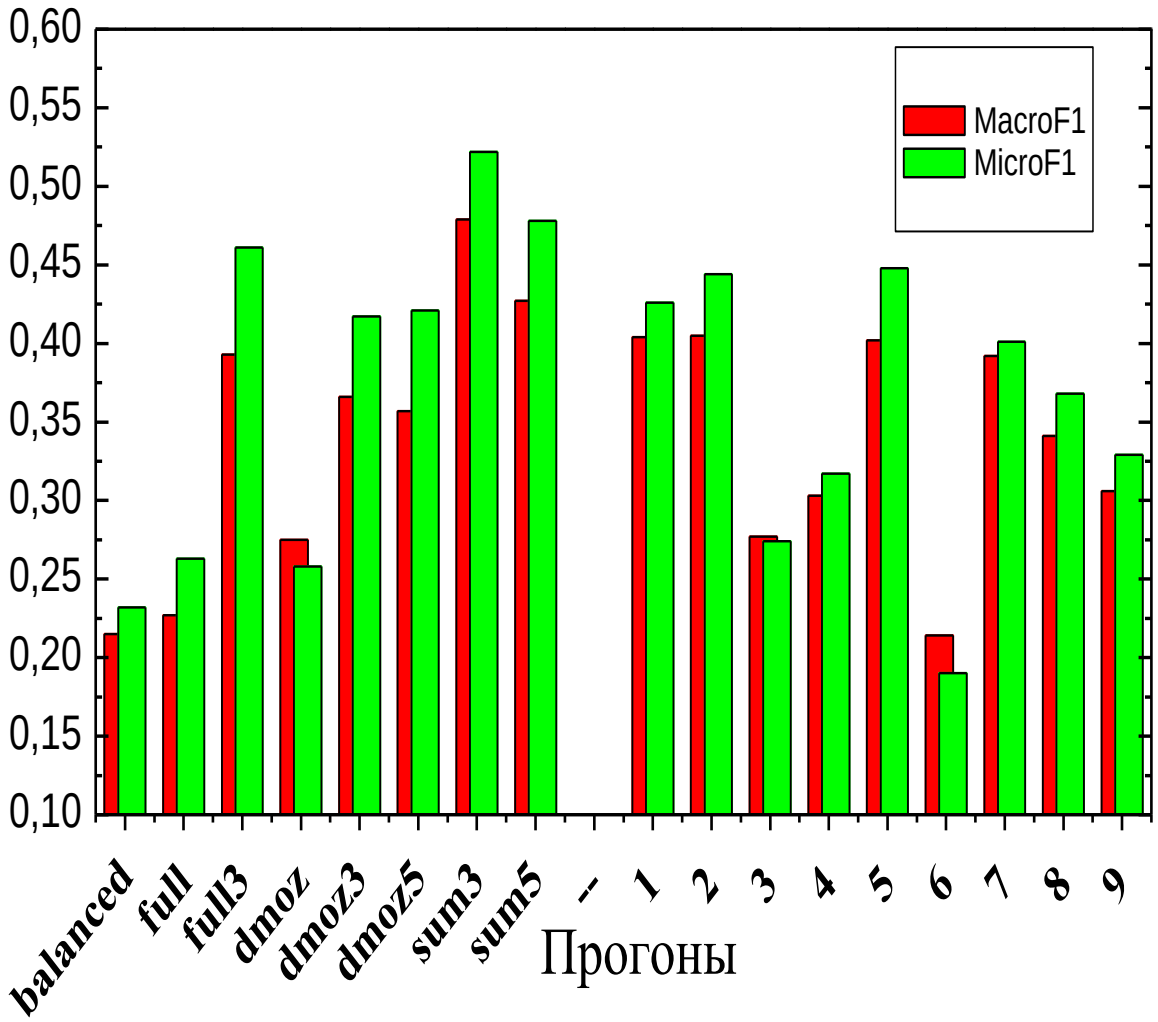
Я

Результаты РОМИП’10 – AND full



Я

Результаты РОМИП’10 – AND judged-only



Я

Хорошее свойство выборок РОМИП

Усредненные показатели сайтов в выборках

	ТИЦ	Число страниц	Показов на СЕРПе
Я.Каталог	160	19 тыс.	150 тыс.
Вне Я.Каталога	11	250	2.5 тыс.
РОМИП-обучение	250	27 тыс.	229 тыс.
РОМИП-тест	10	1400	5.1 тыс.

- Сайты в **обучающей** выборке РОМИП похожи на сайты, **включаемые** в веб-каталоги
- Сайты в тестовой выборке РОМИП похожи на сайты, **не включаемые** в веб-каталоги

Я

Расхождения РОМИП-тест и Dmoz: статистика

- 122 сайтов из оценок РОМИП в 2007-2009 гг. описано в Dmoz
- Из них 34 сайта – вне рубрики Dmoz Европа –> Беларусь
- Для 12 сайтов оценки совпадают, для 22 сайтов оценки различны, т.е. согласие **~35%**

Я

Расхождения РОМИП-тест и Dmoz: анализ

1. Отнесение к ближайшей из увиденных рубрик (9)
   - http://www.billiard.by

   Спорт –> Баскетбол и Спорт –> Новости_и_СМИ вместо Спорт –> Бильярд
   - http://www.warmuseum.by

   Досуг –> Коллекционирование вместо Справочники –> Музеи
2. Классификация коммерческих/некоммерческих сайтов (8)
   - http://www.francemobile.by – клуб любителей французских авто

   Бизнес –>Автомобили вместо Досуг –>Автомобилизм
   - http://www.bms.by – сайт НТЦ "Белмикросистемы"

   Наука –> Технологии вместо Бизнес –> Электроника_и_электротехника
3. «Почти попадание» (4)
   - http://www.otk.by – советы по выбору и использованию товаров

   Домашнее –> Домашнее_хозяйство вместо Домашнее –> Инфо_для_потребителя

Я

Предложения по процедуре оценки

1. Увеличить learn за счет сайтов из подрубрик (в 20 раз)
2. Поменять процедуру оценки
   - Взять весь рубрикатор Dmoz на 2м уровне – 200 рубрик
   - Соблюдать принцип «Один сайт – одна рубрика»
   - Использовать описания рубрик, опубликованные на Dmoz.org
   - В инструменте оценки желателен поиск
     - http://www.swadba.by: Общество –> Отношения –> Свадьба
     - http://www.warmuseum.by: Справочники –> Музеи
   - Поиск – как на Dmoz.org, только с морфологией
   - Метод общего котла не нужен
     - участники дорожки классифицируют всю коллекцию
     - нет ранжирования
     - пропадают сжатые сроки по оценке => можно повысить качество

Я

Вопросы?

Я