RCO на РОМИП 2006

Поляков П.Ю., Плешко В.В.
ООО «Гарант-Парк-Интернет»
rco@metric.ru

Классификация правовых документов

- Метод опорных векторов
- Отбор терминов
  - Однословные / Однословные + многословные
- Веса терминов
  - Бинарные / Частотные / TF*IDF
- Тип ядра
  - Линейное / Гауссово / Полиномиальное

Отбор терминов

- Многословные термины
  - эксплицирование элементов смысла (Ермаков А.Е.)
- Фильтрация
  - Документная частота
  - Информационная значимость

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_i, \bar{t}_i\}} P(t,c) \cdot \log\left(\frac{P(t,c)}{P(t)P(c)}\right)$$

Взвешивание терминов

- Бинарное (0/1)
- Частотное
- TF*IDF

- $||d||=1$

Типы ядра

- SVMLight
- Линейное

$$K(x, y) = \vec{x} \cdot \vec{y}$$

- Полиномиальное

$0.001 \le s \le 10 \qquad 2 \le d \le 5 \qquad 0 \le c \le 10$

$$K(x, y) = (s\vec{x} \cdot \vec{y} + c)^d$$

$s = 10 \quad d = 3 \quad c = 10$

- Гауссово

$0.005 \le g \le 1$

$$K(x, y) = \exp(-g\left\|\vec{x} - \vec{y}\right\|^2)$$

$g = 0.2$

Прогоны

линейное ядро			полиномиальное			гаусово ядро			Линейное ядро без многословных терминов		
TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B
15	16	17	18	19	20	21	22	23	24	25	26

Матрицы 2004, 2005, 2006

F1(micro)

	линейное ядро			полиномиальное			гаусово ядро			лучший
	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	результ.
2004	0.574	0.563	0.569	0.547	0.598	0.557	0.577	0.585	0.577	0.467
2005	0.591	0.577	0.584	0.568	0.610	0.566	0.594	0.599	0.589	0.592
2006	0.490	0.502	0.486	0.447	0.513	0.430	0.485	0.514	0.478	0.514

F1(macro)

	линейное ядро			полиномиальное			гаусово ядро			лучший
	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	результ.
2004	0.484	0.520	0.448	0.411	0.491	0.384	0.466	0.517	0.430	0.349
2005	0.423	0.476	0.397	0.350	0.443	0.326	0.407	0.472	0.377	0.432
2006	0.416	0.470	0.410	0.362	0.443	0.350	0.397	0.455	0.395	0.470

Выводы

- Выбор ядра меньше влияет на результат, чем способ взвешивания терминов
- Линейное vs Нелинейное ядро
  - паритет по результатам (micro/macro)
- Лучший способ взвешивания
  - частотный
- Многословные термины добавляют 4-6%
  - подозрительно мало