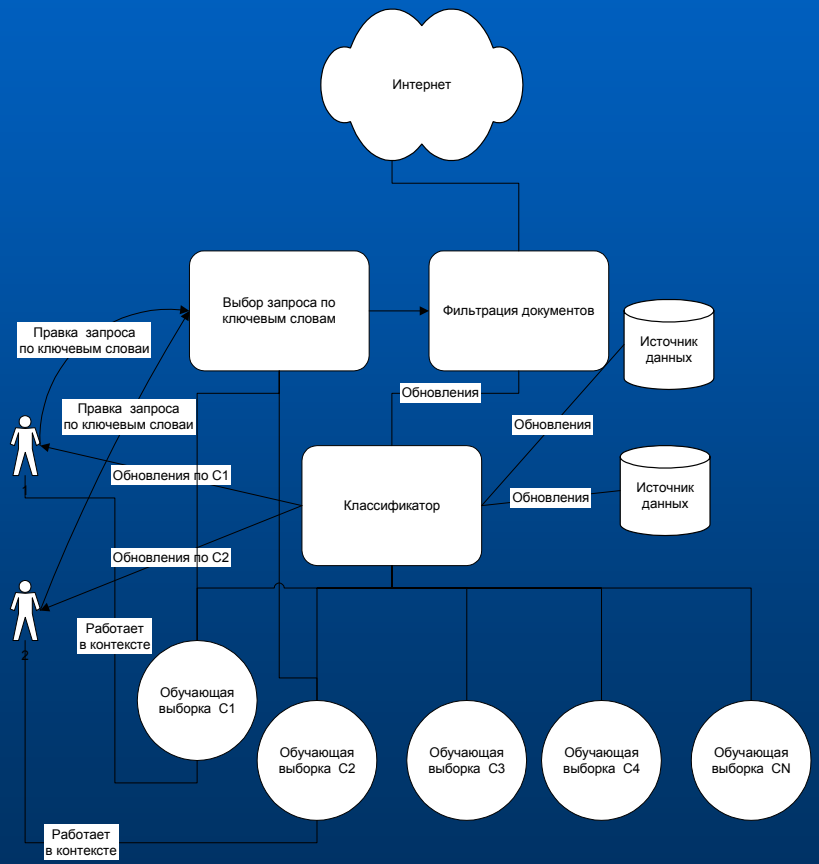
Сравнительный анализ алгоритмов классификации и способов представления Web-документов

Схема работы SPeCS



Этапы рубрикации

Предварительная обработка текста

- **Морфологический анализ**
- **Синтаксический анализ и постморфология**
- **Выбор фраз**
- **Анализ структуры Web-документов**

Классификация (мультиномиальная модель)

- **PrTFIDF**
- **Модифицированый Naive Bayes**
- **SVM**

Синтаксический анализ и пост-морфология

- Существующие решения (Dialing, LinkParser) ориентированы на точный разбор синтаксической структуры предложений для проверки правописания или машинного перевода
- Существующие решения обладают сравнительно низкой производительностью (50-200 секунд на 100 документов)
- В Dialing не учитывается часть возможных в реальных текстах связок лексем.

В данной работе предлагается упрощенный вариант синтаксического анализа, позволяющий увеличить производительность анализатора в несколько раз.

Результаты работы синтаксического анализатора используются:

- Для устранения морфологических неоднозначностей.
- Для построения фраз.

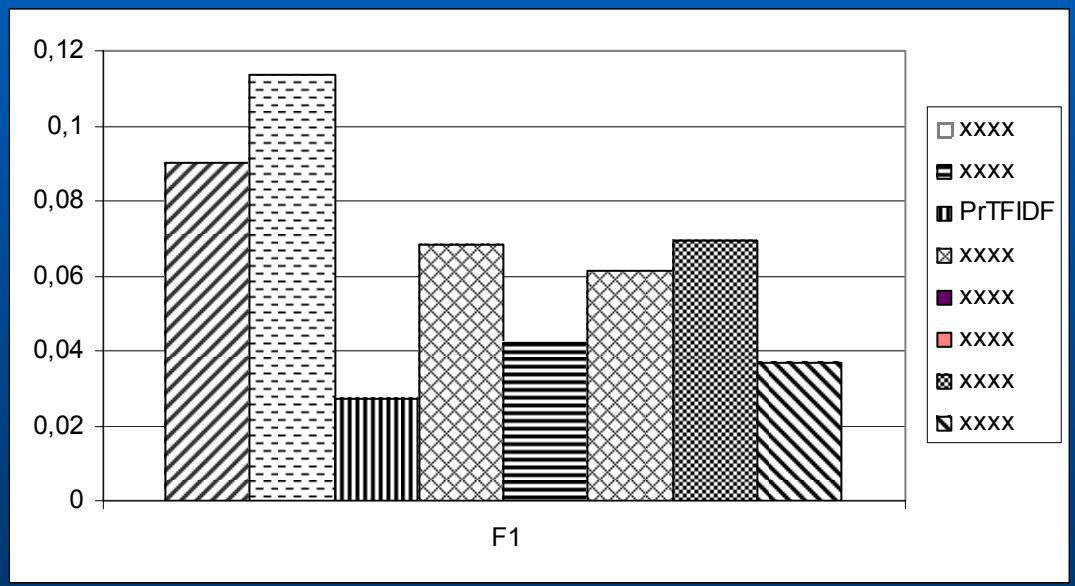
Алгоритм работы синтаксического анализатора.

1. На выходе морфологического анализатора для каждой лексемы получаем набор возможных морфологических вариантов {I}
2. Для каждого варианта предложения (фрагмента):
3. Производим последовательную подстановку синтаксических правил (с заменой лемм, участвующих в правиле на синтаксическую группу)
4. Выбираем вариант предложения, в котором минимальное количество лексем, не включенных в синтаксические группы.
5. В результате индексируются морфологические варианты лексем, соответствующие выбранному варианту предложения

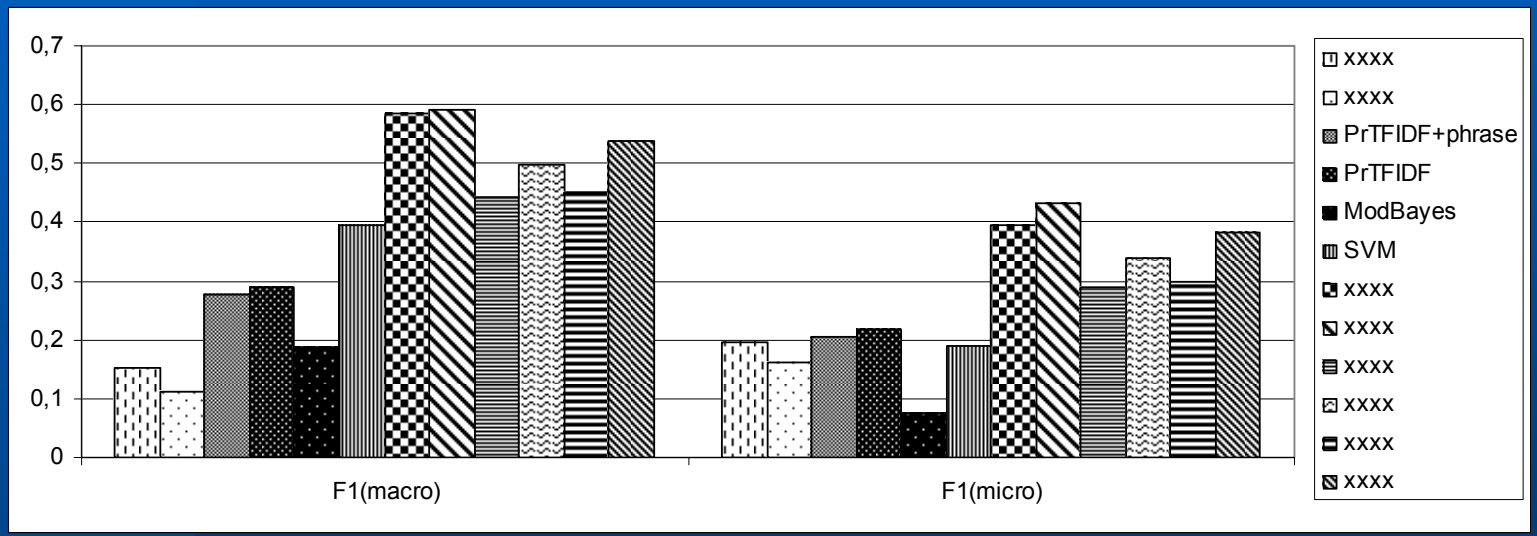
Сопоставление документу признаков

- Получение n-грамм(фраз): синтаксический анализ или статистическая фильтрация
- Использование модификаторов веса на основе инверсной частоты лексемы IDF (IDF'). IDF может быть более адекватно оценен с помощью данных словарей (WordNet)
- Учет разметки Web-документа
- Для алгоритма наивный Байес:
  - Реверсирование
  - Сглаживание (log, по Лапласу)
  - Нормализация по длине документа
  - Нормализация априорной вероятности

Дорожка классификации Web-документов



Дорожка по классификации нормативных документов



Интерпретация результатов

- Статистический выбор фраз не дает выигрыша при используемых параметрах
- Необходима серьезная доработка этапа предварительного анализа текста
- Необходима доработка алгоритмов классификации

Алгоритм ModSimpl

Идея: Вместо построения одной гиперплоскости, разделяющей документы с минимальной ошибкой, строим несколько гиперплоскостей, вдоль нормалей к каждой из которой документы разных классов максимально разнесены, т.е. минимизирован дискриминант Фишера.

1. **Методом адаптивного градиентного спуска находим локальный максимум дискриминанта Фишера.**
2. **Сортируем все документы по их проекции на полученное направление (вектор)**
3. **Отбрасываем область только положительных (отрицательных) документов с краев направления.**
4. **Запоминаем направление, точки отсечения положительных и отрицательных документов, а также оптимальную точку их разделения (по F-мере)**
5. **Если осталось более 2 документов, повторяем от шага 1 с оставшимися документами**

Эксперименты на обучающем наборе коллекции Legal

	NB	PrTFIDF	ModBayes	ModSimpl	SVM
точность	< 10%	< 10%	45,46%	44,54%	47,83%

Алгоритм ModSimpl. Анализ

- Эффективно решает задачу бинарной классификации (сравнимо с SVM)
- Более высокая скорость обучения алгоритма по сравнению с SVM (O(n*log(n)) против $O(n^a)$, где a > 1.2.
- Существенно меньшие требования по оперативной памяти (O(n+m))
- Обладает высокой точностью, но относительно низкой полнотой.
- При наличии большого количества классов точность сравнима с точностью модифицированного алгоритма Байеса (и SVM). Однако ресурсные требования алгоритма Байеса существенно ниже

Дальнейшие направления работы

- Доработка вероятностных алгоритмов для решения задачи рубрикации с большим количеством неравномощных классов.
- Исследование и доработка алгоритма ModSimpl. Данный алгоритм, в отличие от вероятностных, показал хорошие результаты и при решении задачи бинарной классификации
- Совместное использование синтаксического и статистического выбора фраз
- Анализ блоков Web-страниц и устранение шумовых элементов
- Анализ контекста ссылок на данный документ
- Использование словарей синонимов и, возможно, адаптированного вероятностного латентно-семантического анализа

Спасибо за внимание

Классификация текстов. PrTFIDF

$$\Pr(C \mid d, \theta) = \sum_{x} \Pr(C \mid x) \bullet \Pr(x \mid d, \theta) \qquad \theta : d \to x$$

$$\{x\} = \{w\} = F$$

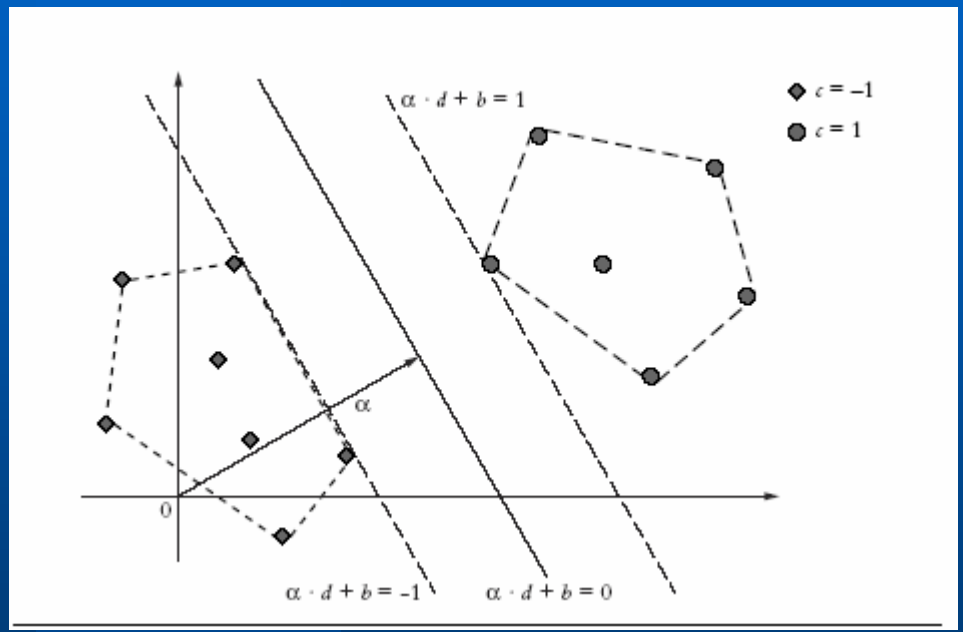
x – представление документа, в данном случае – в виде одного слова

$$\Pr(C \mid w) = \frac{\Pr(w \mid C) * \Pr(C)}{\sum_{C' \in C} \Pr(w \mid C') * \Pr(C')}$$

$$\Pr(x \mid d, \theta) = \Pr(w \mid d, \theta) = \frac{TF(w, d)}{\sum_{w' \in F} TF(w', d)}$$

$$\Pr(C \mid d, \theta) = \sum_{w \in F} \frac{\Pr(w \mid C) * \Pr(C)}{\sum_{C' \in C} \Pr(w \mid C') * \Pr(C')} \cdot \Pr(w \mid d, \theta)$$

Support Vector Machine



минимизировать $\frac{1}{2}\alpha\cdot\alpha(=\frac{1}{2}\|\alpha\|^2)$

при $c_i(\alpha\cdot d_i+b)\geq 1,\forall i=1...n$

минимизировать $\frac{1}{2}\alpha\cdot\alpha+C\sum_i\xi_i$

при $c_i(\alpha\cdot d_i+b)\geq 1-\xi_i,\forall i=1...n$

$\xi\geq 0_i,\forall i=1...n$

Задача сводится к задаче квадратичной оптимизации:

максимизировать $\sum_i\lambda_i-\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j c_i c_j(d_i\cdot d_j)$

при $\sum_i c_i\lambda_i=0$

$0\leq\lambda_i\leq C,\forall i=1...n$