Васильев В.Г.

Комплексная технология автоматической классификации текстов

ИПИ РАН

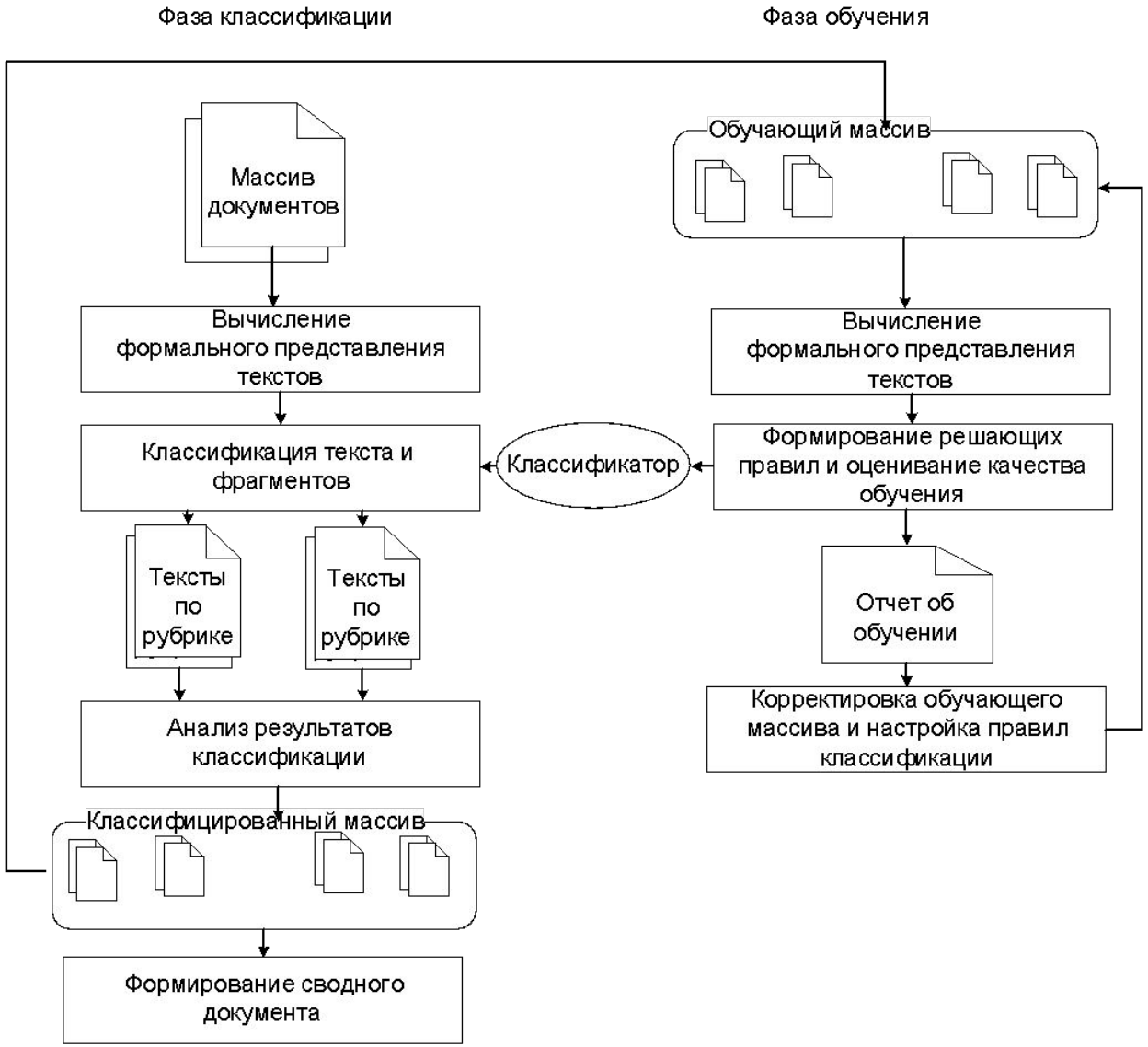
Особенности реальных массивов текстов

- Недостаточное количество обучающих примеров
- Наличие ошибок в эталонной классификации
- Несоответствие обучающих и обрабатываемых данных
- Совместное использование нескольких принципов деления на классы
- Политематический и зашумленный характер текстов
- Сложность интерпретации результатов классификации
- Наличие повторяющейся и дублирующей информации

Организационные проблемы

- Ограниченный доступ разработчиков систем автоматической классификации к исходным данным и массивам текстов
- Выполнение настройки и использования средств классификации пользователями, которые не являются специалистами в области автоматической обработки текстов

Комплексная технология классификации текстов



Недостаточное количество обучающих примеров

Прикладные проблемы:

- невозможность построения правил классификации для большинства методов, основанных на обучении по примерам;
- низкая надежность оценки качества обучения.

Решение: поддержка совместного использования трех типов решающих правил для рубрик:

- статистических (обучаемых на примерах документов),
- логических (задаются экспертами на специальном информационно-поисковом языке),
- шаблонных (задаются экспертами в виде регулярных выражений).

Наличие ошибок в эталонной классификации

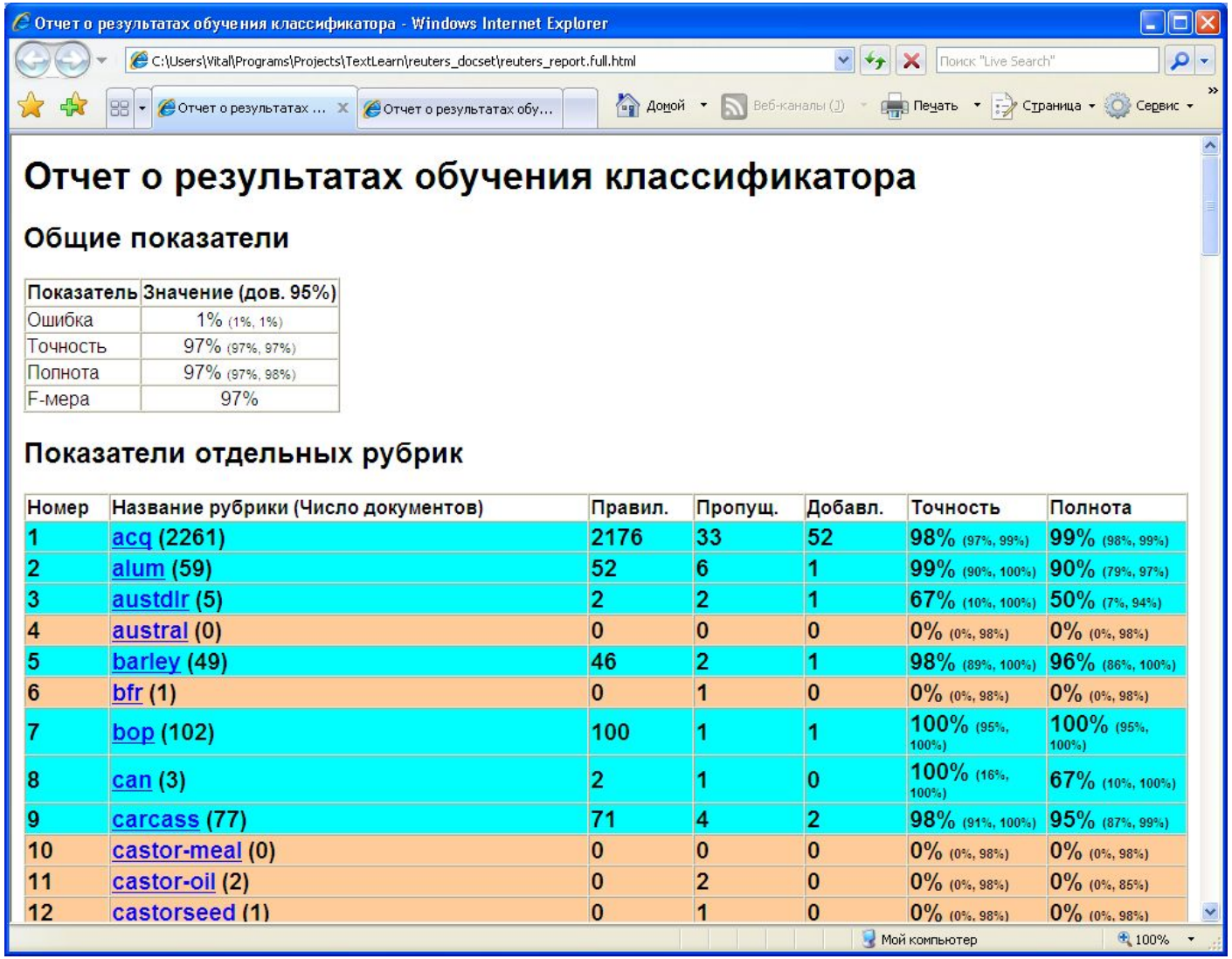
Прикладные проблемы:

- формирование ошибочных правил классификации;
- результаты оценки качества обучения оказываются некорректными.

Решение:

- выполнение при обучении оценки качества классификации и ошибок в эталонном множестве документов;
- учет степени тематической близости рубрик друг к другу;
- реализация интерактивной процедуры обучения классификатора.

Пример оценки эталонного множества документов



Отчет о результатах обучения классификатора

Общие показатели

Показатель	Значение (дов. 95%)
Ошибка	1% (1%, 1%)
Точность	97% (97%, 97%)
Полнота	97% (97%, 98%)
F-мера	97%

Показатели отдельных рубрик

Номер	Название рубрики (Число документов)	Правил.	Пропущ.	Добавл.	Точность	Полнота
1	acq (2261)	2176	33	52	98% (97%, 99%)	99% (98%, 99%)
2	alum (59)	52	6	1	99% (90%, 100%)	90% (79%, 97%)
3	austdlr (5)	2	2	1	67% (10%, 100%)	50% (7%, 94%)
4	austral (0)	0	0	0	0% (0%, 98%)	0% (0%, 98%)
5	barley (49)	46	2	1	98% (89%, 100%)	96% (86%, 100%)
6	bfr (1)	0	1	0	0% (0%, 98%)	0% (0%, 98%)
7	bop (102)	100	1	1	100% (95%, 100%)	100% (95%, 100%)
8	can (3)	2	1	0	100% (16%, 100%)	67% (10%, 100%)
9	carcass (77)	71	4	2	98% (91%, 100%)	95% (87%, 99%)
10	castor-meal (0)	0	0	0	0% (0%, 98%)	0% (0%, 98%)
11	castor-oil (2)	0	2	0	0% (0%, 98%)	0% (0%, 85%)
12	castorseed (1)	0	1	0	0% (0%, 98%)	0% (0%, 98%)

Несоответствие обучающих и обрабатываемых данных

Прикладные проблемы:

- результаты классификации текстов могут быть неопределенными;
- результаты оценки качества обучения являются завышенными.

Решение:

- выполнение оценки качества классификации в процессе обучения;
- обеспечение переобучения в процессе обработки новой информации;
- использование дополнительных словарей квазисинонимов для повышения полноты классификации.

Иерархический характер и использование нескольких принципов деления на классы

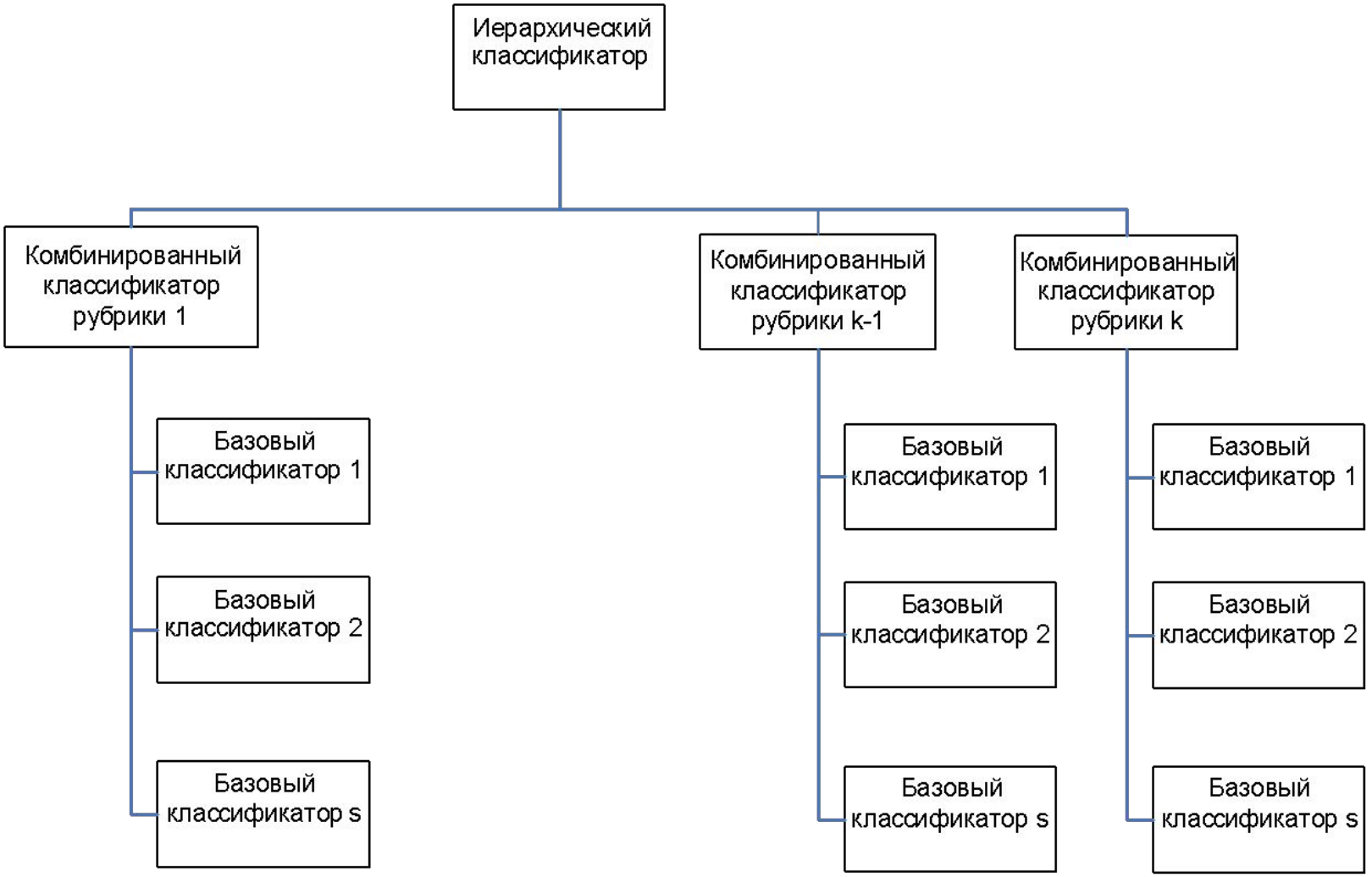
Прикладные проблемы:

- сложность построения эффективных процедур классификации, основанных на использовании одной модели или метода для всех рубрик и уровней классификатора.

Решение:

- поддержка нескольких типов признаков (лексических, грамматических, синтаксических);
- комбинирование различных методов классификации;
- поддержка режима фасетной классификации.

Комбинированный иерархический метод классификации



Базовые методы классификации

Методы байесовской классификации

- Смесь факторных анализаторов
- Смесь полиномиальных распределений
- Смесь многомерных распределений Бернулли
- Смесь распределений фон Мизеса-Фишера

Методы классификации на основе расстояний

- k-ближайших соседей
- SVM
- Роччио

Методы классификации на основе правил

- Деревья решений
- Логические правила

Пример реализации базовых методов

Метод	**К-ближайших соседей**	**Смесь факторных моделей**
Размер обуч. множества	10 – 50000	10 – 50000
Веса признаков	TF-IDF [1]	TF-IDF
Снижение размерности	селекция по частоте документов	селекция по частоте документов, последовательный метод LSI
Оценка параметров	число соседей – 5, максимальное число эталонов – 250, оригинальный алгоритм отбора эталонов на основе кластерного анализа	оригинальный робастный алгоритм [11], размерность пространства факторов – 5
Решающее правило	байесовское правило с откл. вер. уровня 60% [2]	байесовское правило с откл. вер. уровня 60%

Комбинированные классификаторы рубрик

Методы построения комбинированных классификаторов

- построение фиксированного решающего правила;
- оценка качества работы базовых классификаторов;
- использование статистического моделирования (например, boosting и bagging).

Метод наилучшего классификатора

- для каждой рубрики требуется хранить в памяти только одно решающее правило;
- результаты оценки качествамогут использоваться для корректировки состава обучающих примеров.

Интегральная оценка качества работы для массива «Reuters-21578-6»

Метод классификации	Точность (дов. инт.)	Полнота (дов. инт.)	F-мера	Процент обученных рубрик
TREE	81% (80%, 82%)	87% (86%, 87%)	84%	50%
PPCA	94% (93%, 94%)	95% (95%, 95%)	94%	50%
MNS	66% (65%, 67%)	60% (60%, 61%)	63%	65%
KNN	76% (76%, 77%)	73% (72%, 74%)	75%	56%
BERN	81% (80%, 82%)	87% (86%, 87%)	84%	70%
VMF	92% (91%, 92%)	93% (92%, 93%)	92%	56%
ROC	41% (40%, 42%)	36% (35%, 36%)	38%	70%
SVM	95% (95%, 95%)	96% (95%, 96%)	95%	56%
Комбинированный без правил	97% (98%, 99%)	97% (98%, 99%)	97%	70%
Комбинированный с правилами	98% (98%, 99%)	99% (98%, 99%)	99%	100%

Оценка качества работы базовых методов для рубрик «Reuters-21578»

Рубрика (размер)	**TREE**	**PPCA**	**MNS**	**KNN**	**BERN**	**VMF**	**ROC**	**SVM**
acq (2261)	85%	95%	40%	54%	92%	95%	3%	**98%**
alum (59)	91%	85%	59%	83%	73%	89%	54%	**94%**
dmk (15)	76%	88%	88%	92%	64%	**97%**	92%	88%
housing (18)	85%	84%	84%	**94%**	81%	91%	88%	88%
l-cattle (10)	-	-	36%	**95%**	67%	90%	88%	29%
meal-feed (51)	**97%**	93%	63%	80%	81%	94%	77%	94%
palm-oil (42)	85%	**98%**	85%	91%	82%	94%	91%	94%
propane (6)	-	-	-	-	55%	-	**80%**	-
rapeseed (35)	72%	86%	76%	90%	**94%**	91%	75%	90%
sfr (3)	-	-	**100%**	-	80%	-	80%	-
soy-oil (26)	22%	36%	7%	20%	**51%**	40%	33%	26%
strategic-metal (39)	75%	**80%**	32%	61%	60%	77%	51%	73%
zinc (48)	74%	85%	60%	70%	74%	**91%**	78%	81%

Политематический и зашумленный характер текстов

Прикладные проблемы:

- сложность формирования решающих правил из-за негативного влияния посторонней информации и наложение рубрик друг на друга;
- неопределенность расположения в тексте информации, релевантной рубрике.

Решение:

- идентификация форматов, языков и кодировок документов;
- очистка текста документов от элементов оформления;
- исключение из текстов вспомогательной информации;
- использование робастных алгоритмов оценивания параметров;
- выделение значимых фрагментов в текстах.

Выделение значимых фрагментов

Модель представления текста

Текст X представляется в виде множества векторов

$$F = \left\{ \bigotimes_{l=l_1}^{l_2} X_l \mid 1 \le l_1 \le l_2 \le n \right\},$$

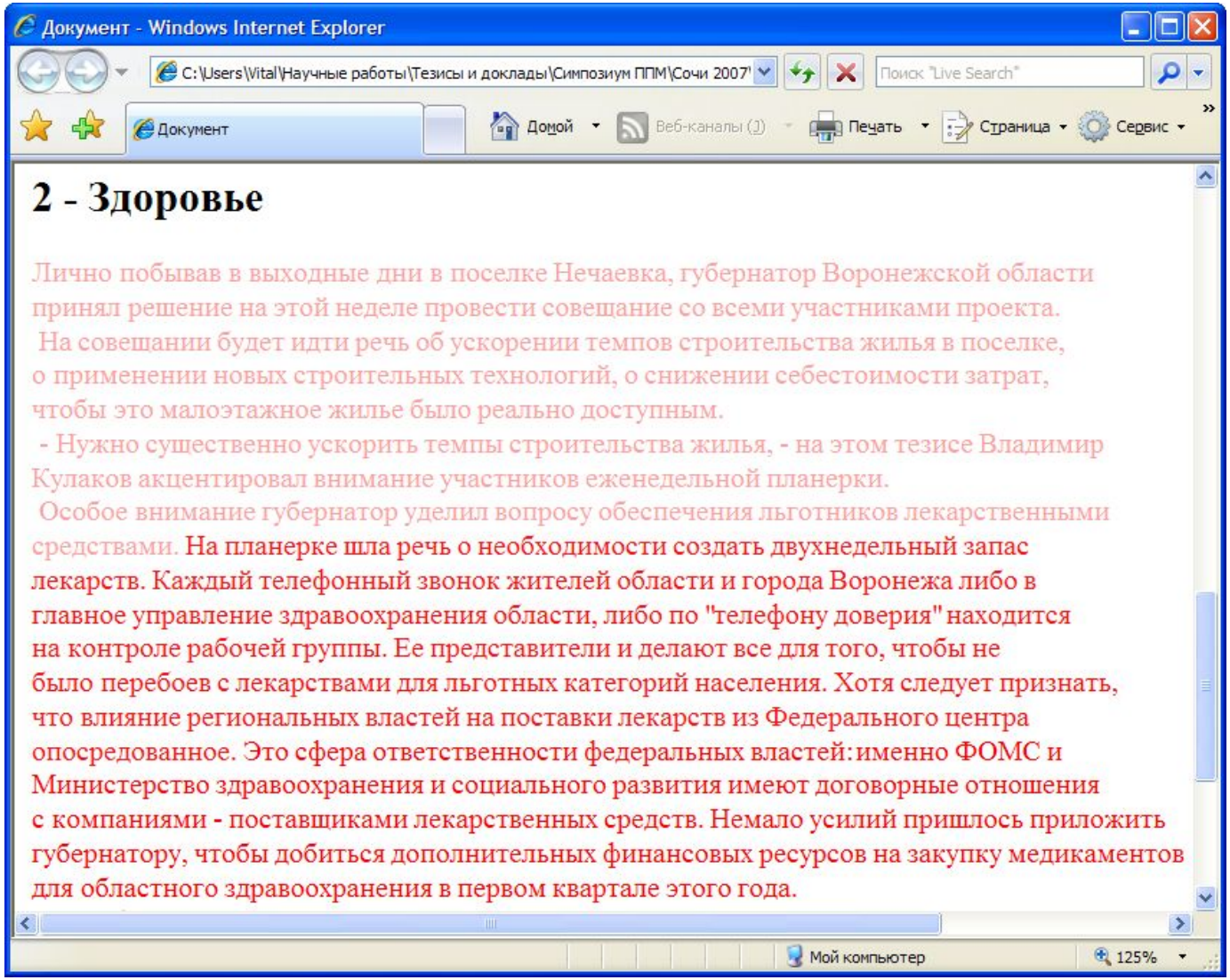
соответствующего множеству всех фрагментов текста (непрерывных последовательностей предложений).

Оценка степени соответствия текста рубрике

$$w_{ij} = \max_{Y \in F, X_i \in Y} g_j(Y),\ i = 1,...,n,\ j = 1,...,k,$$

где $g_j(Y)$ - функция оценки принадлежности текста Y к рубрике ω_j.

Пример разметки текста с помощью иерархического покрытия



Наличие повторяющейся и дублирующей информации во входном потоке текстов

Прикладные проблемы:

- сложность просмотра и анализа результатов классификации.

Решение:

- упорядочение документов в рубриках с учетом их тематической близости друг к другу;
- выявление "почти дубликатов" документов;
- выявление основных тем документов в рубриках;
- автоматическое формирование сводных документов.

Пример выявления основных тем в рубрике при классификации