КОМПЛЕКСНАЯ ТЕХНОЛОГИЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

COMPLEX TECHNOLOGY OF AUTOMATIC TEXT CLASSIFICATION

Васильев В.Г. *(vvg_2000@mail.ru)*
Институт прикладной информатики РАН

В докладе рассматриваются проблемы, которые возникают при построении прикладных систем автоматической классификации текстов. Приводится описание основных элементов комплексной технологии классификации текстов, обеспечивающей полный цикл обработки и анализа данных, начиная с очистки и выделения текста из документов, заканчивая анализом результатов классификации. Особое внимание уделяется вопросам построения комбинированных решающих правил для выполнения иерархической классификации текстов.

Введение

Потребность в автоматизации различных задач, связанных с обработкой и анализом текстовых данных на естественном языке, испытывают как рядовые пользователи средств вычислительной техники, так и крупные государственные и частные организации. В области автоматизированной обработки текстов уже сложился ряд относительно самостоятельных направлений (задач): извлечение объектов и признаков, реферирование, классификация, кластерный анализ, интеллектуальный поиск, фактографический анализ, пространственный (географический) анализ. В настоящей работе указанные базовые задачи анализа текстов рассматриваются не независимо друг от друга, а как элементы комплексной технологии автоматической классификации текстовых данных, обеспечивающей эффективную обработку информации и представление результатов анализа для конечных пользователей (см. рис. 1).

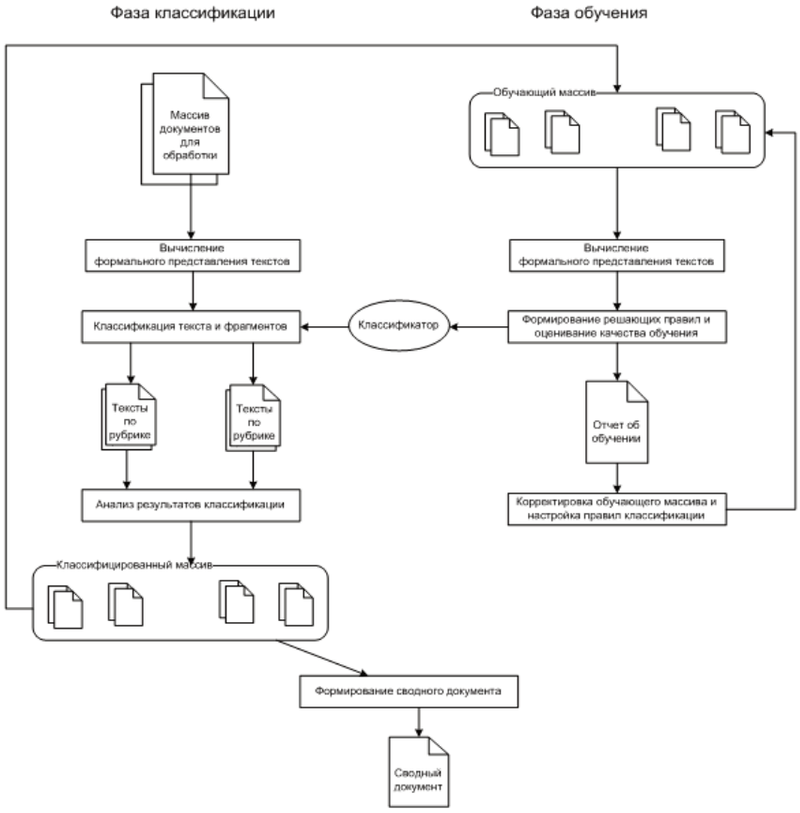

Рис. 1. Общая схема комплексной технологии классификации текстов

Краткое описание задач, решаемых в рамках приведенных на рис. 1 функциональных блоков приводится в табл. 1.

№	Функциональный блок	Задачи функционального блока
1	Вычисление формального представления текстов	1. Получение текста документа (идентификация формата, языка, кодировки документа, очистка текста от элементов оформления, разбиение на составные части). 2. Лингвистический анализ (графематический, морфологический и постморфологический анализ, выделение словосочетаний). 3. Формирование векторного (матричного) представления текстов.
2	Классификация текстов и фрагментов	1. Предварительная обработка текстов (отображение словаря признаков документа в пространство признаков классификатора; оценка адекватности и возможности классификации текста с помощью данного классификатора). 2. Классификация текста и выделение значимых фрагментов в нем (выделение кодов рубрик с помощью регулярных выражений; применение логических правил, построенных экспертами, и статистических решающих правил; корректировка результатов классификации с учетом иерархической структуры рубрик).
3	Анализ результатов классификации	1. Выявление "почти" дубликатов документов. 2. Выявление основных тем документов в рубриках..
4	Формирование сводного документа	1. Формирование сводного документа (упорядочение документов по их релевантности рубрике; упорядочение фрагментов с учетом их близости друг к другу; построение объединенного документа).
5	Формирование решающих правил и оценивание качества обучения	1. Формирование обучающих и тестовых множеств для рубрик (построение разбиения обучающего массива на блоки; анализ взаимосязей и пересечений отдельных рубрик; формирование множеств отрицательных и положительных примеров). 2. Оценивание параметров базовых моделей рубрик (вычисление весов признаков; снижение размерности; оценивание параметров моделей; формирование решающих правил; оценка качества обучения). 3. Построение комбинированных решающих правил для отдельных рубрик и классификатора в целом. 4. Формирование отчета о результатах обучения (описание решающих правил, описание терминологии рубрик, рекомендации по корректировке примеров документов, описание взаимосвязей рубрик).
6	Корректировка обучающего массива и настройка правил классификации	1. Корректировка обучающих примеров для рубрик путем анализа добавленных и пропущенных документов в рубриках, значимых фрагментов, взаимосвязей рубрик. 2. Настройка правил классификации для отдельных рубрик (явное задание предпочтительных статистических моделей, задание необходимых, достаточных и исключающих логических правил на специальном языке).

Таблица 1. Описание функциональных блоков комплексной технологии

Описанная выше технология полностью реализована как в виде отдельного пакета «Text Classification Toolbox« для системы Matlab [9], так и виде элемента ряда заказных информационно-аналитических систем.

При ее построении особое внимание было уделено учету особенностей реальных массивов текстов, которые часто оказывает негативное влияние на итоговое качество классификации текстов. В табл. 2 приведено описание наиболее типичных из них (отмечаются во многих работах и наблюдались автором на практике), а также способы их учета в рамках предлагаемой технологии.

№	Название особенностей и недостатков (ссылки на работы)	Проблемы и средства их решения в рамках комплексной технологии
1	Наличие недостаточного количества или отсутствие обучающих примеров для ряда рубрик ([1], [3], [5], [6]).	Проблемы: невозможность построения правил классификации для большинства методов, основанных на обучении по примерам; низкая надежность оценки качества обучения. Решение: - поддержка совместного использование 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).
2	Наличие ошибок и непоследовательность формирования эталонного распределения текстов по рубрикам ([1], [2], [5], [6]). Односторонность примеров в обучающем массиве текстов ([1]). Наличие рубрик с одинаковым содержанием, но с разными названиями.	Проблемы: формирование ошибочных правил классификации; результаты оценки качества обучения оказываются некорректными. Решение: - автоматическое выполнение при обучении оценки качества классификации и ошибок в эталонном распределении документов по рубрикам; - формирование обучающих примеров для отдельных рубрик с учетом оценки степени тематической близости рубрик друг к другу; - реализация оригинальной интерактивной процедуры обучения классификатора, обеспечивающей корректировку обучающих примеров и задание логических правил.
3	Несоответствие тематики и характера обучающего массива текстов тематике и характеру классифицируемых текстов ([5], [6]).	Проблемы: результаты классификации текстов могут быть неопределенными и зависеть от незначительных случайных факторов; 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.
4	Многоуровневый (иерархический) характер классификаторов ([1], [5]). Совместное использование нескольких оснований (принципов) разделения данных на классы (тема и тональность). Неоднородность характера текстов в рубриках.	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); - использование оригинального комбинированного иерархического метода классификации, обеспечивающего подбор для каждой рубрики наиболее подходящих для нее моделей и методов классификации [4].
5	Одновременная оценка текстов с использованием нескольких классификаторов (например, страна, организация, тема, жанр) [7].	Проблемы: сложность подбора примеров документов и обучения объединенного классификатора, получаемого путем декартового произведения рубрикаторов для отдельных фасетов; сложность или невозможность независимого обучения классификаторов из-за наличия контекстных связей между рубриками классификаторов. Решение: - поддержка режима фасетной классификации, который обеспечивает независимое обучение отдельных фасетов и комбинирование их работы с использованием специальных логических правил.

№	Название особенностей и недостатков (ссылки на работы)	Проблемы и средства их решения в рамках комплексной технологии
6	Политематический и составной характер документов ([1]). Наличие служебных элементов и посторонних блоков текста, не относящихся к основной тематике документа [1]. Наличие в обучающем массиве аномальных документов (пустых, в неизвестных кодировках и т.п.) ([2]).	Проблемы: сложность автоматического формирования решающих правил для рубрик из-за негативного влияния посторонней информации; снижение качества классификации из-за наложения нескольких рубрик друг на друга; сложность интерпретации результатов классификации из-за неопределенности расположения в тексте информации, релевантной рубрике. Решение: - реализация полного комплекса средств для идентификации форматов, языков и кодировок документов; - реализация оригинальных алгоритмов для очистки текста документов от элементов оформления, основанных на оценке распределения плотности текста на странице; - реализация оригинальных алгоритмов для исключения из текстов вспомогательной информации, основанных на сопоставлении лексического состава отдельных предложений; - реализация робастных вариантов алгоритмов оценивания параметров моделей и методов классификации [4, 11]; - реализация эффективного выделения значимых фрагментов в текстах путем использования оригинальной иерархической модели представления текстов (основывается на представлении текста множеством векторов признаков, соответствующих элементам специального иерархически упорядоченного перекрывающегося множества фрагментов текста) [10].
7	Наличие повторяющейся информации во входном потоке текстов [8]. Большие объемы и неоднородность входного потока текстов ([3]).	Проблемы: сложность просмотра и анализа результатов классификации. Решение: - реализация оригинального алгоритма упорядочения документов в рубриках, который учитывает не только релевантность документов рубрике, но и их тематическую близость друг к другу; - реализация оригинального алгоритма для выявления "почти дубликатов" документов, основанного на использовании методов иерархического кластерного анализа и иерархической модели представления текстов; - реализация оригинальных алгоритмов кластерного анализа результатов классификации, основанных на использовании модели смеси вероятностных анализаторов главных компонент и обеспечивающих эффективное формирование графических карт массивов текстов [11].

Таблица 2. Типичные особенности и недостатки реальных массивов текстов

Полное описание всех элементов комплексной технологии сложно провести в рамках одной работы. По этой причине остановимся более подробно на рассмотрении одного из ее базовых элементов – методе комбинированной иерархической классификации.

Описание метода комбинированной иерархической классификации

Разнородный характер текстов в рубриках и использование различных оснований классификации приводит к сложности выделения метода классификации, который был бы эффективным во всех случаях. Возможным решением является совместное использование сразу нескольких методов классификации текстов [4].

В общем виде работу комбинированного иерархического классификатора можно представить следующим образом. На вход поступает вектор весов информационных признаков анализируемого текста или фрагмента, а на выходе формируются два вектора $c = (c_1,...,c_k)$ и $w = (w_1,...,w_k)$, где общее число рубрик в классификаторе, $c_j \in \{0,1\}$ и $w_j \in [0,1]$ – признак и степень принадлежности к рубрике ω_j, соответственно.

Решающие правила для отнесения текстов к рубрикам получаются путем комбинирования результатов работы сразу нескольких базовых методов классификации с помощью метаклассификаторов более высокого уровня. Всего выделяется три уровня: уровень базовых классификаторов, уровень комбинированных классификаторов, уровень иерархического классификатора.

На первом уровне для каждой рубрики ω_j, $j = 1,...,k$ производится построение бинарных решающих с помощью следующих базовых методов [1, 2, 3]:

- методы вероятностной классификации, основанной на представлении рубрик в виде смеси распределений Бернулли (BERN), фон Мизеса-Фишера (VMF), полиномиального (MNS) и анализаторов главных компонент (PPCA);

- методы классификации на основе вычислений расстояний: классификаторы *k* – ближайших соседей (KNN), машин опорных векторов (SVM), Роччио (ROC);

- методы классификации на основе правил: деревья решений (TREE), логические правила на специальном языке.

Все приведенные методы, за исключением логических правил, основаны на обучении на примерах. При этом при обучении для каждого метода реализуется специализированная процедура обработки данных, которая включает проверку достаточности размера обучающей выборки, снижение размерности путем селекции и трансформации признаков, оценивание параметров, построение решающих правил, оценка качества обучения (особенности реализации и параметры по умолчанию для базовых методов показаны в табл. 3).

Метод	**Мин. и макс. размер обуч. множества**	**Веса признаков**	**Снижение размерности**	**Оценка параметров**	**Решающее правило**
BERN	2 - 50000	Бинарные [1]	селекция по частоте документов [1]	байесовское оценивание	байесовское правило с откл. вер. уровня 65% [2]
VMF	5 - 50000	TF-IDF [1]	селекция по частоте документов [1]	Оригинальный робастный алгоритм	байесовское правило с откл. вер. уровня 95% [2]
MNS	2 - 50000	TF-IDF [1]	селекция по частоте документов [1]	Оригинальный робастный алгоритм	байесовское правило с откл. вер. уровня 80% [2]
PPCA	10 - 50000	TF-IDF [1]	селекция по частоте документов, последовательный метод LSI [11]	оригинальный робастный алгоритм [11], размерность пространства факторов – 5.	байесовское правило с откл. вер. уровня 60% [2]
KNN	5 - 50000	TF-IDF [1]	селекция по частоте документов [1]	число соседей – 5, максимальное число эталонов – 250, оригинальный алгоритм отбора эталонов на основе кластерного анализа	байесовское правило с откл. вер. уровня 60% [2]
SVM	5 - 50000	TF-IG [1]	селекция по частоте документов [1]	линейная ядерная функция [2]	линейная решающая функция
ROC	2 - 50000	TF-IDF [1]	селекция по частоте документов [1]	стандартный алгоритм [1]	линейная решающая функция
TREE	10 - 50000	Бинарные - IDF [1]	селекция по частоте документов, метод фильтрации признаков Information Gain [1]	оригинальный алгоритм отбора эталонов на основе кластерного анализа, стандартный алгоритм из пакета matlab	логическое решающее правило

Таблица 3. Особенности реализации базовых методов классификации

Логические правила строятся вручную для уточнения и дополнения статистических решающих правил, а также построения правил для рубрик без примеров документов. Они разбиваются на три типа: достаточные (справедливость достаточна для отнесения текста к рубрике), необходимые (для отнесения текста к рубрике необходима справедливость данного правила и построенного статистического правила), отрицательные (при его справедливости текст не относится к рубрике).

Реализованный язык задания логических правил обеспечивает поиск отдельных слов (с учетом и без учета морфологии, с учетом ошибок, с заданными морфологическими и семантическими характеристиками), задание логических условий, задание условий на расстояние между выражениями в тексте.

На втором уровне для каждой рубрики ω_j, $j = 1,...,k$ осуществляется построение отдельного комбинированного классификатора на основе бинарных классификаторов первого уровня $C_{j1},...,C_{jL}$, построенных для данной рубрики, где L – число различных методов классификации. Для этих целей реализована поддержка нескольких методов, которые условно можно разбить на три группы [2, 4]:

- методы, основанные на построении фиксированного решающего правило, которое не зависит от качества работы отдельных классификаторов (например, правило произведения апостериорных вероятностей, правило суммирования апостериорных вероятностей, правило большинства голосов, правила минимума апостериорных вероятностей классов).

- методы, основанные на построении комбинированного правила классификации, учитывающего оценки качества работы классификаторов первого уровня (например, байесовский метод и метод наилучшего классификатора).

- методы, основанные на использовании статистического моделирования (например, boosting и bagging).

Экспериментальная оценка приведенных групп правил показала, что использование методов первой группы не приводит к улучшению качества классификации, но при этом время работы и требования к памяти заметно возрастают из-за необходимости одновременного использования для каждой рубрики нескольких алгоритмов при классификации. Методы третьей группы требуют выполнения чрезвычайно ресурсоемкого моделирования, которое не позволяет проводить обучение классификаторов за разумное время на практике.

В настоящей работе в качестве основного был выбран метод наилучшего классификатора, в рамках которого комбинированный классификатор для рубрики ω_j, $j = 1,...,k$ получается путем выбора из $C_{j1},...,C_{jL}$, классификатора, обеспечивающего наибольшее значение F-меры на тестовом множестве.

Такой подход обладает следующими преимуществами:

- при классификации для каждой рубрики требуется хранить в памяти только одно решающее правило,

- результаты оценки качества, проводимой для выбора наилучшего метода, могут использоваться для корректировки состава обучающих примеров.

На третьем уровне осуществляется построение иерархического классификатора, объединяющего результаты работы бинарных классификаторов второго уровня, таким образом, чтобы обеспечить отнесение текстов одновременно к нескольким рубрикам с учетом их иерархической структуры. Его работа сводится выполнению серии процедур, которые, например, осуществляют проверку при отнесении документа к рубрике, что он относится к родительской рубрике, производят проверку различных аномальных случаев.

Для оценки эффективности разработанной технологии были проведены эксперименты с различными массивами текстов и рубрикаторами. К сожалению, большинство из рассмотренных массивов не являются общедоступными, что затрудняет публикацию информации по ним. По этой причине в качестве примера приведем результаты экспериментов только с массивом «Reuters-21578», который широко используется в различных работах по автоматизированной обработке текстов [1].

Для проведения экспериментов использовался пакет «Text Classification Toolbox», реализующий описанный выше комбинированный иерархический алгоритм классификации. При обучении классификатора в рамках данного пакета автоматически производится формирование отчета с оценками качества работы базовых методов и классификатора в целом с использованием метода 5-шаговой кросс проверки. Для упрощения проведения экспериментов было решено не использовать эталонное разбиение на обучающую и тестовую выборку, которое имеется в массиве «Reuters-21578». Это может приводить к несколько отличным от других исследователей абсолютным значениям показателей качества классификации, но это не является критичным, так как в данном случае для иллюстрации наибольшее значение имеют относительные значения показателей. Все базовые алгоритмы использовались с приведенными выше стандартными значениями параметрами.

В табл. 4 приводятся оценки качества обучения 13 из 142 рубрик данного массива с помощью 8 методов классификации.

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%

Таблица 4. Качество обучения отдельных рубрик с помощью различных методов

Необходимо отметить, что у ряда рубрик в массиве «Reuters-21578« отсутствуют примеры документов или их количество является недостаточным для оценивания параметров моделей данных применяемых в отдельных методах классификации. Такие случаи отмечены в таблице прочерками.

Из табл. 4 видно, что для каждого метода классификации существуют рубрики, на которых он оказывается значительно предпочтительнее других методов.

В табл. 5 приводятся усредненные оценки качества обучения комбинированного классификатора для следующих случаев: используется только один базовый метод классификации, используются все базовые методы без задания и с заданием экспертных логических правил. Оценки коэффициентов точности, полноты и F-меры вычислялись с использованием микро-усреднения [1, 3].

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 метод (без логических правил)	97% (98%, 99%)	97% (98%, 99%)	97%	70%
Комбинированный метод (с логическими правилами)	98% (98%, 99%)	99% (98%, 99%)	99%	100%

Таблица 5. Оценка качества обучения с использованием микро-усреднения

Из табл. 5 видно, что среди базовых методов наилучшие результаты показал метод SVM, что в целом согласуется результатам экспериментов других авторов с данным массивом текстов [1]. Использование комбинированного метода классификации позволяет дополнительно повысить качество классификации. При этом наилучшие результаты достигаются в том случае, когда для каждой рубрики, в которой имеются ошибки, экспертом задаются логические правила на специальном языке, которые эти ошибки устраняют. Необходимо отметить, что в результате задания экспертных логических правил удается также построить решающие правила для рубрики, у которых отсутствуют обучающие примеры или их меньше 2.

Влияние необученных рубрик и рубрик небольшого размера на показатели качества обучения более отчетливо проявляется при использовании макро-усреднения. В данном случае, обобщенные показатели вычисляются как арифметическое среднее показателей для отдельных рубрик. В результате, значение F-меры для метода SVM (наилучший метод при микроусреднении) становится равным 49%, а для комбинированного метода – 63% (без задания экспертных логических правил) и 85% (с заданием экспертных логических правил).

Значительное отличие микро и макро усреднения связано с тем, что в массиве имеется несколько больших рубрик, на которых достигаются высокие показатели качества классификации (например, для рубрики «acq» F-мера равна 99%), и большое количество пустых и маленьких рубрик, на которых F-мера принимает значения близкие к нулю. Повышенные значения качества работы комбинированного алгоритма с экспертными логическими правилами объясняются тем, что за счет подбора данных правил достаточно легко обеспечить точность и полноту классификации близкую к 100% для рубрик состоящих всего из нескольких документов.

Заключение

Таким образом, в настоящей работе рассмотрены базовые элементы комплексной технологии классификации текстов, которая реализована в виде пакета для системы Matlab. Отличительной особенностью данной технологии является ориентация на совместное применение экспертных и статистических методов классификации текстов, а также интегрированное использование всего множества процедур автоматизированной обработки текстов, начиная с очистки текстов от посторонней информации, заканчивая интерпретацией результатов. В частности, приведенные эксперименты с массивом «Reuters-21578« показали перспективность подхода, основанного на комбинировании методов классификации текстов.

К перспективным задачам можно отнести следующие: разработка эффективных методов обучения фасетных классификаторов при наличии выборок ограниченного объема и большом количестве фасетов; совершенствование процедур комбинирования результатов работы экспертных и статистических процедур классификации за счет задания соответствующих правил на специальном языке; разработка эффективных процедур и методик автоматизированного формирования сводных документов по результатам автоматической классификации.

Список литературы

1. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 34(1), 2002. – pp. 1-47.

2. Webb R.A. Statistical Pattern Recognition. Second Edition. // John Wiley & Sons Ltd., England, 2002. – 515 p.

3. Baldi P., Frasconi P., Smyth P. Modeling the Internet and theWeb. Probabilistic Methods and Algorithms // JohnWiley & Sons Ltd, 2003. – 306 p.

4. Кривенко М.П., Васильев В.Г. Проблемы разработки и внедрения технологий извлечения информации // Системы высокой доступности 3-4, т.2. – М.: Радиотехника, 2006. – с. 6-21.

5. Агеев М.С., Добров Б.В., Лукашевич Н.В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Труды 6-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004, Пущино, Россия, 2004. – 10 с.

6. Dumais S.T., Lewis D.D. & Sebastiani F., Report on the Workshop on Operational Text Classification Systems // SIGIR-02, Tampere, Finland. – 4 p. (http://www.sigir.org/forum/F2002/sebastiani.pdf).

7. Браславский П.И. Использование стилистических параметров документа при поиске информации в Internet // Доклады VI рабочего совещания по электронным публикациям – EL-PUB–2001. Новосибирск: ИВТ СО РАН, 2001. (http://www.ict.nsc.ru/ws/elpub2001/1812).

8. Yang H., Callan. J. Near-Duplicate Detection for eRulemaking // Proceedings of the 5th National Conference on Digital Government Research (DG.O2005), Atlanta, GA, USA, 15-18 May 2005.

9. Васильев В.Г., Кривенко М.П., Ефременкова М.В. Библиотека процедур классификации тексто-вых данных // Редакция «ОПиПМ» Обозрение приклад-ной и промышленной математики. Выпуск 1, том 13. – М., 2006. – с. 743.

10. Васильев В.Г. Автоматическое выделение значимых фрагментов в текстах // Редакция «ОПиПМ» Обозрение прикладной и промышленной математики. Выпуск 3, том 14. – М., 2007. – с. 518.

11. Кривенко М.П., Васильев В.Г. Кластерный анализ массивов текстовых данных // Препринт. – М.: ИПИ РАН, 2004. – 190 с.