УДК 004.85; 004.89

# АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ АНАЛИЗА И РУБРИКАЦИИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

**М. И. Петровский**[1], **В. В. Глазкова**[1]

Методы машинного обучения и интеллектуального анализа данных предназначены для решения задач анализа, классификации и выявления скрытых закономерностей в больших объемах разнородных сложно структурированных данных. К таким задачам относится прикладная задача анализа и рубрикации больших коллекций электронных текстовых и гипертекстовых документов. Для ее решения необходима разработка эффективных по точности и скорости алгоритмов для многотемной классификации (multi-label classification), т.е. классификации в условиях существенно перекрывающихся классов, когда любой объект классификации (документ) может принадлежать более чем одному классу (теме) одновременно, а также разработка формальных моделей представления гипертекстовых данных, эффективных по точности представления исходной информации и занимаемой при этом памяти. В настоящей статье предлагается новая модель представления данных, основанная на выделении частых эпизодов лексем (или N-грамм), и новый метод учета гиперссылок, основанный на классификации с помощью N-граммного классификатора текста адресов гиперссылок и замене их в исходном тексте документа на специальные признаки. Кроме того, исследуется возможность использования подхода на основе декомпозиции "каждый против каждого" для решения задачи многотемной классификации. Предлагается новый метод многотемной классификации, основанный на подходе попарных сравнений с помощью набора бинарных классификаторов, где результирующие вероятности принадлежности документа темам (релевантности классов) вычисляются с помощью обобщенной модели Брэдли–Терри, а нерелевантные классы отсекаются с помощью пороговой функции, заданной в пространстве релевантностей классов. Все разработанные алгоритмы экспериментально проверены на эталонных тестовых наборах данных и показали лучшие результаты по сравнению с традиционными методами. Работа поддержана грантом РФФИ № 06–01–00691, грантом поддержки научных школ № 02.445.11.7427 и грантом Президента РФ МК–4264.2007.9.

Ключевые слова: алгоритмы рубрикации текстовых и гипертекстовых документов, модели представления гипертекстовой информации, алгоритмы многотемной (multi-label) классификации, метод попарных сравнений.

**1. Введение.** В настоящее время во многих прикладных областях важную роль начинают играть методы машинного обучения и интеллектуального анализа данных, предназначенные для анализа, классификации и выявления скрытых закономерностей в больших объемах разнородных сложно структурированных данных. Следует отметить, что поскольку большинство этих методов основаны на решении задач оптимизации и информационного поиска, вопросы разработки эффективных моделей представления исходных данных и эффективных численных алгоритмов являются ключевыми.

Настоящая статья посвящена разработке таких алгоритмов для важной прикладной задачи — классификации текстовых и гипертекстовых документов. Это одна из основных алгоритмических задач, возникающих при реализации многих прикладных систем, в частности, систем анализа и фильтрации Интернет-трафика, систем защиты от несанкционированных электронных почтовых рассылок (спама), систем анализа и рубрикации корпоративных архивов документов и др.

Для эффективного решения указанной задачи необходимо решить две базовые подзадачи:

— разработать эффективную формальную *модель представления* текстовых и гипертекстовых документов с учетом как их содержимого, так и ссылочной структуры;

— разработать эффективный алгоритм решения задачи *многотемной классификации* (*multi-label classification*), т.е. классификации в условиях существенно перекрывающихся классов, когда любой объект классификации (документ) может принадлежать более чем одному классу (теме) одновременно, а некоторые классы могут быть вложенными.

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 119899, Москва; e-mail: mash@cs.msu.ru

При разработке модели и алгоритмов одной из ключевых задач являлось достижение высокой скорости работы на больших объемах данных наряду с высокой точностью классификации.

Настоящая статья имеет следующую структуру. В разделе 2 дан краткий обзор существующих формальных моделей представления гипертекстовых документов и описаны предложенная нами модель представления данных, основанная на выделении частых эпизодов лексем (или N-грамм), и предложенный новый метод учета гиперссылок. В разделе 3 рассматривается постановка задачи классификации многотемных документов, а также разработанный нами новый метод многотемной классификации, основанный на подходе попарных сравнений с помощью набора бинарных классификаторов. В этом подходе результирующие вероятности принадлежности документа темам (релевантности классов) вычисляются с помощью обобщенной модели Брэдли–Терри, а нерелевантные классы отсекаются с помощью пороговой функции, заданной в пространстве релевантностей классов. Раздел 4 посвящен экспериментальному исследованию разработанных методов классификации и моделей представления на эталонных наборах данных.

## 2. Модель представления гипертекстовых данных.

**2.1. Краткий обзор существующих подходов.** Объекты классификации — текстовые и гипертекстовые документы и их фрагменты — являются слабо структурированными разнородными данными. Большинство алгоритмов классификации работают с формальным описанием объектов на основе *векторной модели представления документа* [1]. В этой модели документ описывается числовым вектором $\boldsymbol{a}$ фиксированной длины $n$, где $n$ — число признаков, а $i$-я компонента вектора определяет вес $i$-го признака. Для реализации модели представления необходимо, во-первых, выбрать *признаковое пространство* и, во-вторых, определить *алгоритм вычисления весов*. Качество выбранной модели представления при фиксированном алгоритме классификации и фиксированном эталонном тестовом наборе документов можно оценить по следующим критериям:

— *точность классификации*: основной критерий (зависит также от алгоритма классификации);

— *размерность признакового пространства*: при одинаковой точности предпочтительнее признаковое пространство меньшей размерности;

— *размер получаемой модели классификации*: при одинаковой точности предпочтительнее компактные модели;

— *время обучения и классификации*: важный критерий, зависящий, в том числе, от двух предыдущих;

— *поддержка морфологии языка*: этот критерий тесно связан со всеми предыдущими, в частности, поддержка морфологии ведет к более точным и компактным моделям классификации.

Самым распространенным способом формирования признакового пространства является метод ключевых слов [1, 2]. В качестве признаков в данном методе используются лексемы, входящие в документы, а размерность признакового пространства равна размерности словаря. Однако данный метод, например, не учитывает морфологию языка, а также семантические связи между словами. Поддержку морфологии можно реализовать с помощью методов стемминга [2], основанного на приведении слов к их базовой словоформе. Но тогда для каждого языка необходим морфологический анализатор, что, во-первых, ведет к дополнительной вычислительной нагрузке, во-вторых, возникает задача определения языка документа (если таковой не указан в свойствах документа) и, в-третьих, для некоторых языков построение морфологического анализатора является достаточно сложной задачей.

Проблему зависимости модели представления от языка решает метод N-грамм [3]. В данном методе в качестве признаков берутся подряд идущие буквосочетания фиксированной длины $N$. При этом каждое слово разбивается на набор признаков, а однокоренные слова образуют сходный набор признаков. Основным достоинством данного метода представления является отсутствие необходимости дополнительной лингвистической обработки текста и независимость от конкретного языка. Разбиение на N-граммы гораздо проще, чем выделение базовой словоформы, а из-за ограниченности алфавита во всех языках максимальное число различных признаков также ограничено. В качестве недостатков метода N-грамм можно указать увеличение числа непустых признаков у документа, а также то, что в модели не учитываются семантические связи между признаками.

Помимо представления содержимого документов, важным моментом является учет в модели наличия гиперссылок между документами. Одним из наиболее эффективных современных подходов к данной проблеме является метод учета ссылочной структуры документов, предложенный в [4]. В этом методе сначала классифицируются документы-соседи рассматриваемого документа, а затем каждая гиперссылка в исходном документе заменяется идентификатором класса документа-соседа, на который она указывает. Таким образом, гиперссылки заменяются на специальные идентификаторы (или специальные лексемы), которые характеризуют темы документов, на которые они ссылаются. Такое представление документов основывается на информации, которая является как локальной, так и нелокальной по отношению к

данному документу. Важным недостатком данного подхода является необходимость получения и классификации содержимого документов-соседей, что вносит дополнительную вычислительную нагрузку, а иногда оказывается невозможным, если документ, на который указывает ссылка, недоступен.

**2.2. Модель представления, основанная на выделении частых эпизодов.** Для преодоления недостатков традиционных подходов в настоящем исследовании предложен метод построения модели представления, основанный на *выделении частых эпизодов*. В этом случае множество частых комбинаций лексем (или N-грамм) формирует новое признаковое пространство, где каждому эпизоду, в который входит одна или более лексема (или N-грамма), соответствует одна координата в признаковом пространстве. Тем самым такой метод является расширением традиционной векторной модели, поскольку часто встречающиеся лексемы (или N-граммы) также входят в модель как эпизоды единичного размера. Для этого на этапе обучения из каждого документа выделяются все *структурные единицы*, входящие в его состав. Структурными единицами гипертекстового документа считаем *предложения, заглавия, названия колонок в таблице, элементы списка* и др. В реализации модуля синтаксического разбора документа понятие структурной единицы может настраиваться в зависимости от специфики задачи. Каждая структурная единица представляет собой отдельную *транзакцию* $t$, состоящую из лексем (или N-грамм). Весь тренировочный набор документов представляется в виде множества таких транзакций $\{t\}$ всех документов, собранных вместе. Далее с помощью алгоритма поиска частых эпизодов FP-tree [5] в $\{t\}$ выделяются эпизоды лексем, которые удовлетворяют заданным параметрам — *минимальной частоте встречаемости* и *максимальному размеру эпизода*. Все полученные эпизоды нумеруются и составляют новое признаковое пространство. Каждый документ представляется вектором, элементами которого являются нормированные частоты вхождения построенных эпизодов во все структурные единицы данного документа. Пример формирования признаков для предложенного алгоритма на основе частых эпизодов приведен ниже.

| | |
|---|---|
| Множество транзакций | Верховная Рада Украины лишила льгот и неприкосновенности депутатов и Президента |
| | Суд признал законным запрет гей-парада в Москве |
| | Взрывом на заводе в Томске занялись прокуратура и суд |
| | Немецкий суд в Дюссельдорфе оставил в силе запрет на ношение хиджаба учительницами в школе |
| | Пожар на заводе в Индии привел к серии взрывов |
| | Политическую неприкосновенность депутатов нужно оставить |
| | Взрыв на заводе в Воронежской области унес жизни трех человек |
| Признаки (эпизоды лексем; максимальный размер эпизода равен двум) | суд (3), взрыв (3), завод (3), неприкосновенность (2), депутат (2), запрет (2), Президент (1), закон (1), льготы (1), прокуратура (1), сила (1), ношение (1), учительница (1), школа (1), пожар (1), серия (1), политический (1), область (1), жизнь (1), человек (1); взрыв завод (3), неприкосновенность депутат (2), суд запрет (2) |

Вес $i$-го признака определяется стандартно как нормированная частота встречаемости этого признака в документе: $w_i = \dfrac{f_i}{\sqrt{\sum\limits_{i=1}^{n} f_i^2}}$, где $f_i$ — частота встречаемости $i$-го признака в документе, $n$ — количество непустых признаков в данном документе.

**2.3. Метод учета гиперссылок.** Как было сказано ранее, при представлении гипертекстовых документов важно учитывать не только текст, содержащийся в самих документах, но и наличие гиперссылок между документами. Согласно результатам, представленным в [4], учет гиперссылок позволяет получить более точное (для классификации) представление документа, по сравнению с учетом только локального текста классифицируемого документа. Идея предлагаемого нами подхода для учета гиперссылок состоит в использовании следующего подхода (рис. 1). Сначала обучаем текстовый классификатор при N-граммном представлении документов, учитывая при обучении только их локальный текст. Затем применяем данный классификатор к текстовому представлению каждой гиперссылки, встречающейся в документе, т.е. классифицируем ссылку, представляя ее URL в виде последовательности N-грамм и получая набор релевантных для нее классов только на основе ее URL. При формировании окончательного представления текущего документа каждую гиперссылку заменяем набором специальных идентификаторов, соответствующих идентификаторам предсказанных классов, как в методе Чакрабарти [4].

Таким образом, предлагаемый подход аналогичен [4], но в отличие от него позволяет учитывать встречающиеся в документах гиперссылки без необходимости получения и анализа содержимого документов-соседей, что очень важно для осуществления анализа документов в режиме реального времени.

**3. Задача классификации многотемных (multi-label) документов.**

**3.1. Общая постановка задачи многотемной классификации.** В задаче многотемной (multi-label) классификации в обучающей совокупности $Z = \{x_i, y_i\}_{i=1}^m$ для каждого примера $x_i$ задан не единственный класс, а множество релевантных классов $y_i \subseteq \{1, \ldots, q\}$, а целью алгоритма машинного обучения является построение классификатора $f_Z : X \to 2^q$, предсказывающего все релевантные классы, где $X$ — исходное пространство признаков и $q$ — число классов. Традиционным подходом к решению таких задач является декомпозиция "каждый против остальных" исходной multi-label задачи в $q$ задач бинарной классификации с целью независимо определить, является ли каждый из $q$ классов релевантным $x$. Для каждого класса $l$ формируется тренировочный набор, где $x_i$ помечен как "положительный", если в $Z$ выполнялось $l \in y_i$; иначе $x_i$ помечен как "отрицательный". Далее для каждого класса обучается отдельный бинарный классификатор. Он отделяет примеры, принадлежащие конкретному классу, от остальных примеров, которые этому классу не принадлежат. В результате суперклассы "принадлежащие классу $l$" и "не принадлежащие классу $l$" не перекрываются, и традиционные методы бинарной классификации могут быть успешно применены. При классификации решение о принадлежности объекта конкретному классу принимается независимо от остальных классов. Однако этот подход критикуется, во-первых, за низкую точность, поскольку не учитываются корреляции между классами, и, во-вторых, за ресурсоемкость, поскольку при обучении необходимо решать $q$ задач размера $m$ вместо одной.

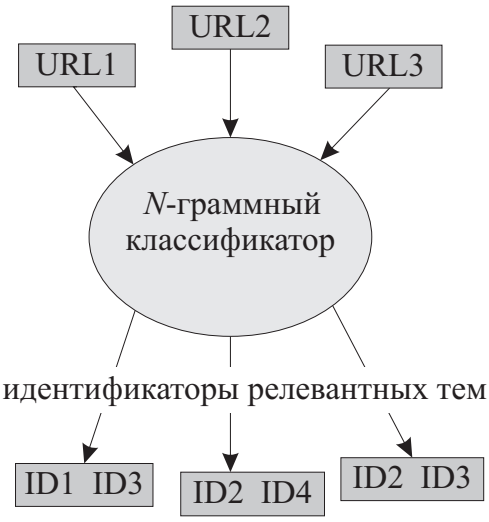


Рис. 1. Иллюстрация предложенного метода учета гиперссылок при представлении документов

В настоящей работе исследуется возможность использования для решения задачи multi-label классификации подхода на основе декомпозиции типа "каждый против каждого" с отсечением наименее релевантных классов. Решение задачи multi-label классификации на основе ранжирования включает в себя два этапа. Первый этап состоит в обучении алгоритма ранжирования, который упорядочивает все классы по степени их релевантности (например, по вероятности принадлежности) для заданного классифицируемого объекта. Второй этап заключается в построении функции multi-label классификации, отделяющей релевантные классы от нерелевантных. Как правило, на втором этапе используются пороговые функции [6], поскольку порог отсечения обычно зависит от классифицируемых объектов и использование фиксированного порога (одинакового для всех объектов) может приводить к низкой точности классификации. Нами предлагается, во-первых, новый алгоритм ранжирования, основанный на модифицированном для случая существенно пересекающихся классов методе попарных сравнений с помощью набора бинарных классификаторов и вычисления степеней принадлежности классам с использованием обобщенной модели Брэдли–Терри с "ничьей"; во-вторых, новый алгоритм построения пороговой функции отсечения нерелевантных классов, строящий пороговую функцию не в исходном пространстве характеристик (как это делает большинство традиционных алгоритмов), а в пространстве релевантностей классов, что позволяет упростить вид пороговой функции, значительно сократить вычисления и в большинстве случаев увеличить точность.

**3.2. Традиционный подход на основе попарных сравнений для взаимно исключающих классов.** Пусть задан обучающий набор $S = \{x_i, y_i\}_{i=1}^m \in X \times Y$, который состоит из $m$ классифицированных примеров $x_i$ из некоторого множества $X$. *Метки классов* $y_i$ принадлежат некоторому множеству $Y = \{1, \ldots, q\}$, $q > 2$. При детерминированной постановке задачи распознавания алгоритм обучения использует обучающий набор $S$ для построения классификатора $f_S : X \to Y$. В вероятностной постановке цель алгоритма обучения — оценить *распределение* $P\big(f(x) = y \,|\, S\big)$. Это означает, что для любого заданного $x$ алгоритм вычисляет апостериорные вероятности классов:

$$\boldsymbol{p}(x) = \big\{p_j(x)\big\}_{j=1}^q, \quad p_j(x) = P\big(y = j|x\big), \quad \sum_{j=1}^q p_j(x) = 1. \tag{1}$$

В традиционной multi-class проблеме сценарий применения подхода на основе попарных сравнений следующий. Сначала необходимо осуществить декомпозицию исходной задачи обучения с тренировочным набором $S$ в $\dfrac{q(q-1)}{2}$ бинарных подзадач — по одной подзадаче для каждой пары классов $k$ и $j$. Каждая подзадача включает только примеры, помеченные $k$ либо $j$. Примеры, помеченные классом $k$, рассматриваются как "положительные" с меткой $+1$; примеры, помеченные классом $j$, — как "отрицательные" и помечаются $-1$. Цель состоит в том, чтобы для каждой подзадачи построить бинарный классификатор, разделяющий "отрицательные" и "положительные" примеры, т.е. отделяющий класс $k$ от класса $j$ и игнорирующий все другие классы. Это идея иллюстрируется на рис. 2.

Примеры, лежащие на "неправильной" стороне разделяющей гиперплоскости, рассматриваются как исключения. При использовании вероятностного бинарного классификатора оцениваются отдельные вероятности:

$$r_{kj} = P\big(f_S(x) = k \,|\, x \in k \cup j\big). \tag{2}$$

Так как классы $j$ и $k$ предполагаются взаимно исключающими, дополнительные вероятности определяются в виде

$$r_{jk}(x) = 1 - r_{kj}(x). \tag{3}$$

Таким образом, мы имеем $\dfrac{q(q-1)}{2}$ обученных классификаторов, которые оценивают вероятности (2) и (3) для любого заданного примера $x$. Чтобы классифицировать новый пример $x$, используя подход попарных сравнений, необходимо применить все обученные классификаторы, оценить вероятности (2) и (3) и затем объединить их, чтобы получить результирующее предсказание. Для реализации этого было предложено много стратегий, включая четкие и нечеткие схемы голосования [7], теоретические игровые модели [8] и статистические модели [9]. Рассмотрим простейшую схему четкого голосования и статистический метод для оценки вероятностей классов (1), использующий модель Брэдли–Терри [9]. Ниже мы расширим эти две методики для multi-label случая. Схема четкого голосования работает следующим образом. Для заданного $x$ и для всех $k$ классов вычисляется функция голосования:

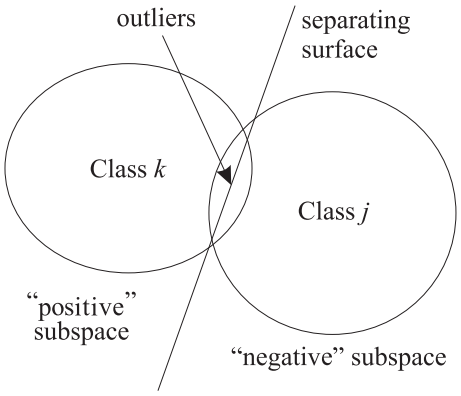


Рис. 2. Разделение взаимно исключающих классов

$$\mathrm{vote}_k(x) = \Big|\big\{j \,|\, j \neq k \wedge r_{kj}(x) \geqslant 0.5\big\}\Big|. \tag{4}$$

Класс, получивший максимальное число голосов (4), присваивается примеру $x$. Хотя эта схема вычислительно эффективна, в ней имеются несколько существенных недостатков: она не является точной; она допускает неопределенности, когда несколько классов получают максимальное количество голосов; она не позволяет оценить вероятности классов (1).

Схема для оценки вероятностей в случае взаимно исключающих классов, основанная на модели Брэдли–Терри [9], не имеет таких недостатков. Эта схема позволяет получить результирующие вероятности (1) на основе подхода попарных сравнений:

$$p(k \,\mathrm{beats}\, j) = \frac{p_k}{p_k + p_j}\,, \tag{5}$$

где приближенные значения $p_k$ могут быть найдены как решение следующей оптимизационной задачи:

$$\min_{\boldsymbol{p}} \left[ -\sum_{k<j} \left( r_{kj} \,\log\, \frac{p_k}{p_k + p_j} + r_{jk} \,\log\, \frac{p_j}{p_k + p_j} \right) \right] \quad \text{при условии} \quad \sum_{j=1}^{q} p_j = 1, \quad p_j \geqslant 0. \tag{6}$$

Для решения задачи (6) было разработано много эффективных итерационных алгоритмов (см. [9 – 11]).

**3.3. Предлагаемый метод ранжирования на основе попарных сравнений для существенно перекрывающихся классов.** Рассмотрим задачу ранжирования на основе попарных сравнений для случая существенно перекрывающихся классов. Следует отметить, что ранее этот подход не давал позитивных результатов для задач многотемной классификации, поскольку не удавалось решить проблему декомпозиции — построить точный классификатор, разделяющий два существенно перекрывающихся (не

взаимно исключающих) класса. Примитивный подход заключается в прямом сведении этой задачи к случаю взаимно исключающих классов. Один из способов — просто исключить из рассмотрения (убрать из тренировочного набора) примеры из перекрывающейся области. Другой способ — рассмотрение примеров из перекрывающейся области как исключений (т.е. такие примеры остаются в тренировочном наборе как положительные примеры для всех классов, к которым они относятся). Однако оба способа не являются корректными и не приводят к приемлемым результатам. Причина ясна: перекрывающаяся область может быть наиболее важной областью одного или обоих классов.

Если примеры из этой области рассматривать как исключения или игнорировать, то структура классов, важная для алгоритма обучения, может быть потеряна. Поэтому в нашем методе каждая пара возможно перекрывающихся (и даже вложенных) классов $j$ и $k$ разделяется с помощью *двух бинарных классификаторов*, которые отделяют *пересекающиеся* и *непересекающиеся* области. Используя их, можно выделить четыре области: область "только класса $j$"; область "только класса $k$"; "перекрывающаяся область" ($j$ и $k$); область, не принадлежащая "ни классу j, ни классу k" (рис. 3). Заметим, что в случае непересекающихся классов обе разделяющие гиперплоскости совпадут. Для построения двух разделяющих поверхностей сформулируем для каждой пары классов $j$ и $k$ две задачи обучения, которые включают в себя только примеры, помеченные в обучающем наборе $Z$ либо $j$, либо $k$, либо обоими классами одновременно (число таких примеров, как правило, существенно меньше $m$). В первой подзадаче примеры, помеченные только классом $k$, рассматриваются как "положительные", все остальные — как "отрицательные". В результате, построенный классификатор предсказывает вероятность $r^+_{kj}(x)$ и дополнительную вероятность $r^-_{kj}(x)$:

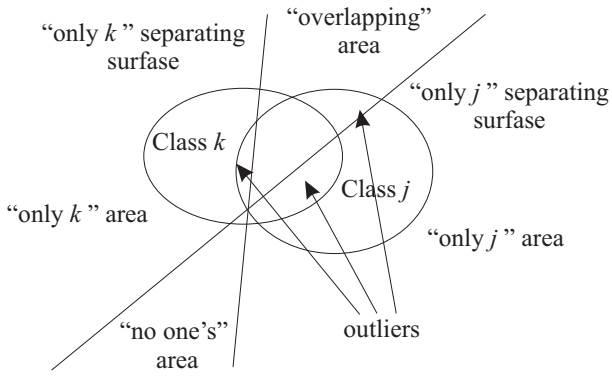


Рис. 3. Разделение существенно перекрывающихся классов

$$\begin{aligned} r^+_{kj}(x) &= P\big(k \in f_S(x) \wedge j \notin f_S(x) \,|\, x \in k \cup j\big), \\ r^-_{kj}(x) &= 1 - r_{kj}{+}(x) = P\big(j \in f_S(x) \,|\, x \in k \cup j\big). \end{aligned} \tag{7}$$

Вторая подзадача формулируется аналогично, но только для класса $j$:

$$\begin{aligned} r^+_{jk}(x) &= P\big(j \in f_S(x) \wedge k \notin f_S(x) \,|\, x \in k \cup j\big), \\ r^-_{jk}(x) &= 1 - r_{jk}{+}(x) = P\big(k \in f_S(x) \,|\, x \in k \cup j\big). \end{aligned} \tag{8}$$

В результате вероятности принадлежности примера $x$ каждой из четырех областей могут быть вычислены следующим образом:

$$\begin{aligned} &P(x \in \text{overlapping } k,j) = r^-_{kj}(x) r^-_{jk}(x), \\ &P(x \in \text{no one's } k,j) = r^+_{kj}(x) r^+_{jk}(x), \\ &P(x \in \text{only } k) = r^+_{kj}(x) r^-_{jk}(x), \\ &P(x \in \text{only } j) = r^-_{kj}(x) r^+_{jk}(x). \end{aligned} \tag{9}$$

Важно отметить, что при такой формулировке оба бинарных классификатора разделяют взаимно исключающие суперклассы, и для решения и оценки попарных вероятностей могут быть использованы стандартные алгоритмы бинарной классификации, например, на основе Support Vector Machines [12] или Kernel Fisher Discriminant [13]. Формулируя и решая таким образом $q(q-1)$ задач бинарной классификации (каждая задача имеет размер меньший, чем $m$), мы получаем попарные вероятности сравнения $r^*_{jk}(x)$.

Затем вероятности принадлежности этим областям, предсказанные всеми бинарными классификаторами, объединяются вместе, чтобы оценить результирующие вероятности принадлежности классам. Для этого мы предлагаем использовать обобщенную модель ранжирования Брэдли–Терри с "ничьей", которую сформулировали Рао и Купер [11]. Эта модель была разработана для анализа результатов попарных

сравнений с возможными связями:

$$\begin{aligned} P(k\,\text{beats}\,j) &= \frac{p_k}{p_k+\theta p_j}\,,\\ P(j\,\text{beats}\,k) &= \frac{p_j}{\theta p_k+p_j}\,,\\ P(j\,\text{ties}\,k) &= \frac{(\theta^2-1)p_jp_k}{(\theta p_k+p_j)(p_k+\theta p_j)}\,, \end{aligned} \tag{10}$$

где $\theta>1$ — вычисляемый параметр модели. Для оценки результирующих вероятностей принадлежности классам на основе этой модели можно использовать итеративный minorization-maximization (MM) алгоритм [10]. На каждой итерации $l$ алгоритм вычисляет вероятности $p_k^{(l+1)}$ и значения параметров $\theta^{(l+1)}$ и $C_l$, используя следующие формулы:

$$p_k^{(l+1)}=\frac{\sum\limits_{j\neq k}s_{kj}}{\sum\limits_{j\neq k}\left(\dfrac{s_{kj}}{\theta^{(l)}p_j^{(l)}+p_k^{(l)}}+\dfrac{\theta^{(l)}s_{jk}}{\theta^{(l)}p_k^{(l)}+p_j^{(l)}}\right)}\,,\quad C_l=\frac{2}{T}\sum_{k=1}^{q}\sum_{j\neq k}\frac{p_j^{(l)}s_{kj}}{\theta^{(l)}p_j^{(l)}+p_k^{(l)}}\,,\quad \theta^{(l+1)}=\frac{1}{2C_l}+\sqrt{1+\frac{1}{4C_l^2}}\,,$$

где $s_{kj}$ — число сравнений и $k$ “не уступает” $j$. В нашем случае имеем

$$\begin{aligned} s_{kj} &= n_{k\cup j}\Big[P\big(x\in\text{“only }k\text{”}\big)+P\big(x\in\text{“overlapping }k,\,j\text{”}\big)+P\big(x\in\text{“no one's }k,\,j\text{”}\big)\Big]=\\ &= n_{k\cup j}\big[r_{kj}^-(x)r_{jk}^-(x)+r_{kj}^+(x)\big], \end{aligned}$$

где $n_{k\cup j}$ — общее число обучающих примеров, помеченных либо $k$, либо $j$, и $T$ — общее число “ничьих” для всех классов. В нашем случае выполнено

$$T=\sum_{k\neq j}n_{k\cup j}P\big(x\in\text{“overlapping }k,\,j\text{”}\cup\text{“no one's }k,\,j\text{”}\big)=\sum_{k\neq j}n_{k\cup j}\big[r_{kj}^-(x)r_{jk}^-(x)+r_{kj}^+(x)r_{jk}^+(x)\big].$$

Необходимо отметить, что MM-алгоритм вычислительно сложен. Поэтому для очень больших наборов данных, таких как Reuters [14], мы рекомендуем выполнять дополнительный шаг, а именно: сначала применяем простейшую схему голосования $\text{vote}_k(x)=\Big|\big\{j\,|\,j\neq k\wedge r_{jk}^-(x)>0.5\big\}\Big|$, а затем выбираем некоторую долю наиболее релевантных классов (например, $1/3$ или $1/2$) и применяем MM-алгоритм только к этим классам, считая вероятности остальных классов нулевыми. Как показали наши эксперименты, такой прием очень незначительно снижает точность предсказания, а скорость классификации при этом значительно возрастает.

**3.4. Отсечение наименее релевантных классов.** Решив с помощью предложенного выше метода на основе попарных сравнений задачу ранжирования, т.е. оценив вероятности принадлежности анализируемого документа всем классам (темам) и упорядочив классы по степени возрастания данной вероятности, мы еще не решили исходную задачу многотемной классификации, поскольку необходимо оставить только наиболее релевантные классы. Для этого используются пороговые функции. В [6] для решения задачи определения порога релевантности предлагается строить нелинейную пороговую функцию в пространстве признаков классифицируемых объектов. Однако для задачи классификации электронных документов временна́я и пространственная сложность вычисления пороговой функции в пространстве признаков очень велика и соизмерима по затратам с решением задачи ранжирования.

Для преодоления этого недостатка нами предлагается новый подход для сведения задачи ранжирования к задаче multi-label классификации, основанный на построении *линейной решающей функции* в пространстве *релевантностей классов*, используя результат работы алгоритма ранжирования как новое множество признаков анализируемого электронного документа. В этом методе предлагается находить пороговую функцию не в исходном пространстве признаков, а в $q$-мерном пространстве векторов релевантностей классов. *Пространство релевантностей классов* — это пространство векторов, координатами которых являются значения релевантностей классов для документов (его размерность равна числу классов и значительно меньше размерности пространства характеристик). Основная мотивация предложенного подхода состоит в упрощении вида решающей функции и в сокращении времени обучения, времени классификации и объема памяти, необходимых для работы алгоритма.

Опишем процесс определения пороговой функции в пространстве векторов релевантностей классов. Для каждого $x$ из исходного тренировочного набора $Z = \{x_i, y_i\}_{i=1}^m$ вычисляются значения степеней принадлежности документа $x_i$ каждому из $q$ классов (с помощью описанного выше метода ранжирования на основе попарных сравнений): $\boldsymbol{p}(x_i) = \big(p_1(x_i), \ldots, p_q(x_i)\big)$. Для каждого из обучающих документов находим приближенное пороговое значение $tr_i = tr\big(\boldsymbol{p}(x_i)\big)$, которое дает наименьшую ошибку классификации. Заметим, что такая ошибка возможна, поскольку алгоритм ранжирования не обязательно находит идеальный порядок степеней релевантностей классов, т.е. некий нерелевантный класс, не принадлежащий $y_i \subseteq \{1, \ldots, q\}$, может быть проранжирован выше (иметь более высокую степень принадлежности), чем релевантный класс из $y_i \subseteq \{1, \ldots, q\}$. Таким образом, ищется значение $tr(x_i)$, при котором достигается минимум целевой функции, т.е. минимум количества неправильно классифицированных релевантных и нерелевантных классов:

$$tr_i = \arg\min_t \left( \Big| \big\{ l \in y_i \,:\, p_l(\boldsymbol{x}_i) \leqslant t \big\} \Big| + \Big| \big\{ l \in \overline{y}_i \,:\, p_l(\boldsymbol{x}_i) > t \big\} \Big| \right).$$

В результате для всего исходного тренировочного набора $Z = \{x_i, y_i\}_{i=1}^m$ получаем набор оценок значений пороговой функции $tr(x_i)$: $Z_{tr} = \{x_i, tr_i\}_{i=1}^m$. Будем искать пороговую функцию в классе линейных:

$$tr\big(\boldsymbol{p}(x)\big) = a_1 p_1(x) + \ldots + a_q p_q(x) + a_{q+1},$$

где $a_1, a_2, \ldots, a_q, a_{q+1}$ — коэффициенты линейной функции, найденные с помощью метода наименьших квадратов:

$$\min_{\boldsymbol{a}} \sum_{i=1}^m \left[ tr_i - \left( \sum_{j=1}^q a_j p_j(x_i) + a_0 \right) \right]^2.$$

В результате на этапе классификации для анализируемого документа $x_{\text{test}}$, для которого необходимо найти набор релевантных классов, мы применяем предложенный выше метод ранжирования на основе попарных сравнений и находим значения релевантностей всех $q$ классов: $p_1(x_{\text{test}}), \ldots, p_q(x_{\text{test}})$. Далее применяем пороговую функцию $tr(x_{\text{test}}) = a_1 p_1(\boldsymbol{x}_{\text{test}}) + \ldots + a_q p_q(\boldsymbol{x}_{\text{test}}) + a_{q+1}$ и находим порог отсечения. Решением исходной задачи многотемной классификации считаем только те классы, которые были проранжированы выше найденного порога:

$$y_{\text{test}} = \big\{ l \,|\, p_l(x_{\text{test}}) > tr(x_{\text{test}}),\, l \in Y \big\}.$$

Таблица 1

Характеристики эталонных наборов тестовых данных

| Набор данных | Число классов | Число признаков | Среднее число классов на образец | Число обучающих образцов | Число тестовых образцов |
|---|---|---|---|---|---|
| BankResearch | 11 | 16139 – 40614 | 1 | 7500 | 2500 |
| Reuters-2000 | 101 | 47236 | 3.0 | 3000 | 3000 |

**4. Результаты экспериментального исследования.** В настоящем разделе приведено описание экспериментального исследования предложенных методов на эталонных тестовых наборах данных. Результат данной работы, по сути, состоит из двух частей: новой предложенной модели представления гипертекстовых данных и нового разработанного метода многотемной классификации. К сожалению, на настоящий момент нет общепринятого эталонного тестового набора, содержащего многотемные гипертекстовые данные. Поэтому тестирование модели представления и алгоритма многотемной классификации пришлось проводить отдельно на двух различных тестовых наборах данных Reuters-2000 (тестовый набор многотемных документов, не содержащих гиперссылок) [14, 18] и BankResearch (тестовый набор однотемных гипертекстовых документов, содержащих гиперссылки) [15] (см. табл. 1).

Для чистоты эксперимента необходимо продемонстрировать, что и предложенная модель представления, и разработанный алгоритм позволяют улучшить качество классификации сами по себе, а не только в совокупности. Поэтому тестирование модели представления и сравнение ее с существующими популярными моделями проводилось на наборе BankResearch на базе традиционного алгоритма классификации типа на основе "$k$ ближайших соседей" [16]. Данный алгоритм был выбран как наиболее чувствительный

к модели представления. Тестирование предложенного алгоритма многотемной классификации проводилось на наборе Reuters-2000 на базе традиционной модели представления с помощью ключевых слов, а результат сравнивался с ведущими современными алгоритмами многотемной классификации. Следует отметить, что в случае многотемной (multi-label) классификации традиционные оценки точности, такие как процент правильно предсказанных классов или ошибки первого и второго рода, не могут быть использованы, поэтому мы использовали следующие общепринятые оценки точности многотемных классификаторов: Hamming Loss, Coverage и Ranking Loss [16]. Их определения приведены ниже (здесь $m$ — размер тестового набора).

Значение ошибки Coverage отражает, как далеко необходимо "спуститься" по отранжированному списку тем, чтобы найти все релевантные темы:

$$\mathrm{Coverage}_{\mathrm{test}} = \frac{1}{m}\sum_{i=1}^{m} C(\boldsymbol{x}_i) - 1, \quad \text{где} \quad C(x_i) = \Big|\big\{l \in Y \,:\, p_l(x_i) \geqslant p_s(x_i)\big\}\Big|, \quad s = \arg\min_{k \in y_i} p_k(x_i).$$

Чем меньше Coverage, тем лучше качество ранжирования. Значение ошибки Ranking Loss (RL) представляет среднюю долю пар тем, которые некорректно упорядочены:

$$\mathrm{RL}_{\mathrm{test}} \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|y_i|\,|\overline{y}_i|}\,\mathrm{RL}(x_i), \quad \text{где} \quad \mathrm{RL}(x_i) = \Big|\big\{(l,s) \in y_i \times \overline{y}_i \,:\, p_l(x_i) \leqslant p_s(x_i)\big\}\Big|, \quad \overline{y}_i = \frac{Y}{y_i}\,.$$

Чем меньше значение $\mathrm{RL}_{\mathrm{test}}$, тем лучше качество ранжирования. Для оценки *точности multi-label классификации* обычно используется критерий Hamming Loss [16], который оценивает различие между предсказанным и истинным наборами тем:

$$\mathrm{HL}_{\mathrm{test}}(f) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|Y|}\,\Big|\big(f(x_i)\big)\nabla(y_i)\Big|,$$

где $a\nabla b = (a \cup b) \setminus (a \cap b)$, $a \subseteq Y$, $b \subseteq Y$ и $f(x)$ — функция multi-label классификации. Чем меньше значение $\mathrm{HL}_{\mathrm{test}}(h)$, тем лучше точность multi-label классификатора. Заметим, что Ranking Loss и Coverage не зависят от порога отсечения и оценивают только качество ранжирования, а Hamming Loss зависит и оценивает также и точность отсечения.

**4.1. Сравнение моделей представления.** Данный раздел посвящен экспериментальному сравнению предложенных и существующих моделей представления гипертекстовых данных. Было проведено сравнение следующих моделей.

1. Выделение признаков на основе ключевых слов с частотной мерой сходства.
2. Выделение признаков на основе $N$-грамм с частотной мерой сходства.
3. Выделение признаков на основе эпизодов ключевых слов с частотной мерой сходства.
4. Выделение признаков на основе эпизодов $N$-грамм с частотной мерой сходства.

Таблица 2

Результаты сравнения различных моделей представления гипертекстовых данных без учета гиперссылок

| Мера сходства | Частотная | | | |
|---|---|---|---|---|
| Выделение признаков | Ключевые слова | $N$-граммы ($N = 3$) | Эпизоды ключевых слов | Эпизоды $N$-грамм (N=3) |
| Обозначение модели | A1 | A2 | A3 | A4 |
| Критерий | | | | |
| Hamming Loss | $0.0246 \pm 0.0005$ | $0.0231 \pm 0.0003$ | $\mathbf{0.0169 \pm 0.0005}$ | $0.0216 \pm 0.0004$ |
| Coverage | $0.592 \pm 0.018$ | $0.506 \pm 0.017$ | $0.349 \pm 0.006$ | $0.397 \pm 0.005$ |
| Ranking Loss | $0.0539 \pm 0.0017$ | $0.0460 \pm 0.0015$ | $0.0317 \pm 0.0005$ | $0.0361 \pm 0.0004$ |
| Average Precision | $0.8812 \pm 0.0036$ | $0.8939 \pm 0.0014$ | $0.9253 \pm 0.0017$ | $0.9112 \pm 0.0016$ |
| Размерность пространства | 46603 | **16139** | 17302 | 34085 |

Таблица 3

Результаты сравнения различных моделей представления гипертекстовых данных с учетом гиперссылок

| Мера сходства | Частотная | | | |
|---|---|---|---|---|
| Выделение признаков | Ключевые слова | $N$-граммы ($N = 3$) | Эпизоды ключевых слов | Эпизоды $N$-грамм (N=3) |
| Обозначение модели | A5 | A6 | A7 | A8 |
| Критерий | | | | |
| Hamming Loss | $0.0228 \pm 0.0005$ | $0.0212 \pm 0.0001$ | $0.0172 \pm 0.0010$ | $0.0202 \pm 0.0004$ |
| Coverage | $0.482 \pm 0.025$ | $0.409 \pm 0.008$ | $\mathbf{0.347 \pm 0.022}$ | $0.356 \pm 0.003$ |
| Ranking Loss | $0.0438 \pm 0.0022$ | $0.0371 \pm 0.0007$ | $\mathbf{0.0316 \pm 0.0020}$ | $0.0323 \pm 0.001$ |
| Average Precision | $0.9008 \pm 0.0029$ | $0.9096 \pm 0.0009$ | $\mathbf{0.9259 \pm 0.0045}$ | $0.9182 \pm 0.002$ |
| Размерность пространства | 46614 | **16150** | 17563 | 34274 |

Для каждой модели было проведено сравнение предложенного метода представления для учета гиперссылок (путем замены ссылок на релевантные им классы) с представлением, основанным только на текстовом содержимом документа. Для сравнения этих моделей представления гипертекстовых данных был использован эталонный тестовый набор HTML-документов BankResearch. Этот набор состоит из 11 000 документов на английском языке, которые разделены на 11 различных тематик, по 1000 документов на каждую. Объем набора составляет около 300 мегабайт. Набор содержит HTML-документы размером от 2 до 900 килобайт. Все документы являются копиями реальных Интернет-страниц. Каждый документ принадлежит только к одной тематике. В качестве базового метода классификации для сравнения моделей представления был выбран алгоритм ML-$k$NN ($k = 5$) [16], поскольку этот алгоритм наиболее зависим от модели представления данных. Параметры были выбраны методом кроссвалидации. Для оценки достоверности результатов тестирования был выбран $k$-раздельный парный $t$-тест перекрестной проверки [17]. Результаты экспериментов представлены в виде сводных таблиц по всем основным критериям точности. В строке "размерность пространства" приведена размерность пространства признаков построенной модели. Результаты сравнения приведены в табл. 2 и 3.

Значение, следующее за "$\pm$", показывает стандартное отклонение; наилучшие результаты по каждому критерию показаны жирным шрифтом. Для наглядности сравнения точности моделей представления вводится частичный порядок "$>$" на множестве всех сравниваемых алгоритмов для каждого критерия оценки. Иными словами, A1 $>$ A2 означает, что модель A1 статистически лучше, чем модель A2 для конкретного критерия (основываясь на $k$-раздельном парном $t$-тесте перекрестной проверки). Частичный порядок для всех сравниваемых моделей представления в терминах каждого критерия приведен в табл. 4.

Заметим, что частичный порядок "$>$" показывает только относительную разницу между моделями A1 и A2 для конкретного критерия оценки. Однако при этом возможна ситуация, что A1 показывает лучшие результаты, чем A2, по одному критерию, но худшие по другому. В этом случае трудно оценить, какая модель действительно лучше. Поэтому для общей оценки производительности с каждой моделью связывается значение, которое учитывает ее относительную производительность по всем метрикам, а именно: для каждого критерия оценки и для каждой возможной пары моделей A1 и A2 модель A1 получает положительную оценку "$+1$", а модель A2 получает "$-1$", если выполнено A1 $>$ A2. Основываясь на совокупных результирующих значениях каждой модели по всем критериям, определяется общий порядок "$\gg$" на наборе всех моделей, что показано в последней строке табл. 4.

Как следует из этой таблицы, предложенная модель представления на основе частых эпизодов с частотной мерой сходства кардинально превосходит традиционные подходы по всем основным критериям. Из таблицы также следует, что предложенный метод включения информации о гиперссылках в модель представления позволяет получить улучшение точности практически для любой базовой модели.

**4.2. Сравнение алгоритмов multi-label классификации.** Тестирование всех алгоритмов проводилось на эталонном наборе многотемных данных Reuters-2000, содержащем новостные статьи (документы) агентства Рейтер за период с августа 1996 г. по август 1997 г. Эти статьи были рубрицированы экспертами по темам. Число различных тем, к которым классифицированы документы набора Reuters-2000, равно 101. Каждый документ в этом наборе помечался набором релевантных для него тем. Разработанный

Таблица 4

Относительная точность между моделями представления на наборе BankResearch

| Критерий оценки | Модель представления |
|---|---|
| | A1, A2, A3, A4, A5, A6, A7, A8 |
| Hamming Loss | A3 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6, A3 > A7, A3 > A8,<br>A7 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6, A3 > A8<br>A8 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6<br>A6 > A1, A3 > A2, A3 > A4, A3 > A5<br>A4 > A1, A3 > A2, A3 > A5<br>A5 > A1, A3 > A2<br>A2 > A1 |
| Coverage | A7 > A1, A3 > A2, A3 > A3, A3 > A4, A3 > A5, A3 > A6, A3 > A8,<br>A3 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6, A3 > A8<br>A8 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6<br>A4 > A1, A3 > A2, A3 > A5, A3 > A6<br>A6 > A1, A3 > A2, A3 > A5<br>A5 > A1, A3 > A2<br>A2 > A1 |
| Ranking Loss | A7 > A1, A3 > A2, A3 > A3, A3 > A4, A3 > A5, A3 > A6, A3 > A8,<br>A3 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6, A3 > A8<br>A8 > A1, A3 > A2, A3 > A4, A3 > A5, A3 > A6<br>A4 > A1, A3 > A2, A3 > A5, A3 > A6<br>A6 > A1, A3 > A2, A3 > A5<br>A5 > A1, A3 > A2<br>A2 > A1 |
| Общий порядок | A7 $\gg$ A3 $\gg$ A8 $\gg$ A4 $\gg$ A6 $\gg$ A5 $\gg$ A2 $\gg$ A1 |

нами метод многотемной классификации сравнивался со следующими существующими методами:

— Multi-class Multi-label Perceptron (MMP) [19]: метод ранжирования многотемных документов, основанный на алгоритме персептрона и пошаговой модификации весов (оценивается ошибка ранжирования тем для каждого обучающего документа) с пороговой функцией в пространстве характеристик;

— AdaBoost.HM [20]: метод multi-label классификации, основанный на boosting-методах минимизации Hamming Loss;

— ML-KNN [16]: метод multi-label классификации, основанный на методе $k$ ближайших соседей и принципе максимизации апостериорных вероятностей принадлежности объекта к классам;

— 1-vs-all-SVM [21]: метод multi-label классификации, основанный на декомпозиции multi-label проблемы в набор независимых бинарных проблем, в котором в качестве метода бинарной классификации применяется метод опорных векторов.

Результаты наших экспериментов приведены ниже в табл. 5. Как следует из таблицы, разработанный метод превосходит существующие по всем основным характеристикам, особенно по качеству ранжирования (Coverage и Ranking Loss).

Таблица 5

Экспериментальные результаты работы алгоритмов классификации на наборе Reuters-2000

| Algorithm | Coverage | Ranking Loss | Hamming Loss |
|---|---|---|---|
| MMP | 5.65 | 0.0137 | 0.0121 |
| AdaBoost.HM | 6.08 | 0.0169 | 0.0146 |
| ML-KNN | 5.72 | 0.0167 | 0.0135 |
| 1vsAll SVM | 6.53 | 0.162 | **0.0095** |
| Наш метод | **4.59** | **0.096** | **0.0095** |

**5. Заключение.** В настоящей статье рассмотрены основные алгоритмические задачи, которые необходимо решать при реализации систем, связанных с анализом больших объемов разнородной текстовой

и гипертекстовой информации. Эти задачи сводятся к необходимости разработки эффективных моделей представления исходных гипертекстовых данных с учетом их текстового содержания и структуры гиперссылок, а также разработки эффективных вычислительных методов решения задачи рубрикации и многотемной классификации.

В результате предлагается новое представление гипертекстовых данных, являющееся расширением традиционного векторного представления. Оно состоит из базовых текстовых признаков (лексемы, если есть поддержка стемминга для языка, или N-граммы для морфологически сложных языков без стемминга), базовых нетекстовых признаков (идентификаторов классов гиперссылок) и составных признаков, вычисляемых с помощью алгоритма поиска частых эпизодов. Для классификации гиперссылок предложен новый подход, основанный на использовании N-граммного классификатора, применяемого для анализа структуры гиперссылки и не требующего загрузки содержимого документа, на который указывает гиперссылка. Предложенное представление обеспечивает высокую точность классификации наряду с высокой скоростью работы.

В работе также исследуется возможность использования подхода на основе декомпозиции "каждый против каждого" для решения задачи multi-label классификации. Предлагается новый алгоритм, основанный на модифицированном методе попарных сравнений с помощью набора бинарных классификаторов. Ранее этот подход не имел позитивных результатов для задач multi-label классификации, поскольку не удавалось решить проблему декомпозиции — построить точный классификатор, разделяющий два существенно перекрывающихся (не взаимно исключающих) класса. Поэтому в предложенном нами методе каждая пара возможно пересекающихся классов разделяется с помощью двух бинарных классификаторов, которые отделяют перекрывающиеся и не перекрывающиеся области. Затем вероятности принадлежности этим областям, предсказанные всеми бинарными классификаторами, объединяются вместе, чтобы оценить результирующие вероятности принадлежности классам. Для этого используется обобщенная модель ранжирования Брэдли–Терри с "ничьей". Далее нами предлагается новый подход для сведения проблемы ранжирования к проблеме multi-label классификации, основанный на построении линейной решающей функции в пространстве релевантностей классов с использованием результата работы алгоритма ранжирования как нового множества признаков анализируемого электронного документа. Предложенный подход позволяет достичь компромисса между точностью и скоростью работы алгоритмов, что очень важно для применения этого подхода в задаче фильтрации Интернет-трафика.

Все разработанные модели представления и методы многотемной классификации были экспериментально проверены на эталонных тестовых наборах данных и показали лучшие результаты по сравнению с традиционными моделями и методами.

## СПИСОК ЛИТЕРАТУРЫ


1. Information Retrieval Tutorials: Document Indexing Tutorial: представление документов для информационного поиска (http://www.miislita.com/information-retrieval-tutorial/indexing.html).
2. Vector Theory and Keyword Weights: выделение признаков из документов (http://www.miislita.com/information-retrieval-tutorial/indexing.html).
3. *Cavnar W.B., Trenkle J.M.* Ngram-based text categorization // Proc. of SDAIR-94 (the 3rd Annual Symposium on Document Analysis and Information Retrieval). Las Vegas, 1994. 161–175.
4. *Chakrabarti S., Dom B.E., Indyk P.* Enhanced hypertext categorization using hyperlinks // Proc. of the ACM Int. Conf. on Management of Data (SIGMOD). New York: ACM Press, 1998. 307–318.
5. *Pei J.* Pattern-growth methods for frequent pattern mining. Ph.D. Thesis. Simon Fraser University. Vancouver, 2002.
6. *Elisseeff A., Weston J.* A kernel method for multi-labelled classification // Proc. of the Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2002.
7. *Abe S., Inoue T.* Fuzzy support vector machines for multiclass problems // Proc. of the European Symposium on Artificial Neural Networks. Bruges (Belgium), 2002. 113–118.
8. *Petrovskiy M.* Probability estimation in error correcting output coding framework using game theory // Proc. of the 18th ACS Australian Joint Conf. on Artificial Intelligence. Lecture Notes in Artificial Intelligence. Vol. 3809. Berlin: Springer, 2005. 186–196.
9. *Huang T.-K., Weng R., Lin C.-J.* A generalized Bradley–Terry model: from group competition to individual skill // Proc. of the Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2004. 85–115.
10. *Hunter D.R.* MM-algorithms for generalized Bradley–Terry models // Annals of Statistics. 2004. **32**, N 1. 384–406.
11. *Rao P.V., Kupper L.L.* Ties in paired-comparison experiments: a generalization of the Bradley–Terry model // Amer. Statist. Assoc. 1967. **62**. 194–204.
12. *Platt J.* Probabilistic outputs for support vector machines and comparison to regularized likelihood methods // Advances in Large Margin Classifiers. Cambridge: MIT Press, 1999. 61–74.

13. *Zheng W., Zhao L., Zou C.* A modified algorithm for generalized discriminant analysis // Neural Computation. 2004. **16**, N 6. 1283–1297.
14. *Lewis D.D., Yang Y., Rose T.G., Li F.* RCV1: A new benchmark collection for text categorization research // Machine Learning. 2004. **5**. 361–397.
15. Bank Research Dataset: Набор данных BankResearch (http://lib.stat.cmu.edu/datasets/bankresearch.zip).
16. *Zhang M.-L., Zhou Z.-H.* A k-nearest neighbor based algorithm for multi-label classification // Proc. of the First IEEE Int. Conf. on Granular Computing. Beijing (China), 2005. 718–721.
17. *Everitt B.S.* The analysis of contingency tables. London: Chapman and Hall, 1977.
18. *Chang C.-C., Lin C.-J.* LIBSVM: a library for support vector machines (http://www.csie.ntu.edu.tw/c̆jlin/libsvm).
19. *Crammer C., Singer Y.* A family of additive online algorithms for category ranking // Machine Learning Research. 2003. **3**. 1025–1058.
20. *Schapire R.E., Singer Y.* BoosTexter: a boosting-based system for text categorization // Machine Learning. 2000. **39**. 135–168.
21. *Boutell M.R., Luo J., Shen X., Brown C.M.* Learning multi-label scene classification // Pattern Recognition. 2004. **37**. 1757–1771.

Поступила в редакцию
12.10.2007