*Andrey Igorevich Shalabay,post-graduate student, institute of mathematics and fundamental informatics, Syberian federal university*

*We consider the overall situation in the field of conducting distance learning. To improve the efficiency of this process is proposed to consolidate directories of learning and teaching materials and simultaneous distributed storage of full texts in educational institutions, leading distance learning network. For this purpose the use of Grid-technologies that can serve as the basis of the resource organization technology platform network of distance learning.*

*Keywords: Network and distance education, Grid, associations of distance education, distributed storage.*

**УДК 005**

# КЛАССИФИКАЦИЯ ТЕКСТОВ, СОДЕРЖАЩИХ ФОРМУЛЫ

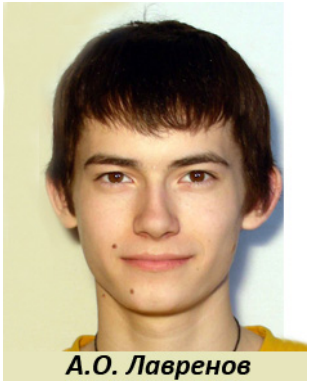
***Антон Олегович Лавренов**, аспирант базовой кафедры вычислительных и информационных технологий*
*Тел.: 908 325 8225,e-mail: lavrton@gmail.com*
***Борис Васильевич Олейников**, к. филос. наук, доц., доц. базовой кафедры вычислительных и информационных технологий*
*Тел.: 902 990 2597, e-mail: oleynik48@mail.ru*
*Институт математики и фундаментальной информатики*
*ФГАОУ ВПО «Сибирский федеральный университет»*
*http://math.sfu-kras.ru*

*В работе.указывается на важность учёта формул при классификации математических и иных документов, особенно для целей их поиска. Рассматривается подход к классификации текстов, содержащих формулы, на основе преобразования исходного документа в tex-формат, и создания блочной его структуры. Приводятся первоначальные результаты проведенных тестовых расчетов на основе предлагаемого подхода.*

*Ключевые слова: классификация текста, tex формат, блочная структура текста, классификация текстов с формулами.*


Б.В. Олейников

В настоящее время в связи со взрывным характером порождения цифровых текстовых документов (интернет, автоматизированный документооборот, цифровые библиотеки, образовательные сайты и порталы и т.п.) все более насущной является проблема их поиска. Основополагающую роль при построении тематического полнотекстового поиска документов играет классификация. В настоящее время разработано достаточно много методов классификации текстов, но практически все из них не учитывают (точнее игнорируют) наличие формул в текстах. Это может привести к значительному искажению классификации, например, в таких предельных случаях, когда основной объем текста занимают формулы, что бывает характерным для математических текстов. Поэтому необходимо разработать такие подходы к классификации, которые бы могли учитывать любое состояние текста включая


А.О. Лавренов

и предельные его состояния (от «текст содержит только слова из некоторого допустимого словаря» и до «текст содержит только формулы»).

Основная модель представления текста — это вектор в дискретном пространстве выделенных и нормированных, т. е. специальным образом приведенных [1] слов некоторого словаря: $\vec{d_i} = (w_1, w_2, \ldots, w_n)$, где $d_i$— векторное представление i-го документа, $w_j$— вес (см. далее) j-го слова в документе, n — общее количество различных слов во всех документах коллекции [2, 3]. Следуя [1] будем называть такие слова термами.

Выделение слов может производиться с применением различных фильтров. Например, могут исключаться все не значащие слова, такие как местоимения, союзы, предлоги и т.п. Для научных текстов дополнительно могут исключаться слова общей лексики. Нормировка слов (приведение их к определенному виду) зависит от задачи и может включать в себя следующие этапы: для существительных – приведение к именительному падежу, для глаголов – к инфинитивной форме, для прилагательных – выделение корневых форм, для синонимов – использование словаря синонимов и т.п.

На данной модели базируются практически все известные методы классификации текстов независимо от весовых предпочтений и подходов к агрегированию.

Все эти манипуляции с текстом обычно производятся на этапе его предобработки. На подготовительном этапе также могут рассматриваться преобразование признаков с помощью некоторых функций, определение наиболее информативных признаков (термов), или наоборот отбрасывание неинформативных признаков, задачи уменьшения признакового пространства. Некоторые из этих процедур описываются в работах [4, 5]

Во многих алгоритмах классификации текстов используется функция сравнения векторов или функция нахождения «расстояния» между векторами, которая может базироваться на метриках, например:

1. Евклидова $r(\vec{a}, \vec{b}) = \sum_{i=1}^{n} (a_i - b_i)^2$;
2. Манхэттенская (или метрика городских кварталов) $r(\vec{a}, \vec{b}) = \sum_{i=1}^{n} | a_i - b_i |$.
3. Хэмминга $r(\vec{a}, \vec{b}) = \sum_{i=1}^{n} sign | a_i - b_i |$, где $a_i$ и $b_i$– двоичные вектора

коэффициент $e$ корреляции:

1. $r(\vec{a}, \vec{b}) = \frac{cov(\dot{a}, \dot{b})}{\delta(\dot{a}) * \delta(\dot{b})}$, где $\dot{a}$, $\dot{b}$ – случайные величины, полученные из векторов $\vec{a}, \vec{b}$;

$\delta(\dot{a})$ - среднеквадратическое отклонение; cov - корреляционный момент.

при различных коэффициентах подобия, например:

1. Косинус $r(\vec{a}, \vec{b}) = \cos(g)$, где g — угол между векторами.
2. Номинальный коэффициент подобия [6]:

$$coef(\vec{a}, \vec{b}) = \frac{R * (n_{IJ} + \alpha * n_{ij})}{R * (n_{IJ} + \alpha * n_{ij}) + n_{iJ} + n_{ji}}$$

где $R, \alpha$-произвольные константы;

$n_{IJ}$ - количество совпадающий ненулевых элементов;

$n_{ij}$ - количество совпадающий нулевых элементов;

$n_{iJ}$ - количество совпадающий ненулевых элементов вектора $\vec{a}$ и нулевых элементов вектора $\vec{b}$;

$n_{iJ}$ - количество совпадающий нулевых элементов вектора $\vec{a}$ и ненулевых вектора $\vec{b}$.

3. Обобщённый коэффициент подобия [6]

Так же возможно применение других коэффициентов или их комбинаций.

Взвешивать слова в тексте можно различными способами, от которых зависит «качество» представления текста и которые могут существенно повлиять на результат классификации текста. Для каждого терма (выделенного нормированного слова) в документе могут определяться различные числовые показатели. Наиболее популярные из них:

1. Частота встречаемости терма в документе - «tf»;
2. Частота встречаемости терма в других документах - «df». Например, если слово встречается в каждом четвёртом документе коллекции, то df=1/4;
3. Длина слова;
4. Показатель важности термов, с которыми используется данный терм. Например, если некоторое существительное в тексте используется с прилагательными имеющими большой вес, можно так же говорить об его важности.

Из этих числовых показателей можно получить функции веса термов:

1. Булево значение веса $w = sign(tf)$, то есть 1 – если слово встретилось в документе, 0 – иначе;
2. $w = tf$ - стандартная частота слова;
3. $w = \frac{tf}{df}$. Такой коэффициент часто называют «tf-idf», то есть произведение частоты слова (tf), на величину, обратную величине частоты встречаемости слова во всех документах коллекции (inverse df). Часто применяется «сглаженная» вариация этот формулы - $w = tf * \log(\frac{1}{df})$.При употреблении формул tf-idf решается проблема общеупотребительных слов - когда слова не несущие большой смысловой нагрузки имеют большой вес. Так же частично решается проблема большой размерности [7].
4. Могут быть определены и другие алгоритмы назначения весов

В статье [13] приводится подробное исследование различных подходов к выбору весов-признаков.

Для проведения классификации существует достаточно большое множество различных алгоритмов. Наиболее популярные из них:

***Наивный байесовский алгоритм***

Простой вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости [8]. Класс принадлежности $c$ документа $d$ считается по формуле:

$$c = argmax_{c \in C} \log(\frac{D_c}{D}) + \sum_{i \in Q} \log(\frac{W_{ic}+1}{n+L_c}),$$

где

- C – множество классов
- $D_c$- количество документов в обучающей выборке принадлежащих классу c,
- $D$ - общее количество документов в обучающей выборке,
- $L_c$ - суммарное количество слов в документах класса $c$ в обучающей выборке,
- $W_{ic}$ - сколько раз $i$-ое слово встречалось в документах класса c в обучающей выборке,
- $Q$- множество слов классифицируемого документа.

***Алгоритм к-ближайших соседей***

Алгоритм классификации, основанный на оценивании сходства объектов, классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки [8].

***Классификатор Роше (Rocchio classifier)***

Для каждого класса вычисляется взвешенный центроид («центральный» элемент) [9] по формуле

$$\vec{g}_c = \frac{1}{|D_c|}\sum_{d\in D_c}\vec{d}_i - h\frac{1}{|D_{c,k}|}\sum_{d\in D_{c,k}}\vec{d}_i,$$

где $D_{c,k}$- k документов, не принадлежащих классу, наиболее близких к центроиду $\frac{1}{|D_c|}\sum_{d\in R_c}\vec{d}$;

$h$- некоторый числовой коэффициент.После вычисления взвешенных центроидов для каждого класса, классификатор Роше определяет принадлежность документа рубрике при помощи вычисления расстояния между вектором обрабатываемого документа и центроидом каждого класса.

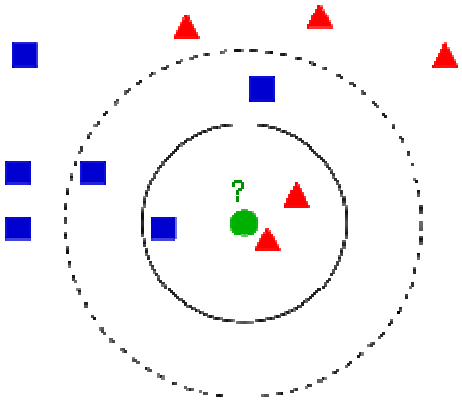



**Рис. 1**

***Метод опорных векторов SV***

Основная идея метода опорных векторов [10] - поиск разделяющей гиперплоскости с максимальным зазором в пространстве признаков. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

1. Алгоритм деревьев принятия решений

Деревья решений (decision trees) [11] разбивают данные на классы на основе значений переменных пространства признаков, в результате чего возникает иерархия операторов «ЕСЛИ-ТО», которые классифицируют данные. На основе обучающего множества строится дерево, узлами которого являются термины документов, листьями – метки классов, а ребра помечены весами терминов. Тестовый документ прогоняется по дереву, выбираются ветви, соответствующие терминам документа. В результате документу присваивается класс, соответствующий достигнутому листу.

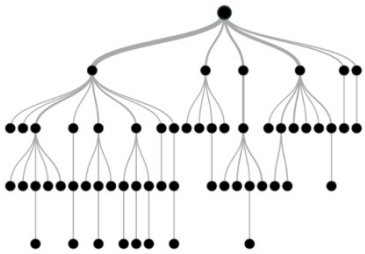


**Рис. 3**

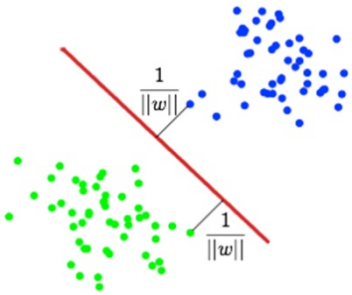



**Рис. 2**

При обучении используют следующую стратегию: рассматривают множество документов, проверяют, все ли документы данного множества имеют одинаковую метку класса (категорию); если нет, то ищут термин, обладающий наибольшей различительной способностью для разделения этих документов на классы; получают два подмножества документов и строят их поддеревья, повторяя всё сначала, пока не получат подмножество документов одного класса, тогда добавляют в соответствующее поддерево лист с меткой этого класса.

Так же существуют другие методы [8, 12, 13].

Все вышеперечисленные методы используют одинаковый подход к математическому моделированию текста – в виде вектора.

К основным проблемам проведения классификации в рамках классической модели текста следует отнести

1. Трудоёмкость:

1. Очистки (фильтрации) текста [1] - из текста необходимо удалить вспомогательные символы, а также слова, несущие вспомогательную смысловую нагрузку: предлоги, союзы, местоимения, частицы и т.п.;
2. Приведение слов текста к нормированному виду;
2. Невозможность классификации специальных текстов (например, математических, содержащих много формул);

Для решения второй из указанных проблем в настоящей работе предлагается подход, который позволяет, не меняя классическую модель документа, расширить информацию о формульном содержимом с помощью tex-представления документа [14], в котором формулы описываются с помощью специального языка разметки с использованием тегов и команд.

TeX – система компьютерной вёрстки, разработанная американским профессором информатики Дональдом Кнутом в целях создания компьютерной типографии. В неё входят средства для секционирования документов, для работы с перекрёстными ссылками. TeX – один из лучших способов для набора сложных математических формул. В частности, благодаря этим возможностям, TeX популярен в академических кругах, особенно среди математиков и физиков.

Например, формула

$$\int_0^R \frac{2x\,dx}{1+x^2} = \log(1+R^2)$$

в tex-формате будет представлена в виде:

\[ \int_0^R \frac{2x\,dx}{1+x^2} = \log(1+R^2).\]

При использовании такого подхода необходимо решить две проблемы:

1. Выделение текстовых блоков документа
2. Выделение формульных блоков документа

Информацию о формулах, представленную в tex-формате, можно добавить к текстовым блокам и воспользоваться вышеописанными методами.

Так же можно добавить коэффициент в определении веса терма, который бы учитывал количество формул в тексте:

$$\dot{w}_i = \begin{cases} \alpha * w_i, & \text{если i – ый элемент – это текстовый терм} \\ (1-\alpha) * w_i, & \text{если i – ый элемент – это формульный терм} \end{cases},$$

где

$$\alpha = \frac{\sum sign(w_i^{text})}{\sum sign(w_i)}$$

- это отношение объёма текстовых данных к объёму всего документа.

Недостатком данного подхода является невозможность корректного и качественного перевода свёрстанного документа (например, в pdf или doc форматах) в tex-представление. Поэтому для решения проблемы создания обучающей и тестовой выборки необходимы оригинальные текстовые документы в tex-формате. Такие текстовые документы имеются в различных открытых ресурсах сети Интернет, в частности, в [15].

На основе данного подхода были проведены тестовые расчеты. Для проведения расчётов была собрана обучающая выборка размером 3480 документов - 29 категорий математической тематики (Algebraic Geometry, Algebraic Topology, Analysis of PDEs, Category Theory, Classical Analysis and ODEs, Combinatorics и т.д.) по 120 документов в каждой.

Ниже представлены результаты вычисления коэффициента похожести с использованием косинуса в качестве метрики близости различных пар документов тестового множества:

1. Из документа взяты только текстовые данные:

Документы одной категории: [0.5, 0.46, 0.12, 0.22, 0.3, 0.26, 0.0, 0.16, …]

Документы разных категорий: [0.02, 0.2, 0.1, 0.14, 0.16, 0.1, 0.24, 0.08, …]

2. Из документа взяты только формульные данные:

Документы одной категории: [0.0, 0.0, 0.64, 0.26, 0.44, 0.32, 0.3, 0.14, …]

Документы разных категорий: [0.0, 0.0, 0.0, 0.0, 0.0, 0.32, 0.22, 0.16, …]

3. Из документа взяты и текстовые и формульные данные:

Документы одной категории: [0.54, 0.3, 0.3, 0.18, 0.18, 0.22, 0.08, 0.38, …]

Документы разных категорий: [0.0, 0.0, 0.0, 0.32, 0.14, 0.08, 0.24, 0.3, 0.26, …]

Сводная таблица 1 анализа полученных результатов представлена ниже:

Таблица 1

| | Среднее значение коэффициента для одной категории | Прирост среднего значения коэф-та для одной категории | Среднее значение коэф-та для разных категорий | Прирост среднего значения коэф-та для разных категорий |
|---|---|---|---|---|
| Только текст | 0,256 | - | 0,177 | - |
| Только формулы | 0,331 | 29,29 % | 0,264 | 49,2 % |
| Текст и формулы | 0,306 | 19,53 % | 0,211 | 19,2 % |

Примечание: прирост среднего значения коэффициента указан в сравнении с чистой классической моделью, т.е. когда из документа извлекается информация только о текстовом содержимом (первая строка).

Так же проведён анализ сравнения коэффициентов похожести документов в зависимости от удельного количества формул в тексте см. Таблица 2

Таблица 2

| | Среднее значение коэф-та для одной категории | Прирост среднего значения для одной категории | Среднее значение коэф. для разных категорий | Прирост среднего значения для разных категорий |
|---|---|---|---|---|
| В обоих документах < 20% формул | 0,188 | - | 0.149 | - |
| В обоих документах >20% формул | 0,256 | 36 % | 0,209 | 40 % |
| В одном документе < 20% формул, а в другом > 20% формул | 0.199 | 19,53 % | 0.167 | 19,2 % |

Примечание: прирост среднего значения коэффициента указан в сравнении со средним значением коэффициента похожести документов, в которых преобладает текстовое содержимое (первая строка).

Вычисление точности классификации см. таблица 3

Таблица 3

| Данные, получаемые из документа, для классификации | Точность классификации |
|---|---|
| только текстовое содержание | 54 % |
| только формульное содержание | **40 %** |
| Текст и формулы | 51 % |

Примечание: классификация проводилась по всему множеству документов со средним удельным весом формул в них около 20%.

По полученным данным можно сделать вывод, что информация о формульном содержимом документа может быть использована для классификации документов. При этом влияние формул на результат классификации текстовых документов, содержащих формулы, зависит от их удельного веса (процентного содержания в документе). Поэтому зная этот параметр, можно принимать решение об их учёте или не учёте при проведении классификации. В частности, очевидно, что для некоторых предельных

случаев, когда удельный вес формул в документе достаточной большой, их учёт может оказать решающее значение при классификации документов. В настоящее время авторами ведутся работы по уточнению и улучшению полученных результатов (увеличение обучающей выборки, улучшение точности определения формульных блоков и их предобработка, поиск оптимальных метрик, механизмов навешивания весов, приемлемых методов классификации, оптимальное встраивания формульной составляющей в классическую векторную модель текста и др.).

**Литература**

1. *Sebastiani F.* Machine Learning in Automated Text Categorization // ACM Computing Surveys. V. 34, No. 1. 2002. P. 1-47.

2. *Шабанов В.И., Власова А.Е.* Алгоритм формирования ассоциативных связей и его применение в поисковых системах. // Диалог-2003: труды междунар. конф. – М.: Наука, 2003. С. 603-608.

3. *Харин Н.П.* Метод ранжирования выдачи, учитывающий автоматически построенные ассоциативные отношения между терминами // НТИ. Сер. 2. 1990. №9. С. 19-23.

4. *К.В.Воронцов* Лекции по методам оценивания и выбора моделей. 2007. [Электронный ресурс]. URL: http://www.ccas.ru/voron/download/Modeling.pdf

5. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.

6. *Олейников Б.В.* Обобщенный коэффициент подобия для биоценотических исследований // – Красноярск: КрГУ, 1984. – 23 с. – Деп. в ВИНИТИ 13.12.84 г., № 7978-84 Деп.

7. *Manning K.D., Raghavan P., Schutze H.* Introduction to information retrieval // Cambridge university press. 2011. – 512 p.

8. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. – М.: Наука, 1974. – 416 с.

9. *Joachims T.A* probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization // Proceedings of ICML-97, 14th International Conference on Machine Learning. Morgan Kaufmann Publishers. P. 143-151.

10. *Joachims T.* Making large-scale SVM learning practical// Advances in kernel methods: support vector learning. – Cambridge: MIT Press, 1999. P. 42-49.

11. *Alsabti K., Ranka S. and Sing. V.* // CLOUDS: A Decision Tree Classifier for Large Dataset. In Proc.Of the 4th Intl. Conf. on Knowledge Discovery and Data Mining. – NY, 1998. P. 5-8.

12. Классификация и кластер/под ред. Дж. Вэн Райзина. ─ М.: Мир, 1980. ─ 390 с.

13. *Zhang. G. P.* Neural Networks for Classification: A Survey // Ieee transactions on systems, man, and cybernetics. Part C: applications and reviews. Vol. 30. No. 4. November. 2000. P. 451-460.

14**.** TeX Users Group (TUG) home page. URL: http://tug.org/ (дата обращения: 25.11.2013)

15. arXiv - highly-automated electronic archive and distribution server for research articles. URL: http://arxiv.org/ (дата обращения: 25.11.2013)

**Classification Of Texts Containing Formulas**

*Anton Olegovich Lavrenov, graduate student of the department of basic computing and information technology.*
*Boris Vasilevich Olejnikov, doctor of philosophy and associate professor of the department of basic computing and information technology.*

*The paper deals with approaches to the classification of documents that contain formulas. Indicates the importance of accounting for the classification of mathematical formulas and other documents. In addition, the classical methods of classification proposes new approaches that can improve the accuracy of classification.*

*Keywords: text classification, tex format, the block structure of the text, text classification with formulas.*