УДК 004

А.В. Мордвинов

МЕТОДИКА АВТОМАТИЧЕСКОЙ КАТЕГОРИЗАЦИИ ТЕКСТОВ

Нижегородский государственный технический университет им. Р.Е. Алексеева

Рассматриваются вопросы разработки методики по автоматической категоризации текстов, основанной на использовании модели текста, разработанной с помощью применения словарно-ориентированного алгоритма сжатия данных. Представленная модель текста позволяет учитывать не только лексическую, но и композиционную семантику документа, а также предоставляет возможность упростить этап индуктивного построения классификатора документов.

Ключевые слова: моделирование, текст, категоризация, автоматизация, алгоритм, сжатие.

За последние 10-15 лет задачи управления документами на основании их содержимого (обобщенное название – «извлечение информации», ИИ) приобрели особенно важное значение в области информационных систем ввиду постоянно повышающейся доступности документов в цифровой форме и вытекающей отсюда необходимости получать к ним доступ максимально быстрыми и удобными способами. Одной из таких задач является категоризация текста (КТ) – задача распределения текстов на естественном языке по тематическим категориям из заранее определенного набора. Появление задачи категоризации текстов относится к началу 60-х годов прошлого века, но только в 90-х она приобрела свою истинную значимость благодаря возросшему прикладному интересу и доступности более мощных аппаратных средств. КТ сейчас применяется во многих контекстах, начиная от индексирования документов на основе контролируемого словаря, заканчивая фильтрацией документов, автоматической генерацией метаданных, заполнением иерархических каталогов Интернет-ресурсов, атрибуцией текстов неизвестных авторов, а также в любых приложениях, требующих автоматизированной организации или диспетчеризации документов.

До конца 80-х наиболее популярным подходом к КТ, по крайней мере, в сообществе, занимающемся прикладными исследованиями, была инженерия знаний (ИЗ). Этот подход состоит в ручном задании набора правил на основании знаний экспертов о том, как классифицировать документы по заданным категориям. В 90-х годах XX века этот подход стремительно утратил популярность (особенно в исследовательском сообществе) в пользу парадигмы машинного обучения (МО). В соответствии с этим подходом производится индуктивное автоматическое построение текстового классификатора с помощью обучения на наборе заранее классифицированных документов.

КТ – задача присвоения булевого значения каждой паре $(d_j, c_i) \in D \times C$, где D – домен документов, а $C = \{c_1, ..., c_{|C|}\}$ – множество заранее заданных категорий. Значение T (*True*), присвоенное (d_j, c_i), обозначает решение классифицировать документ d_j в категорию c_i; тогда как значение F (*False*) обозначает решение не классифицировать документ d_j в категорию c_i. Более формально, ставится задача аппроксимировать неизвестную целевую функцию $\breve{\Phi} : D \times C \to \{T, F\}$ (которая описывает, как документы должны быть классифицированы) посредством функции $\Phi : D \times C \to \{T, F\}$, называемой классификатором, такой, чтобы она максимально совпадала с $\breve{\Phi}$.

Тексты в обычном представлении не могут быть интерпретированы классификатором или алгоритмом построения классификатора. Поэтому к документам должна быть заранее применена процедура индексирования, которая ставит в соответствие каждому тексту ком-

пактное представление его содержимого. Выбор этого представления зависит от того, что считать значимыми элементами текста и какие правила естественного языка считать значимыми для комбинирования этих элементов. В задаче КТ вторая проблема обычно игнорируется, и текст представляется вектором весов элементов, выбранных в качестве текстообразующих. Типичным выбором на сегодняшний день является представление текста в виде вектора слов. Подобное представление текста представляется сильно ограниченным, соответственно исследования в области моделирования текста продолжаются и являются актуальными с точки зрения развития методик КТ.

При построении любой автоматизированной системы по категоризации текстов необходимым является решение трех основных задач:

- индексирования документа и уменьшения размерности пространства элементов в полученном представлении;
- индуктивного построения классификатора документов;
- оценки эффективности созданной системы.

На сегодняшний день при решении задач индексирования документа и оценки эффективности системы ТК практически полностью полагается на базовые механизмы ИИ. Причина в том, что КТ – задача управления документами на основе их содержимого, поэтому имеет схожие характеристики с другими ИИ задачами, такими как, например, текстовый поиск. С одной стороны, это обоснованное решение, так как КТ является подзадачей ИИ, причем последняя дисциплина располагает гораздо более серьезной базой теоретических и практических исследований. С другой стороны, получается, что различные подходы к категоризации различаются в основном тем, как они решают проблему индуктивного построения текстового классификатора. Причем простота применяемой на первом этапе модели текста обусловливает необходимость использования сложных методик построения классификаторов, таких, например, как ансамбли классификаторов или boosting-техники. По этой же причине для обеспечения приемлемой эффективности системы приходится создавать выборки обучающих документов достаточно большого объема: по несколько десятков текстов для каждой категории.

В данной статье предлагается методика по категоризации текстов, которая позволяет:

- во-первых, упростить этап построения классификатора за счет использования классификаторов данных общего назначения вместо специализированных классификаторов текстов и их ансамблей;
- во-вторых, обеспечить высокую эффективность разрабатываемой системы при небольших объемах обучающей выборки.

Эти цели достигаются путем использования специально разработанной модели текста вместо общепринятого вектора весов слов.

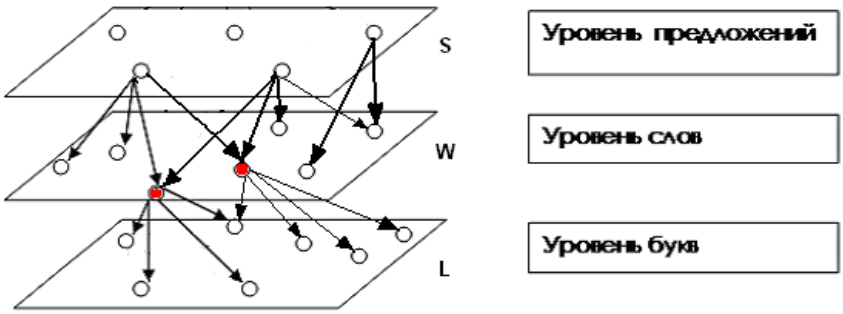


Рис. 1. Иерархия текстовой системы

С точки зрения системного подхода, текст обладает всеми признаками стационарной системы: с одной стороны, это целостный объект, в то же время мы можем выделить наи-

меньшие элементы текста, взаимодействие которых по определенным правилам порождает текст как систему. При этом потенциал текстовой системы (по такому параметру, как "смысловая нагрузка") больше суммы потенциалов составляющих ее элементов: $P(A) > [P(a_1) + P(a_2) + ... + P(a_n)]$. Кроме того, текст обладает четкой иерархической структурой, где каждый элемент более низкого уровня входит в состав какого-либо элемента более высокого уровня иерархии (рис. 1).

Таким образом, статистически каждый уровень текстовой системы может быть описан множеством соответствующих элементов, каждому из которых соответствует определенный вес – число ссылок на этот элемент (или частота встречаемости элемента) на более высоком уровне иерархии: $\vec{S} = \{s_1,...,s_n\}$, $\vec{W} = \{w_1,...,w_n\}$, $\vec{L} = \{l_1,...,l_n\}$.

Общепринятый в задаче категоризации подход состоит в том, что модель строится на множестве слов $\vec{W} = \{w_1,...,w_n\}$, из которого с помощью специальных техник выбираются элементы с максимальными весами. Для вычисления весов слов чаще всего используется функция $tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#Tr(t_k)}$, где $\#(t_k, d_j)$ обозначает частоту элемента t_k в документе d_j, а $\#Tr(t_k)$ – число документов в обучающей выборке *Tr*, в которых встречается t_k.

Предлагаемый в данной статье способ моделирования текста основывается на использовании в качестве базы для модели данных, которые создаются словарно-ориентированным алгоритмом LZ78 в процессе кодирования текста и уничтожаются при завершении работы алгоритма (так называемый словарь). Используя эти промежуточные данные, мы можем с помощью ряда разработанных алгоритмов представить модель текста в виде дерева, которое содержит в себе те подстроки, которые кодировщик счел статистически значимыми для данного текста. Дерево объединяет в себе максимум М поддеревьев, где М – число символов входного алфавита (рис. 2).

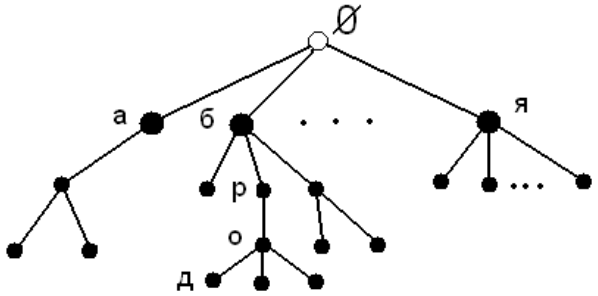


Рис. 2. Модель текста в виде дерева

Вес конкретной подстроки, содержащейся в дереве, $P(X) = N + 1$, где N – число нижележащих узлов дерева. Практически N равняется числу подстрок модели, образованных из данной. Если обозначить за m_X число вхождений подстроки в исходный текст, то будем иметь соотношение: $P(X) < m_x$.

Деревья применяются для визуального отображения модели и хранения ее информации в памяти ЭВМ, такое представление позволяет эффективно использовать рекурсию для извлечения и обработки данных модели. Однако в таком виде модель имеет слишком большую размерность для того, чтобы быть обработанной классификатором. Для снижения размерности и нормализации сравниваемых моделей каждая из них представляется в виде спектра подстрок заданной длины, которая может варьироваться от 1 до до K, где K – число уровней дерева (рис. 3 и рис. 4). Подобная возможность динамически регулировать максимальную длину подстрок модели, считающихся значимыми при проведении обучения классификатора, позволяет использовать в системах по категоризации текстов один классифика-

тор с заранее настроенными параметрами, а настройку системы на особенности входных данных производить, изменяя уровень детализации спектров.

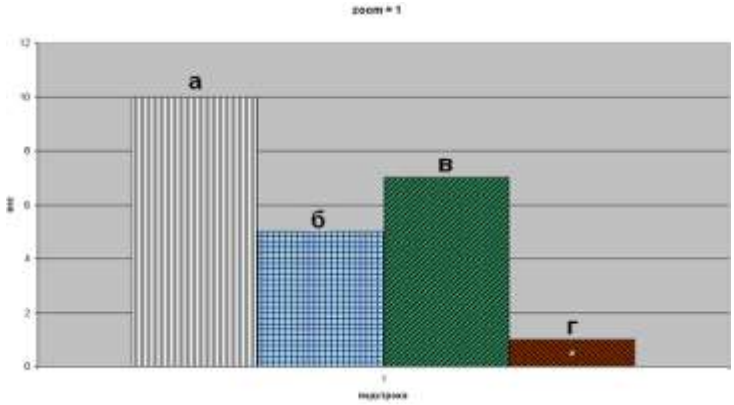


Рис. 3. Спектр модели текста с уровнем детализации 1

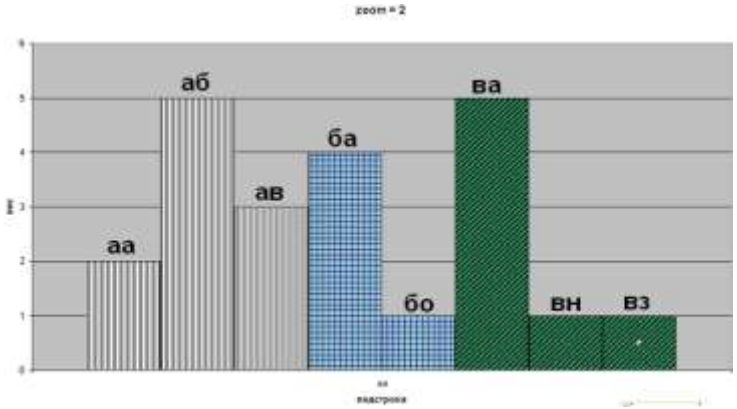


Рис. 4. Спектр модели текста с уровнем детализации 2

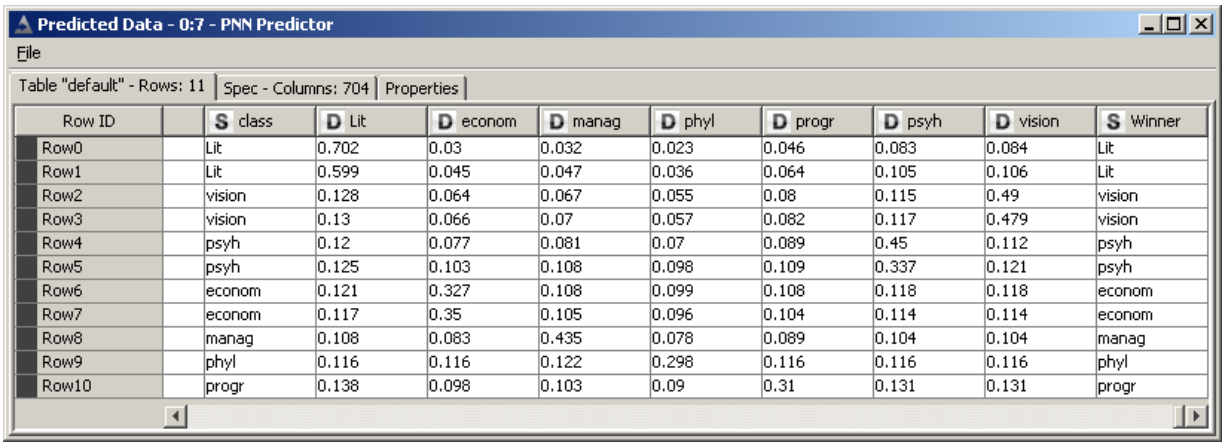
Рассмотрим теперь методику по категоризации текстов, использующую разработанную модель текста. Будем решать задачу в рамках МО подхода, основывающегося на том, что изначально доступна обучающая выборка документов $Tr = \{d_1,..., d_{|Tr|}\} \subset D$, уже классифицированная по категориям $C = \{c_1,..., c_{|C|}\}$.

Последовательность шагов:

- с помощью разработанных алгоритмов для каждого обучающего документа $d_j \in Tr$ строим модель m_j;
- выбираем уровень точности k и для каждой модели получаем спектр подстрок s_j, заранее уменьшая таким образом размерность векторов для классификации;
- зная для каждой $c_i \in C$ подмножество $Tr_{|ci|} \subset Tr$ обучающих документов, рассчитываем для каждой категории профиль P_i, который содержит подстроки спектров документов из $Tr_{|ci|}$, наиболее значимые для данной категории. Ценность подстроки будем оценивать по ее весу $w_i = \sum_i \#(t_k, c_i) \cdot \ln \frac{|Tr|}{\sum_{j \neq i} \#(t_k, c_j)}$, где $\sum_i \#(t_k, c_i)$ – количество раз, которое подстрока t_k встречается в спектрах категории c_i; $|Tr|$ – количество документов во всех категориях; $\sum_{j \neq i} \#(t_k, c_j)$ – количество раз, которое подстрока t_k встречается во всех категориях, кроме данной;

- производим дополнительное снижение размерности, убирая из каждого профиля P_i подстроки, вес которых меньше определенного порогового значения τ, определяемого экспериментально;
- в общем случае имеем $|P_i| \neq |P_j|, \forall i, j$, поэтому нормируем профили, добавляя в каждый из них отсутствующие подстроки из других профилей, присваивая им нулевой вес, собственным же подстрокам профиля назначаем вес 1; таким образом мы переходим к бинарным весам;
- имея для каждой категории $c_i \subset C$ профиль $P_i : |P_i| = \text{const}, \forall i$ и спектры обучающих документов из $Tr_{|ci|} \subset Tr$, вычисляем для каждого документа d_j вектор $\vec{V}_j : v_{ji} = \begin{cases} 1, v_{ji} \in P_i \,\&\&\, v_{ji} \in S_j, \\ 0, v_{ji} \in P_i \,\&\&\, v_{ji} \notin S_j; \end{cases}$
- обучаем классификатор на множестве векторов $V = \{\vec{V}_1, ..., \vec{V}_j\}$;
- система по категоризации текстов по категориям $C = \{c_1, ..., c_{|C|}\}$ настроена и готова к работе.

Предложенная методика по категоризации текстов была реализована программно, ее эффективность проверена с использованием нескольких широкоизвестных обучающихся алгоритмов классификации данных: Decision Trees, Fuzzy Rules, SVM, MLP, PNN. При объеме обучающей выборки всего в три документа процент верно классифицированных документов составил: Decision Trees – 73%, Fuzzy Rules – 82%, MLP – 91%, PNN и SVM – 100%. Пример результатов классификации для алгоритма PNN представлен на рис. 5.


Predicted Data - 0:7 - PNN Predictor

File

Table "default" - Rows: 11 | Spec - Columns: 704 | Properties

Row ID	S class	D Lit	D econom	D manag	D phyl	D progr	D psyh	D vision	S Winner
Row0	Lit	0.702	0.03	0.032	0.023	0.046	0.083	0.084	Lit
Row1	Lit	0.599	0.045	0.047	0.036	0.064	0.105	0.106	Lit
Row2	vision	0.128	0.064	0.067	0.055	0.08	0.115	0.49	vision
Row3	vision	0.13	0.066	0.07	0.057	0.082	0.117	0.479	vision
Row4	psyh	0.12	0.077	0.081	0.07	0.089	0.45	0.112	psyh
Row5	psyh	0.125	0.103	0.108	0.098	0.109	0.337	0.121	psyh
Row6	econom	0.121	0.327	0.108	0.099	0.108	0.118	0.118	econom
Row7	econom	0.117	0.35	0.105	0.096	0.104	0.114	0.114	econom
Row8	manag	0.108	0.083	0.435	0.078	0.089	0.104	0.104	manag
Row9	phyl	0.116	0.116	0.122	0.298	0.116	0.116	0.116	phyl
Row10	progr	0.138	0.098	0.103	0.09	0.31	0.131	0.131	progr

Рис. 5

Библиографический список

1. **Прангишвили, И.В.** Системный подход и общесистемные закономерности / И.В. Прангишвили. – М.: СИНТЕГ, 2000. – 528 с.
2. **Гаврилова, Т.А.** Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2000.
3. **Акимов, С.В.** Четырехуровневая интегративная модель для автоматизации структурно-параметрического синтеза // Труды учебных заведений связи. СПб.: СПбГУТ. 2004. № 171. С. 165–173.
4. **Айвазян, С. А.** Прикладная статистика: основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1983.
5. **Айвазян, С. А.** Прикладная статистика: классификация и снижение размерности / С. Айвазян [и др.]. – М.: Финансы и статистика, 1989.

6. **Androutsopoulos, I.** 2000. An experimental comparison of naive Bayesian and keyword based anti-spam filtering with personal e-mail messages / I. Androutsopoulos [at al.] // In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval (Athens, Greece, 2000). P. 160–167.
7. **Cohen, W.W.** 1995b. Text categorization and relational learning. // In Proceedings of ICML- 95, 12th International Conference on Machine Learning (Lake Tahoe, CA, 1995). P. 124–132.
8. **Cohen, W. W.** 1999. Context sensitive learning methods for text categorization / W. W. Cohen, Y.and Singer // ACM Trans. Inform. Syst. 17, 2. P. 141–173.
9. **Dagan, I.** 1997. Mistake driven learning in text categorization / I. Dagan, Y. Karov, D. and Roth // In Proceedings of EMNLP-97, 2nd Conference on E mpirical Methods in Natural Language Processing (Providence, RI, 1997). P. 55–63.
10. **Drucker, H.** 1999. Automatic text categorization and its applications to text retrieval / H. Drucker, V. Vapnik, D. and Wu // IEEE Trans. Neural Netw. 10, 5. P. 1048–1054.
11. **Fuhr, N.** 1994. Probabilistic information retrieval as combination of abstraction inductive learning and probabilistic assumptions / N. Fuhr, U. and Pfeifer // ACM Trans. Inform. Syst. 12, 1. P. 92–115.
12. Lewis, D. D. 1992b. Representation and Learning in Information Retrieval. Ph. D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA.
13. YANG, Y. 1999. An evaluation of statistical approaches to text categorization. Inform. Retr. 1, 1–2. P. 69–90.
14. Joachims, T. 2002. Guest editors' introduction to the special issue on automated text categorization / T. Joachims, F. and Sebastiani // J. Intell. Inform. Syst. 18, 2/3 (March-May). P. 103–105.

Дата поступления
В редакцию 15.10.2010

A. V. Mordvinov

AUTOMATED TEXTS CATEGORIZATION METHOD

Aspects of development of an automated texts categorization method based on the text model built using word-oriented compression algorithm are examined. The presented text model allows to take into consideration not only lexical, but also composition document semantics. It also allows to simplify the phase of inductive construction of text classifiers.

Key words: modeling, text, categorization, automation, algorithm, compression.