Сентябрь, № 6 **Технические науки** **2013**

УДК 519.226.3

АРТЕМ ВИКТОРОВИЧ СКАБИН
аспирант кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
artb00g@gmail.com

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ
доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
rogov@psu.karelia.ru

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РАСПОЗНАВАНИЯ СИМВОЛОВ*

Описывается математическая модель распознавания символов при расшифровке исторических рукописных стенограмм. В качестве объекта исследования взяты исторические стенограммы XIX века. Приводится описание математической модели, основанной на Байесовском подходе. Описывается метод оценки точности распознавания символа и алгоритм решения трудозатратной задачи построения матриц большой размерности для вычисления оценки вероятности вхождения фрагмента в текст. Приводятся результаты оценки точности распознавания для обучающей выборки и оценки вероятности появления пяти наиболее встречающихся в тексте слов. Описывается алгоритм для реализации предложенной математической модели в информационной системе для распознавания исторических рукописных документов.

Ключевые слова: математическая модель, распознавание символа, рукописные документы, Байесовский подход

ВВЕДЕНИЕ

В настоящее время большое внимание при введении в научный оборот рукописных документов уделяется их оцифровке, которая подразумевает не только сканирование или фотографирование, но и перевод на машинный язык. Существует большое количество программ для оцифровки документов, но часть из них работает только с печатным текстом либо же в системе реального времени. В данной статье рассматривается математическая модель распознавания рукописных стенографических символов, в качестве начальных данных используются стенографические записи Анны Григорьевны Снитkиной, обучавшейся по учебнику Ольхина [3]. Данная модель будет использована в создаваемой информационной системе по расшифровке исторических стенограмм.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РАСПОЗНАВАНИЯ СИМВОЛА

Обозначим через $x_1, \ldots, x_n$ последовательность стенографических символов. К сожалению, очень часто стенографические символы определяются неоднозначно. Для символа x_k обозначим через $x_1^k, \ldots, x_{l_k}^k$ множество его возможных распознаваний. Каждому распознанному символу определяются его возможные трактовки $y_1^{ki}, \ldots, y_{m_{ki}}^{ki}$. Тогда распознанный текст примет вид $y_{j_1}^{1i_1}, \ldots, y_{j_n}^{ni_n}$. Ставится задача найти такой набор индексов, чтобы вероятность правильного распознавания была максимальной. $P\left(y_{j_1^*}^{1i_1^*} \ldots y_{j_n^*}^{ni_1^*}\right) = \max P\left(y_{j_1}^{1i_1} \ldots y_{j_n}^{ni_n}\right)$, где максимум берется по всем $1 \le i_1 \le l_1$, $1 \le j_1 \le m_{1i_1}$, …, $1 \le i_n \le l_n$, $1 \le j_n \le m_{ni_n}$. Оценим вероятность $P\left(y_{j_1}^{1i_1} \ldots y_{j_n}^{ni_n}\right)$.

На основании формулы Байеса она равна

$$P\left(y_{j_1}^{1i_1} \ldots y_{j_n}^{ni_n}\right) = P\left(y_{j_1}^{1i_1}\right) \ldots P\left(y_{j_n}^{ni_n} \middle|_{y_{j_1}^{1i_1} \ldots y_{j_{n-1}}^{(n-1)i_{n-1}}}\right). \quad (1)$$

Оценка k-го ($k > 3$) сомножителя в правой части формулы (1) имеет вид:

$$P\left(y_{j_k}^{ki_k} \middle|_{y_{j_1}^{1i_1} \ldots y_{j_{k-1}}^{(k-1)i_{k-1}}}\right) = aP\left(x_{i_k}^k\right) + (1-a)P\left(y_{j_k}^{ki_k} \middle|_{y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_{k-1}}^{(k-1)i_{k-1}}}\right). \quad (2)$$

Оценка k-го сомножителя при $k \le 3$ производится аналогично. Коэффициент a настраивается в зависимости от качества распознавания стенограммы.

ОЦЕНКА ПЕРВОГО СЛАГАЕМОГО МАТЕМАТИЧЕСКОЙ МОДЕЛИ

Первое слагаемое в правой части формулы (2) характеризует точность распознавания стенографического символа. Оно вычисляется как:

$$P(x_{i_k}^k) = e^{-\alpha_i \rho(x_{i_k}^k, y_i)^{\beta_i}}, \quad (3)$$

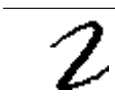
где $P(x_{i_k}^k, y_i)$ – это расстояние от текущего символа до эталона y_i класса символов. Для каждого класса находим $\rho^i_{\text{точ}}$ – минимальное расстояние от эталона класса до элементов из другого класса и $\rho^i_{\text{пол}}$ – максимальное расстояние от эталона класса до элементов класса. На наших данных оказалось, что $\rho^i_{\text{точ}} < \rho^i_{\text{пол}}$. Параметры формулы (3) подбираются как решение следующей системы:

$$\begin{cases} \frac{2}{3} = e^{-\alpha_i} (\rho^i_{\text{точ}})^{\beta_i} \\ \frac{1}{3} = e^{-\alpha_i} (\rho^i_{\text{пол}})^{\beta_i} \end{cases}.$$

В табл. 1 представлены результаты поиска расстояний для обучающей выборки.

Таблица 1

Расчет коэффициентов для символов обучающей выборки

Символ			α	β
	400	1200	0,001766	0,907297
	400	1000	0,00059	1,08782
	500	1000	0,000053	1,4380

ОЦЕНКА ВТОРОГО СЛАГАЕМОГО МАТЕМАТИЧЕСКОЙ МОДЕЛИ

Второе слагаемое математической модели (2) является вероятностью появления данного фрагмента в тексте. Она оценивается как

$$P\left(y_{j_k}^{ki_k} \left| y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_{k-1}}^{(k-1)i_{k-1}} \right. \right) = \frac{N\left(y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_k}^{ki_k} \right)}{N\left(y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_{k-1}}^{(k-1)i_{k-1}} \right) + 1}, \quad (4)$$

где $N\left(y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_k}^{ki_k} \right)$ – частота появления фрагмента текста $y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_k}^{ki_k}$. Данные числовые характеристики вычисляются на основе аналогичных произведений, а лучше принадлежащих одному автору. Приведем пример вычисления вероятности на основании произведений Ф. М. Достоевского. Для вычисления было использовано 28 произведений общим количеством слов более 80 тысяч.

Для вычисления оценки вероятности (4) строились пятерки слов $y_{j_{k-5}}^{(k-5)i_{k-5}} \ldots y_{j_k}^{ki_k}$, так как они дают более точное указание авторства, нежели пары или тройки. Для вычисления таких вероятностей необходимо было построить матрицы вероятностей встречи данного слова после всех возможных четверок слов в данных произведениях. Учитывая, что общее количество слов порядка 80 тысяч, оценок значений вероятностей включения данных четверки и пятерки в тексте будет порядка 512 триллионов. Однако, несмотря на то что большое количество из всех возможных пятерок слов не встречаются в тексте и матрица пятерок будет сильна разреженной, ее хранение и построение довольно трудозатратно.

Для построения данной матрицы использовался следующий алгоритм:

1. Пронумеровать все слова, используемые в тексте, так, чтобы они получили следующие координаты: идентификатор текста, в котором встречается данное слово, порядковый номер предложения в тексте, содержащего данное слово, и порядковый номер слова в данном предложении. В нашем случае была использована база знаний Smalt [2], в которой данная операция была произведена раньше.

2. Далее строятся всевозможные пятерки слов $N\left(y_{j_{k-3}}^{(k-3)i_{k-3}} \ldots y_{j_k}^{ki_k} \right)$ с таким условием, что у слов в данном словосочетании равны идентификаторы текста и предложения, а порядковые номера в словосочетании идут по возрастанию.

3. Из полученных пятерок выбираются уникальные и высчитывается количество включения их в текст. В 28 рассматриваемых произведениях из 80 тысяч слов было составлено порядка 69 тысяч уникальных пятерок слов.

4. Для тех пятерок, которые не были построены, можно считать, что вероятность включения их в текст равна 0.

Этот алгоритм обладает хорошей скоростью построения данных матриц, и позволяет избежать избыточности данных, вызваной сильной разреженностью матриц. В табл. 2 приведены оценки вероятности наиболее часто встречающихся пятерок слов в текстах автора.

Таблица 2

Частота и вероятность появления пятерок слов

Пятерка слов	Частота появления	Оценка вероятности
Не смотря на то что	11	0,8461
Ни съ того ни съ	3	0,75
Корпорація студентовъ какъ особое званіе	2	0,66
Теплѣй человѣчеству нежели отъ словъ	2	0,66
Этихъ ошибокъ этихъ примѣровъ всякаго	2	0,66

РЕАЛИЗАЦИЯ МАТЕМАТИЧЕСКОЙ МОДЕЛИ В ИНФОРМАЦИОННОЙ СИСТЕМЕ

Разрабатываемая информационная среда для дешифровки исторических стенограмм будет предлагать различные варианты распознавания текста. Более подробно о разрабатываемой системе и ее отдельных модулях говорится в работах [1], [4], [5]. После того как пользователь системы выделил символы, система разбила их на строки, выделив надстрочные и подстрочные символы, происходит расшифровка стенограммы согласно следующему алгоритму.

1. Обозначим символ, подвергаемый дешифровке, через S_i. Процесс дешифровки происходит со строкой, в которой находится текущий символ, слева направо, то есть все символы левее искомого уже дешифрованы. Обозначим как l длину строки, в которой находится искомый символ.

2. Пусть k – количество символов левее искомого в рассматриваемой строке. Если $k < 5$, то при расшифровке используется группа символов из k. Иначе рассматриваются пятерки слов.

3. Обозначим $y_i^1, y_i^2, ..., y_i^j$ – возможные расшифровки символа S_i. Исходя из всех возможных комбинаций расшифровок $\left\{y_{l-k}^{j1}, y_{l-k+1}^{j2}, ..., y_l^{jk}\right\}$ вероятность появления данного фрагмента в тексте рассчитывается по формуле (4).

4. Используя все вероятности, полученные в пункте 3, и оценки вероятностей, полученных в пункте 9, находим максимальную вероятность правильной расшифровки по формуле (2), умножив при этом второе слагаемое на вероятность того, к какому типу (основной, подстрочный, надстрочный) относится данный символ.

5. Пользователю предлагаются различные варианты дешифровки символа, упорядоченные по убыванию оценки вероятности их появления.

ЗАКЛЮЧЕНИЕ

Разрабатываемая информационная система по дешифровке стенограмм в данный момент проходит опытную проверку. После ее завершения будет разработан интернет-ресурс.

* Работа выполнена при поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

СПИСОК ЛИТЕРАТУРЫ

1. Гиппиев М. Б., Жуков А. В., Рогов А. А., Скабин А. В. Распознавание строк в стенографических документах // Современные проблемы науки и образования. 2013. № 5 (49).
2. Котов А. А., Некрасов М. Ю., Седов А. В., Рогов А. А. Информационная система для создания размеченных корпусов малой размерности // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2012. № 8 (129). Т. 1. С. 108–112.
3. Ольхин П. Руководство к русской стенографии. СПб.: Тип. доктора М. Хана, 1866. 187 с.
4. Рогов А. А., Скабин А. В., Штеркель И. А. Автоматизированная информационная система распознавания исторических рукописных документов // Информационная среда ВУЗА XXI века. Куопио, 2012.
5. Скабин А. В., Рогов А. А. Бинаризация и выделение символов исторической стенограммы // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2013. № 4 (133). С. 110–115.

Skabin A. V., Petrozavodsk State University (Petrozavodsk, Russian Federation)
Rogov A. A., Petrozavodsk State University (Petrozavodsk, Russian Federation)

MATHEMATICAL MODEL OF CHARACTER RECOGNITION

This article describes a mathematical model of character recognition for interpretation of historical handwritten documents. Historical shorthand records of the XIX century are taken as objects. The article describes a mathematical model based on the Bayesian approach. A method for estimating the accuracy of character recognition, algorithm solutions, and labor-intensive task of building high dimensional matrices to calculate the estimates of the probability of entering fragments in to the test is described. The results of estimating the accuracy of recognition for the training sample and estimating the probability of occurrence of the five most frequently used words in the test are given. An algorithm to implement the proposed mathematical model of the information system for the recognition of handwritten historical documents is described.

Key words: mathematical model, character recognition, handwritten documents, the Bayesian approach

REFERENCES

1. Gippiev M. B., Zhukov A. V., Rogov A. A., Skabin A. V. Recognition of lines in the historical handwritten documents [Raspoznavanie strok v stenograficheskikh dokumentakh]. *Sovremennye problemy nauki i obrazovaniya*. 2013. № 5 (49).
2. Kotov A. A., Nekrasov M. Yu., Sedov A. V., Rogov A. A. Information system for marked small dimension corpuses' development [Informatsionnaya sistema dlya sozdaniya razmechennykh korpusov maloy razmernosti]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural & Engineering Sciences]. 2012. № 8 (129). Vol. 1. P. 108–112.
3. Ol' khin P. *Rukovodstvo k russkoy stenografii* [Guide to the Russian shorthand]. St. Petersburg, Tip. doktora M. Khana, 1866.
4. Rogov A. A., Skabin A. V., Shterkel' I. A. The automated information system of recognition of handwritten historical documents [Avtomatizirovannaya informatsionnaya sistema raspoznavaniya istoricheskikh rukopisnykh dokumentov]. Informatsionnaya sreda VUZA XXI veka. Kuopio, 2012.
5. Skabin A. V., Rogov A. A. Binarization and isolation of historical manuscripts' symbols [Binarizatsiya i vydelenie simvolov istoricheskoy stenogrammy]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural & Engineering Sciences]. 2013. № 4 (133). P. 110–115.

Поступила в редакцию 16.07.2013