

Повышение достоверности обработки данных на основе избирательного избыточного кодирования семантических единиц текста

© С.В. Минаков, О.А. Финько

Краснодарское высшее военное училище (военный институт)
имени генерала армии С.М.Штеменко
ofinko@yandex.ru

Аннотация

Рассматриваются пути повышения достоверности текстовых данных на основе использования гибридной семантико-кодовой избыточности применительно к семантическим единицам текста наиболее подверженных ошибкам.

1 Введение

Текстовый тип данных в ближайшем будущем все еще будет составлять основу документированной информации. Для повышения достоверности электронных документов в настоящее время успешно применяются всевозможные избыточные коды [2], а в необходимых случаях и электронная цифровая подпись (ЭЦП). Как известно избыточность текста носит крайне неравномерный характер. В одних случаях морфологическую ошибку легко распознать и исправить, руководствуясь здравым смыслом и грамматикой языка. А в других - распознать ошибку, основываясь только на избыточности естественного языка, невозможно (цифровые последовательности, слова, например, такие как «июль» и «июнь», буквенные литеры, пароли в системах разграничения доступа и т.п.).

Текстовая и файловая обработки информации могут существенно отличаться друг от друга. Текст необходимо создавать и редактировать. В некоторых наиболее ответственных случаях текст даже обрабатывается и передается побуквенно (криптография, шифр гаммирования). Более того, при хранении и передаче особо ценной информации [5] распределенный принцип обработки просто необходим. В этих случаях «Шенноновский» отрыв от смысловой нагрузки на слово не всегда положителен. Поэтому

применения традиционных методов повышения достоверности файлов, основанных на ЭЦП и избыточном кодировании недостаточно.

2 Неравномерность избыточности текста

Текст неравномерен по смысловой нагрузке, следовательно, требования к повышению достоверности семантических единиц (СЕ) текста, влияющих на смысл, должны быть разными. Целесообразно повысить достоверность СЕ вероятность искажения, которых максимально влияет на смысл текста, и применять к ним методы повышения достоверности.

Одним из известных методов повышения достоверности текстовых данных является внесение семантической и лингвистической избыточностей (повторы текстовых единиц, применение синонимов, перефразировки, повторы цифровых данных словами и т.д.). Другим путем повышения достоверности текстовых данных является помехоустойчивое кодирование [2], достоинством которого является возможность обнаруживать или исправлять ошибки на выходе конечных устройств. Недостаток – кодирование не учитывает смысл (ценность) передаваемой информации.

В основе предлагаемого решения повышения достоверности СЕ лежит автоматическое определение СЕ минимальное искажение которых может привести к трансформации в другое существующее (семантически близкое) слово текста на каком-либо естественном языке. Причем предлагается СЕ сравнивать не с множеством СЕ, составляющих текст, а с множествами образованными онтологическими рядами относительно анализируемой СЕ. Для выявления семантически близких СЕ применим метод динамического программирования [6]. В процессе вычислений значения $d_{i,j}$ (операции удаления, вставки, замены) записываются в массив

$(m+1)(n+1)$, вычисляются с помощью следующего рекуррентного соотношения:

$$d_{i,j} = \min \{d_{i-1,j} + w(a_i, \varepsilon), d_{i,j-1} + w(\varepsilon, b_j), d_{i-1,j-1} + w(a_i, b_j)\},$$

где:

a_i - SE текста множества A соответствующего вводимому тексту;

b_j - слова, содержащиеся в базе данных множества B (эти слова объединяются в онтологические ряды путем разбиения единой базы данных по тематическим признакам);

$w(a_i, b_j)$ - цена преобразования символа a_i в символ b_j .

Ниже приведен массив (табл. 1), полученный при вычислении расстояния Левенштейна между строками «Моховое» и «Меховое». Из него видно, что расстояние между этими строками, то есть $d_{7,7}$, равно 1.

Таблица 1
Пример вычисления расстояния Левенштейна

	j	0	1	2	3	4	5	6	7
i			м	е	х	о	в	о	е
0		0	1	2	3	4	5	6	7
1	м	1	0	1	2	3	4	5	6
2	о	2	1	1	2	3	4	5	6
3	х	3	2	2	1	2	3	4	5
4	о	4	3	3	2	1	2	3	4
5	в	5	4	4	3	2	1	2	3
6	о	6	5	5	4	3	2	1	2
7	е	7	6	5	5	4	3	2	1

3 Избирательное избыточное кодирование семантических единиц текста

Для выбора SE текста (рис. 1) возьмем $d = 1$.

Данному условию удовлетворяют SE выделенные рамкой (все цифры; название месяцев «июль / июнь»; фамилии «Рыбалко / Рыбалка / Рыбалков», «И. Сталин / В. Сталин / И. Салин», название населенных

пунктов «Моховое / Меховое / Мохово»; «р. Ока / р. Оса / р. Ака» и т.п.).

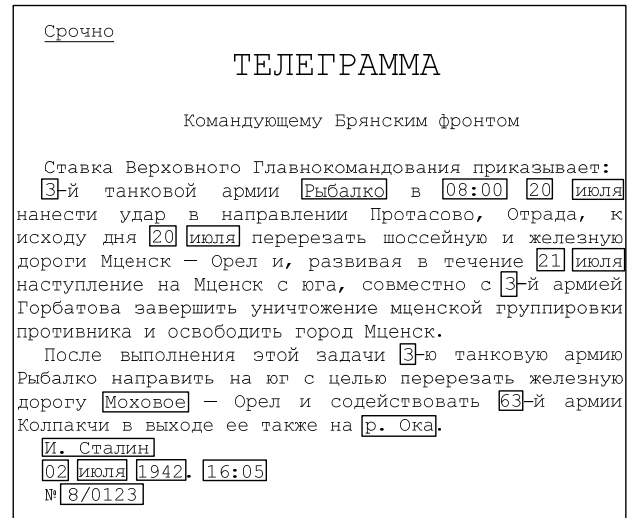


Рис. 1. Пример выделения SE в тексте телеграммы

Присвоим выделенным SE текста (рис. 1) кодовые обозначения (табл. 2).

Таблица 2
Пример кодирования алфавита

А	-	0	Т	-	18	(-	40
Б	-	1	У	-	19)	-	41
В	-	2	Ф	-	20	,	-	44
Г	-	3	Х	-	21	—	-	45
Д	-	4	Ц	-	22	.	-	46
Е	-	5	Ч	-	23	/	-	47
Ж	-	6	Ш	-	24	:	-	58
З	-	7	Щ	-	25	;	-	59
И	-	8	Ъ	-	26	0	-	48
Й	-	9	Ы	-	27	1	-	49
К	-	10	Ь	-	28	2	-	50
Л	-	11	Э	-	29	3	-	51
М	-	12	Ю	-	30	4	-	52
Н	-	13	Я	-	31	5	-	53
О	-	14	!	-	33	6	-	54
П	-	15	?	-	34	7	-	55
Р	-	16	№	-	36	8	-	56
С	-	17	¶	-	37	9	-	57
						пробел	-	32

Вычислим для каждой полученной, таким образом, числовой последовательности проверочное число. Например, суммируем каждую последовательность по модулю 64:

$$M = \left(\sum_{i=1}^n a_i \right) \bmod 64,$$

где n – количество символов в SE.

Пример базы онтологических рядов для анализируемого текста

Тематические признаки	Элементы множества B
годы	..., 1939, 1940, 1941, 1942, 1943, 1944, 1945 ..., 2008, 2009, ...
месяцы года	январь, февраль, март, апрель, май, июнь, июль, август, сентябрь, октябрь, ноябрь, декабрь
даты	01, 02, 03, 04, 05, ..., 29, 30, 31
время	00:00, 00:01, ..., 01:00, 01:01, 01:02, ..., 23:58, 23:59
военачальники ВОВ (маршалы)	Василевский, Говоров, Жуков, Конев, Малиновский, Мерецков, Рокоссовский, Сталин, Тимошенко, Толбухин
военачальники ВОВ (генералы арий)	Антонов, Апанасенко, Василевский, Еременко, Жуков, Конев, Малиновский, Мерецков, Павлов, Попов, Рокоссовский, Соколовский, Тюленев, ...
орган управления	отделение, взвод, ..., дивизия, корпус, армия, ..., Генеральный Штаб.
виды и рода ВС	авиация, ВВС, сухопутные, ..., морские, ВМФ, связь
воинские звания	рядовой, ефрейтор, младший сержант, ..., генерал-армии, маршал
вооружение (танки)	Т-34, Т-34-57, ОТ-34, ТО-34, КВ-1с, КВ-85, ИС-1, ИС-3, ИСУ-152, ИСУ-122, ИСУ-122С, ...
...	...
географические наименования (реки)	Большая Чернава, Быстрая Сосна, Кшень, Нерусса, Общерица, Ока, Олым, Орлик, Семенек
географические наименования (города)	..., Мценск, Мымрино, Мыцкое, Навесное, Навля, ..., Шатилово, Шахово, Шашкино, Щербово, Юшково, Яковлево, ...
...	...

Полученные проверочные числа добавим к исходным выделенным СЕ. Для отличия их от текста введем маркер – ## (редко встречающееся сочетание символов) и двухзначные проверочные цифры, соответствующие числу M избыточного кода, скрытые от пользователя (рис. 2).

Принимающая сторона вычисляет для соответствующих СЕ проверочные символы и сравнивает их значение с прикрепленными. При несовпадении результатов адресат делает вывод об искажении данной СЕ текста.

Элементами b_j единой базы данных B являются семантические единицы передаваемых сообщений. Пример таких СЕ представлен в таблице 3.

4 Заключение

Введение семантико-кодовой избыточности позволит повысить достоверность данных при соблюдении принципа *распределенной* (посимвольной) обработки, хранения и передачи. Причем предложенный метод может использоваться и для обработки документированной информации на «твердых» (бумажных) носителях.

В рассмотренном случае был применен достаточно простой избыточный код с контролем по модулю. Однако предполагается использовать более подходящие для данной задачи

многозначные коды [1, 3, 4, 7, 8], избыточность которых будет устанавливаться адаптивно в соответствии с избыточностью и ценностью СЕ текста.

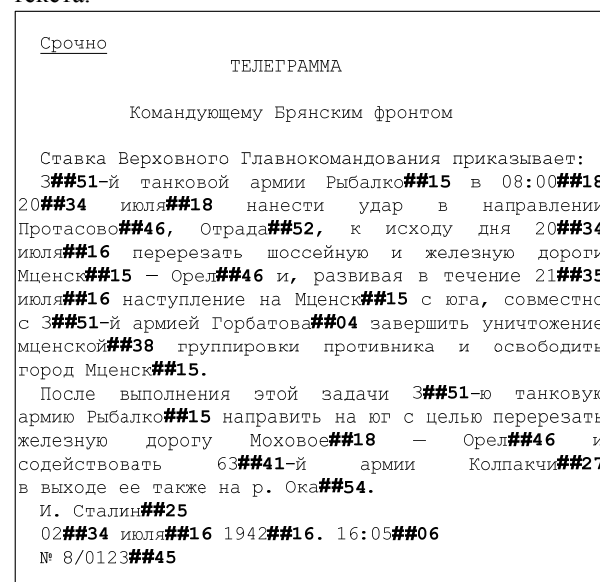


Рис. 2. Текст документа с введенной кодовой избыточностью

Литература

- [1]. Кирилов, А.А. Коды с произвольным основанием, исправляющие одиночные ошибки/ А.А. Кирилов. – В кн.: Проблемы кибернетики. - М.: Наука, 1970, Вып. 22, С. 282-287.
- [2]. Питерсон, У. Коды, исправляющие ошибки/ У. Питерсон, Э. Уэлдон. – М.: Мир, 1976. – с. 596.
- [3]. Тимофеев, Б.Б. Методы обнаружения ошибок в алфавитно-цифровых последовательностях на этапе подготовки и ввода данных в ЭВМ/Б.Б. Тимофеев, В.А. Литвинов. – УСиМ, 1977. №4, С. 20-27.
- [4]. Четвериков, В.Н. Подготовка и телеобработка данных в АСУ/ В.В. Четвериков. – М.: Высшая школа, 1981.
- [5]. Шанкин, Г.П. Ценность информации. Вопросы теории и приложений/ Г.П. Шанкин. – М.: Филоматис, 2004. – с. 128.
- [6]. Graham, Stephen. String Search/, Stephen A. Graham. – UK. School of Electronic Engineering Science University College of North Wales, 1992. – p. 103.
- [7]. Herr, J.R. Self-checking number system/ J.R. Herr. – Comput. Des., 1974, vol. 13, N 1, p. 85-91.
- [8]. Sethi, A.S. An error-correcting coding scheme for alphanumeric data/ A.S. Sethi, V. Rajaraman, P.S. Kenjale. – Inform. Process. Lett., 1978, N 2, p. 72-77.

Increasing reliable processing data on base of the electoral surplus coding semantics text units

S.V. Minakov, O.A. Finko

Considering ways of increasing the validity of textual data on base the use of hybrid semantics-code redundancy with reference to semantics text units most subject to errors.