УДК 004.9

И.Н. Егорова, С.В. Егоров

Харьковский национальный университет радиоэлектроники, Харьков

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СЕМАНТИЧЕСКОГО СЖАТИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

Проведено исследование современных методов сжатия текстовой информации. Разработана математическая модель сжатия текста, позволяющая осуществить как физическое сжатие текста, так и сохранение его семантической составляющей. Модель дает возможность пользователю самостоятельно задавать уровень сжатия текста.

Ключевые слова: *текст, семантическое сжатие, уровень сжатия, модель, множество, кортеж.*

Введение

Развитие информационных технологий порождает необходимость обработки все больших объемов информации, в том числе текстовой. Эта проблема, в свою очередь, требует разработки новых эффективных моделей и методов сжатия текста. Среди них особое место занимает семантическое сжатие текстовой информации, которое требует не просто физического сжатия, но и сохранения семантической составляющей текста.

Разработана модель семантического сжатия текстовой информации, которая позволяет задавать уровень сжатия и при этом обеспечивает включение в аннотацию предложений, наиболее полно отражающих смысл исходного текста.

Анализ последних исследований. Все существующие методы сжатия текста условно можно разделить на две группы: сжатие с потерями и без потерь. Метод семантического сжатия текста отличается от существующих методов сжатия тем, что, с одной стороны, должен обеспечивать само сжатие текста, а с другой стороны – гарантировать сохранение смысла документа.

В основу многих современных моделей сжатия текстовой информации положены принципы теории Шеннона [1]. Эти модели являются вероятностными и используют понятие энтропии. Шенноном разработаны модели нулевого, первого, второго и третьего порядка, а также общая модель [2]. Ключевым понятием моделей является понятие коэффициента энтропии H.

Так, для общей модели коэффициент энтропии H определяется формулой

$$H = \lim_{n\to\infty} \frac{1}{n} \sum p(B_n) \log_2 p(B_n) \text{bits/character}, \quad (1)$$

где B_n – первые n-букв алфавита.

Таким образом, в общей модели (1) сумма представляет собой все m^n значений B_n, при этом m – размерность алфавита.

Модели нулевого – общего порядка были положены в основу теоремы кодирования источника без потерь Шеннона. Впоследствии, принцип кодирования теоремы был использован для построения модели сжатия, в том числе, текстовой информации.

Так, фонетическая модель, основанная на словаре, использует модель общего порядка Шеннона [3]. В этой модели реализована идея замены последовательности символов их звуковой транскрипцией. Модель требует создания кодовой таблицы, по которой осуществляется кодирование текста, и используется для сжатия текста без потерь.

Модель семантического сжатия «Слово-за-словом» также базируется на модели общего порядка Шеннона [3]. Модель требует наличия тезауруса, который ставит в соответствие каждому слову его синоним, в качестве кодовой таблицы. В модели осуществляется замена исходных слов текста на более короткие эквиваленты, взятые из тезауруса.

Упомянутые модели, как и модель общего порядка, являются итеративными. В этих моделях осуществляется пошаговое сжатие слов исходного текста. Следует заметить, что общее количество слов является неизменным, поэтому по достижении определенного шага итерации (обычно, шестого-седьмого), возможность последующего сжатия стремится к нулю. Это объясняется тем, что на этом шаге алгоритм не может подобрать более короткого синонима слову из тезауруса.

Недостатком модели «Слово-за-словом» является необходимость сохранить общее количество слов исходного текста, что не позволяет достичь существенных показателей сжатия. Еще одним недостатком модели является искажение смысла исходного текста при многократной замене слов текста их более короткими эквивалентами. При этом восстановление исходного варианта текста не представляется возможным.

Таким образом, в представленной модели реализован метод сжатия текста с потерями.

Среди современных моделей по обработке текстовой информации следует отметить модель «Мешок слов» (bag of words model). Модель предполагает, что каждому термину, обнаруженному в документе, присваивается вес, зависящий от количества появлений этого термина в документе [4]. Схема присваивания весов определена как «частота термина».

Недостатком модели «Мешок слов» является игнорирование точного порядка следования терминов в документе, поскольку основное значение придается количеству вхождений каждого термина в документ. Таким образом, использовать модель в неизменном виде для семантического анализа текста не представляется возможным.

Постановка задачи исследования. Рассмотренные модели не являются универсальными и обладают рядом ограничений, которые не позволяют использовать их для семантического сжатия текстовой информации. Семантическая составляющая процедуры сжатия текста требует фундаментальных изменений в методологии исследования процесса сжатия текста. В связи с этим возникает необходимость разработки модели, позволяющей, с одной стороны, корректно осуществлять сжатие текста, а с другой – гарантировать сохранение смысловой составляющей исходного текста.

Разработка математической модели семантического сжатия текста

Процесс сжатия текста включает ряд этапов. Исходный текст подвергается ряду преобразований на пути его сжатия. Весь процесс поэтапного преобразования исходного текста в аннотацию может быть представлен в виде модели. Модель семантического сжатия текста должна включать ряд компонентов, таких как: исходные и выходные данные, а также характеристики функциональных преобразований исходных данных в аннотацию. В разрабатываемой модели предлагается ввести новый компонент, характеризующий уровень сжатия текста. Этот компонент даст возможность пользователю самостоятельно определять размер аннотации, в то время как модель должна гарантировать корректное формирование самой аннотации.

Таким образом, модель семантического сжатия (semantic compression) текста представлена в виде кортежа

$$M^{SC} = \langle X, Y, F, CR \rangle, \qquad (2)$$

где X – множество исходных данных; Y – множество выходных данных; F – множество функций преобразования исходных данных в выходные; CR – множество значений уровня сжатия (compression rate).

Модель семантического сжатия текста (2) дает возможность, во-первых, поэтапно представить процесс преобразования исходных данных в аннотацию, а во-вторых – позволяет пользователю задавать уровень сжатия и при необходимости – изменять его значение. Следует отметить, что пользователь может влиять только на размер аннотации, а сама модель гарантирует включение в аннотацию предложений, которые наиболее полно отражают смысл исходного текста, и не зависит от уровня сжатия.

Анализ модели основывается на исследовании ее компонентов. В модели (2) множество исходных данных X представлено в виде упорядоченного набора (кортежа) предложений ST_x (sentences), а также слов исходного текста W_x (words)

$$X = \langle ST_x, W_x \rangle \qquad (3)$$

Такое описание исходных данных позволяет сохранить порядок следования слов в каждом предложении, порядок следования предложений в тексте, а, следовательно, – сохранить семантическую составляющую модели (2).

Кортеж предложений в исходном тексте ST_x представляет собой упорядоченный набор предложений исходного текста, каждому из которых присвоен номер

$$ST_x = \langle st_{x1}, st_{x2}, ..., st_{xi}, ..., st_{xN} \rangle, \overline{i = 1, N}, \qquad (4)$$

где st_{xi} – i-е предложение в исходном тексте; N – количество предложений исходного текста.

В свою очередь, набор слов в каждом i-м предложении исходного текста также может быть представлен в виде кортежа

$$W_{xi} = \langle w_{x1}, w_{x2}, ..., w_{xij}, ..., w_{xiM} \rangle, \overline{j = 1, M}, \qquad (5)$$

где w_{xij} – j-е слово в i-м предложении исходного текста; M – количество слов в i-м предложении.

Весь набор слов исходного текста с учетом местоположения каждого слова в предложении и предложения в тексте может быть представлен в виде

$$W_x = \sum_{i=1}^{N} \sum_{j=1}^{M} w_{xij}. \qquad (6)$$

В модели (2) множество выходных данных Y представлено в виде упорядоченного набора предложений ST_y, и кортежа слов W_y, вошедших в аннотацию

$$Y = \langle ST_y, W_y \rangle. \qquad (7)$$

Кортеж предложений, вошедших в аннотацию, ST_y, как и в случае с предложениями исходного текста (4), представлен в модели (2) в виде упорядоченного набора, однако количество предложений аннотации может быть равным или меньше количеству предложений в исходном тексте.

Аналогично (5) описан кортеж слов аннотации W_y с той разницей, что привязка осуществляется к порядку следования слов и номерам предложений в аннотации.

В модели (2) все множество слов исходного текста W_x включает множество ключевых слов KW_x

(keywords), множество стоп-слов SW_x (stop-words) и множество обычных слов OW_x, которые не вошли ни в одно из первых двух множеств

$$W_x = \{KW_x, SW_x, OW_x\}. \quad (8)$$

Следует ввести базовые понятия.

Определение 1. Ключевым словом называется слово, которое входит в аннотацию, и имеет высокое значение ранга.

Определение 2. Ранг слова обозначает количество повторений слова в исходном тексте.

Стоп-словом в информационном поиске считается слово, не несущее самостоятельной смысловой нагрузки (предлоги, союзы, местоимения). Они могут быть исключены из исходного текста без ущерба для его семантической составляющей. Модель (2) предполагает исключение стоп-слов только на этапе обработки исходного текста.

Что касается аннотации, то включенные в нее предложения, содержат полный набор слов, в том числе и стоп-слова.

Множество слов исходного текста W_x, с учетом (8), может быть представлено как сумма множеств

$$W_x = KW_x \cup SW_x \cup OW_x. \quad (9)$$

Аналогично (8) и (9) может быть представлено множество слов аннотации W_y, которое включает множество ключевых слов KW_y, множество стоп-слов SW_y и множество обычных слов OW_y, вошедших в аннотацию.

Модель (2) предполагает, что все элементы множества выходных данных (7) входят в множество элементов входных данных (3). Другими словами, множество предложений аннотации ST_y является частью множества предложений исходного текста ST_x:

$$ST_y \subset ST_x, \quad (10)$$

а множество слов аннотации входит в множество слов исходного текста

$$W_y \subset W_x. \quad (11)$$

В модели (2) предусмотрено поэтапное преобразование исходных данных в аннотацию. Внутреннее представление текста в модели кардинально отличается от исходного. Такой подход позволяет осуществить более точную и быструю обработку данных. Он дает возможность удалить из текста «шумовую» составляющую в виде стоп-слов и осуществить группировку слов по их основам.

Процесс преобразования исходных данных описан в модели (2) множеством F, которое представляет собой множество функций преобразования и может быть представлен в общем виде как

$$F = \{F_1, F_2, ..., F_i, ..., F_T\}, \overline{i = 1, T}, \quad (12)$$

где T – общее количество функциональных преобразований, предусмотренных моделью (2). Следует заметить, что общее их количество T будет изменяться динамически, поскольку пользователь имеет возможность не только задавать уровень сжатия, но и изменять это значение [5].

Множество значений уровня сжатия CR определено в работе как

$$CR = \{cr_1, cr_2, ..., cr_i, ..., cr_B\}, \overline{i = 1, B}, \quad (13)$$

где B – количество значений коэффициента сжатия.

Уровень сжатия CR является базовым понятием модели (2) и количественно определяет, насколько должен быть сжат исходный текст. В модели предусмотрено дискретное задание пользователем коэффициента сжатия в процентном отношении с точностью до десятой доли процента, что позволяет обрабатывать тексты больших и сверхбольших объемов.

Для детального определения набора функций (12) необходимо поэтапно рассмотреть все функциональные преобразования исходного текста в аннотацию, предусмотренные моделью (2).

Так, на первом этапе модель предполагает разбиение исходного текста на предложения и присваивание каждому предложению номера ID_s (sentence identifier). Результатом работы этапа является упорядоченный набор предложений ST_x (4).

Второй этап предполагает представление каждого i-го предложения упорядоченным набором слов W_{xi} (5), а всего исходного текста – общим набором слов W_x (6). Таким образом, модель (2) учитывает местоположение каждого слова в предложении и, соответственно, расположение каждого предложения в исходном тексте.

Третий этап предусматривает удаление стоп-слов из исходного текста. В модели (2) такая процедура позволяет упростить и ускорить последующий семантический анализ исходного текста.

Функциональное преобразование F_1 из множества преобразований (12) подразумевает процедуру удаления стоп-слов и может быть представлено в виде

$$F_1 : W_x \to W', \quad (14)$$

где W' – набор слов исходного текста, из которого удалены стоп-слова.

На четвертом этапе текст, полученный в результате преобразований первых трех этапов, подвергается процедуре стемминга (приведение слова к его основе). В результате набор слов W', полученный после третьего этапа, будет преобразован и представлен в виде

$$F_2 : W' \to W'', \quad (15)$$

где W'' – набор слов, полученный в результате приведения слов набора W' к их основам.

На пятом этапе полученный набор основ слов W'' будет преобразован к виду модифицированного инвертированного индекса E

$$E = \{e_1, e_1, ..., e_i, ..., e_{LI}\}, \overline{i = 1, LI}, \quad (16)$$

где LI – количество элементов в модифицированном инвертированном индексе.

Определение 3. Модифицированный инвертированный индекс представляет собой инвертированный индекс, построенный в порядке уменьшения рангов слов и содержащий ссылки на номер предложения в документе.

Функциональное преобразование основ слов в элементы модифицированного инвертированного индекса может быть представлено в виде отображения

$$F_3 : W'' \to E . \quad (17)$$

На шестом этапе осуществляется формирование аннотации. Этот процесс проводится поэтапно и состоит из отбора ключевых слов и последующего отбора предложений, содержащих наибольшее количество таких слов. Такие предложения и будут представлять конечную аннотацию. Функциональное преобразование элементов модифицированного инвертированного индекса в аннотацию может быть представлено отображением

$$F_4 : E \to ST_y . \quad (18)$$

Следует заметить, что функциональное преобразование (18) непосредственно зависит от заданного пользователем значения уровня сжатия текста CR:

$$F_4 : F(CR) . \quad (19)$$

В модели (2) предусмотрена возможность задания значения уровня сжатия текста CR в соответствии с (13), а, следовательно, размер аннотации будет зависеть от заданного значения.

Выводы

В работе проведено исследование современных моделей сжатия текстовой информации, основанных на моделях нулевого - общего порядка Шеннона, таких как: фонетическая модель и модель «Слово-за-словом». Модели являются итеративными и по достижении определенного шага итерации эффективность последующего сжатия стремится к нулю. Общее количество слов в этих моделях остается неизменным, что не позволяет существенно сократить исходный текст. Кроме того, при многократной замене слов на более короткие эквиваленты, искажается смысл исходного текста.

Проведенный в работе анализ модели «Мешок слов» позволил установить, что основное значение в ней придается количеству вхождений каждого термина в документ, а порядок следования терминов в документе игнорируется.

Исследованные модели не могут быть использованы для целей семантического сжатия текста в неизменном виде.

Авторами разработана модель семантического сжатия текстовой информации, которая позволяет эффективно осуществлять сжатие, гарантируя при этом сохранение смысла исходного текста. Модель предоставляет пользователю возможность самостоятельно задавать уровень сжатия текста. Модель может быть использована как для целей семантического сжатия текстовой информации любых объемов (вплоть до сверхбольших), так и для целей информационного поиска [6].

Список литературы

1. Shannon C.E. A mathematical theory of communication / C.E. Shannon // Bell System Tech. J, 1948, 27. – P. 379-423.

2. Shannon C.E. Claude Elwood Shannon: collected papers / edited by N.J.A. Sloane, A.D. Wyner // IEEE Press, NJ, 1993. – 923 p.

3. Witten I. Semantic and generative models for lossy text compression / I. Witten, T. Bell, A. Moffat // Computer Journal 37(2), 1994. – P. 83-87.

4. Кристофер Д. Маннинг. Введение в информационный поиск: Пер. с англ. / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. – М.: ООО "И.Д. Вильямс", 2011. – 528 с.

5. Патент на корисну модель №82942 Україна, МПК51 G06F 17/21 (2006.01). Спосіб семантичної компресії тексту із заданим рівнем стислості / Винахідники: Єгоров С.В., Єгорова І.М.; Власник Єгоров С.В. – № и2013 00978; заявл. 28.01.13; опубл. 27.08.13, Бюл. № 16.

6. Егоров С.В. Информационный подход в семантическом сжатии текста / С.В. Егоров // Вестник молодых ученых Санкт-Петербургского государственного университета технологий и дизайна: в 3 вып. Вып. 1: Естественные и технические науки / С. Петербургск. гос. ун-т. технологий и дизайна. – СПб.: ФГБОУВПО «СПГУТД», 2013. – С. 22-27.

Поступила в редколлегию 1.11.2013

Рецензент: д-р техн. наук, проф. А.М. Синотин, Харьковский национальный университет радиоэлектроники, Харьков.

МАТЕМАТИЧНА МОДЕЛЬ СЕМАНТИЧНОГО СТИСНЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

І.М. Єгорова, С.В. Єгоров

Проведено дослідження сучасних методів стиснення текстової інформації. Розроблено математичну модель стиснення тексту, що дозволяє здійснити як фізичне стиснення тексту, так і збереження його семантичної складової. Модель дає можливість користувачу самостійно задавати рівень стиснення тексту.

Ключові слова: *текст, семантичне стиснення, рівень стиснення, модель, множина, кортеж.*

MATHEMATICAL MODEL FOR SEMANTIC COMPRESSION OF TEXTUAL INFORMATION

I.N. Iegorova, S.V. Iegorov

Research for modern methods of compression of textual information was conducted. Mathematical model for text compression was developed which allows to make both physical compression of text and preserve it's semantic part. Model gives possibility for the user to set compression level by himself.

Keywords: *phototypograph semantic compression, level of compression, model, great number, cortege.*