УДК 004.9

С.В. ЕГОРОВ, аспирант, ХНУРЭ, Харьков;
И.Н. ЕГОРОВА, канд. техн. наук, проф., ХНУРЭ, Харьков

МЕТОД СЕМАНТИЧЕСКОГО СЖАТИЯ ТЕКСТА

Предложен метод семантического сжатия текста, позволяющий пользователю самостоятельно определять конечный размер аннотации путем задания коэффициента сжатия исходного текста. С помощью данного метода пользователь получает аннотацию, которая полностью отражает смысл исходного текста и имеет заданный объем. В методе установлены зависимости между значением коэффициента сжатия и объемом аннотации. При этом уровень сжатия не влияет на семантическое представление исходного текста. Метод является универсальным и не зависит от формата и языка исходного документа.

Ключевые слова: семантическое сжатие, уровень сжатия, ключевое слово, ранг слова, ранг предложения, аннотация, инвертированный индекс.

Введение. Стремительное развитие современных информационно-поисковых систем приводит к постоянному росту объемов обрабатываемой ими информации, в том числе, текстовой. Огромные объемы такого рода информации, подлежащей обработке и хранению, требуют разработки новых методов сжатия при условии сохранения смысла исходного текста.

Предложенный в работе метод предоставляет новые возможности в области семантического сжатия текста, позволяя пользователю самостоятельно задавать объем конечной аннотации [1].

Метод может найти применение в области информационного поиска, оптимизации хранения информации в базах данных, а также с целью аннотирования любого рода текстовых документов.

Анализ последних исследований. Современные методы сжатия информации условно делятся на две большие группы: сжатие без потерь (обратимое) и сжатие с потерями (необратимое). Автором теории информации является *Шеннон*, изложивший математические основы сжатия информации [2]. Однако впоследствии выяснилось, что решение проблемы сжатия информации, в частности, в системах обработки текста не завершено.

Последующие работы в этой области позволили определить, что методы сжатия информации без потерь наиболее эффективны для сокращения объёма аудио- и видеоданных, а также цифровых фотографий. Однако, область применения таких методов для сжатия текстовой информации ограничивается архиваторами.

Методы сжатия информации с потерями, в свою очередь, позволяют достичь более высокой степени сжатия за счет отбрасывания некоторых данных и могут быть применены для обработки текстовой информации. Извест-

ны *методы свертывания регистра*, *стемминга* и *исключения стоп-слов*. Аналогично, модель векторного пространства и методы уменьшения размерности, такие как *латентно-семантическое индексирование*, позволяют создать компактное представление, по которому невозможно восстановить исходную коллекцию [3].

Применение сжатия информации с потерями целесообразно в системах поиска, когда их использование не сопровождается риском снижения эффективности работы таких систем.

Наибольший интерес среди современных методов семантического сжатия текста представляют: *метод упрощения текста* для информационно-поисковых приложений [4] и *способ автоматического обобщения текста* [5].

Суть метода упрощения текста для информационно-поисковых приложений заключается в формировании из сложных предложений, содержащих различные обороты, так называемых *легкодоступных предложений* (Easy Access Sentences). Такие предложения содержат только один глагол и имеют максимально упрощенную синтаксическую структуру.

Создание таких легкодоступных предложений необходимо с целью упрощения анализа текста системами машинного перевода и извлечения информации, а также системами аннотирования текстов.

Способ автоматического обобщения текста основан на принципе количества кода и понятии фразы-существительного. Принцип количества кода предполагает, что наиболее важная информация в тексте будет содержать больше лексических элементов и, следовательно, будет выражена большим числом единиц (слов, предложений). Фраза-существительное представлена в способе как достаточно общая единица, предоставляющая большую гибкость в плане числа элементов, которые она может содержать, способная нести большее или меньшее количество информации в зависимости от потребностей пользователя. Для нахождения фраз-существительных в предложениях был использован инструмент BaseNP Chunker, разработанный в университете г. Пенсильвания, США. Существует один важный момент, который необходимо учитывать – использование BaseNP Chunker (равно как и любого другого инструмента обработки естественного языка) может вносить ряд ошибок.

Учитывая ограничения, существующие для каждого из рассмотренных методов, следует заметить, что актуальной представляется задача разработки методов аннотирования текстовой информации, способных не только эффективно сжимать исходный текст, но и гарантировать его семантическую сохранность.

Постановка задачи. Исходный текст, подлежащий сжатию, может быть представлен в разных форматах, таких как doc, docx, pdf, txt и других. Кроме того, тексты могут быть написаны на разных языках. Особый интерес представляет задача разработки универсального метода аннотирования текста, не

зависящего от формата и языка исходного документа.

Метод семантического сжатия. Суть метода заключается в том, чтобы формируемая аннотация включала в себя предложения, содержащие слова, несущие наибольшую смысловую нагрузку. Работа метода требует введения ряда новых понятий, таких как: *уровень сжатия*, *ранг слова*, *ключевое слово*, *ранг текста*, *относительный ранг слова* и *относительный ранг аннотации*.

Уровень сжатия, заданный в методе коэффициентом K, показывает насколько должен быть сокращен (сжат) исходный текст. Коэффициент определен как отношение количества предложений, исключенных из аннотации в результате сжатия, к общему количеству предложений в тексте

$$K = \frac{P - P_{sum}}{P} \times 100\%, \quad (1)$$

где P – общее количество предложений в исходном тексте; P_{sum} – количество предложений, включенных в аннотацию.

Метод дает возможность пользователю самостоятельно задавать коэффициент сжатия. Коэффициент может принимать значение

$$K = \left(0;\ 0.1;\ ...;\ 1\right] x100\%.$$

Ранг i-го слова R_i определен в методе как число повторений i-го слова в тексте.

Под ключевым словом подразумеваются слова, имеющие наивысшие значения рангов R_i.

Соответственно, под рангом текста подразумевается значение R, определенное как сумма рангов входящих в него слов

$$R = \sum_{i=1}^{N} R_i, \quad (2)$$

где N – количество слов, приведенных к словарной форме, которые содержатся в исходном тексте.

Для того, чтобы привести слова исходного текста к словарной форме, следует, прежде всего, применить процедуру удаления стоп-слов. Эта процедура не приводит к потере семантической составляющей текста, поскольку стоп-слова не несут смысловой нагрузки.

Последующее применение процедуры стемминга/лемматизации позволяет сформировать словарь словопозиций. Под стеммингом (stemming) обычно подразумевают приближенный эвристический процесс, в ходе которого от слов отбрасываются окончания в расчете на то, что в большинстве случаев это себя оправдает. Стемминг часто подразумевает удаление производных аффиксов. Под *лемматизацией* (lemmatization) подразумевают точный процесс с использованием лексикона и морфологического анализа слов, в результате которого удаляются только флективные окончания и возвращается основная, или словарная, форма слова, называемая леммой [3].

В результате, после применения названных процедур получаем словарь, не содержащий стоп-слов, в котором хранятся словоформы с привязкой к рангу соответствующего слова и номеру предложения, в котором оно встречается.

На следующем этапе необходимо упорядочить словоформы по убыванию значений их рангов, сохранив привязку к номеру предложения. Одновременно вводим понятия относительного ранга слова $R_{отні}$ как отношение ранга i -го слова к общему рангу текста:

$$R_{отн.i} = \frac{R_i}{R}. \tag{3}$$

На этапе формирования аннотации следует учитывать два аспекта:

- первый – количественный. В аннотацию должно войти столько предложений P_{sum}, чтобы, исходя из (1), выполнялось условие

$$P_{sum} \geq P \times \left(1 - \frac{K}{100\%}\right); \tag{4}$$

- второй – семантический. В аннотацию должны войти предложения, наиболее полно отражающие смысл исходного текста. Другими словами, эти предложения должны включать наибольшее количество ключевых слов.

Формирование аннотации осуществляем пошагово, начиная с ключевого слова, имеющего наивысшие показатели ранга и относительного ранга. На каждом шаге следует рассчитывать относительный ранг аннотации $R_{отн.sum}$ как сумму относительных рангов ключевых слов:

$$R_{отн.sum} = \sum_{i=1}^{N_{sum}} R_{отн.i}, \tag{5}$$

где N_{sum} –- количество ключевых слов в аннотации, зависящее от уровня сжатия, заданного пользователем.

Включение в аннотацию каждого нового слова должно осуществляться с проверкой следующего условия: относительный ранг аннотации должен быть больше или равен разнице между единицей (наивысшее значение коэффициента сжатия) и коэффициентом сжатия, выраженным в относительных единицах:

$$R_{отн.sum} \geq \left(1 - \frac{K}{100\%}\right). \tag{6}$$

Таким образом, формируем аннотацию, в которой количество ключевых слов будет изменяться в зависимости от заданного коэффициента сжатия.

То есть, установлена зависимость между количеством ключевых слов в аннотации N_{sum} и уровнем сжатия с заданным коэффициентом K,

$$N_{sum} = f(K). \tag{7}$$

Формирование аннотации осуществляется посредством включения

предложений, которые содержат ключевые слова, имеющие наивысший ранг, то есть несущие наибольшую смысловую нагрузку. При этом на каждом шаге проверяется условие (4). Таким образом, количество предложений, вошедших в аннотацию, также как и количество ключевых слов, зависят от заданного пользователем уровня сжатия.

Последовательность включения выбранных предложений в аннотацию соответствует их расположению в исходном тексте, то есть в соответствии с возрастанием номера предложения. Таким образом, соблюдается последовательность изложения текста.

Анализ полученных результатов. Предложенный метод семантического сжатия текста позволяет формировать аннотацию требуемого объема при условии сохранения смысла исходного текста. Метод является интерактивным и дает возможность пользователю самостоятельно задавать уровень сжатия исходного текста.

Перспективы дальнейших исследований. Метод семантического сжатия текста может быть расширен в области многодокументного поиска, а также в применении к базам данных, используемым в информационно-поисковых системах.

Выводы. В работе предложен метод семантического сжатия текста, объединяющий две парадигмы: количественную и семантическую, между которыми установлены функциональные зависимости. Метод позволяет формировать аннотацию в объеме, соответствующем заданному пользователем коэффициенту сжатия, которая полностью отражает смысл исходного текста. Метод является универсальным, то есть не зависящим от языка и формата исходного текста.

Список литературы: 1. Патент на корисну модель №82942 Україна, МПК[51] G06F 17/21 (2006.01). Спосіб семантичної компресії тексту із заданим рівнем стислості/ Винахідники: *Єгоров С.В., Єгорова І.М.;* Власник *Єгоров С.В. – № u2013 00978;* заявл. 28.01.13; опубл. 27.08.13, Бюл. № 16. **2.** *Shannon C. E.*, A mathematical theory of communication. Bell System Tech. J, 1948, 27, 379—423 **3.** *Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение* в информационный поиск : Пер. с англ. — М. : ООО "И.Д. Вильямс", 2011. - 528 с. **4.** *Klebanov B. B., Knight K., Marcu D*. Text Simplification for Information-Seeking Applications. - Springer Verlag, 2004. – 13 p **5.** *Lloret E., Palomar M*. Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. - Department of Software and Computing Systems, University of Alicante, Spain, 2009. – 8 p.

Поступила в редколлегию 27.09.2013

УДК 004.9

Метод семантического сжатия текста / С.В. Егоров, И.Н. Егорова // Вісник НТУ «ХПІ». Серія: Математичне моделювання в техніці та технологіях. – Харків: НТУ «ХПІ», 2013. – №54

(1027). – С. 118 – 123. Бібліогр.: 5 назв.

Запропоновано метод семантичного стиснення тексту, що дозволяє користувачу самостійно визначати кінцевий розмір анотації шляхом задання коефіцієнту стиснення вихідного тексту. За допомогою даного методу користувач отримує анотацію, що повністю відображує сенс вихідного тексту та має заданий об'єм. У методі встановлені залежності між значенням коефіцієнту стиснення та об'ємом анотації. При цьому рівень стиснення не впливає на семантичне представлення вихідного тексту. Метод є універсальним та не залежить від формату та мови вихідного документу.

Ключові слова: семантичне стиснення, рівень стиснення, ключове слово, ранг слова, ранг речення, анотація, інвертований індекс.

Suggested method of semantic text compression allows the user to determine final annotation size singlehanded by means of compression rate adjustment of the source text. By means of this method user obtains annotation which fully reflects meaning of the source text and has the given volume. Dependencies between the value of compression rate and annotation volume have been established in the method. At the same time compression rate doesn't affect semantic representation of the source text. Method is universal and doesn't depend on format and language of the source document.

Key words: semantic compression, compression rate, keyword, word rank, sentence rank, annotation, inverted index.

УДК 004.652.6

Т.М. ЗАГОРОДНЯ, аспірант СумДУ, Суми

ОПТИМІЗАЦІЯ ПАРАМЕТРІВ НАВЧАЛЬНИХ ЗАНЯТЬ ЗА ДОПОМОГОЮ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

Запропоновано інформаційну технологію та програмну реалізацію системи підтримки прийняття рішень, яка дозволяє оптимізувати параметри навчальних занять при створенні відповідної структури навчального процесу та оптимально розподіляти час між різними навчальними завданнями та різними видами навчальних занять з метою забезпечення максимально високого рівня компетенцій майбутнього інженера.

Ключові слова: компетенції, процес підтримки прийняття рішень, оптимізація.

Аналіз літератури та постановка проблеми. Для формування набору *компетенцій* [1] – [4], необхідних для майбутніх інженерів, використовується спеціальна організація навчального процесу, розроблення обґрунтованих логічно-завершених модулів, завданням яких є формування загальних і професійних компетенцій (залежно від дисципліни, напрямку підготовки), їх відповідне наповнення та зв'язок з уже вивченим матеріалом, об'єктивна диференціація навчального матеріалу, адаптація навчальних і навчально-методичних матеріалів до сучасної моделі студента, розроблення і впровадження нових інформаційний технологій для можливості підбору більш гну-