УДК 004.9

С.В. Егоров

Харьковский национальный университет радиоэлектроники, Харьков

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ СЕМАНТИЧЕСКОГО СЖАТИЯ ТЕКСТА

Разработана информационная технология семантического сжатия текстовой информации, базирующаяся на модели и методе семантического сжатия текста. Реализована информационная система, которая дает возможность пользователю самостоятельно задавать уровень сжатия текстовой информации и может быть использована для сжатия текстов любых объемов.

Ключевые слова: *текст, семантическое сжатие, уровень сжатия, информационная система, информационная технология, модель, метод.*

Введение

Развитие Интернет – технологий, а также баз данных, порождает необходимость хранения и обработки все больших объемов информации. В масштабах Сети хранение информации требует огромных вычислительных и временных затрат.

С целью сокращения объемов информационных хранилищ, ускорения выдачи необходимой информации и решения ряда других актуальных проблем, целесообразно применять и совершенствовать техники сжатия информации.

Работа посвящена созданию информационной технологии сжатия текста, позволяющей получать из исходного текста аннотацию заданного объема, которая полностью отражает его смысл.

Анализ проблемы. Создание информационной технологии сжатия текстовой информации базируется на разработке соответствующих методов.

Существующие методы сжатия информации разделяют на две основные группы: сжатие с потерями и сжатие без потерь.

Сжатие информации с потерями (lossy compression) позволяет достичь более высокой степени сжатия за счет отбрасывания некоторых данных [1]. Примерами сжатия информации с потерями являются свертывание регистра, стемминг и исключение 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.

Сжатие с потерей информации целесообразно, когда "потерянная" информация в дальнейшем не используется.

При использовании сжатия без потерь возможно полное восстановление исходных данных.

Сжатие без потерь обычно используется для передачи и хранения текстовых данных, компьютерных программ, реже – для сокращения объёма аудио- и видеоданных, цифровых фотографий и т. п., в случаях, когда искажения недопустимы или нежелательны.

Особый интерес для целей сжатия текстовой информации представляют методы, позволяющие учесть семантическую составляющую. Такие методы должны позволять, с одной стороны, эффективно сжимать текст, а с другой – полностью сохранять смысл исходного текста.

Следует заметить, что существующие модели, такие как фонетическая, модель «Слово-за-словом»,

модель «Мешок слов» [1, 2], – не могут быть использованы без изменения для целей семантического сжатия текста.

В связи с этим возникает необходимость разработки новых методов, а на их основе – информационной технологии, реализующей эффективный подход к семантическому сжатию текстовой информации.

Постановка задачи исследования

Актуальность темы сжатия текстовой информации постоянно возрастает в связи с ростом объемов хранилищ данных, а также увеличением числа Web-сервисов, которые оперируют текстовыми данными. В связи с этим разработка информационной технологии семантического сжатия текста, основанной на новых эффективных методах, представляется актуальной.

В работе использован метод семантической компрессии текста с заданным уровнем сжатия, который предоставляет пользователю возможность самостоятельно задавать уровень сжатия текстовой информации и тем самым задавать размер аннотации. При этом сам метод гарантирует включение в аннотацию предложений, наиболее полно отражающих смысл исходного текста [3].

Метод семантического сжатия текста в работе положен в основу программной реализации в информационной системе [4].

Разработка информационной технологии семантического сжатия текста

Разработка информационной технологии подразумевает создание и использование соответствующей информационной системы. В свою очередь, разработка информационной системы предполагает выполнение ряда этапов, таких как проектирование, разработка архитектуры информационной системы, выбор инструментальных и программных средств разработки, программная реализация, тестирование и внедрение.

Проектирование информационной системы семантического сжатия текста осуществляется в работе с учетом современных подходов. Основными этапами проектирования являются: определение типа приложения, выбор стратегии развертывания, а также выбор решения о путях реализации сквозной функциональности [5].

Тип программного приложения, реализованный в системе, сориентирован на наиболее современную операционную систему (ОС) Microsoft Windows 8 и представляет собой насыщенное клиентское приложение, предназначенное для развертывания из Интернет.

Выбор стратегии развертывания осуществлен в работе с учетом компромисса между требованиями приложения и ограничениями, которые среда накладывает на варианты развертывания [5] Проведен анализ моделей развертывания, таких как модель распределенного и нераспределенного развертывания. С учетом ограничений целевой ОС, которая позволяет развертывать приложения из инсталляционных пакетов прямо на клиентский компьютер или мобильное устройство, была выбрана модель нераспределенного развертывания.

Сквозная функциональность представляет собой набор функций, от которых зависят все слои или уровни приложения.

В разрабатываемой информационной системе реализованы такие аспекты сквозной функциональности как логгирование (централизованное ведение журнала служебной информации) и централизованная обработка исключений.

При разработке архитектуры информационной системы учтены наиболее эффективные архитектурные решения и шаблоны проектирования.

Так, архитектура информационной системы семантического сжатия текста реализует несколько архитектурных стилей, а именно: многослойной и компонентной архитектур, а также архитектуры, основанной на предметной области.

Многослойная архитектура системы реализована в виде трех слоев, а именно: UI Layer, Business Logic Layer, Data Access Layer, а также шаблона проектирования MVVM (рис. 1)

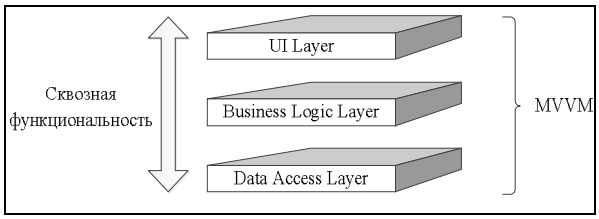

Рис. 1. Архитектура информационной системы

Слой UI Layer содержит компоненты, реализующие пользовательский интерфейс (User Interface, UI).

Слой Business Logic Layer, в свою очередь, содержит компоненты, которые реализуют как бизнес-логику системы, так и сам метод семантического сжатия текста [4]. При разработке этого слоя использован архитектурный стиль, основанный на предметной области.

Слой Data Access Layer содержит компоненты для доступа к файловой системе устройств пользователей. При разработке слоя в работе использована компонентная архитектура.

Для эффективной реализации информационной системы в работе были использованы шаблоны проектирования, которые представляют собой оптимальные подходы к решению различного спектра задач проектирования и программной реализации.

Так, для реализации взаимодействия между слоем представления и слоем бизнес-логики в рамках шаблона MVVM в работе был применен шаблон Command, который позволяет инкапсулировать запросы в качестве объектов. Команды в приложении используются для отправки запросов со слоя представления на слой бизнес-логики и отслеживания состояния доступности команд.

Для проектирования класса-интерфейса слоя бизнес-логики в работе использован шаблон Façade, который определяет высокоуровневый интерфейс, позволяющий упростить использование сложной подсистемы клиентами. За счет использования шаблона Façade в приложении достигается высокая степень инкапсуляции (сокрытия) деталей внутренней реализации на уровне слоя.

Шаблон Factory Method использован при разработке информационной системы с целью улучшения инкапсуляции данных, упрощения взаимодействия клиента с компонентами системы, а также для обеспечения контроля над процессом создания и инициализации объектов.

Наряду с названными в работе использован шаблон Memento, позволивший реализовать объектно-ориентированный подход к сериализации / десериализации (сохранению и восстановлению внутреннего состояния) объектов. Такой подход использован в работе с целью сохранения и загрузки конфигурационных файлов.

Шаблон Singleton реализован в классе бизнес-логики информационной системы для обеспечения условия уникальности существования одного экземпляра бизнес-логики в приложении.

Алгоритм работы информационной системы, реализующий метод семантического сжатия текста [4], в общем виде может быть представлен в виде схемы (рис. 2)

Алгоритм, реализующий метод семантического сжатия текста [4], выполняет последовательно ряд этапов. На первом этапе после считывания исходного текста осуществляется процедура удаления стоп-слов, позволяющая очистить текст от слов, которые не несут смысловой нагрузки (предлоги, союзы, местоимения). Метод дает возможность пользователю самостоятельно задавать уровень сжатия, таким образом влияя на размер аннотации. Следует отметить, что пользователь может неоднократно изменять это значение и, тем самым, в случае необходимости, получать аннотации разного размера. На следующем этапе метод предусматривает применение процедуры стемминга, которая позволяет привести слова к их основам. В дальнейшем основы слов группируются в инвертированный индекс, который представляет собой список, содержащий основы слов со значениями их рангов, и номерами предложений, в которых они встречаются. Следующий этап предусматривает упорядочивание элементов инвертированного индекса в порядке убывания их рангов и, таким образом, осуществляется построение модифицированного инвертированного индекса. В дальнейшем происходит формирование инвертированного индекса аннотации, который формируется поэтапно в зависимости от заданного пользователем значения уровня сжатия. При этом осуществляется отбор предложений, содержащих наибольшее количество ключевых слов, то есть слов, имеющих максимальное значение ранга, и занимающих в модифицированном инвертированном индексе наивысшие позиции. Заключительным этапом работы алгоритма является формирование аннотации, которая содержит предложения, включающие максимальное количество ключевых слов, наиболее полно отражающих смысл текста. Таким образом, метод семантического сжатия текста, реализованный в алгоритме, позволяет, с одной стороны, пользователю задавать уровень сжатия, а с другой стороны, – гарантировать сохранение семантической составляющей исходного текста.

Рис. 2. Схема алгоритма семантического сжатия текста

Программная реализация алгоритма осуществлена в работе в среде программирования Microsoft Visual Studio на языке C#.

Проведенное тестирование интерфейса пользователя показало пригодность информационной системы для использования на устройствах с различ-

ным разрешением экранов и различной плотностью пикселей. Тестирование программной части подтвердило работоспособность системы при эксплуатации на реальных устройствах.

Таким образом, в работе реализована информационная технология семантического сжатия текста, позволяющая получать аннотацию заданного пользователем размера и максимально полно отражающую смысл исходного текста.

Выводы

В работе дано описание информационной технологии семантического сжатия текста, которая основана на методе семантической компрессии текста с заданным уровнем сжатия. Приведен алгоритм, реализующий данный метод. Отличительной особенностью метода является предоставление пользователю возможности задавать уровень сжатия, а также неоднократно изменять его при необходимости. В методе предусмотрено формирование аннотации из предложений, содержащих наибольшее количество ключевых слов, которые максимально полно отражают смысл исходного текста. Таким образом, метод гарантирует сохранение семантической составляющей при сжатии текстовой информации.

Разработана информационная система, рассмотрены основные этапы ее проектирования, приведена архитектура системы. При проектировании особое внимание было уделено выбору стратегии развертывания, определению типа приложения и выбору решения о путях реализации сквозной функциональности. Архитектура информационной системы разработана с учетом особенностей нескольких архитектурных стилей, которые позволяют наиболее эффективно реализовать функциональные задачи системы. Так, многослойная архитектура используется в работе для обеспечения возможности масштабирования приложения, а также для разделения границ ответственности компонентов каждого слоя. Компонентный тип архитектуры позволяет обеспечить в работе унифицированность отдельных компонентов. Тип архитектуры, основанной на предметной области, позволил учесть в работе, с одной стороны, все особенности предметной области, а с другой – обеспечил гибкость ядра информационной системы.

При разработке информационной системы с целью обеспечения унифицированного подхода к решению стандартных задач проектирования и разработки программных продуктов, а также их сопровождению были использованы шаблоны проектирования Command, Façade, Factory Method, Memento, Singleton.

В работе осуществлена программная реализация информационной системы на языке C# в среде разработки Microsoft Visual Studio.

Разработанная информационная технология может быть использована для сжатия текстовой информации различных объемов, а также для целей информационного поиска.

Список литературы

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

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

3. Патент на корисну модель №82942 Україна, МПК51 G06F 17/21 (2006.01). Спосіб семантичної компресії тексту із заданим рівнем стислості/ Винахідники: С.В. Єгоров, І.М. Єгорова; Власник С.В. Єгоров. – № u2013 00978; заявл. 28.01.13; опубл. 27.08.13, Бюл. № 16.

4. Егоров С.В. Метод семантического сжатия текста / С.В. Егоров, И.Н. Егорова // Вестник НТУ «ХПИ»: вып. 9' 54 (1027)'2013: Математическое моделирование в технике и технологиях. – 2013. – С. 118-123.

5. Руководство Microsoft по проектированию архитектуры приложений, 2-е издание. – Майкрософт пресс, 2009. – 529 с.

Поступила в редколлегию 5.11.2013

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

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ СЕМАНТИЧНОГО СТИСНЕННЯ ТЕКСТУ

С.В. Єгоров

Розроблено інформаційну технологію семантичного стиснення текстової інформації. що базується на моделі та методі семантичного стиснення тексту. Реалізовано інформаційну систему, що надає можливість користувачу самостійно задавати рівень стиснення текстової інформації, та може бути використана для стиснення текстів будь-якого обсягу.

Ключові слова: *текст, семантичне стиснення, рівень стиснення, інформаційна система, інформаційна технологія, модель, метод.*

INFORMATIONAL TECHNOLOGY FOR SEMANTIC TEXT COMPRESSION

S.V. Iegorov

Informational technology for semantic compression of textual information was developed which is based on the model and method of semantic text compression. Information system was implemented which allows user to set compression rate by himself. The information system may be used for compression of documents of any volume.

Keywords: *text, semantic compression, level of compression, informative system, information technology, model, method.*