Grishunov Stepan Sergeevich, assistant, stepangrishunov@yandex.com, Russia, Kaluga, Kaluga Branch of Moscow Bauman State Technical University,

Chukhraev Igor Vladimirovich, candidate of technical science, docent, head of chair, igor.chukhraev@mail.ru, Russia, Kaluga, Kaluga Branch of Moscow Bauman State Technical University

УДК 004.4'414

ИЗВЛЕЧЕНИЕ ФАКТОВ ИЗ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА С ПРИМЕНЕНИЕМ КОНЦЕПТУАЛЬНЫХ ГРАФОВЫХ МОДЕЛЕЙ

М.Ю. Богатырев

Рассматривается применение методов концептуального моделирования в технологии извлечения фактов из текстовых данных. Текстовые данные представлены неструктурированными текстами естественного языка, образующими текстовый корпус. В качестве концептуальных моделей применяются концептуальные графы и решётки понятий. Применение концептуальных графов позволяет эффективно решать на текстах задачи извлечения именованных сущностей и отношений между ними. Эти решения используются при построении решёток понятий, которые служат источником данных в системе извлечения фактов. Извлечение фактов выполняется путём обработки запросов пользователей и нахождения соответствующих им понятий в решётке понятий .

Ключевые слова: концептуальное моделирование, концептуальные графы, решётки понятий, анализ формальных понятий.

В настоящее время возрастает интерес к созданию информационных систем обработки текстовых данных. Прежде всего, это связано с развитием сети Интернет, в которой тексты являются основным информационным ресурсом. В направлении исследований, известном как Text Mining [3], создано значительное число методов обработки текстовых данных, составляющих основу технологий автоматической рубрикации текстов, автоматического аннотирования текстов, классификации текстов и т.д.

Среди технологий обработки текстовых данных технологии извлечения фактов из текстов естественного языка имеют большое практическое значение и востребованы не только в Интернет-системах информационного поиска, но также в системах поддержки принятия решений, вопросно-ответных системах и им подобных.

Главной проблемой при разработке любой технологии извлечения фактов является проблема формализации понятия «факт». Сложность данной проблемы обусловлена среди прочего тем, что одна и та же информация, извлечённая из текста, например, дата, может считаться или не считаться фактом в зависимости от конкретных условий. Все эти условия составляют контекст запроса к системе извлечения фактов, который должен быть учтён в методах извлечения фактов.

Несмотря на обилие подходов, термин «факт» в современных работах, посвящённых анализу естественного языка, трактуется достаточно вольно. Под фактом понимают явление, событие, понятие, которое извлекается из текста, и представляет интерес для пользователя. При этом часто термины «факт» и «событие» не различаются, хотя с событиями логично связывать временные интервалы. Факт может представлять собой слово или множество слов. Многие системы извлечения фактов используют лексико–грамматические шаблоны, содержащие ключевые слова или словосочетания, с которыми сравниваются результаты обработки анализируемого текста. В этом случае факт задаётся шаблоном, а наличие факта в тексте диагностируется как соответствие текста такому шаблону [13].

В самом общем виде можно определить факт как отношение на множестве слов. Под такое достаточно общее определение подпадают все известные определения фактов в виде ключевых слов, словосочетаний, лексико-грамматических и лексико-семантических шаблонов. Но для его применения на практике необходимо исследовать конкретные варианты отношений в контексте решаемой задачи и реализовать способ использования данных отношений в извлечении фактов. Это являлось одной из целей данной работы.

В разрабатываемом методе извлечения фактов применяются две концептуальные модели: концептуальные графы и решётки понятий. Трактовка фактов как отношений на множестве слов позволяет корректно применить эти модели.

Все методы TextMining можно классифицировать по двум направлениям: методы, использующие статистики встречаемости слов в текстах, и методы, основанные на применении семантических моделей текста [3]. Такое же деление характерно и для методов решения задач извлечения фактов. Именно ко второму «семантическому» направлению в методах решения задач извлечения фактов относится рассматриваемый здесь метод.

Концептуальное моделирование. Концептуальное моделирование является достаточно широким направлением в моделировании [4]. Концептуальная модель в общем виде представляет собой множество объектов, связанных отношениями. Эти объекты, называемые «концепт», «концепция», «понятие», могут иметь различную природу. В концептуальных моделях чаще всего применяются бинарные отношения, но в них могут

входить и отношения произвольной арности. Такая универсальность позволяет применять концептуальные модели в самых разных областях [5, 6].

Концептуальная модель представляет собой граф, вершинами которого являются понятия, а стрелками или дугами – связи (отношения) между понятиями.

Одним из направлений концептуального моделирования является *анализ формальных понятий* (АФП) [1]**.** Концептуальной моделью здесь является *решётка понятий*.

Решётки понятий строятся следующим образом. Пусть имеются два множества: множество объектов G и множество принадлежащих им атрибутов M. Эти множества частично упорядочены некоторыми отношениями, обозначаемыми как ϕ и P соответственно: $G = (G, \phi)$, $M = (M, \mathrm{P})$. На данных множествах определяется *формальный контекст* $\mathbf{K} = (G, M, I)$, в котором связь между объектами их атрибутами задаётся отношением $I \subseteq G \times M$, которое представляет собой набор кортежей $< g, m > \in I$.

Связи между объектами и атрибутами задаются отображениями $A' : A \to B$ и $B' : B \to A$ со следующими свойствами полноты: $A' := \{y \in Y \mid \forall x \in A < x, y > \in I\}$, $B' := \{x \in X \mid \forall y \in B < x, y > \in I\}$. Пара подмножеств ($A$, B), таких, что $A' = B$, $B' = A$, называется *формальным понятием* контекста **K**. Как следует из условий полноты, в матрице контекста понятия (A, B) задаются максимальными по вложению подматрицами со всеми ненулевыми элементами. Множества A и B замкнуты в силу композиции отображений: $A'' = A$, $B'' = B$. Множество A образует *объем* формального понятия (A, B), а множество B – его *содержание*. Отношения частичного порядка ϕ, P на множествах G и M индуцируют отношение частичного порядка $\leq$ на множестве понятий.

Если для понятий (A_1, B_1) и (A_2, B_2) $\mathrm{A}_1 \subseteq \mathrm{A}_2$, что эквивалентно $\mathrm{B}_2 \subseteq \mathrm{B}_1$, то ($A_1$, B_1) $\leq$ (A_2, B_2). В этом случае логично считать понятие (A_1, B_1) менее общим, чем понятие (A_2, B_2). Формальный контекст имеет представление в виде матрицы инцидентности отношения I, в которой ненулевые элементы обозначают факт принадлежности атрибута $m \in M$ объекту $g \in G$.

Согласно основной теореме АФП частично упорядоченное по вложению объемов множество формальных понятий контекста **K** образует математический объект – *решётку* [15], которая называется «*решётка понятий*».

Решётка понятий, построенная на формальном контексте, является инструментом представления и извлечения знаний из данных контекста. В роли знаний выступают понятия, организованные иерархично. При этом граф решетки понятий не является деревом, что характерно для графов

многих концептуальных моделей, а имеет более общую структуру – структуру решётки. Это позволяет представлять знания, выражающиеся понятиями, характеризующимися меньшей и большей общностью, меньшими и большими объемом и содержанием.

В задаче извлечения фактов решётки понятий служат хранилищем фактов. Понятия – узлы решётки – интерпретируются как множество фактов определённого уровня (тематики), которое связано с другими фактами.

Построение формального контекста и решёток понятий на текстах требует выявления на них отношений принадлежности «объект – атрибут». Для этого применяются концептуальные графы [14] – двудольные графы, моделирующие семантику отдельного предложения в виде концептов и концептуальных отношений.

В работе [9] рассмотрены особенности построения концептуальных графов на текстах и применение их для генерации формальных контекстов и решёток понятий. Здесь, используя указанные результаты, рассмотрим в целом технологию извлечения фактов из текстов естественного языка и ее реализацию на примере исследования биотопов бактерий.

Метод извлечения фактов. Разрабатываемая технология извлечения фактов основана на следующем методе.

1. На предложениях обрабатываемых текстов строится множество концептуальных графов. Это множество строится методом, рассмотренным в [9], с применением стандартного морфологического анализатора и решения задачи разметки семантических ролей [2].

2. На множестве концептуальных графов решается задача их агрегирования. Агрегирование необходимо для исключения избыточной размерности концептуальных моделей, не связанной с полезной информацией. В качестве средства агрегирования применяется кластеризация концептуальных графов. Методы кластеризации и их параметры рассмотрены в работе [12].

3. На агрегированном множестве концептуальных графов строится формальный контекст. Формальный контекст представляет собой отношение на множествах объектов и их атрибутов и задаётся матрицей. Построение формального контекста на текстах является самой сложной задачей метода. Для ее решения необходимо построить алгоритм отбора лексических элементов текста в формальный контекст и выполнить исследование эффективности такого алгоритма.

4. На формальном контексте выделяются формальные понятия и строится другая концептуальная графовая модель – решётка понятий. Имея решётку понятий, можно выявлять связи между понятиями по принципу «общее – частное». Понятия – узлы решётки – интерпретируются как множество потенциальных фактов определённого уровня (тематики), которое связано с другими фактами.

5. Пункты 1 – 4 предлагаемого метода предназначены для подготовки данных, которые используются системой извлечения фактов, реализующей данный метод в виде информационной технологии. Извлечение фактов из текстов выполняется с использованием построенного фактографического интерфейса к решётке понятий и программной оболочки, позволяющей управлять диалогом пользователя с системой.

Реализация метода извлечения фактов. Разработанный метод опробован в решении задач классификации сообщений в системах технической поддержки [5], моделирования требований к информационным системам по текстам технических заданий [6], классификации запросов к биомедицинским системам [7].

Рассмотрим новое приложение данного метода в задаче исследования биотопов бактерий.

Биотопом называется область (ареал), занятая определённым биоценозом, например, бактериями. Задача нахождения биотопов бактерий сводится к идентификации названий бактерий, упоминаемых в текстах, и определении связей (отношений) между названиями бактерий и другими сущностями, обозначающими ареал обитания бактерий (почва, вода, внутренние органы людей и животных), а также отношений их к известным заболеваниям людей и животных. Тексты, содержащие информацию о бактериях, составляют корпус, доступный в сети Интернет.

В современном анализе данных извлечение новой нетривиальной информации из текстовых данных называется Text Mining. Применительно к текстам биомедицинской тематики сформировалось целое направление: Biomedical Text Mining [10]. Актуальность обработки текстовых данных в биоинформатике обусловлена естественными причинами: биомедицинские данные в большинстве своём носят описательный характер и представлены текстами. В современных информационных системах тексты хранятся в полнотекстовых базах данных или корпусах. Корпус представляет собой базу данных, дополненную *системой разметки*. Разметка представляет собой метаданные для данных корпуса – это может быть, например, сопоставление каждому слову его части речи, а также другая специальная информация.

Исследуемый корпус биотопов бактерий размечен так, что каждому тексту описаний бактерий соответствует не только название бактерии, но также её кодовое обозначение и указание на возможное семейство бактерий, к которым она относится. Данная информация используется при обработке текста и построении решётки понятий.

Рассмотрим реализации основных положений предлагаемого метода извлечения фактов на примере текстов англоязычного корпуса биотопов бактерий [11].

Построение формальных контекстов и решётки понятий. Как уже отмечалось, в методе используются концептуальные графы в качестве семантической модели предложений текста. Особенности применения концептуальных графов для формирования формальных контекстов иллюстрируются следующим примером.

Рассмотрим предложение из текста описания бактерии, известной как Burkholderia phytofirmans: «*Burkholderia phytofirmans belongs to the beta-Proteobacteria and was isolated from surface-sterilized Glomus vesiculiferum-infected onion roots*». Перевод приложения: «*Бактерия Burkholderia phytofirmans принадлежит к бета-протеобактериям и была выделена со стерилизованной поверхности клубней лука, инфицированного бактериями вида vesiculiferum*».

Концептуальный граф данного предложения показан на рис. 1.

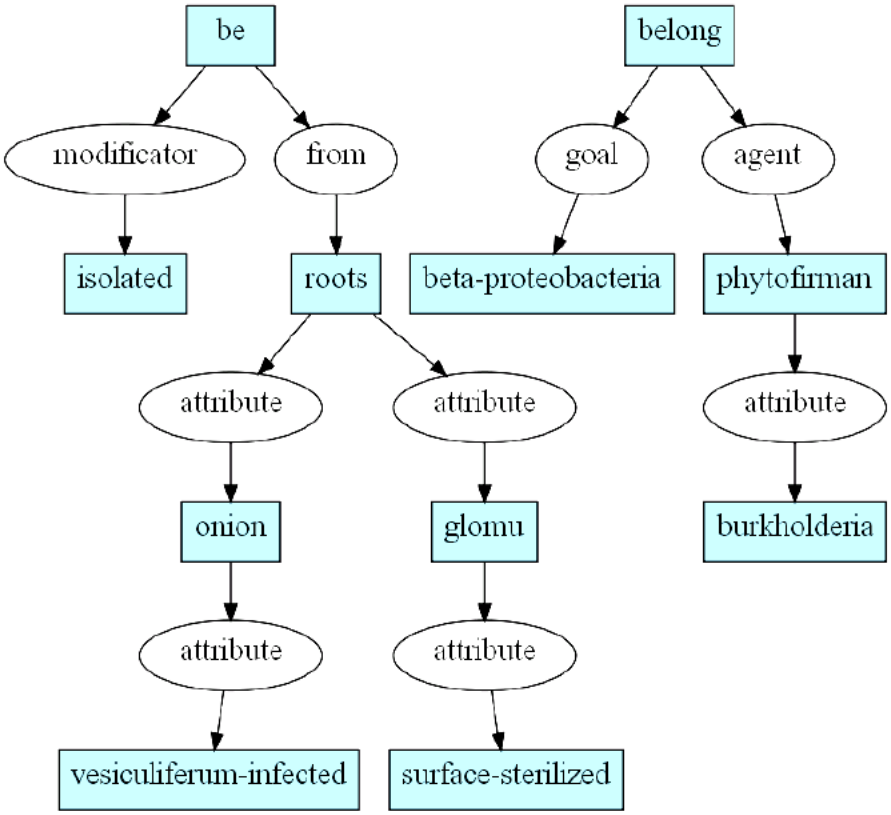

Рис. 1. Концептуальный граф предложения, описывающего бактерию

В графе на рис. 1 пять отношений «атрибут», но не все из них пригодны для включения в формальный контекст. Алгоритм построения концептуальных графов связал слова *Burkholderia phytofirmans* отношением «атрибут», однако эта пара слов является собственным именем бактерии и должна применяться как одно целое. Используя разметку корпуса текстов о бактериях, осуществляется необходимая коррекция подобных решений.

Формальный контекст по определению содержит объекты и их атрибуты. Концептуальные графы позволяют найти отношения связывающие пары слов в предложениях, как это показано на рис. 1. Формально отношения «атрибут» в концептуальных графах должны учитываться при построении контекста и соответствующие им пары слов (например, «roots» – «onion» на рис. 1) входить в формальный контекст.

Имея концептуальные графы, есть возможность рассматривать и другие отношения в качестве отношений «объект – атрибут» с целью включения их в формальный контекст. Выполненные авторами исследования [8] демонстрируют невысокую эффективность данного решения: применение других известных из лингвистики отношений («генитив», «локация» и т.д.) не вносит в формальный контекст существенной информации.

Принципиальным решением здесь является использование концептуальных графов целиком либо их подграфов. В самом деле, как следует из характерного примера на рис.1, информация, содержащаяся в тексте, его смысл, могут быть заданы предикатными формами, включающими глаголы, и представляющими собой подграфы концептуального графа. Так, основной смысл предложения, моделируемого графом на рис. 1, состоит в том, что: а) бактерия Burkholderia phytofirmans принадлежит бета-протобактериям; б) эта бактерия заражает клубни лука.

В результате исследования различных способов формирования формальных контекстов на множестве концептуальных графов были сформулированы следующие правила.

1. В концептуальных графах фиксируются предикатные формы вида <*субъект*> - < *предикат*> - <*объект*>. В качестве предиката, как правило, выступает глагол. В концептуальном графе может быть несколько таких форм – подграфов. Каждый подобный подграф формирует элементы формальных контекстов, доставляя в них предикаты в качестве объектов и связанные с ними слова в качестве атрибутов.

2. При обработке концептуальных графов учитывается иерархия их элементов. Верхние элементы графа, близкие к основному глаголу, отражают основной смысл фразы, соответствующей графу, а далее идущие элементы – детали. В целом, это соответствует принципу построения предложений обычного текста. На рис. 1 левый подграф концептуального графа имеет иерархию концептов. Основной смысл фразы, соответствующей данному подграфу, состоит в том, что бактерия Burkholderia phytofirmans выделена из корней – клубней.

3. При построении концептуальных графов и их обработке для создания формального контекста поддерживаются типы концептов, описанные в [9]. Кроме поддержки типов, при построении формального контекста используются ключевые слова и словосочетания. Например, это названия бактерий, их аббревиатуры и кодовые обозначения. Эти слова являются обязательными для включения их в контекст. Другие слова включаются,

если они связаны с данными словами отношениями из другого заданного набора – набора отношений. Среди подобных слов выделяются глаголы, которые используются в формальном контексте согласно предикатной модели фактов.

Извлечение фактов. Система извлечения фактов строит решётку понятий на исследуемых текстах фиксированной тематики и использует решётку в качестве хранилища потенциальных фактов. Каждый новый текст обрабатывается системой с учётом существующей решётки и может изменять ее структуру. Извлечение фактов выполняется путём обработки запросов пользователей. Запрос представляет собой либо произвольный текст, либо текст с элементами, взятыми из списков, доступных через интерфейс системы. Списки формируются на основе тематики текстов и отражают востребованность определённых тем в запросах. Если запрос представляет собой произвольный текст (достаточно короткий), то он преобразуется в концептуальный граф и далее его элементы – концепты и отношения – используются в качестве запроса. В общем случае пользовательский запрос определяет тематику факта. Этой тематике соответствует одно или несколько понятий решётки понятий.

Навигация в решётке понятий позволяет определять своеобразные макро-факты, которыми являются ее понятия, и более конкретные факты внутри понятий. Ценность представленных данных как фактов определяется пользователем. Пользователь управляет глубиной и широтой поиска фактов с помощью интерфейса. Широта поиска задаётся определённым уровнем решётки. Глубина поиска определяется количеством уровней решётки, используемых при поиске фактов.

Анализ результатов. Для проверки разработанного метода были выполнены вычислительные эксперименты на текстовом корпусе, содержащем описания бактерий. В экспериментах обрабатывались тексты описаний 130 наиболее известных бактерий при помощи системы моделирования решёток понятий [8].

Рассмотрим пример извлечения фактов с помощью решётки понятий. Визуально извлечение фактов выполняется при помощи видов, строящихся в результате обработки запроса пользователя. На рис. 2 показан пример вида как фрагмента решетки понятий, отражающего свойство грам-отрицательности шести бактерий: *Borrelia turicatae, Frankia, Legionella, Clamydophila, Thermoanaerobacter tengcongensis, Xanthomonas oryzae*. Известно, что бактерии со свойством грам-отрицательности резистентны к обычным антибиотикам, поэтому выявление такого свойства весьма важно.

Из рис. 2 извлекаются следующие факты:

только три бактерии из рассматриваемых, а именно Thermoanaerobacter tengcongensis, Clamydophila и Xanthomonas oryzae, являются грам-отрицательными;

две грам-отрицательные бактерии – Thermoanaerobacter tengcongensis и Xanthomonas oryzae – имеют форму палочки;

одна грам – отрицательная бактерия Clamydophila является существенно патогенной.

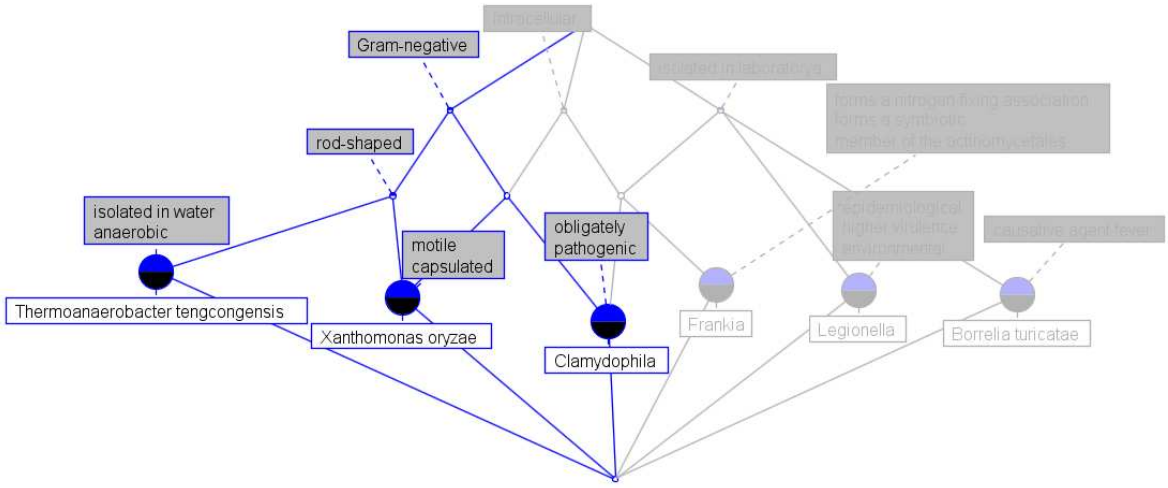

Рис. 2. Вид в решётке понятий, соответствующий свойству грам-отрицательности бактерий

При этом термин «существенно патогенная» сформирован из двух слов – obligately и pathogenic, являющихся ключевыми.

Для визуализации решёток понятий используется программное средство [16].

Главным преимуществом применяемого метода извлечения фактов является возможность не только поиска отдельных ключевых слов, но также извлечения понятий, соответствующих используемой в текстовых данных терминологии.

Недостатком метода является необходимость его настройки на тексты определённой тематики. Применение корпусов текстов позволяет использовать их разметку при настройке, что сокращает объем операций, выполняемых с данными.

Работа выполнена при поддержке РФФИ, грант № 15-07-05507.

Список литературы

1. Bernhard G., Gerd S., Rudolf W., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag. 2003. No. 3626.

2. Gildea D., Jurafsky D. Automatic labeling of semantic roles // Computational Linguistics. 2002. Vol. 28. P. 245 – 288.

3. Kao A. and Poteet S. Natural Language Processing and Text Mining. London: Springer-Verlag, 2007.

4. Olivé Antoni. Conceptual Modeling of Information Systems. Springer-Verlag. Berlin, Heidelberg, 2007.

5. Bogatyrev M. and Kolosoff A. Using Conceptual Graphs for Text Mining in Technical Support Services. Pattern Recognition and Machine Intelligence // Lecture Notes in Computer Science. 2011. Vol. 6744/2011, P. 466 – 471.

6. Bogatyrev M., Nuriahmetov V. Application of Conceptual Structures in Requirements Modeling // Proc. of the International Workshop on Concept Discovery in Unstructured Data (CDUD 2011) at the Thirteenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - RSFDGrC 2011. M., 2011. P. 11-19.

7. Богатырев М.Ю., Вакурин В.С. Концептуальное моделирование в исследовании биомедицинских данных // Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 340 – 349.

8. Bogatyrev M., Samodurov K. Framework for Conceptual Modeling on Natural Language Texts // Proc. of the International Workshop on Concept Discovery in Unstructured Data (CDUD 2016) at the Thirteenth International Conference on Concept Lattices and Their Applications. M., 2016. P. 13 – 24.

9. Богатырев М.Ю., Нуриахметов В.Р., Вакурин В.С. Методы анализа формальных понятий в информационных системах технической поддержки // Известия Тульского государственного университета. Технические науки. Тула: Изд-во ТулГУ, 2013. Вып. 2. С. 25 – 36.

10. Shatkay H, Craven M. Biomedical Text Mining. Cambridge, Massachussets: MIT Press, 2007.

11. Bossy BioNLP Shared Task - The Bacteria Track // BMC Bioinformatics. 2012. Vol. 13 (Suppl. 11). P. 1 – 15.

12. Богатырев М.Ю., Латов В.Е., Столбовская И.А. Применение концептуальных графов в системах поддержки электронных библиотек. Электронные библиотеки: перспективные методы и технологии, электронные коллекции // Труды Девятой Всероссийской научной конференции RCDL’2007. Т. 2, С. 104 – 110.

13. Jiang Jing, Information Extraction from Text // Mining Text Data. Springer. 2012. 524 p.

14. Sowa J.F. Conceptual Structures: Information Processing in Mind and Machine. London: Addison-Wesley, 1984.

15. Биркгоф Г. Теория решеток. М.: Наука, 1984. 284 с.

16. Евтушенко С.А. Система анализа данных "CONCEPT EXPLORER"// КИИ-2000 // Труды конференции. М.: Изд-во физ.- мат. литературы. 2000.

Богатырев Михаил Юрьевич, д-р техн. наук, проф., okkambo@mail.ru, Россия, Тула, Тульский государственный университет

FACT EXTRACTION FROM NATURAL LANGUAGE TEXTS WITH CONCEPTUAL GRAPH MODELS

M.Yu. Bogatyrev

Applications of methods of conceptual modeling in the fact extraction technology from textual data are considered. Textual data is presented as non-structured natural language texts, forming a text corpus. Conceptual graphs and concept lattices are used as conceptual models. The use of conceptual graphs allows to solve effectively the problems of named entity recognition and relations extraction. These solutions then applied for building concept lattices which serve as a data source in the fact extraction system. Extraction of facts is doing by processing user requests and finding the corresponding concepts in the concept lattice..

Key words: conceptual modeling, conceptual graphs, conceptual lattices, formal concept analysis.

Bogatyrev Mikhail Yurievich, doctor of technical sciences, professor, okkambo@mail.ru, Russia, Tula, Tula State University

УДК 621.391:519.72

РАЗРАБОТКА МОДЕЛИ СИСТЕМЫ ПЕРЕДАЧИ ДАННЫХ С МНОГОПОРОГОВЫМ ДЕКОДЕРОМ САМООРТОГОНАЛЬНЫХ КОДОВ С ПРИМЕНЕНИЕМ ТЕХНОЛОГИИ OPENCL

Д.С. Демидов, Г.В. Овечкин

Выполнен анализ особенностей моделирования систем передачи данных с многопороговым декодированием МПД самоортогональных кодов. Показано, что для ускорения процесса моделирования можно использовать вычислительные ресурсы GPU. Разработана компьютерная модель системы передачи данных с МПД с использованием GPU, рассмотрены возможности ее ускорения. Показано, что применение предложенной модели позволяет увеличить скорость работы модели до 40 раз по сравнению с аналогичной моделью на CPU.

Ключевые слова: системы передачи данных, помехоустойчивое кодирование, многопороговое декодирование, компьютерное моделирование, OpenCL, GPU, CPU.

Быстрый рост объемов обработки данных, развитие цифровых систем вещания и вычислительных сетей предъявляют высокие требования к минимизации ошибок в используемых цифровых данных. Поэтому одной из важнейших задач является обеспечение высокой достоверности передачи данных [1]. Решением задачи обеспечения высокой достоверности передачи данных занимается помехоустойчивое кодирование.