УДК 004.912

М.А. Павленко

*Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков*

# АНАЛИЗ МЕТОДОВ РЕШЕНИЯ ЗАДАЧИ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ

*Одной из ключевых задач при разработке систем поддержки принятия решений является синтез и анализ текстовых сообщений. Однако, решение данной задачи также необходимо и для выявления необходимых данных, поиска новой информации, установления неявных связей, семантического сжатия и ряда других задач. На сегодняшний день разработано большое количество методов обработки текстовых сообщений, позволяющих решать задачи такого класса, однако остаются не решенными задачи машинного понимания текста, оценки его значимости, эмоциональной оценки и другие. В статье сделана попытка выработки подхода к решению задач, связанных с информационным анализом текста.*

***Ключевые слова:*** *анализ текста, аппарат формализации, обработка текста.*

## Введение

На сегодняшний день одним из перспективных направлений обработки текстовой информации является извлечение информации из текста. Извлечение информации является интеллектуальным процессом [1 – 4] и не может быть решено в полной мере существующими информационными методами [5 – 8]. Решение данной задачи требует привлечения аппарата искусственного интеллекта и эвристических методов обработки информации [9 – 14]. Под извлечением информации будем понимать автоматическое извлечение структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов [15].

Извлечение информации является разновидностью информационного поиска, связанного с обработкой текста на естественном языке. Главная цель извлечения информации – возможность анализа изначально «хаотичной» информации, представленной в виде текста с помощью стандартных методов обработки данных [1 – 15]. Более узкой целью

может служить, например, задача выявления логических закономерностей, логических соответствий, неполноты и других структур на основании данных, представленных в текстах.

Большой объем информации, циркулирующей в информационном пространстве, а также ее низкая структурированность обуславливают повышение значимости роли такой процедуры, как извлечение информации из текста. Задача преобразования текста в более структурированную форму посредством преобразования в реляционную форму или добавлением XML разметки [14 – 16] является перспективной и потребует своего отдельного решения с учетом темпов роста объемов информации в общедоступном виде [14].

Типичной задачей извлечения информации может служить следующая: сканирование набора документов, написанных на естественном языке, и наполнение базы данных выделенной полезной информацией. Современные процедуры извлечения информации, использующие методы обработки текстов на естественном языке, как правило, направлены лишь на решение только очень узкого класса задач (отбор ограниченного набора тем (вопросов, проблем), а зачастую и только на одну тему). Это привело к тому, что разработаны эффективные процедуры обработки текста, но их наличие не позволяет применить полученные результаты в качестве универсального метода обработки текста [1 – 14].

В классе задач обработки текстовой информации отдельное место занимает проблема генерации текстов. Для систем логического вывода были предприняты попытки разработки методов генерации текстовых пояснений результатов вывода [16 – 18]. Но, в общем случае, эту проблему решить не удается из-за высокой сложности задачи [19 – 21]. Кроме того, была попытка реализации функции генерации текста [18], однако предложенный метод не позволяет получить читаемый текст на произвольную тему, а лишь генерирует ограниченный по объему и содержанию текст заданной тематики.

**Цель работы.** В статье предпринята попытка формулирования задачи разработки универсального метода анализа текстовых сообщений для выявления сложных структур и поиска неявных данных.

**Анализ литературы** свидетельствует о том, что в настоящее время активно ведется разработка программных средств, позволяющих автоматизировать процессы обработки текстовой информации [1 – 26]. В [8] предложена классификация существующих программных средств по их назначению, а в [8, 11, 15, 21, 22, 24 – 26] представлен систематизированный анализ методов, позволяющих обнаруживать и извлекать из текста конструктивные элементы. Работы [21 – 26] доказывают, что именно использование лингвистических методов позволяет существенно улучшить качество автоматического анализа текста.

Традиционно в системах анализа текстов для представления знаний используется четыре типа моделей: продукционная, формально-логическая, фреймовая и семантико-сетевая модели. На базе этих моделей описываются решения и основные перспективы их использования [21 – 26].

Следует отметить перспективность использования функциональных систем управления для решения задач анализа текста. При этом ядром таких систем могут стать интегрированные интеллектуальные информационные системы, включающие элементы искусственного интеллекта, основанные на методах и средствах теории интеллекта [11 – 18].

## Основная часть

Существующее многообразие частных задач обработки текстовой информации позволяет сгруппировать их в следующие крупные классы, связанные с анализом текстовой информации [1 – 7]:

1. Распознавание именованных элементов (сущностей), например: имён людей, названий организаций, географических названий и пр.
2. Разрешение анафоры и кореференций: поиск связей, относящихся к одному и тому же объекту. Типичный случай таких ссылок – местоименная анафора.
3. Выделение терминологии: нахождение для данного текста ключевых слов и словосочетаний (коллокаций).
4. Автореферирование: выделение из текста смысловой, эмотивной, оценочной и пр. информации. Бывает генеративным и декларативным.
5. Корпусная лингвистика, создание и использование электронных корпусов текстов
6. Создание электронных словарей, тезаурусов, онтологий. Словари используют, например, для автоматического перевода, проверки орфографии.
7. Автоматический перевод текстов.
8. Автоматическое извлечение фактов из текста (извлечение информации).
9. Построение систем управления знаниями.
10. Создание вопросно-ответных систем.
11. Информационный поиск.

Для решения отмеченных классов задач применяются следующие основные методы [1 – 21]:

- Data Mining;
- ассоциативные правила;
- продукционная модель;
- формально-логическая модель;
- фреймовая модель;
- семантико-сетевая модель;
- дерево решений;
- кластеры;
- математические функции и др.

Анализ приведенных методов показывает, что каждый из них обладает хорошо разработанной формальной системой, которая позволяет достаточно полно описывать сущности и процессы различных предметных областей. Однако все они имеют существенный недостаток при своем использовании. Они не позволяют описывать нелогичность, неполноту и противоречивость текста, которые являются отражением естественного языка. Возникает парадокс, связанный с ограниченными возможностями логических формальных систем и необходимостью описания как логичных особенностей, так и не логичных знаний и данных, содержащихся в текстах.

При этом часто получают результат, предсказанный Лотманом М.Ю. в книге "Мандельштам и Пастернак: (попытка контрастивной поэтики)": "*Методы анализа текста дают результаты, которые в большей степени характеризуют сами методы, нежели тот текст, который при их помощи якобы описывается. Анализ текста есть перевод его на метаязык исследователя, сопоставление его с идеальной моделью, порождённой этим метаязыком. А всякий перевод связан, помимо всего прочего, с определённым искажением оригинала: метаязык (как и любой другой язык) что-то навязывает языку объекту, а чего-то не замечает в нём*" [31].

Таким образом, результаты, получаемые с помощью перечисленных выше методов, ограничены возможностью самих методов. Было бы интересно рассмотреть проблему полноты совокупности данных методов по решению проблемы анализа текстов и выявлению тех областей знаний, которые имеют место быть, но не подкреплены существующими формальными системами и методами обработки.

Решение подобной задачи возможно при описании всей совокупности задач анализа текста и проведения всестороннего анализа возможностей существующих методов по их решению. Это, в свою очередь, позволит сформулировать задачи разработки методов для решения тех задач, которые не разрешимы существующими методами.

Кроме того, при решении задачи создания среды для информационной обработки текстов необходимо помнить о необходимости создания специфических баз данных и знаний, направленных на хранение информации и методов ее обработки для реализации процедур анализа текстов. Это, в свою очередь, требует разработки стандартов таких баз данных и знаний, что позволяет сконцентрировать усилия на разработке эффективных процедур и алгоритмов анализа данных, а не на разработке уникальных структур баз знаний и данных.

С ростом объемов баз данных и знаний потребуется разработка эффективных процедур их использования и обработки, что будет являться отдельным направлением исследований в данной области. Полученные результаты могут приводить к изменению принципов и стандартов построения баз данных и знаний, что повысит их эффективность и усовершенствует процесс обработки текстов. Систематизируя изложенное, можно сказать, что существующие на сегодняшний день походы к реализации процедуры анализа текстов могут быть реализованы в рамках структуры, приведенной на рис. 1.

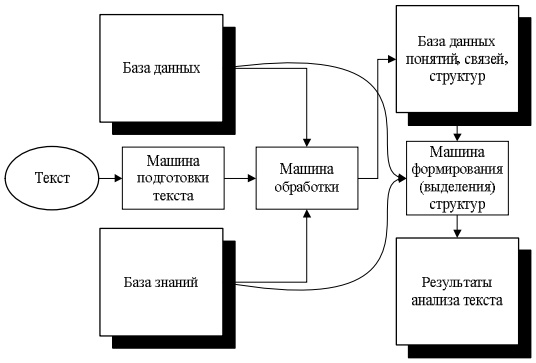


Рис. 1. Структура среды обработки текста

Исходя из представленной среды обработки текстов, можно более полно описать структуру блоков базы данных и знаний. Такая детализация помогает понять возникающие проблемы и сформировать саму структуру базы знаний и данных.

Несмотря на то, что на сегодняшний день для разработки баз данных разработаны и опробованы стандартные методы, разработка баз знаний требует своего отдельного рассмотрения. Исходя из перечня задач, которые должна решать система обработки текстов, перспективным видится создание гибридных баз знаний. Такие базы знаний могут позволить объединить разнотипные методы формализации знаний в гибридную систему. Однако это породит большое количество проблем, связанных с их построением и организацией логического вывода на них [27 – 29].

Преодоление проблем, связанных с обработкой гибридных моделей знаний, может быть реализовано путем перехода к обобщенному методу формализации, который может позволить описать модели знаний гибридной модели. В качестве такого обобщенного аппарата формализации может служить теория нечетких множеств. При этом накладываются ограничения на используемые модели знаний, в данном случае это должны быть: исчисление предикатов, структуры целевых установок (частный случай семантической сети) и модальная логика. Кроме того, если рассматривать методы, на основании которых решаются задачи обработки текста, станет видно, что и теория нечетких множеств также входит в обобщенную гибридную модель.

Для решения задачи анализа текста предлагается воспользоваться другим подходом. Принципиальным отличием предлагаемого подхода является

то, что на первом этапе обработки предлагается разработать аппарат преобразования текста из естественно-языковой формы в формальную модель, являющуюся отражением реального текста в виде формальной структуры. Суть и необходимость данного преобразования аналогичны преобразованию Лапласа в математике (одной из особенностей которого является то, что многим соотношениям и операциям над оригиналами соответствуют более простые соотношения над их изображениями) [30].

В таком случае структура среды обработки текста приобретет следующий вид (рис. 2).

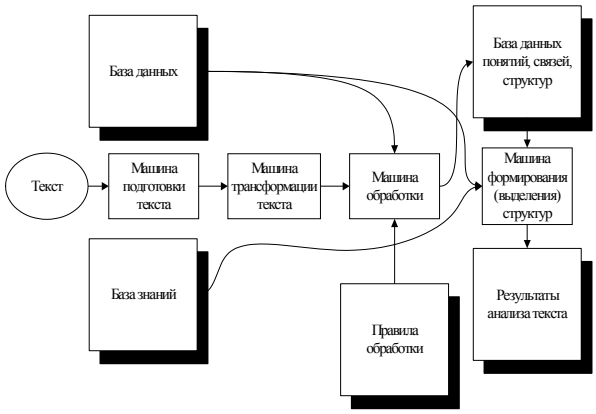


Рис. 2. Перспективная структура среды обработки текста

В данной структуре машина трансформации текста должна преобразовать текст в некую формальную структуру, над которой возможно проведение простейших операций сложения, умножения, деления и вычитания. Такой структурой может быть, например, векторное представление структурных единиц текста: буквы, слова, фразы и т.д.

Нахождение адекватного физического описания (представления) позволит наглядно представлять результаты преобразования и обработки текста. Использование такого подхода может позволить существенно упростить процессы анализа текстов, применять единый подход к анализу текстов любой направленности и содержания, проводить алгебраические операции вместо процедур логического вывода.

Однако использование данного метода не снимает вопросов хранение огромных объемов информации в процессе обработки текстов и хранения полученных результатов. Это оставляет актуальной задачу разработки структур баз данных для решения задач обработки текстов.

## Выводы

Обработка текстов на естественном языке представляет собой сложную и противоречивую задачу. Текст является лишь отражением нашего знания и сознания возможностями языка. За формами текста скрывается и подтекст, и временная привязка, и скрытые знания, и пласты культуры, и многие другие факторы, которые отражают и, в тоже время, влияют и на автора, и на лицо, воспринимающее текстовую информацию. Нахождение единого формального аппарата для решения задачи анализа текста позволило бы решать задачи реферирования, анализа, перевода и многие другие. Однако, разработка гибридных моделей знаний для решения таких задач – это, скорее всего, тупиковое направление исследований. Наиболее перспективным видится разработка новых подходов к решению данной задачи. Одним из таких подходов может быть перенос задачи из сферы синтаксического и семантического анализа текста в сферу алгебраических преобразований над текстом при условии переноса его в новую формальную форму. Что может быть такой формой? Одной из возможных форм может быть векторное представление элементов текста, что позволит использовать новые подходы к его обработке. Однако детальная проработка этого вопроса требует проведения целого ряда дополнительных исследований.

## Список литературы

*1. Baeza-Yates R. Modern Information Retrieval / R. Baeza-Yates, B. Ribeiro-Neto. — Addison-Wesley, 1999. – ISBN 0-201-39829-X.*

*2. Manning C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schütze. – Cambridge University Press, 2008. – ISBN 0-521-86571-9.*

*3. Ландэ Д.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. – М.: Либроком (Editorial URSS), 2009. – 264 с. — ISBN 978-5-397-00497-8.*

*4. AI3's Inaugural State of Tooling for Semantic Technologies / Adaptive Information Adaptive Innovation Adaptive Infrastructure // URL ISSN 2079-0031. Вестник НТУ "ХПИ", 2012, № 62 (968). [Электронный ресурс]. – Режим доступа к ресурсу: http://www.mkbergman.com/991/ the-state-of-tooling-for-semantic-technologies.*

*5. Ермаков А.Е. Автоматизация онтологического инжиринга в системах извлечения знаний из текста /А.Е. Ермаков // Труды Международной конференции Диалог'2008. – Москва: Наука, 2008. – С. 136-140.*

*6. Corcho O. Methodologies, tools, and languages for building ontologies. Where is their meeting point? / O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez // Data & Knowledge Engineering. – 2003. – 46. – P. 41-64.*

*7. Buitelaar P. Ontology Learning from Texts: An Overview / P. Buitelaar, P. Cimiano, B. Magnini // In Ontology Learning from Text: Methods, Evaluation and Applications, 2005. – Vol. 123. – P. 234-265.*

*8. Simperl E. Achieving Maturity: the State of Practice in Ontology Engineering / E. Simperl, M. Mocho // In International Journal of Computer Science and Applications, Technomathematics Research Foundation. – 2010. – Vol. 7. – № 1. – P. 45-65.*

*9. Makki J. Semi Automatic Ontology Instantiation in the domain of Risk Management / J. Makki, A.-M. Alquier, V. Prince // In IFIP, Advances in Information and Communication Technology. – 2008. – Vol. 288. – P. 254.*

*10. Builеaar P. Topic extraction from scientific literature for competency management / P. Builеaar, T. Eigner // In*

*The 7th International Semantic Web Conference PICKME 2008, 27 octobre Karlsruhe, Germany. – P. 55-67.*

*11. Zhou L. Ontology Learning: State of the Art and Open Issues / L. Zhou // Information Technology and Management. – 2007. – 8 (3). – P. 241-252.*

*12. O'Brien I.A. Management Information Systems: A Managerial End User Perspective / I.A. O'Brien. – Boston: IRVIN, 1990. – 650 p.*

*13. Уотермен Д. Руководство по экспертным системам / Д. Уотермен. – М.: Мир, 1989. – 388 с.*

*14. Белоногов Г.Г. Компьютерная лингвистика и перспективные компьютерные технологии / Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов. – М.: "Русский мир", 2004. – 248 с.*

*15. Дубровин А.Д. Интеллектуальные информационные системы: Учеб. пособие. Ч.1 / А.Д. Дубровин. – М.: МГУКИ, 2008. – 232 с.*

*16. Бажин И.И. Информационные системы менеджмента / И.И. Бажин. – М.: ГУ – ВШЭ, 2002. – 688 с.*

*17. Гаскаров Д.В. Интеллектуальные информационные системы / Д.В. Гаскаров. – М.: ФГУП "Издательство Высшая школа", 2003. – 464 с.*

*18. Семенов Н.А. Интеллектуальные информационные системы / Н.А. Семенов. – Тверь: Издательство ТГТУ, 2004. – 168 с.*

*19. Business Process Reengineering: The Oracle Perspective. ORACLE CONSULTING, 1994.*

*20. Hammer M. Reengineering the Corporation. A Manifesto for Business Revolutions / M. Hammer, J. Champy. – HarperBusiness, 1993.*

*21. Шабанов-Кушнаренко Ю.П. Компараторная идентификация лингвистических объектов: Монография / Ю.П. Шабанов-Кушнаренко, Н.В. Шаронова. – К.: ИСДО, 1993. – 116 с.*

*22. Бондаренко М.Ф. Теория интеллекта. Учебник / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Х.: Изд-во СМИТ, 2007. – 576 с.*

*23. Шаронова Н.В. Использование метода компараторной идентификации для разбиения семантического пространства предметной области знание-ориентированых систем / Н.В. Шаронова, Н.Ф. Хайрова // Вестн. Херсон. гос. техн. ун-та. – Херсон, 2007. – № 4 (27). – С. 39–42.*

*24. Шаронова Н.В. Метод персонификации интеллектуального корпоративного ресурса компании / Н.В. Шаронова, В.А. Тарловский, Н.Ф. Хайрова // Вісник Національного технічного університету "Харківський політехнічний інститут". Збірник наукових праць. Тематичний випуск: Інформатика і моделювання. – Х.: НТУ "ХПІ", 2009. – № 43. – С. 191-197.*

*25. Оробинская Е.А. Языковая компетенция информационных систем / Е.А. Оробинская, О.И. Король, Н.В. Шаронова // Вісник Національного технічного університету "Харківський політехнічний інститут". Збірник наукових праць. Серія: Інформатика та моделювання. – Х.: НТУ "ХПІ". – 2012. – № 62 (968). – С. 148-154.*

*26. Шаронова Н.В. Использование логических сетей для формирования баз знаний логического типа / Н.В. Шаронова, В.А. Тарловский, Н.Ф. Хайрова // Вісник Херсонського національного технічного університету. – Херсон: ХНТУ. – 2011. – № 41. – С. 184-188.*

*27. Павленко М.А. Разработка метода многоэтапной формализации знаний о процессе распознавания оперативно-тактических ситуаций / М.А. Павленко, П.Г. Бердник, С.В. Кукобко, Ю.В. Данюк // Системи обробки інформації. – Х.: ХУ ПС, 2012. – Вип. 5(103). – С. 60-64.*

*28. Павленко М.А. Разработка процедуры формализации модальных знаний с использованием теории нечетких множеств для экспертных систем реального времени / М.А. Павленко, А.И. Тимочко, А.Н. Бесчасный, В.П. Докучаев // Збірник наукових праць ХУ ПС. – Х.: ХУ ПС, 2012. – Вип. 3(32). – С. 122-125.*

*29. Павленко М.А. Метод формализации знаний о процессе распознавания ситуаций нарушения правил движения воздушными судами / М.А. Павленко// Системи управління, навігації і звя'зку. – К.: ДП «ЦНДІ НіУ», 2012. – Вип. 2(22). – С. 86-92.*

*30. Диткин В.А. Интегральные преобразования и операционное исчисление / В.А. Диткин, А.П. Прудников. – М.: Главная редакция физико-математической литературы издательства «Наука», 1974. – 544 с.*

*31. Лотман М.Ю. Мандельштам и Пастернак: (попытка контрастивной поэтики) / М.Ю. Лотман. – Таллинн: Александра, 1996. – 175 с.*

*Поступила в редколлегию 19.08.2013*

**Рецензент:** д-р техн. наук, доц А.И. Тимочко, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

## АНАЛІЗ МЕТОДІВ РІШЕННЯ ЗАВДАННЯ ДОБУВАННЯ ІНФОРМАЦІЇ З ТЕКСТІВ

М.А. Павленко

*Одним з ключових завдань при розробці систем підтримки прийняття рішень є синтез і аналіз текстових повідомлень. Однак, рішення даної задачі також необхідно і для виявлення даних, пошуку нової інформації, встановлення неявних зв'язків, семантичного стиснення і ряду інших завдань. На сьогоднішній день розроблено велику кількість методів обробки текстових повідомлень, що дозволяють вирішувати завдання такого класу, однак залишаються не вирішеними завдання машинного розуміння тексту, оцінки його значимості, емоційної оцінки та інші. У статті запропонована спроба вироблення підходу до вирішення завдань, пов'язаних з інформаційним аналізом тексту.*

***Ключові слова:*** *аналіз тексту, апарат формалізації, обробка тексту.*

## ANALYSIS METHODS FOR SOLVING PROBLEMS INFORMATION EXTRACTION

M.A. Pavlenko

*One of the key objectives in the development of decision support systems is the synthesis and analysis of text messages. However, the solution of this problem is also needed to identify the required data, the search of new information, the establishment of implicit links, semantic compression and various other tasks. To date, developed a number of methods for processing text messages, allowing to solve problems of this class, however, remain unsolved problem of machine understanding of the text, evaluate its worth, emotional evaluation, and others. The article is an attempt to develop an approach to solving problems related to the information text analysis.*

***Keywords:*** *text analysis, machine formalization of word processing.*