

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ  
НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
«ХАРЬКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ»

На правах рукописи



**АДЖИТ ПРАТАП СИНГХ ГАУТАМ**

УДК 004.912:007.51

**ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ  
ЭКСТРАКЦИИ БИЗНЕС ЗНАНИЙ ИЗ ТЕКСТОВОГО КОНТЕНТА  
ИНТЕГРИРОВАННОЙ КОРПОРАТИВНОЙ СИСТЕМЫ**

Специальность 05.13.06 – информационные технологии

Диссертация на соискание ученой степени

кандидата технических наук

Научный руководитель  
Шаронова Наталья Валерьевна,  
доктор технических наук,  
профессор

Харьков – 2016

## СОДЕРЖАНИЕ

СПИСОК ИСПОЛЬЗУЕМЫХ СОКРАЩЕНИЙ.....	4
ВВЕДЕНИЕ.....	5
РАЗДЕЛ 1	
АНАЛИЗ ПРОБЛЕМЫ ЭКСТРАКЦИИ ЗНАНИЙ ИЗ ТЕКСТОВОГО КОНТЕНТА И ПОСТАНОВКА ЗАДАЧ ИССЛЕДОВАНИЯ .....	
1.1. Обзор исследований в области задач автоматизации интегрированных корпоративных систем.....	13
1.2. Анализ основных проблем обработки текстов в интегрированных корпоративных информационных системах.....	19
1.3. Сравнительный анализ существующих моделей представления знаний в интеллектуальных информационных системах .....	25
1.4. Аналитический обзор существующих технологий извлечения знаний из текстового контента.....	30
1.5. Обзор исследований по экстракции и идентификации знаний из текстового контента и постановка задач исследования.....	36
РАЗДЕЛ 2	
МАТЕМАТИЧЕСКИЙ АППАРАТ ДЛЯ МОДЕЛИРОВАНИЯ ПРОЦЕССОВ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ В ИНТЕГРИРОВАННЫХ КОРПОРАТИВНЫХ СИСТЕМАХ .....	
2.1. Использование алгебры предикатов и теории категорий в качестве формальных средств моделирования интеллектуальных процессов.....	41
2.2. Использование алгебры предикатов и предикатных операций для представления знаний в интеллектуальных системах.....	49
2.3. Особенности использования метода компараторной идентификации при моделировании интеллектуальной деятельности .....	56
2.4. Реализация метода компарации при работе с информационными объектами текстового контента .....	62
2.5. Формальная модель использования наивного байесовского классификатора для определения типов сущностей предметной области .....	69

Выводы по второму разделу.....	73
РАЗДЕЛ 3	
МЕТОДЫ ЭКСТРАЦИИ И ИДЕНТИФИКАЦИИ БИЗНЕС ЗНАНИЙ ИЗ ТЕКСТОВОГО КОНТЕНТА.....	
3.1. Описание процедуры формирования информационного пространства интегрированной корпоративной системы .....	75
3.2. Метод выявления актуального множества классифицированных сущностей предметной области.....	82
3.3. Логико-лингвистическая модель генерации фактов из текстовых потоков информационной корпоративной системы .....	87
3.4. Метод построения информационного пространства фактов интегрированной корпоративной системы .....	92
Выводы по третьему разделу .....	97
РАЗДЕЛ 4	
ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ.....	
4.1. Особенности экстракции и идентификации знаний текстового web- контента.....	100
4.2. Разработка концептуальной модели представления единого информационного пространства ресурсов бизнес деятельности корпорации .....	106
4.3. Решение проблемы формальной оценки эффективности для технологии экстракции знаний из текстового контента.....	112
4.4. Практическое использование модели извлечения знаний из слабоструктурированных текстов в системе формирования фактографической базы научной библиотеки .....	117
Выводы по четвертому разделу .....	123
ЗАКЛЮЧЕНИЕ .....	126
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ .....	128
Приложение_А .....	143
Приложение_Б.....	150