

# СОВРЕМЕННЫЕ МЕТОДЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Засекан Маргарита  
4 курс ПО группа  
«Б»

**Реферирование** – это сокращенное изложение  
сути

основного текста научной работы (статьи,  
монографии и т.д.) и извлечение из него требуемой  
информации или основного содержания.

**Типы рефератов:**

Информативные

Индикативные

Монографические

Обзорные

# Классификация методов автоматического реферирования

Экстракция – извлечение из исходного текста наиболее важных и существенных информационных блоков (абзацев, предложений)

## *Реферат-экстракт*

*Реферат-экстракт* – это вторичный документ, полученный путем извлечения из текста первичного документа наиболее важных в смысловом отношении предложений, обладающих формальными текстовыми признаками (маркерами и индикаторами).

*Структура реферата-экстракта* включает следующие элементы:

- библиографическое описание реферируемого текста,
- текст реферата, полученный путем экстрагирования, т.е. извлечения из первичного текста маркированных предложений-
- дополнительные сведения (указание на количество библиографических ссылок, иллюстраций, таблиц в следующей последовательности: «ил., ..., табл..., библиограф. назв...»; ссылки на опубликованные ранее документы, если реферируемый документ является продолжением; примечания референта, если в реферируемом документе допущены явные ошибки).

Абстракция – генерация реферата с порождением нового текста, содержательно обобщающего первичный документ или документы. Данный подход характеризуется тремя неотъемлемыми этапами: анализ исходного текста с генерацией внутреннего представления, семантическое сжатие внутреннего представления и порождение нового текста (реферата).

Абстракция

```
graph TD; A(Абстракция) --> B(Абстракция на основе лингвистического сжатия); A --> C(Абстракция с опорой на знания);
```

Абстракция на  
основе  
лингвистическо  
го сжатия

Абстракция с  
опорой на  
знания

# Абстракция на основе лингвистического сжатия

Этап анализа входного текста представляет собой синтаксический разбор и формирование синтаксического дерева разбора. Процедуры сжатия манипулируют непосредственно деревьями с целью удаления и перегруппировки частей, например, путем сокращения ветвей на основании некоторых структурных критериев, таких, как скобки или встроенные условные или подчиненные предложения. После такой процедуры дерево разбора существенно упрощается, становясь, по существу, структурной «выжимкой» исходного текста. Заключительным этапом является формирование текста итогового реферата.

## Абстракция с опорой на знания

Формируются концептуальные репрезентативные структуры всей исходной информации, которые аккумулируются в базе знаний. В качестве структур могут быть использованы формулы логики предикатов или такие представления, как семантическая сеть и набор фреймов. В процессе преобразования концептуальное представление претерпевает несколько изменений. Избыточная и не имеющая прямого отношения к тексту информация устраняется путем удаления поверхностных суждений или отсечения концептуальных подграфов. Затем информация подвергается дальнейшему агрегированию путем слияния графов (или шаблонов) или обобщения информации, например, при помощи таксономических иерархий отношений подклассов. В результате преобразования формируется концептуальная репрезентативная структура реферата, по существу, концептуальные «выжимки» из текста.

Одним из первых коммерческих и наиболее используемых в настоящее время приложений систем автоматического реферирования является аннотирование результатов информационного поиска. Так, модуль автоматического реферирования документов входит в состав каждой крупной ИПС (Google, Yandex, Rambler, Yahoo). Существующие системы автоматического реферирования общего назначения, как правило, имеют развитый графический интерфейс, богатые возможности визуализации и манипулирования с данными, предоставляют доступ к различным источникам данных.

# Системы автоматического реферирования ТЕКСТОВ

Intelligent Miner for Text	Extractor
TextAnalyst	ОРФО
OracleText RCO	Либретто
MS Office AutoSummarise	Пакет «МедиаЛингва Аннотатор SDK 1.0»
Inxight Summarizer	Поисковая система «Следопыт»
Summarist	Поисковая машина «Золотой Ключик»
Prosum	Columbia NewsBlaster
Copernic summarizer	Ultimate Research Assistant

В таблице представлены не все системы. Их существует намного больше. Рассмотрим некоторые из них поподробнее.



## Наиболее известные системы автоматического реферирования в России

- • Intelligent Miner for Text (IBM) – объединяет мощную совокупность инструментов, базирующихся в основном на механизмах поиска информации, одним из которых является Annotation Tool – утилита составления рефератов-аннотаций к исходным текстам.
- • TextAnalyst («Мегапьютер Интеллидженс») решает задачи глубокого анализа текста, такие, как создание семантической сети большого текста, подготовка реферата текста, поиск по тексту, автоматическая классификация и кластеризация текстов.
- • OracleText RCO – программный комплекс, интегрированный в СУБД, позволяющий эффективно работать с запросами, относящимися к неструктурированным текстам.

# Вывод

Практически все из существующих программных продуктов автоматического реферирования являются коммерческими разработками, отсутствует какая-либо подробная информация об алгоритмах и методах, используемых авторами, все продукты ориентированы на лексические особенности какого-то конкретного языка, решение узкоспециализированной задачи реферирования, например, сжатия новостного кластера; ряд продуктов отличается относительно низким качеством получаемых рефератов и аннотаций. Растущая потребность информационно-аналитического направления в более совершенных системах сводного реферирования определяет необходимость развития новых подходов к решению этой задачи с применением более эффективных методов и алгоритмов.

**Спасибо за внимание!**