# Методы анализа текстов в технологиях «Big Data»

О.Н. Пошатаев, А.А. Хорошилов

# Технологии интеллектуальной обработки текстов документов

- Формализация смысловой структуры текстов
- Тематическое рубрицирование текстов
- Кластеризация документов
- Установление в тексте персон, организаций, объектов
- Семантический поиск
- Машинный перевод

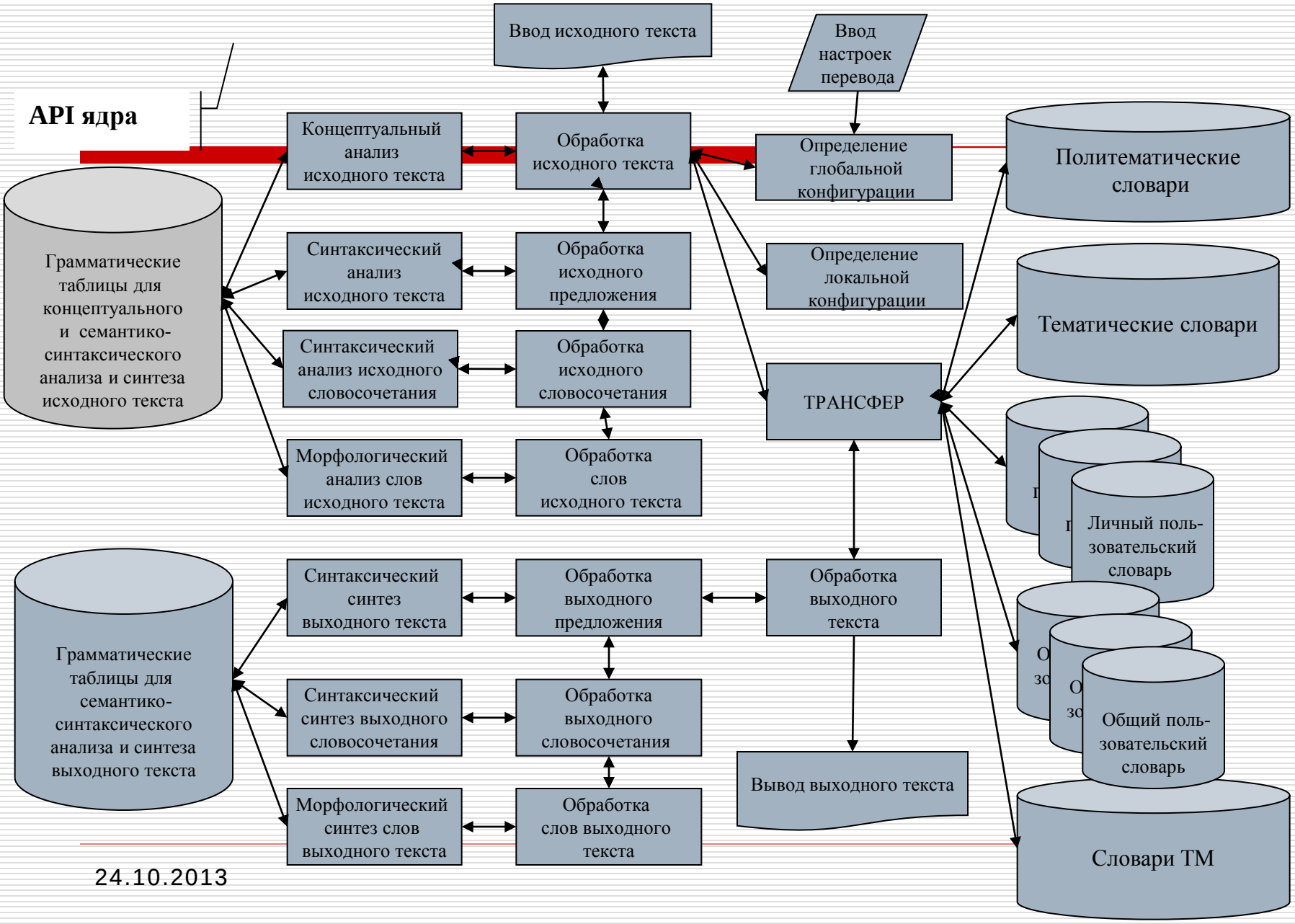
# Программные продукты

- Система оперативного мониторинга СМИ
- Система мониторинга и анализа проектной документации
- Автоматизированная система формирования и анализа фондов и реестра научно-исследовательских и опытно-конструкторских разработок
- Отраслевая автоматизированная система стандартизированного машинного перевода
- Автоматизированная система обеспечения перевода научных публикаций

# Проблемы обработки больших объемов тектовой информации

- сложная и многоступенчатая обработка
- использование  словарей больших обьемов
- большая длительность обработки

# Общая схема систем машинного перевода



# Графематический анализ текста

| CGraphemAnalys |
| --- |
| **МЕТОДЫ**<br>• Определение языка текста;<br>• Разделение входного текста на слова, разделители и т.д.<br>• Выделение дат в цифровых форматах;<br>• Выделение электронных адресов;<br>• Выделение предложений;<br>• Выделение абзацев, заголовков, примечаний; |
| **ДАННЫЕ**<br>• Исходный текст (Plain Text);<br>• Длина текста (в символах);<br>• Язык текста;<br>• Массив адресов слов;<br>• Массив адресов предложений;<br>• Массив адресов абзацев; |

# Морфологический анализ текста

| CMorphoAnalys |
| --- |
| **МЕТОДЫ**<br>• **Поиск в словаре слов-исключений;**<br>• **Поиск в словаре конечных буквосочетаний слов;**<br>• **Поиск в таблице супплетивных форм слов;**<br>• **Установления наличия чередований в основах слов;**<br>• **Определение флективного класса слов;**<br>• **Назначение слову грамматической информации;** |
| **ДАННЫЕ**<br>• **Буквенный код слова;**<br>• **Длина слова;**<br>• **Длина окончания;**<br>• **Лексикограмматический класс слова;**<br>• **Флективный класс слова;**<br>• **Наборы грамматических признаков (род, число, падеж, лицо);** |

# Семантико-синтаксический анализ текста

| CSyntAnalys |
| --- |
| **МЕТОДЫ**<br>• Членение на простые предложения;<br>• Определение главных членов предложения;<br>• Определение второстепенных членов предложения;<br>• Установление однородных второстепенных членов предложения;<br>• Построение дерева зависимостей;<br>• Разрешение многозначности грамматической информации слов; |
| **ДАННЫЕ**<br>• Адрес предложения;<br>• Длина предложения;<br>• Массив адресов простых предложений;<br>• Массив адресов словосочетаний;<br>• Массив адресов слов; |

# Концептуальный анализ текста

| CConceptAnalys |
| --- |
| **МЕТОДЫ**<br>• Выявление наименований понятий в тексте;<br>• Установления смысловых отношений между понятиями;<br>• Разрешение анафорических ссылок;<br>• Приведение понятий к их каноническим формам;<br>• Построение таблицы связей между понятиями; |
| **ДАННЫЕ**<br>• Массив адресов наименований понятий в тексте;<br>• Массив длин наименований понятий в тексте;<br>• Массив адресов наименований понятий-отношений в тексте; |

# Быстродействие процедур семантического анализа текстов

| Название процедуры семантического анализа текста | Скорость обработки (слов / сек) | Длительность процесса (в %) | Скорость обработки (слов / сек) | Длительность процесса (в %) |
|---|---|---|---|---|
| | Текущая версия | | Перспективная версия | |
| Графематический анализ | 50000 | 5% | 24000 | 4% |
| Морфологический анализ | 65000 | 6% | 34000 | 3% |
| Семантико-синтаксический анализ | 4500 | 23% | 34000 | 15% |
| Концептуальный анализ | 780 | 56% | 324 | 78% |
| Система семантического анализа (Общее быстродействие) | 537 | 100% | 134 | 100% |

# Развитие методов обработки текстовой информации

- Более сложные процедуры обработки
- Использование словарей большего объема

# Пути повышения быстродействия

- Многопоточность
- Кластерные решения
- Технологии “big data”

# Технология Hadoop

- Технология Hadoop разработана в рамках вычислительной парадигмы MapReduce, согласно которой приложение разделяется на большое количество одинаковых элементарных заданий, выполнимых на узлах кластера и естественным образом сводимых в конечный результат.
- Ядром этой технологии является распределённая файловая система HDFS (Hadoop Distributed File System) [

# Алгоритм MapReduce

- MapReduce состоит из двух шагов: Map и Reduce.
- На Map-шаге происходит предварительная обработка входных данных
- Главный узел получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам) для предварительной обработки.

# Алгоритм MapReduce
## (продолжение)

- На Reduce-шаге происходит свертка предварительно обработанных данных

- Главный узел получает ответы от рабочих узлов и на их основе формирует конечный результат — решение поставленной задачи.

# Алгоритм MapReduce (продолжение)

- MapReduce реализует операции предварительной обработки и свертки
- Эти операции работают независимо друг от друга и выполняются параллельно
- Множество рабочих узлов могут осуществлять свертку — для этого необходимо только чтобы все результаты предварительной обработки с одним конкретным значением ключа обрабатывались одним рабочим узлом в один момент времени

# Принципы реализации технологического процесса обработки текста

- Технологический процесс необходимо разделить на элементарные семантические процедуры
- Текст и результаты его обработки должны быть разделены на фрагменты.
- На каждом узле должен выполняться определенный этап обработки конкретного фрагмента текста

# Принципы реализации технологического процесса обработки текста

- В начале каждого этапа обработки текста или результатов его обработки производится операция MAP – разделения на фрагменты

- При завершении каждого этапа обработки текста или результатов его обработки производится операция REDUCE – объединения частных результатов в окончательный

# Принципы реализации технологического процесса обработки текста

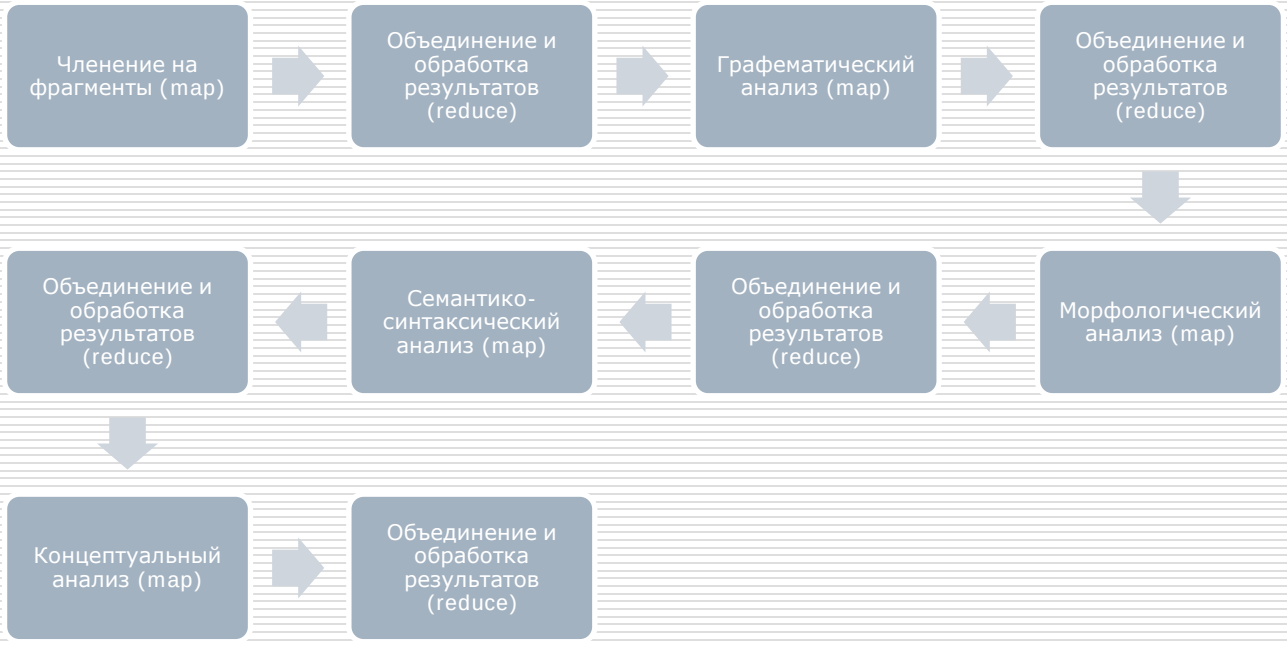
- Текст недопустимо произвольным образом делить на фрагменты. Это может привести к разрушению его смысловой структуры. Необходимо разработать процедуры, выполняющие такое деление текста на основе его упрощенного анализа.
- Необходимо также разработать процедуры корректного разделения и объединения частных или конечных результатов анализа текста.

# Принципы реализации технологического процесса обработки текста

- Все файлы, содержащие информацию об частных или конечных результатов анализа текста должны сопровождаться идентификатором

- Идентификатор файла должен содержать необходимая информация  для его обработки в распределённой файловой системе HDFS.

# Технологический процесс семантического анализа текста



# Спасибо за внимание