Logo

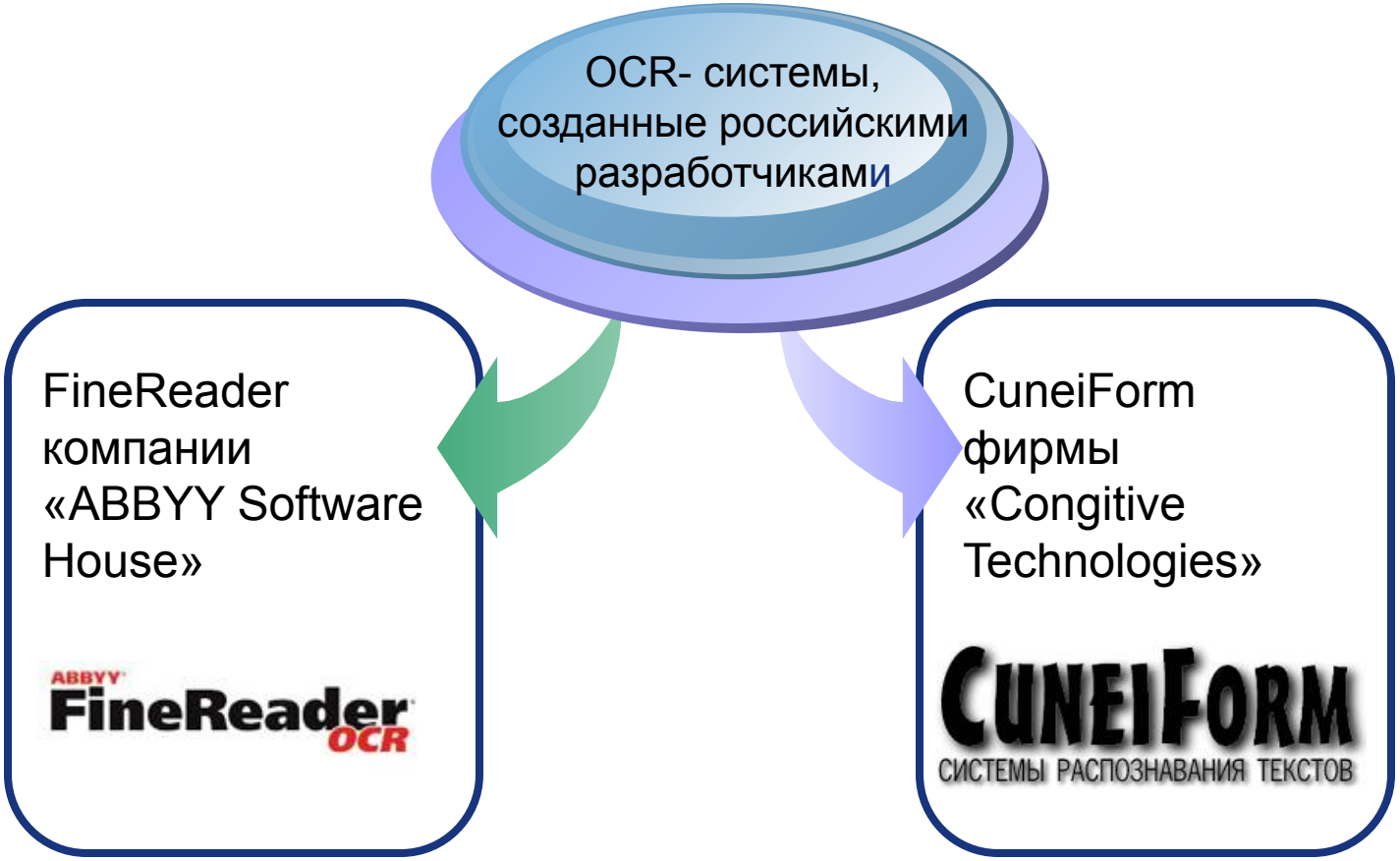
ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ОБРАБОТКЕ ТЕКСТОВ АВТОМАТИЧЕСКОЕ ЧТЕНИЕ ТЕКСТА

Система автоматического чтения текста (OCR- система — Optical Character Recognition).

— это компьютерная программа, позволяющая преобразовать текст с бумажного носителя в электронный текстовый файл, который может быть прочитан средствами обработки текстов.

Сканер

работает по принципу фотоаппарата, позволяя ПК «увидеть» текст. Для того чтобы «понять» его содержание, т.е. перевести графическое (точечное) изображение символов в пригодную для дальнейшей обработки (редактирования, реферирования, перевода и т.д.) текстовую форму, необходима система автоматического чтения текста


OCR- системы, созданные российскими разработчиками
FineReader компании «ABBYY Software House»
ABBYY FineReader OCR
CuneiForm фирмы «Congitive Technologies»
CUNEIFORM
СИСТЕМЫ РАСПОЗНАВАНИЯ ТЕКСТОВ

возможности систем автоматического чтения текста огромны:

1. позволяют распознавать печатные символы почти двух сотен языков
2. хорошо распознаются рукопечатные символы, написанные от руки печатными буквами с небольшим интервалом между ними
3. узнают все используемые шрифты без предварительного обучения, воспринимают полужирный, курсивный, слипшийся текст
4. способны самообучаться и распознавать плохо пропечатанные символы или символы незнакомых программе языков

5 поддерживают все модели сканеров и любые графические форматы.

6 широко используются сетевые версии программ автоматического чтения текста

7 поддерживают публикацию бумажных документов в глобальной сети Интернет

8 точность распознавания OCR-систем на текстах хорошего и среднего качества достигает 97—99 %.



Logo

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ И АННОТИРОВАНИЕ ТЕКСТА



Logo

- ***Реферат*** — связный текст, который кратко выражает не только тему или предмет какого-либо документа, но и цель, применяемые методы, основные результаты описанного исследования или разработки.

*Процесс составления реферата называется **реферированием***

Аннотация — краткое изложение содержания документа, дающее общее представление о его теме.

*Процесс составления аннотации называется **аннотированием**.*

Реферирование и аннотирование текста являются довольно сложными и трудными видами интеллектуальной деятельности и занимают много времени.



Logo

Выход есть!!!



Logo

Автоматическое реферирование и аннотирование

Этапы построения человеком реферата (аннотации)

Подготови-тельный

референт определяет тематическую направленость текста и пытается понять и осмыслить документ в целом

Аналитический

текст делится на фрагменты (абзацы, аспекты и т.п.)., в нем выделяют основные смысловые единицы (предложения, словосочетания, слова), составляется план аннотации (реферата)

Построение аннотации (реферата)

выделенные ранее смысловые единицы (их комбинации или преобразования) располагаются в единый вторичный текст в соответствии с планом реферата или аннотации.

Компьютер должен уметь выполнять те же действия, которые осуществляет человек:

1. находить в тексте ключевые слова, словосочетания и предложения
2. находить в тексте менее значимые единицы
3. составлять из текстовых единиц двух первых типов смысловые единицы реферата или аннотации
4. составлять из таких единиц текст реферата или аннотации

Ключевое (опорное) слово *— это термин, относящийся к основному содержанию текста и повторяющийся в нем несколько раз (с учетом всех возможных синонимов).*

Ключевое словосочетание *— это сочетание слов, среди которых есть одно или несколько ключевых.*

Ключевое предложение *- предложение, содержащее два и более ключевых слова или ключевых словосочетания.*

Смысловые еденицы реферата

1. полные (без изменения) ключевые предложения исходного текста;
2. перефразированные ключевые предложения исходного текста;
3. предложения, составленные из ключевых слов или словосо четаний исходного текста с помощью специальных связующих элементов;
4. предложения, обобщающие несколько предложений исход ного текста (не обязательно ключевых).

Смысловые еденицы аннотации

1. ключевые слова или словосочетания исходного текста с предшествующими им специальными фразами — реляторами типа: «В статье рассматриваются следующие вопросы:...», «Книга посвящена следующим проблемам: ...» и т.п.

2. специальные предложения, содержащие оценочные элементы: «Рассматривается важная проблема...», «Статья посвящена актуальной теме...» и т.д.;

3. специальные предложения, содержащие клише, т.е. специ ализированные словесные штампы, фиксирующие внимание читателя на определенных аспектах содержания: «Недостаток... заключается», «Цель публикации...», «Ставится задача...», «Делается попытка...» и т.д.

4. предложения, обобщающие несколько предложений исходного текста (не обязательно ключевых).

Logo

Методы автоматического реферирования

Статисти-ческие

1) ключевыми словами считаются такие знаменательные слова текста, которые с учетом всех синонимов встречаются в тексте наибольшее число раз;
2) ключевым предложением считается предложение текста, которое:
а) имеет несколько ключевых слов;
б) содержит ключевые слова на небольшом расстоянии друг от друга.

Позицио н-ные

ключевым предложением считается предложение, входящее в заголовок, подзаголовок, начало или конец какой-то части текста или всего текста. и содержат информацию о целях, методах, выводах и результатах исследования. Важность тех или иных предложений с указанной точки зрения определяется экспертами путем изучения семантической структуры первичных документов определенного типа.

Логико-семантические

опираются на исследование структуры и семантики текстов. Существует несколько вариантов этих методов, но цель их одна — выделить из конкретного текста предложения с наибольшим функциональным весом.

Logo

Системы автоматического реферирования

t

«Либретто»

«Inxight Summerizer»

«Extractor»

TextAnalyst