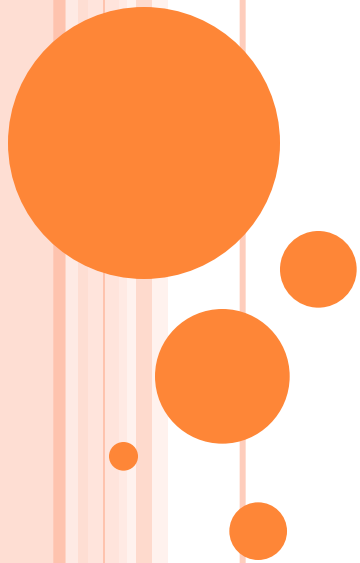
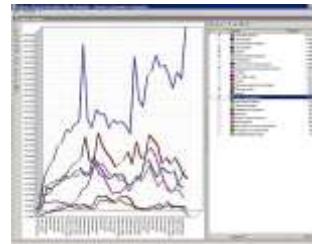


- **Методы автоматического установления
смысловой близости документов на
основе их концептуального анализа**



АКТУАЛЬНОСТЬ ПРОБЛЕМЫ

1. Аналитические системы



2. Поисковые системы



3. Системы классификации



4. Системы выявления заимствований



ПОДХОДЫ К СРАВНЕНИЮ ДОКУМЕНТОВ

Методы, основанные на
сравнении символьного
представления текстов

Методы, основанные на
сравнении семантической
структуры текстов



ИСПОЛЬЗУЕМЫЕ ПРОЦЕДУРЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ



ТЕХНОЛОГИЯ СОСТАВЛЕНИЯ ЧАСТОТНЫХ СЛОВАРЕЙ ПО ТЕКСТАМ ДОКУМЕНТОВ



МОДЕЛИ ПРЕДСТАВЛЕНИЯ СМЫСЛОВОГО СОДЕРЖАНИЯ ДОКУМЕНТОВ

- Концептуальный образ документа (КОД)
- Формализованная предикатно-актантная структура документа (ФПАС)
- Сверхфразовое единство

КОНЦЕПТУАЛЬНЫЙ ОБРАЗ ДОКУМЕНТА

- Единицы смысла - понятия.
- Концептуальный образ документа - совокупность выявленных в тексте наименований.
- Каждый элемент КОДа сопровождается весовым коэффициентом.

ФОРМАЛИЗОВАННАЯ ПРЕДИКАТНО-АКТАНТНАЯ СТРУКТУРА ДОКУМЕНТА

- Модель опирается на единицу смысла - предложение.
- В основе и предложений на естественном языке, и формализованных логических высказываний лежит предикатно-актантная структура
- Предикатно-актантные структуры - смысловые инварианты, которые позволяют осуществлять автоматический перевод текстов с естественных языков на формализованные и с формализованных на естественные.
- Формализованные предикатно-актантные структуры (ФПАС) создаются на основе автоматического концептуального анализа текстов.
- ФПАС состоит из совокупности наименований понятий-актантов, связанных между собой смысловыми отношениями – предикатами.
- Модель может быть представлена в виде графа, вершинами которого являются актанты, а дугами – отношения – предикаты.

СВЕРХФРАЗОВОЕ ЕДИНСТВО

- Модель основывается на формализации единиц смысла - сверхфразовых единств.
- В роли сверхфразовых единств выступают текстовые фрагменты, состоящие из двух и более предложений и связанных общей темой.
- Формализованное представление текстовых фрагментов можно обеспечить путем их пословной нормализации и представления их в виде упорядоченной совокупности значимых слов текстового фрагмента

МЕТОДЫ УСТАНОВЛЕНИЯ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ

МЕТОД УСТАНОВЛЕНИЯ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ ПО КОДАМ

- Концептуальный образ анализируемого документа (его КОД) попарно сравнивается с концептуальными образами (КОДами) документов массива.
- Для каждой пары сравниваемых КОДов определяется коэффициент их смысловой близости.
- В упрощенной модели коэффициент можно положить равным отношению суммы весов наименований понятий КОДа анализируемого документа, совпавших с наименованиями понятий КОДа из массива, к сумме весов всех наименований понятий КОДа документа.

МЕТОД УСТАНОВЛЕНИЯ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ ПО ФПАСАМ

- Для установления смысловой близости документов определяются меры подобия графов ФПАСа анализируемого документа и ФПАСов документов массива.
- Определение меры подобия, основывается поиске на максимального общего подграфа

МЕТОД УСТАНОВЛЕНИЯ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ ПО ТЕКСТОВЫМ ФРАГМЕНТАМ

- Если в двух текстах имеются такие текстовые фрагменты, в которых содержатся одинаковые значимые слова и их доля превышает заданный порог, то эти фрагменты находятся в смысловой связи.
- Степень их смысловой близости будет определяться отношением совпавших значимых слов к их общему числу.
- С целью обеспечения установления смыслового тождества значимых слов на уровне словоизменения все слова текстового фрагмента должны быть нормализованы.

ОТЛИЧИЕ ДАННОГО ПОДХОДА ОТ СУЩЕСТВУЮЩИХ

Комплексный подход включающий в себя:

- Автоматизированное выявление наименований понятий в тексте методами семантико-синтаксического и концептуального анализа
- Методы отождествления понятий
- Автоматическое составление по документу его формализованного смыслового описания.
- Автоматическое приведение наименований понятий к каноническому виду (с учетом синонимов и гипонимов)
- Установление коэффициентов смысловой значимости понятий в тексте
- Возможность установление связей между разноязычными документами
- Использование концептуальных словарей, позволяющих увеличивать значения коэффициентов значимости наименований понятий принадлежащим к определенным областям
- Каждый из методов ориентирован на конкретные задачи, и все методы будут использоваться в комплексе

ОСОБЕННОСТИ ПОДХОДА: РАБОТА СЛОВАРНОЙ СЛУЖБЫ

- Автоматическое составление словарей синонимов и гипонимов
- Автоматическое составление концептуальных словарей предметных областей
- Наличие двуязычных словарей объемом 3.5 млн. словарных статей
- Наличие фразеологического словаря синонимов более 400 тыс.

ПРОВЕДЕННЫЕ ЭКСПЕРИМЕНТЫ

- Эксперимент по установлению заимствований в научно-технических документах (тематика судостроение) (1, 3 методы, выборка 187 документов)
- Эксперимент по установлению смысловой близости на массиве 1000 документов общественно-политических документов из базы системы мониторинга сми (система www.мониторингсми.рф)

. ЭКСПЕРИМЕНТЫ ПО УСТАНОВЛЕНИЮ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ: ЭКСПЕРИМЕНТ №1

Экспертами создаются
группы документов,
похожих на документ-
образец

Документы
ранжируются по степени
смысловой близости к
документу-образцу.

Затем в автоматическом
режиме ранжирование
производится с помощью
программного
обеспечения,
реализующего методы,
предложенные в
настоящей статье.

Сравнение результатов
работы экспертов и
системы.

. ЭКСПЕРИМЕНТЫ ПО УСТАНОВЛЕНИЮ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ: ЭКСПЕРИМЕНТ №2

Экспертами
выбирается в
массиве
документов
несколько
документов-
образцов.

Подбираются
документы,
похожие на них по
смысловому
содержанию на
документы-
образцы.

С помощью
программного
обеспечения
документы в
автоматическом
режиме
соотносятся с
документами-
образцами

Сравнение
результатов
работы экспертов
и системы

СПАСИБО ЗА ВНИМАНИЕ