ТЕХНОЛОГИИ СЕМАНТИЧЕСКОГО ПОИСКА В МАССИВАХ НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ

НТУУ «КПИ» ННК «ИПСА»
кафедра СП-САПР
Шумейко Юрий Дмитриевич

ЦЕЛЬ РАБОТЫ

- Сравнительный анализ существующих моделей информационно-поисковых систем по критериям сложности реализации и релевантности полученного результата.
- Расширение базовой модели поиска возможностями семантического анализа.
- Адаптация типовой архитектуры информационно-поисковой системы к возможностям расширенной семантической модели поиска.

Модель информационного поиска

- Способ представления документов (образы документов)
- Способ представления поисковых запросов
- Критерии релевантности документов

Модели информационного поиска

- Теоретико-множественные модели
- Алгебраические модели
- Вероятностные модели

СРАВНИТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ МОДЕЛЕЙ ИНФОРМАЦИОННОГО ПОИСКА

Характеристика	Теоретико-множественные модели	Алгебраические модели	Вероятностные модели
Способ представления документов	Множество термов (+) простота реализации	Вектор в многомерном пространстве термов (+) учет весов термов (-) не учитывается взаимозависимость термов	Множество термов без учёта частоты встречаемости терма в документе
Способ представления поискового запроса	Булевская формула (+) формальный способ задания запроса (-) сложность использования (-) все компоненты имеют одинаковый вес	Вектор в многомерном пространстве термов (-) сложность выбора метрик в многомерном пространстве	Множество термов
Критерии релевантности документа	Если формула истинна (выполнена на документе), документ соответствует запросу (+) точность соответствия запросу (-) невозможность ранжирования по степени релевантности (-) нет учета синонимии, омономии и т.п.	Релевантность рассчитывается как скалярное произведение векторов документа и запроса (+) простота расчета (+) широкий спектр способов классификации документов коллекции (-) расчеты массивов высокой размерности	Вероятность того, что данные документ может быть интересен пользователю. Рассчитывается на основании соотношения встречаемости термов в релевантном наборе и в остальной коллекции (-) документы, не содержащие слова запроса, не будут найдены (-) необходимость постоянного

НЕДОСТАТКОМ МОДЕЛЕЙ ИНФОРМАЦИОННОГО ПОИСКА

При оценке релевантности текст документа рассматривается как простой набор взаимно независимых слов,

многозначность слов, синонимия, омонимия, как правило, не учитываются

СЕМАНТИЧЕСКИЙ ПОИСК

– вид полнотекстового информационного поиска, учитывающего смысловую нагрузку текстов документов при определении релевантности документов поисковому запросу

Устранение выявленных недостатков

- Предлагается использовать векторную модель информационного поиска, расширенную семантической с применением онтологий для работы в едином поисковом механизме

ПРЕДЛАГАЕТСЯ

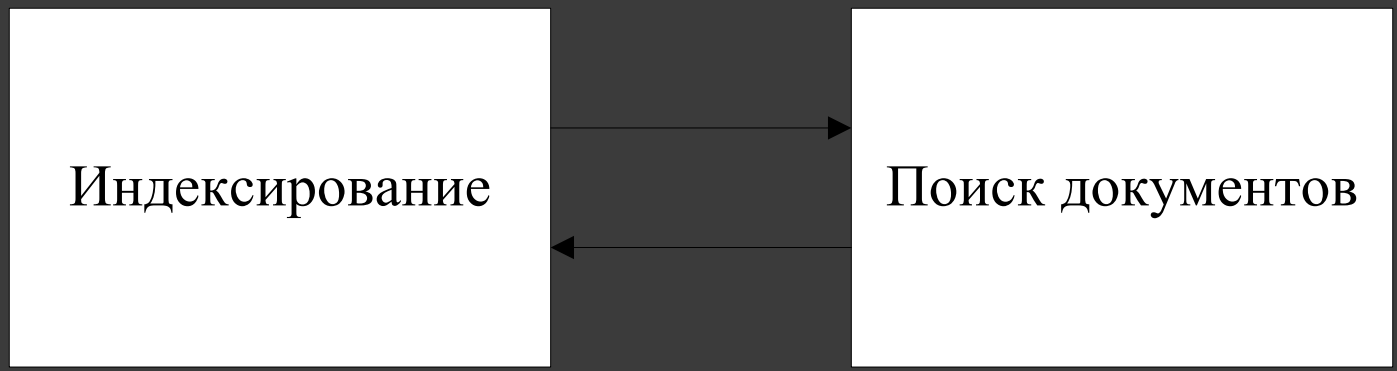
использовать онтологии предметных областей для:

- Выбор термов для представления документа
- Классификация документов коллекции
- Определение релевантности документа путём сравнения онтологий документов с онтологиями предметных областей
- Структурирование результатов
- Автоматическая генерация ответа

ВЕКТОРНО-СЕМАНТИЧЕСКАЯ МОДЕЛЬ

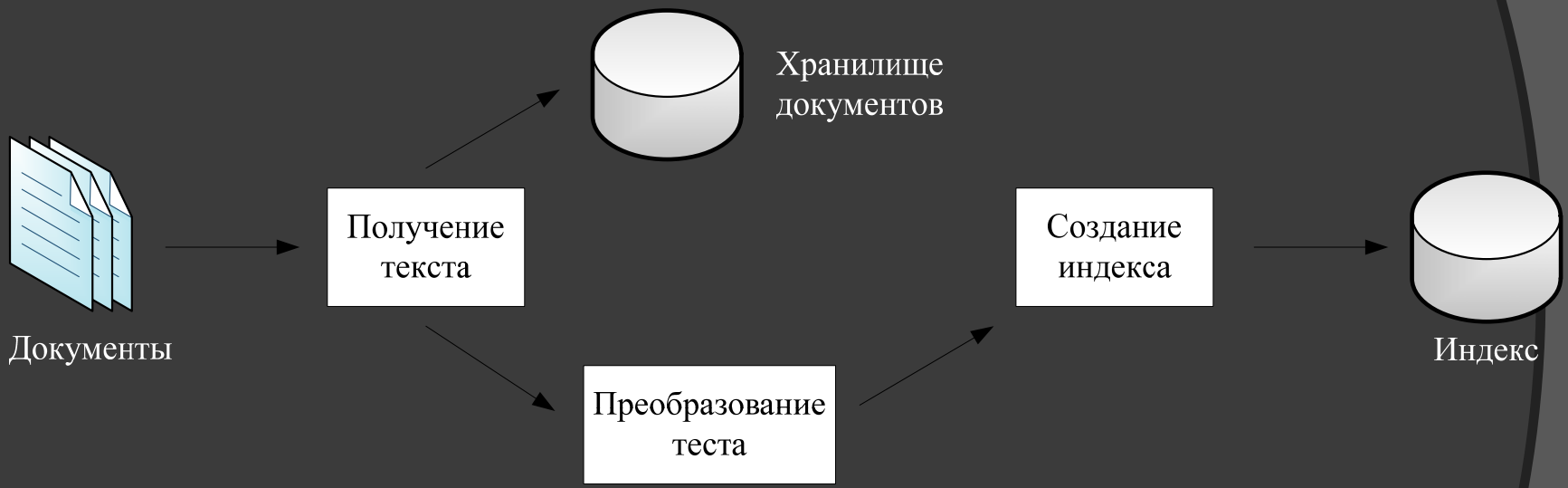
Характеристика	Векторно-семантическая модель
Способ представления документов	Вектор в многомерном пространстве термов. Поисковый образ документа в виде онтологий. (+) учет весов термов (+) учитываются взаимозависимость термов
Способ представления поискового запроса	Вектор в многомерном пространстве термов. Образ поискового запроса в виде онтологии.
Критерии релевантности документа	Релевантность определяется на основании результатов сравнения онтологий документов с онтологией предметной области.

АРХИТЕКТУРА ПОИСКОВОЙ СИСТЕМЫ



- Индексирование
- Поиск

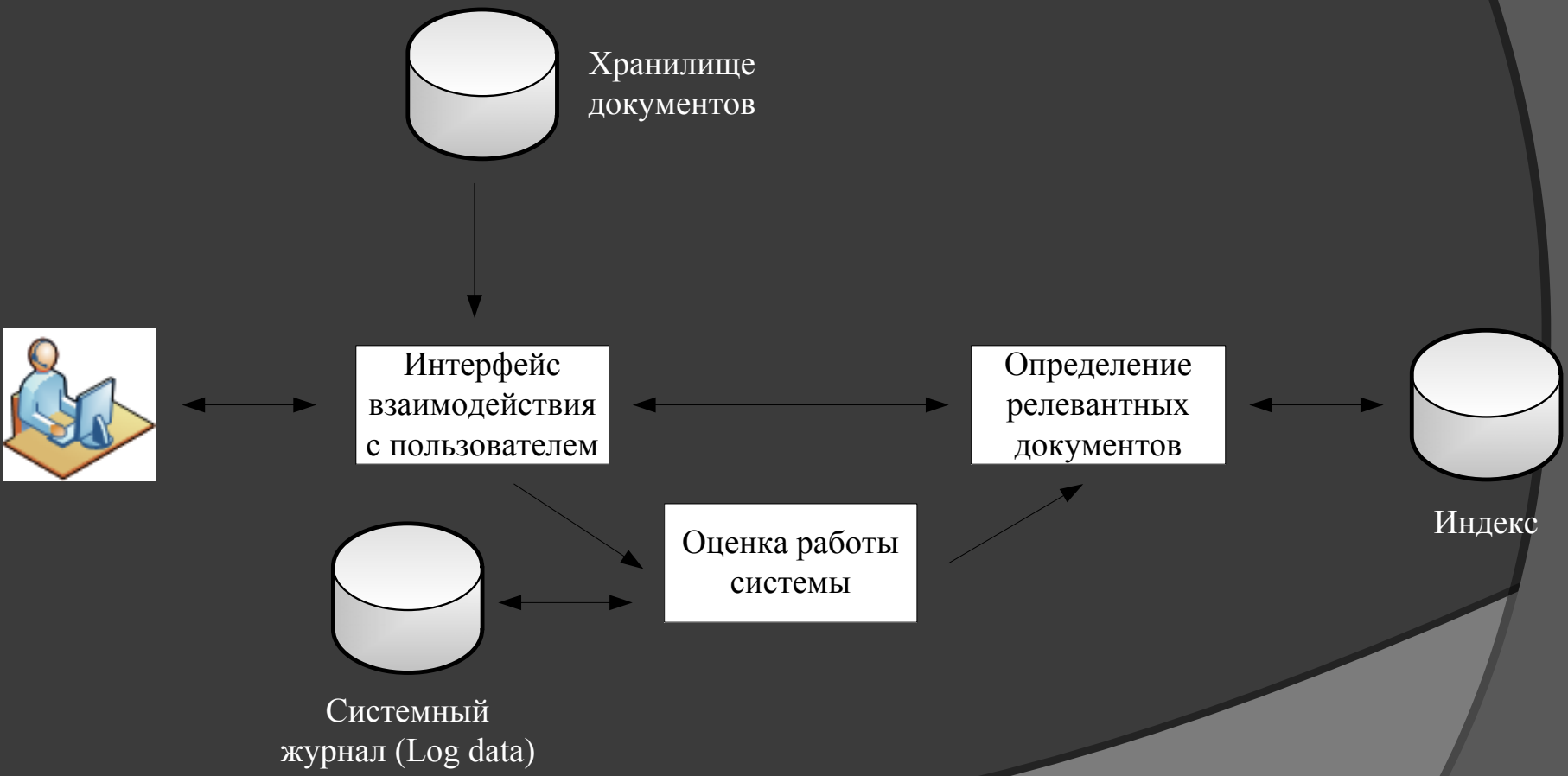
ИНДЕКСИРОВАНИЕ



ПРОЦЕСС ИНДЕКСИРОВАНИЯ ДОПОЛНЯЕТСЯ

- Модуль автоматического создания онтологий – построение поисковых образов документов в виде онтологий

ПОИСК


Хранилище документов
Интерфейс взаимодействия с пользователем
Определение релевантных документов
Индекс
Оценка работы системы
Системный журнал (Log data)

ПРОЦЕСС ПОИСКА ДОПОЛНЯЕТСЯ

- Модулем сравнения онтологий
- Модулем классификации документов с применением онтологий
- Модулем структурирования результатов поиска с применением онтологий
- Модулем автоматической генерации ответа на поставленный вопрос

ВЫВОДЫ

- Выполнен сравнительный анализ моделей информационно-поисковых систем по критериям сложности реализации и релевантности полученного результата.
- Векторная модель расширена семантической с использованием онтологий для работы в едином поисковом механизме
- Адаптирована типовая архитектура информационно-поисковой системы к возможностям расширенной семантической модели поиска.

СПАСИБО ЗА ВНИМАНИЕ

ВЕКТОРНАЯ МОДЕЛЬ ПОИСКА

выбрана в силу:

- Возможности гибкой модификации модели в виду удобства представления документов коллекции матрицей весов терм-документ
- Широкий выбор методов оценки релевантности документов поисковому запросу

НЕДОСТАПТКИ ТЕОРЕТИКО-МНОЖЕСТВЕННЫХ МОДЕЛЕЙ

- Крайняя жёсткость и непригодность для ранжирования найденных документов по степени релевантности, поскольку отсутствуют критерии её оценки.
- Не будет найден документ, в котором встречаются только синонимы слова, указанного в запросе, в случае, когда само слово не встречается.
- Сложность использования – далеко не каждый пользователь может свободно оперировать булевскими операторами при формулировке своих запросов.

HITS

- Алгоритм HITS обеспечивает выбор из информационного потока лучших «авторов» (первоисточников) и «посредников» (документов от которых идут ссылки цитирования).
- Понятно, что страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим первоисточником, если она упоминается хорошими посредниками.
- Для каждого документа рекурсивно вычисляется его значимость как первоисточника a_p и посредника h_p по формулам:
- $a_p = \Sigma h_q$, $h_p = \Sigma a_q$

PAGE RANK

- В отличие от литературного индекса цитирования не все ссылки считаются равнозначными.
- PageRank подсчитывает общий "авторитет" документа, в то время как HITS определяет "авторитет" документа для конкретной темы.
- Вероятность того, что блуждающий в Сети пользователь перейдет на некоторую определенную Web-страницу - это ее ранг - PageRank. PageRank Web-страницы тем выше, чем больше других страниц ссылается на нее, и чем эти страницы популярнее.

АЛГЕБРАИЧЕСКИЕ МОДЕЛИ

$$R(Q, D_j) = \cos\alpha = \frac{QD_j}{|Q|\,|D_j|} = \frac{\sum_{k=1}^{n} w_k^Q \cdot w_k^j}{\sqrt{(\sum_{k=1}^{n} w_k^{Q\,2}) \cdot (\sum_{k=1}^{n} w_k^{j\,2})}} \qquad (1.4)$$

Где $R(Q, D_j)$ – мера релевантности поискового запроса Q документу D_j;

Q и D_j – вектора, представляющие соответственно поисковый зап
рассматриваемый документ D_j;

n – количество различных термов коллекции документов;

w_k^Q и w_k^j – веса термов, в векторах представляющих соответственно п
запрос Q и анализируемый документ D_j.

АЛГЕБРАИЧЕСКИЕ МОДЕЛИ

– Коэффициент Дайса (Dice's coefficient)

$$R(D_i D_j) = 2 \cdot \frac{\sum_{k=1}^{n} w_k^i \cdot w_k^j}{\sum_{k=1}^{n} w_k^i + \sum_{k=1}^{n} w_k^j} \qquad (1.5)$$

Где $R(D_i, D_j)$ – мера релевантности документа D_i документу D_j;

D_i и D_j – вектора, представляющие соответственно i и j документы;

n – количество различных термов коллекции документов;

w_k^i и w_k^j – веса термов, в векторах представляющих соответственно документ D_i и анализируемый документ D_j.

– Коэффициент Жаккара (Jaccard's coefficient)

$$R(D_i D_j) = \frac{\sum_{k=1}^{n} w_k^i \cdot w_k^j}{\sum_{k=1}^{n} w_r^i + \sum_{k=1}^{n} w_r^j - \sum_{k=1}^{n} w_r^i \cdot w_r^j} \qquad (1.6)$$