Исследование средств сбора и поиска информации на заданном множестве Web-сайтов (на примере информационных порталов EGEE и НТУУ «КПИ»)

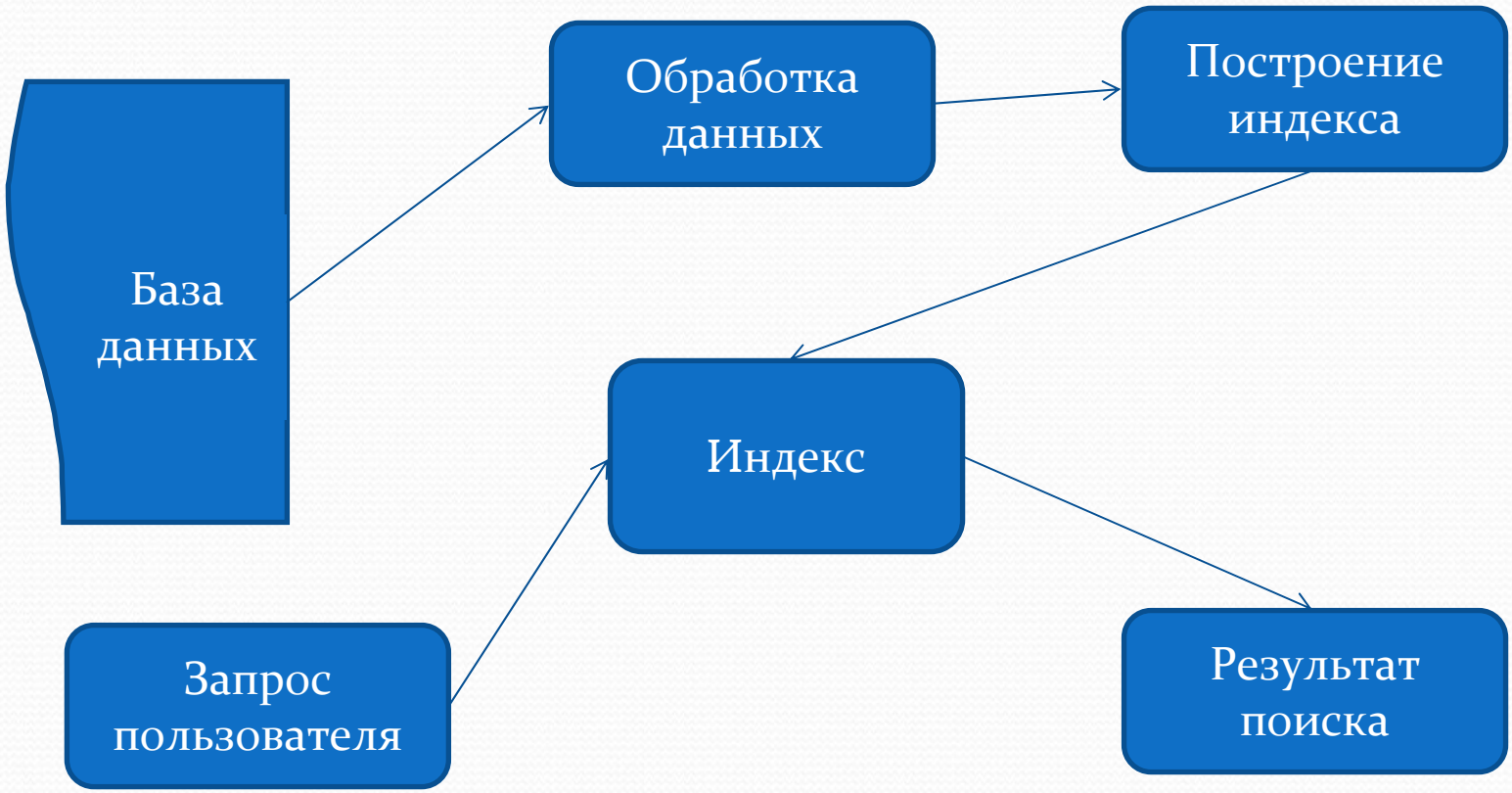
работа студента группы ДА-42м

Шафрая Андрея Юрьевича

Модуль поиска

- Задачи
  - Обработка текста на странице
  - Индексация данных
  - Поиск по собранным данным

Модуль поиска в общей системе



Критерии оценки систем индексирования и поиска

- скорость индексирования
- скорость переиндексации
- размер базы и скорость поиска
- работа с разными языками и стемминг

Сравнительные характеристики систем индексирования и поиска

	Sphinx	**Apache Lucene**	**Xapian**
Лицензия	GPL 2 или коммерческая	Apache License 2.0	GPL
Тип	отдельный сервер или MySQL storage engine	отдельный сервер или сервлет, встраиваемая библиотека	встраиваемая библиотека
Платформа	C++	Java (реализован также на PHP, C#/.NET, Perl, Ruby, Python)	C++
Варианты поиска	булевый поиск, поиск по фразам, учёт близости слов	булевый поиск, поиск по фразам, учёт близости слов, поиск по маске	булевый поиск, поиск по фразам, поиск с ранжированием, поиск по маске, поиск по синонимам

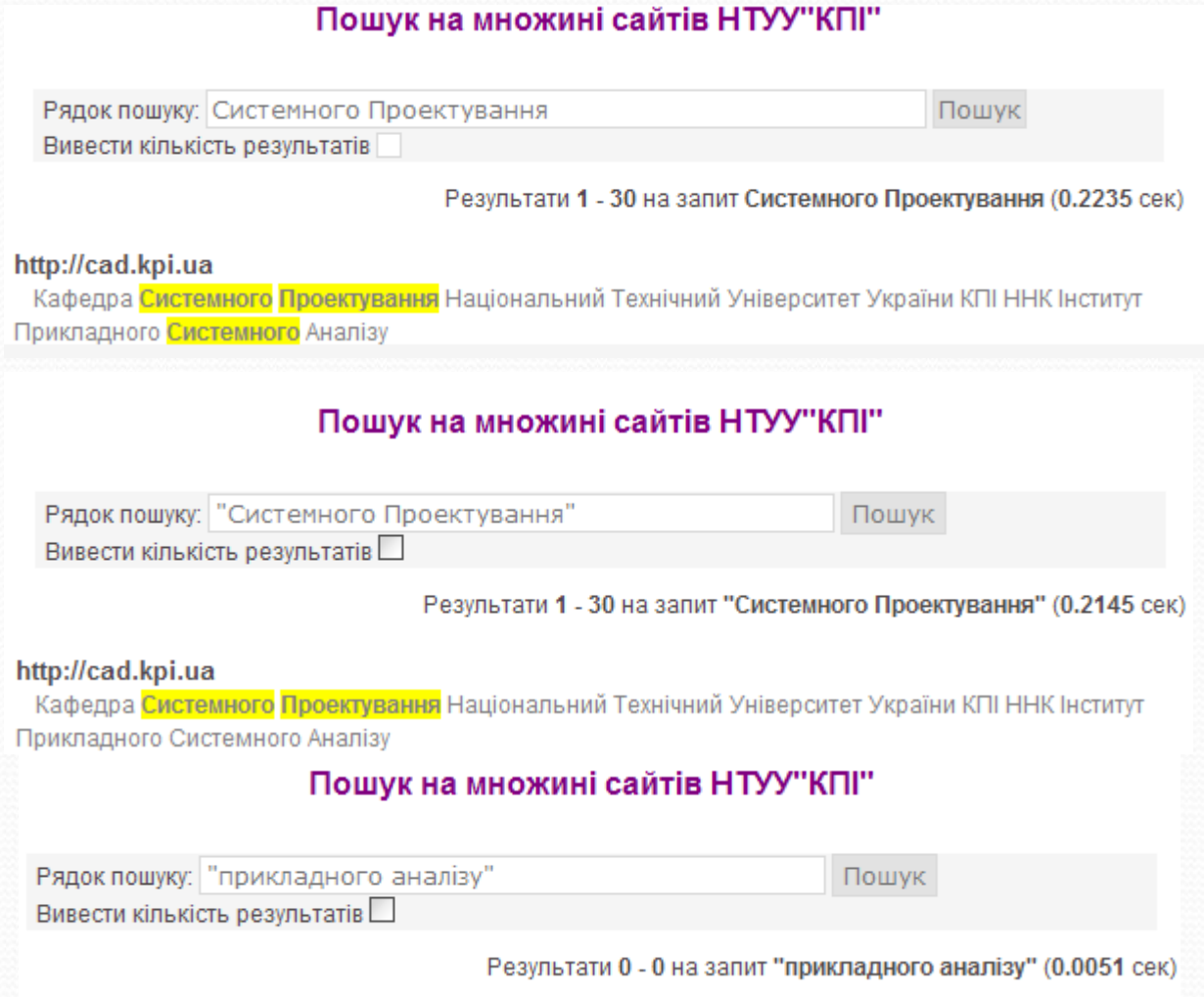
Сравнительные характеристики систем индексирования и поиска

	Sphinx	Apache Lucene	Xapian
Поддержка языков	встроенный английский и русский стемминг, soundex для реализации морфологии	отсутствует морфология, есть стемминг (Snowball) и анализаторы для ряда языков (включая русский)	отсутствует морфология, есть стемминг для ряда языков (включая русский), проверка правописания в поисковых запросах
Размер индекса/скорость	индексация около 10 Мб/сек (зависит от CPU), поиск около 0.1 сек/~2 – 4 Гб индексе, поддерживает размеры индекса в сотни Гб и сотни миллионов документов.	около 20 Мб/минута, размер индексных файлов ограничен 2 Гб (на 32-bit ОС). Есть возможности параллельного поиска по нескольким индексам и кластеризация	тесты скорости на офф. сайте отсутствуют. Известны работающие инсталляции на 1.5 Тб индекса

Сравнительные характеристики систем индексирования и поиска

	MySQL	Lucene	Sphinx
Индексация, min	1627	176	84
Индекс, MB	3011	6328	2850
Match all, ms/q	286	30	22
Match phrase, ms/q	3692	29	21
Match bool top-20, ms/q	24	29	13

Результаты работы модуля поиска



Поиск по структурам Semantic Web

Поисковый запрос:
author:Ivanov

id	fileId	attribute	value
1	1	author	Ivanov I.
2	1	ISBN	978-5-17-063541-2
3	1	Name	Integrated Circuits Architechture

1. Ivanov I. Integrated Circuits Architecture - 978-5-17-063541-2

Выводы

- Исследованы различные системы поиска информации на множестве заданных сайтов.
- Разработан и реализован модуль поиска информации для комплексной системы.
- Решены следующие задачи:
  - Выбор механизма индексации и поиска, удовлетворяющего соотношению эффективности и скорости поиска и вычислительных мощностей, которые для этого необходимы.
  - Осуществлен полнотекстовый поиск информации, собранной с заданных веб-сайтов.
  - Достижение высокой скорости поиска информации.
  - Построение пользовательского интерфейса для поиска.

Спасибо за внимание