

Исследование, разработка и внедрение  
оценки заданного набора Webometrics  
параметров на заданном множестве  
Web-сайтов  
(на примере информационных  
порталов EGEE и НТУУ «КПИ»)

---

работа студента группы ДА-42м  
Ушеренко Романа Борисовича

# Модуль оценки

- Задачи
  - **Обработка текста на странице**
    - Определение объема
    - Определение объема текста на иностранных языках
  - **Обработка прикрепленных файлов**
    - Определение прикрепленных файлов
    - Определение типов прикрепленных файлов
  - **Обработка структур Semantic Web**
    - Определение структур Semantic Web
    - Извлечение данных из структур Semantic Web

# Модуль оценки в общей системе



# Webometrics и Scopus

- Критерии оценки
  - **Webometrics**
    - Объем текстовой информации на страницах веб-сайта
    - Объем файлов формата PDF и DOC
    - Количество таких файлов
  - **Scopus**
    - Объем текстовой информации, представленной на английском языке

# Обработка страниц

Робот    Операції з базою    Пошук    **Статистика**    Параметри    Інформація

**Збір статистики по сайтам**

Зібрати статистику    Показати статистику    Коефіцієнти

close or Esc Key

Параметр	Коефіцієнт
volume	<input type="text" value="0.2"/>
fullTextDoc	<input type="text" value="0.4"/>
pdf	<input type="text" value="0.4"/>

# Обработка страниц

Збір статистики по сайтам

Оновити статистику

Видалити відмічене

Показати Семантичний Веб

Експорт в CSV

Дата	Ім'я	Адреса	Об'єм		Об'єм документів		Об'єм PDF		Видалити
			(Kb)	Місце	(Kb)	Місце	(Kb)	Місце	
Інститути									
2010-05-24 12:46:29	ВПІ	<a href="http://vpi.kpi.ua">http://vpi.kpi.ua</a>	42500	30	3610	61	3610	48	<input type="checkbox"/>
2010-05-24 12:46:29	ВІТІ	<a href="http://viti.kpi.ua">http://viti.kpi.ua</a>	50788	26	47471	20	47471	17	<input type="checkbox"/>
2010-05-24 12:46:29	ФТІ	<a href="http://pti.kpi.ua">http://pti.kpi.ua</a>	859	102	0	98	0	82	<input type="checkbox"/>
2010-05-24 12:46:29	ММІ	<a href="http://mmi.kpi.ua">http://mmi.kpi.ua</a>	1826	92	200	83	0	81	<input type="checkbox"/>
2010-05-24 12:46:29	ІПО	<a href="http://ipo.kpi.ua">http://ipo.kpi.ua</a>	4676	76	228	82	0	78	<input type="checkbox"/>
2010-05-24 12:46:29	ІТС	<a href="http://www.its.kpi.ua">http://www.its.kpi.ua</a>	166738	8	11681	41	6091	41	<input type="checkbox"/>
2010-05-24 12:46:29	ІПСА	<a href="http://ipsa.kpi.ua">http://ipsa.kpi.ua</a>	1469	99	0	97	0	79	<input type="checkbox"/>
2010-05-24 19:35:05	ІПСА	<a href="http://ipsa.kpi.ua">http://ipsa.kpi.ua</a>	1507	98	0	100	0	80	<input type="checkbox"/>

# Обработка структур Semantic Web



# Обработка структур Semantic Web

```
<div typeof="a:Book">  
  <p property="a:ISBN">  
    978-5-17-063541-2  
  </p>  
  <p property="a:author">  
    Ivanov I.  
  </p>  
  <a rel="a:source"  
href="1234.pdf">Download</a>  
</div>
```

```
<div typeof="a:Book">  
  <p property="a:ISBN">  
    978-5-17-063541-2  
  </p>  
  <p property="a:title">  
    Integrated Circuits Architecture  
  </p>  
  <a  
href="5678.pdf">Download</a>  
</div>
```

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

# Обработка структур Semantic Web

Робот

Статистика

Збір статистики по сайтам

Оновити статистику

Видалити відмічене

Показати Семантичний Веб

Експорт в CSV

Дата	Ім'я	Адреса	RDF	RDFa
2010-05-26 23:54:31	Semantic Future	http://semanticfuture.net/	+	
2010-05-26 23:56:07	Semantic Web Health Care and Life Sciences (HCLS) Interest Group	http://www.w3.org/2001/sw/hcls/	+	
2010-05-26 23:56:54	RDFa Wiki	http://rdfa.info/wiki/RDFa_Wiki	+	+
2010-05-26 23:58:59	The London Gazette	http://www.london-gazette.co.uk/		+
2010-05-26 23:59:20	JRS Semantic Web Utils	http://sw.joanneum.at/		+

# Выводы

- Проведен анализ проблемы формирования рейтинга ВУЗа в общем и проведения оценки заданных веб-сайтов по заданному набору критериев.
- Разработано программное обеспечение для проведения автоматической оценки. Приложение оформлено в виде модуля для комплексной системы контроля веб-сайтов.
- В процессе разработки были выполнены такие задачи:
  - Сбор информации со страниц целевых веб-сайтов, определение их объема, определение объема англоязычных разделов сайта
  - Выделение прикрепленных файлов текстового формата, определение их объема.
  - Выявление структур Semantic Web, извлечение информации из таких структур.
  - Сравнение информации, представленной на страницах, по заданным наборам характеристик.
- В результате экспериментов было показано, что разработанное приложение позволяет с высокой точностью определять объем содержимого страниц веб-сайтов

Спасибо за внимание