Алгоритмические основы разработки поисковой системы

Трегубов А.А., Кононова Т.С.

Таганрогский Государственный Радиотехнический университет

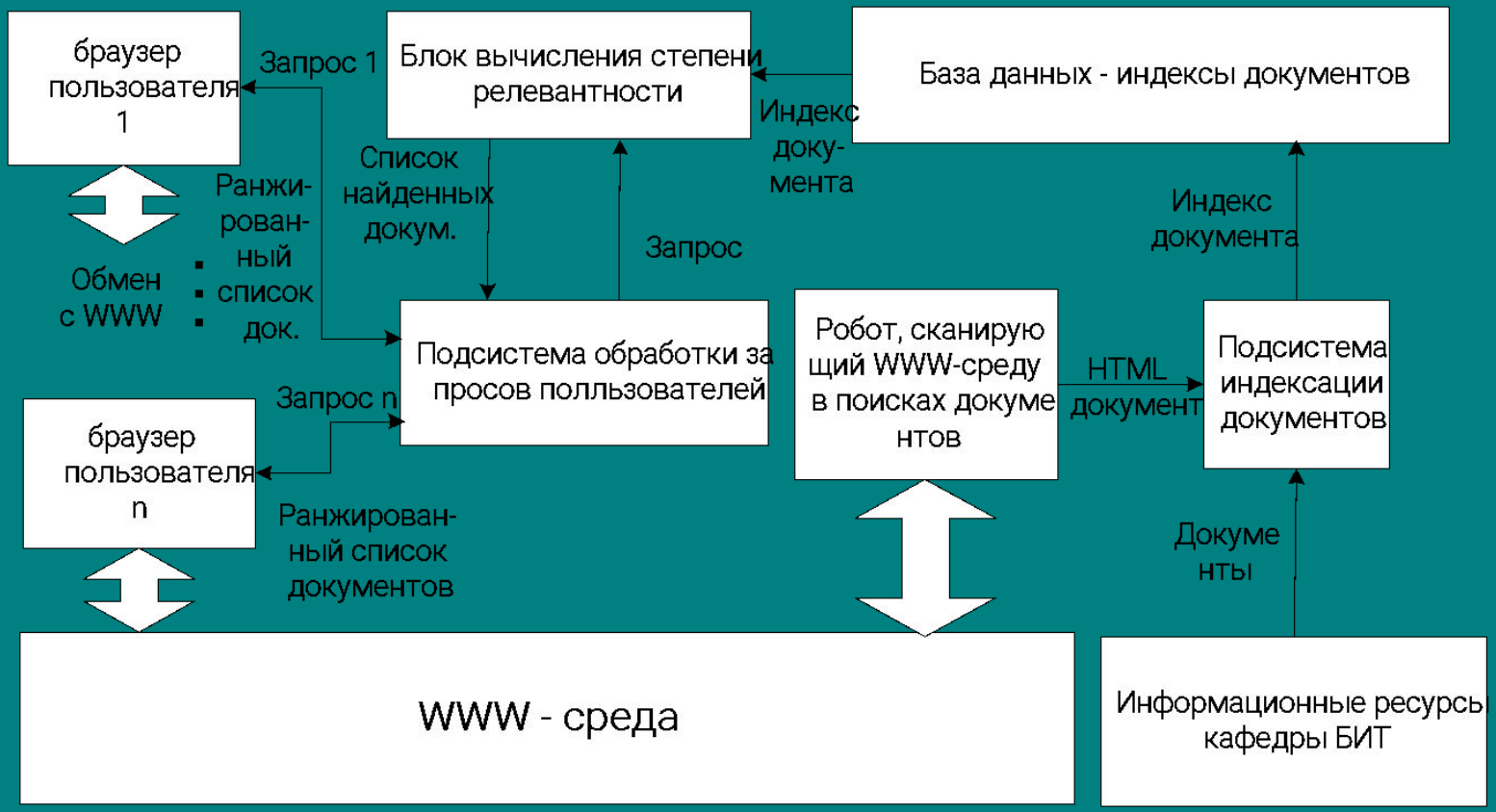
Факультет информационной безопасности,

кафедра БИТ

Росия, г. Таганрог, ул. Чехова 2

E-mail: taa_trtu@mail.ru

Структура организации поисковой системы



Пример таблицы парадигм склонения русских существительных

Код скл.	Падеж					
	им.	род.	дат.	вин.	твор.	пред.
единственное число						
1	Ø	Ø	Ø	Ø	Ø	Ø
2	ка	ки	ке	ку	кой	ке
3	Ø	а	у	а	ом	е
4	Ø	а	у	Ø	ом	е
5	Ø	а	у	а	ем	е
6	Ø	а	у	Ø	ем	е
7	-	у	-	-	-	-
8	-	ю	-	-	-	-
9	-	-	-	-	-	у
10	-	у	-	-	-	у

Всего для существительных:

- 36 флективных парадигм в единственном числе
- 46 флективных парадигм во множественном числе

Пример таблицы типов машинного склонения русских существительных.

Коды склонений		
машинное склонение	Ед. ч.	Мн. ч.
0001	4	42
0002	3	41
0003	4	-
0004	35	66
0005	4	49
0006	30	47
0007	32	-
0008	35	-
0009	14	57
0010	14	-

Общее количество типов машинного склонения для существительных - 97

Организация словарной статьи для слова:

модель

Машинная основа слова: модел

В единственном числе данная основа имеет следующий набор флексий: ь-и-и-ь-ью-и

Во множественном числе: и-ей-ям-и-ями-ях

Из таблицы парадигм:

- в единственном числе код склонения - 17
- во множественном числе код склонения - 57

Из таблицы типов машинного склонения:
код машинного склонения - 0018

Словарная статья в автоматическом словаре основ:

модел 0018 ь

Статистический метод индексирования

Относительная частота появления термина ti :

$$f_{ij} = \frac{Nt}{N}$$

где Nt – число встречаемости термина в документе,
N – число всех терминов в документе.

Инверсная частота появления термина:

$$\log\left(\frac{N}{df_i}\right)$$

где *dfi* - количество документов в коллекции,
содержащих термин *ti,*
N – число всех терминов в документе.

Комбинированный метод индексации:

$$w_{ij} = f_{ij} \cdot \log\left(\frac{N}{df_i}\right)$$

Алгебраический метод определения релевантности

- Представление множества индексов документов коллекции набором векторов в векторном пространстве индексируемых терминов;
- Представление запроса вектором в векторном пространстве индексируемых терминов;
- Определение степени релевантности как меры расстояния между векторами индекса документа и запроса по формуле Хемминга:

$$d\left(\overline{x},\overline{C}\right)=\left|x_1-C_1\right|+\ldots+\left|x_n-C_n\right|$$

где x – вектор индекса документа,
C – вектор запроса.