Полнотекстовой поиск

в сети Интернет

Павелко Павел, ИУ7-61 Ломовской И.В., ИУ7

МГТУ им. Баумана

Москва, 2016

Цель

Разработка информационной системы для сбора информации в сети Интернет и последующего поиска по ней.

Задачи

1. Анализ предметной области;
2. Разработка БД для хранения информации;
3. Разработка поискового робота для сбора информации;
4. Разработка программы для поиска;
5. Разработка поисковой страницы;

Полнотекстовой поиск

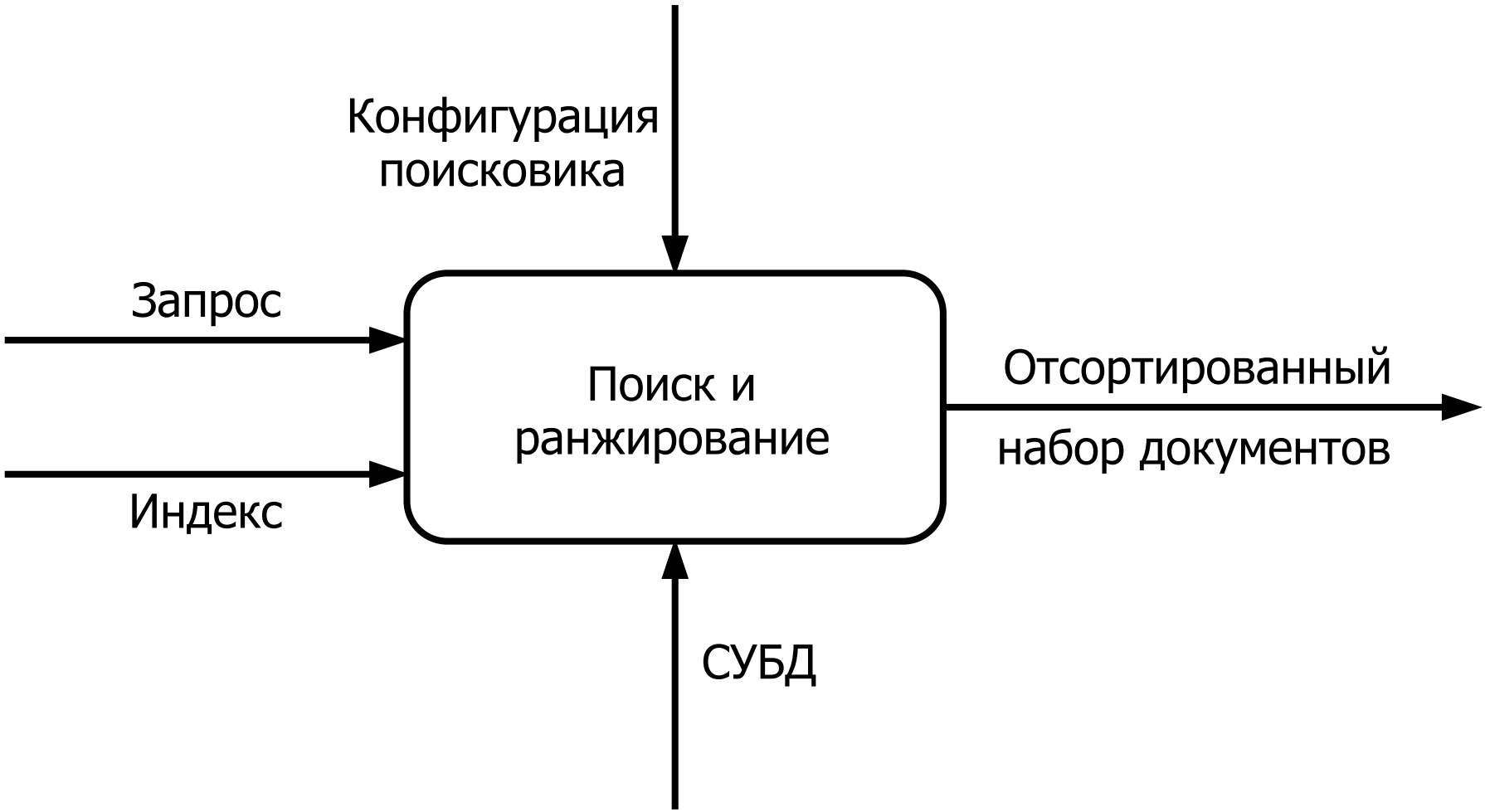
Задача

Для заданного пользовательского запроса найти и отранжировать релевантные документы из выборки страниц сети Интернет.

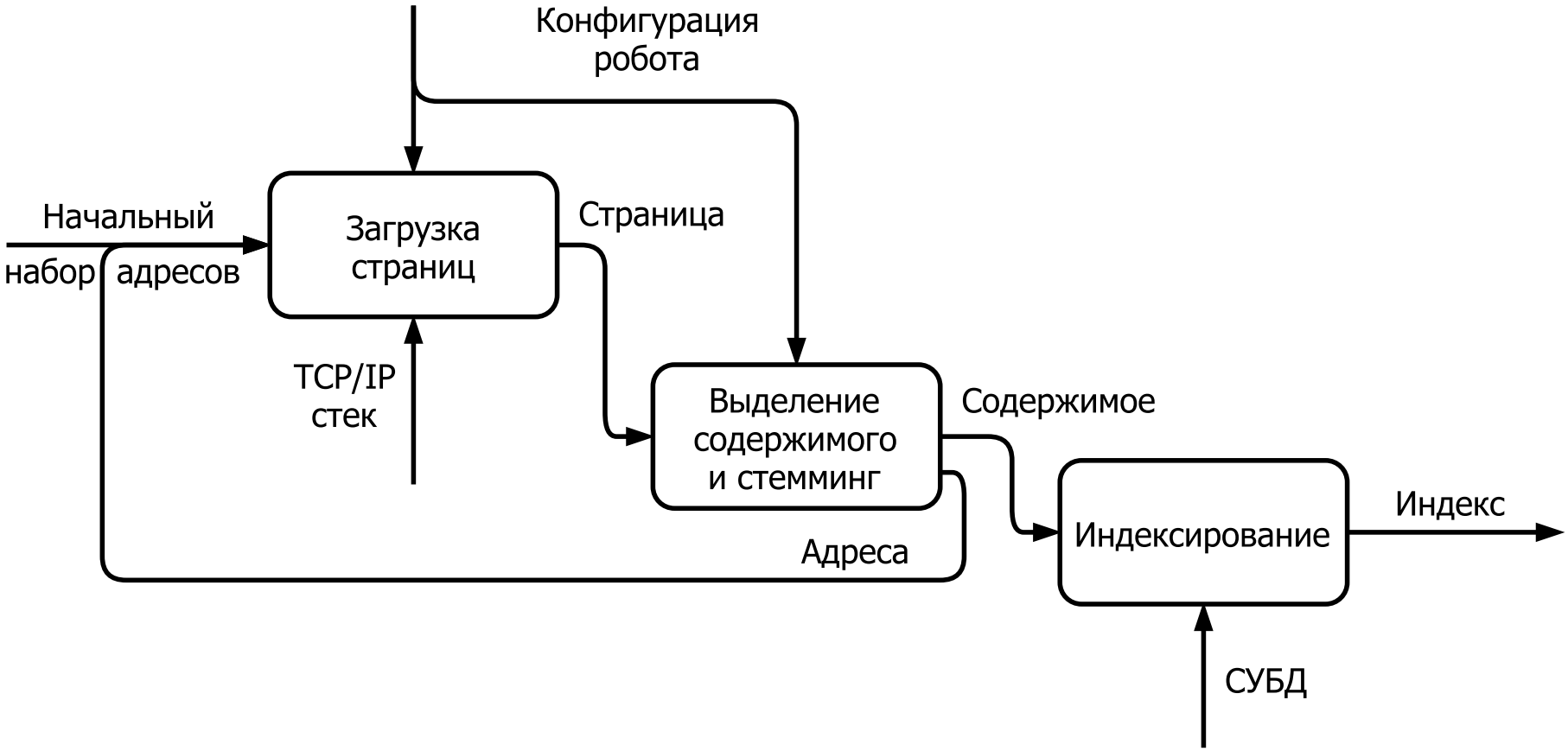
Этапы решения

- Сбор информации
- Индексирование
- Поиск

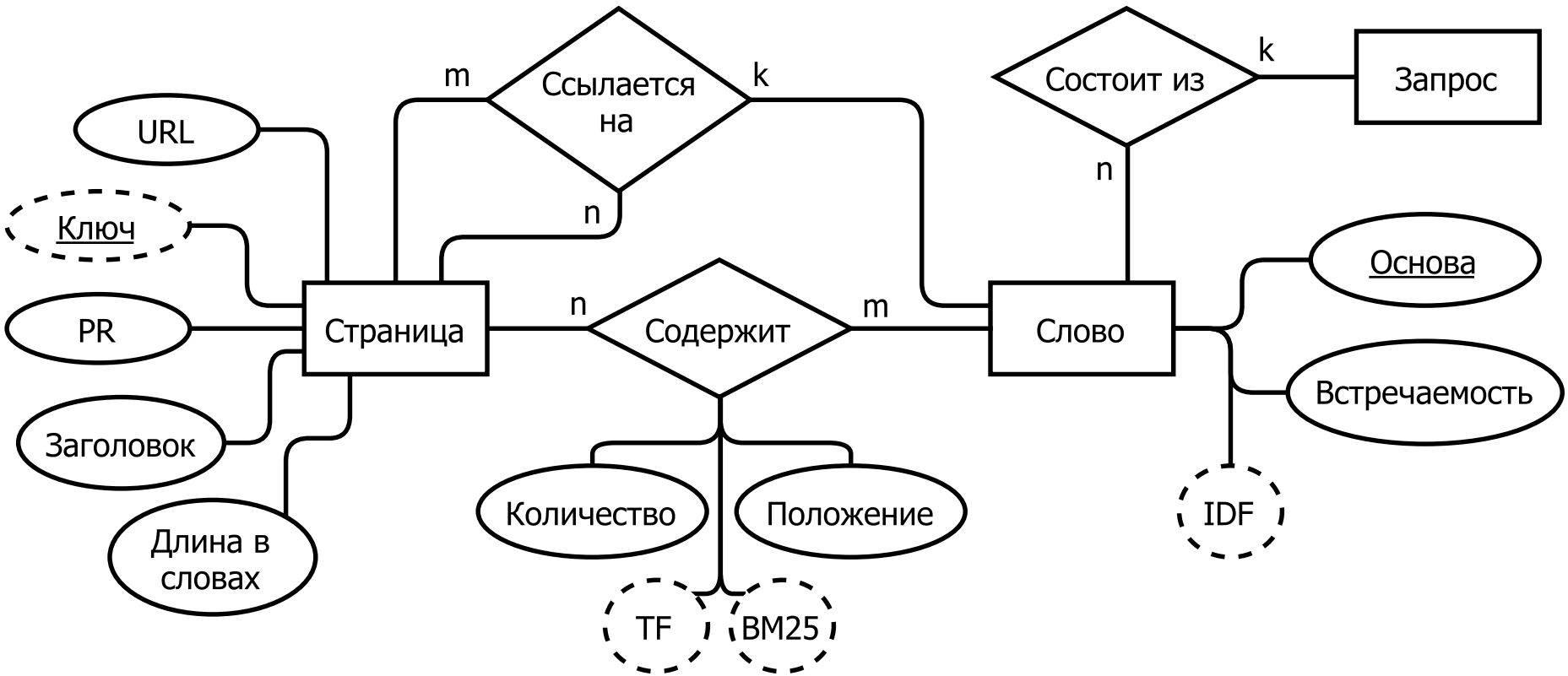
IDEF0 поисковика



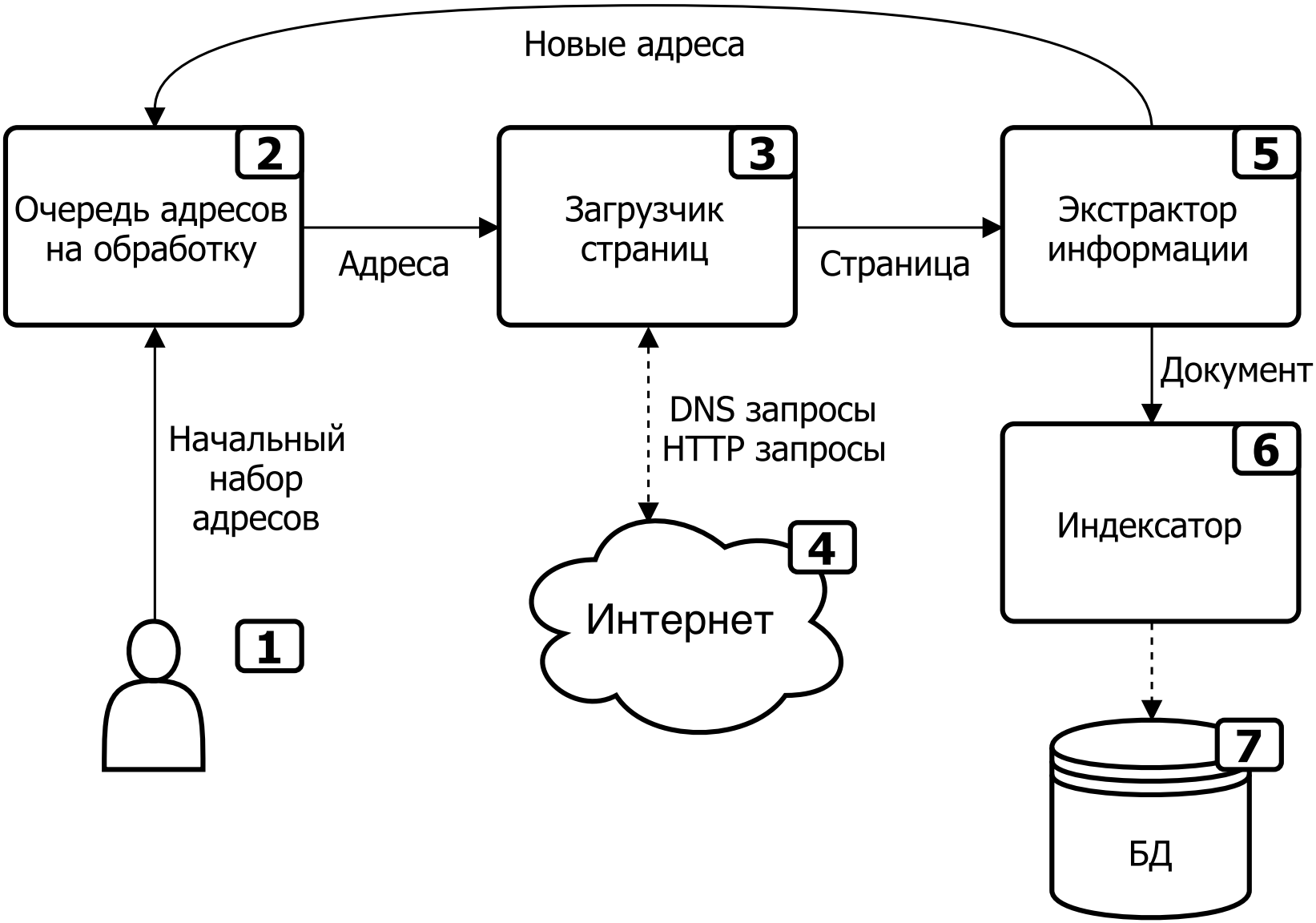
IDEF0 поискового робота



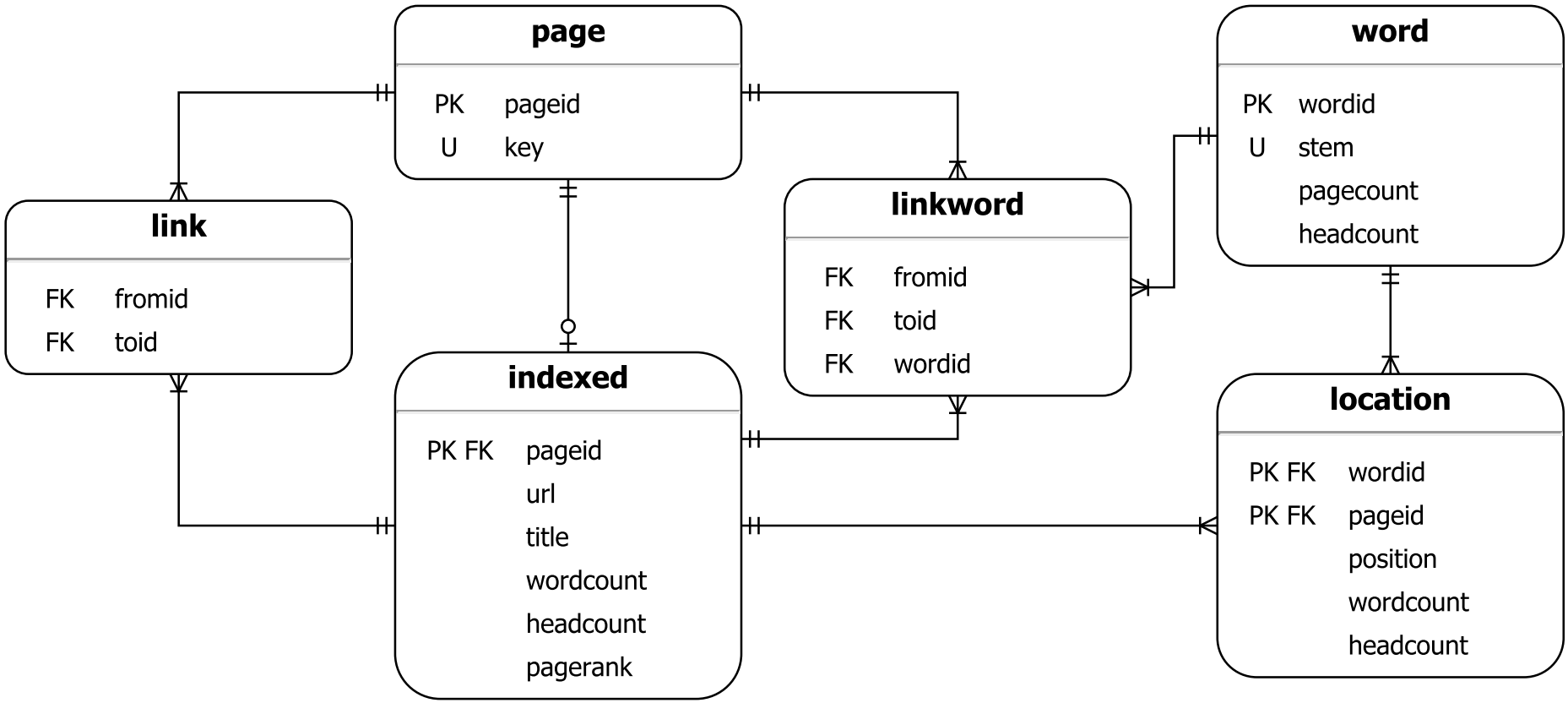
ER-диаграмма предметной области



Архитектура поискового робота



База данных


page
PK pageid
U key
word
PK wordid
U stem
pagecount
headcount
link
FK fromid
FK toid
linkword
FK fromid
FK toid
FK wordid
indexed
PK FK pageid
url
title
wordcount
headcount
pagerank
location
PK FK wordid
PK FK pageid
position
wordcount
headcount

Поиск

Этапы

1. Извлечение из БД релевантных документов
2. Подсчёт факторов
3. Нормализация факторов
4. Обработка выбросов
5. Итоговое ранжирование

Функция ранжирования

Итоговая функция ранжирования определяется как средневзвешенная сумма рассмотренных далее факторов.

Ранжирование по содержимому

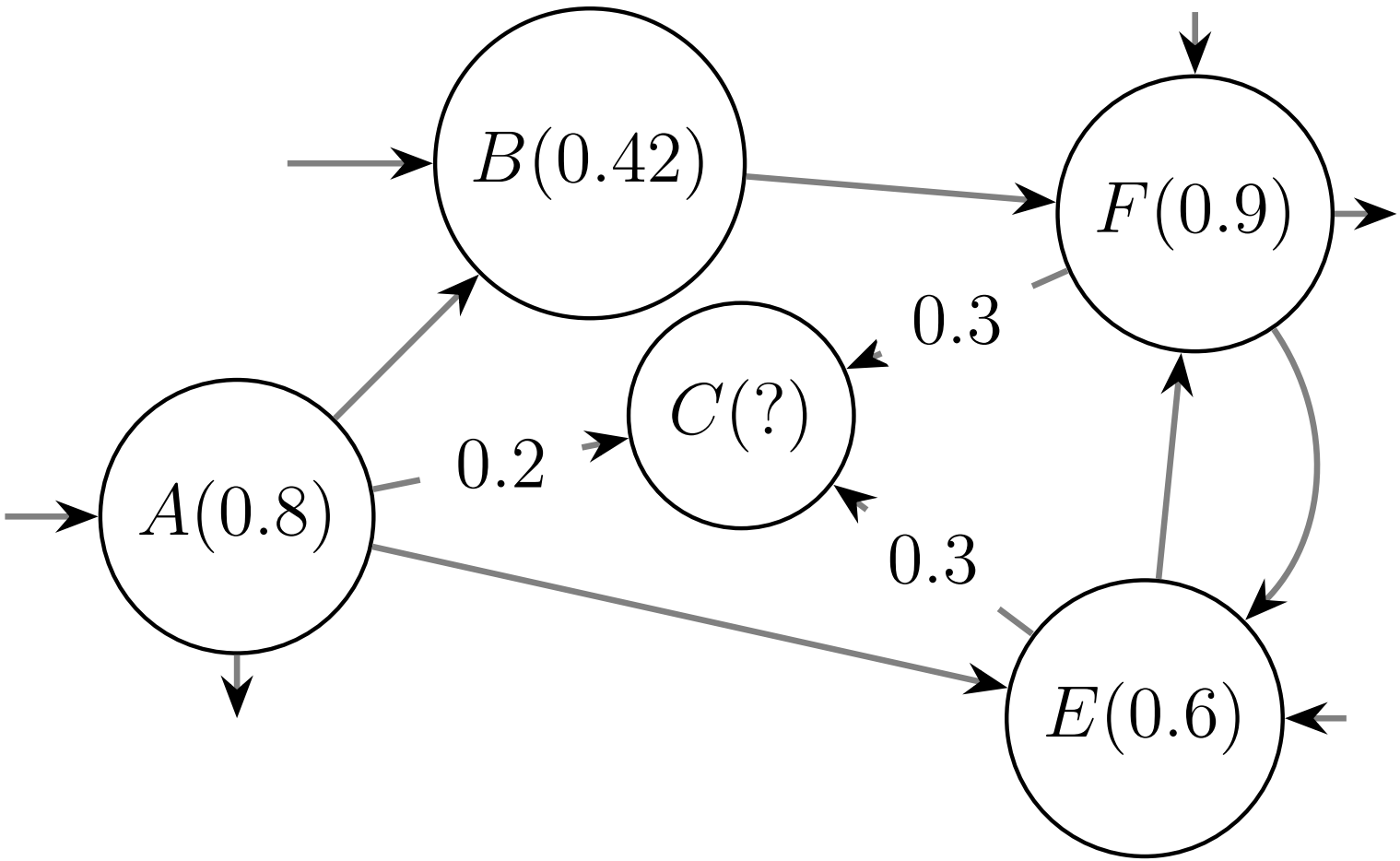
Данный класс факторов основывается на информации, которую можно получить непосредственно из документа.

Факторы основываются на

- Общем количестве слов в документе
- Частоте вхождения слов в документ
- Частоте вхождения слов в заголовки
- Положении запрошенных слов
- Расстоянии между запрошенными словами

Ссылочное ранжирование

1. Использование текста ссылок
2. PageRank



Используемые технологии

- Язык программирования: ECMAScript 2015 (JavaScript);
- Платформа: Node.js;
- Архитектура: событийно-ориентированная, асинхронная;
- СУБД: SQLite 3;

Спасибо за внимание!

Московский государственный
технический университет им
Н.Э. Баумана

Москва, 2016