Лекция 7.

Процессы поиска информации

1. Информационно-поисковые системы.
2. Алгоритмы поиска.
3. Состав и принципы работы поисковой системы

Модели информационного процесса поиска

$$IP_{def} =< D, Q, R, D' >, \qquad D' \subset D$$

D – некоторое множество документов или библиотека (поисковый массив);
Q – множество информационных запросов;
R – множество отношений, свойств, при наличии которых любому запросу $q_i \subset Q$ ставится в соответствие подмножество *D';*
D' – ответ на информационный запрос.

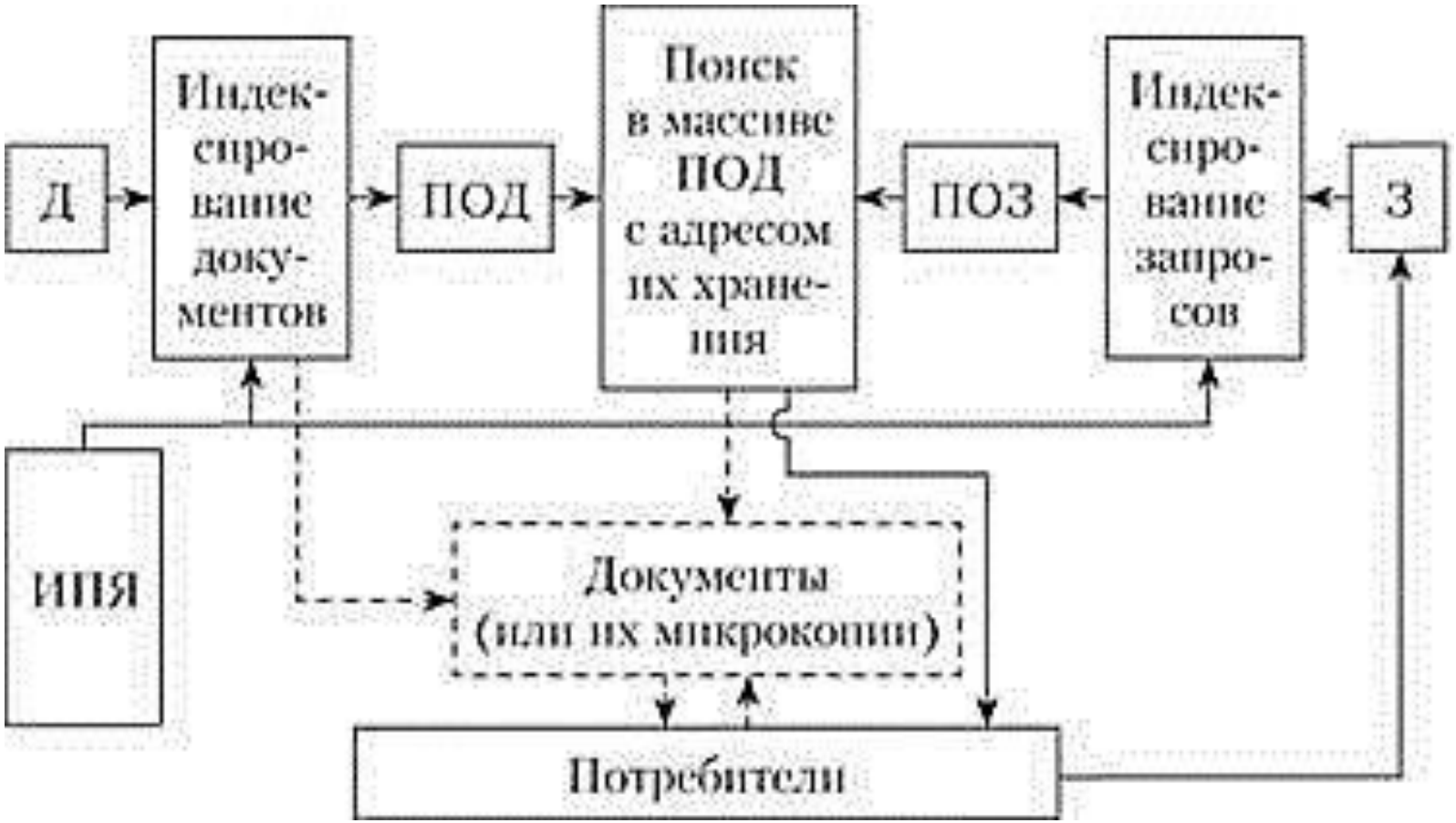
$$IPS_{def} =< LS, D, TS, N >$$

LS *– логико-семантический аппарат* ;
D *– поисковый массив*;
TS *– технические средства*;
N *– люди, взаимодействующие с системой.*

$$IPS_{def} =< RL, IND, KSS >$$

RL – ИПЯ;
IND – правила индексирования;
KSS – критерий выдачи или критерий смыслового соответствия.

Структура функционирования ИПС



Состав логико-семантического аппарата ИПС

Модель информационно-логической поисковой системы

$$IPS_{def} =< RL, IND, KSS >$$

RL – ИПЯ;
IND – правила перевода с естественного языка на информационный, т.е. правила индексирования;
LV – правила логического вывода, которые предназначены для алгоритмического получения новой информации I_n.

Модель информационно-семантической поисковой системы

$$IPS_{def}S =< a, St, tp_{iss}, co, t_j >$$

a – цель;
St – структура;
$tp_{iss} \subseteq TP$ – подмножество технологических процессов для данной *ISS,*
co – условия;
t_i – время.

$$tp_{iss} =< met, re, SemSI >$$

met – методы;
re – средства;
SemSI – семантическая переработка семантической информации.

Характеристики ИПС

Два аспекта полноты

$$П=(N/P)\times 100\%.$$

$$P = (N1/N) \times 100\%$$

$$S = (N2/N)\times 100\% = 100\% - P$$

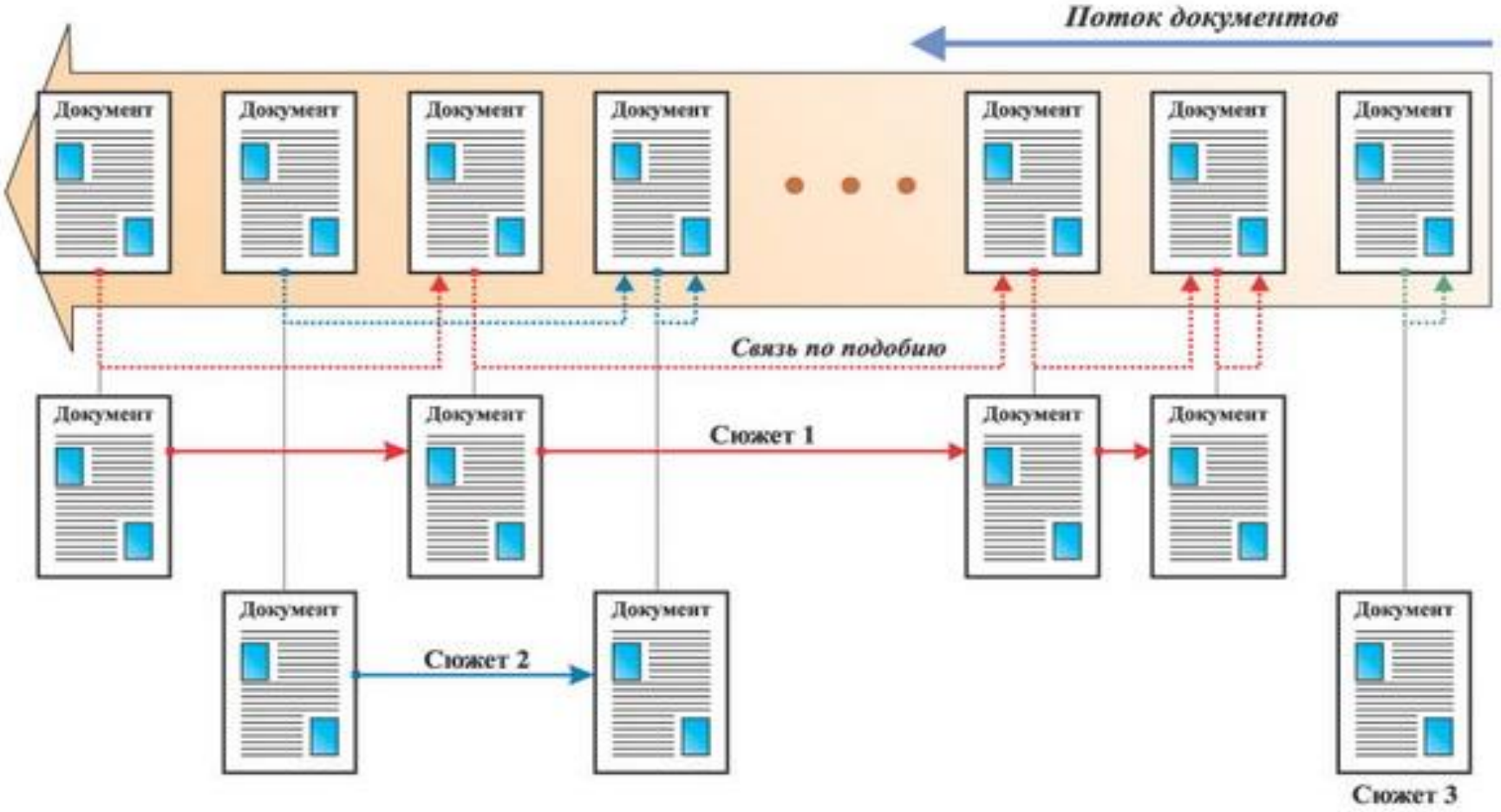
Релевантность и пертинентность

Профессиональный запрос к системе «Интегрум» по теме «Услуги связи»

«услуги связи» ИЛИ *«междугородные переговоры»* **ИЛИ** «телефонные переговоры» **ИЛИ «мобильная связь» ИЛИ** *«фиксированная связь»* или «сотовая связь» **ИЛИ «сотовый оператор» ИЛИ** *«средства связи»* **ИЛИ** «телефонная связь» **ИЛИ «спутниковая связь» ИЛИ** *«космическая связь»* **ИЛИ** GPS **ИЛИ ростелеком ИЛИ** *связьинвест* **ИЛИ** госкомсвязь **ИЛИ госкомтелеком ИЛИ** *госсвязьнадзор* **ИЛИ** телекоммуникации **ИЛИ электросвязь ИЛИ** *АТС* **ИЛИ** ГТС **ИЛИ минсвязи ИЛИ** *«министерство связи»* **ИЛИ** «волоконно-оптическая линия связи» **ИЛИ ВОЛС**

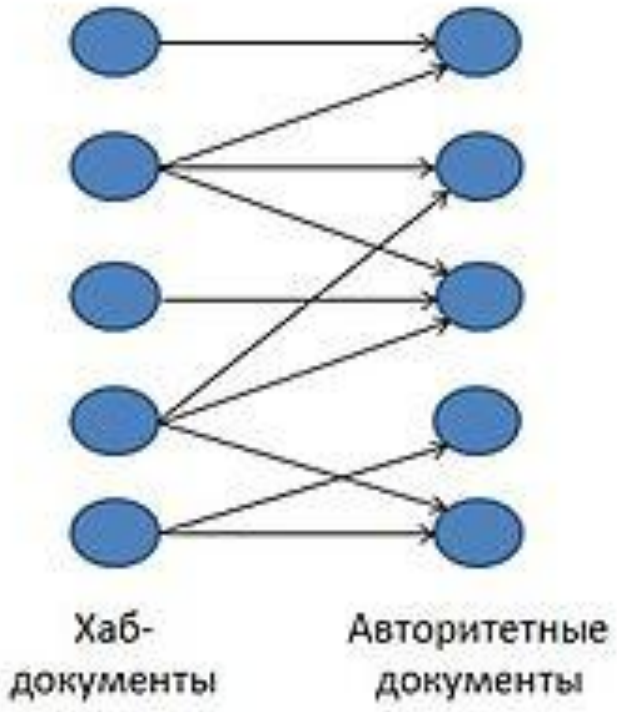
Рассылка сообщений по теме «Мобильная связь»

(((мобильн~связ) | (мобільн~зв?яз) | (сотов~связ) | (стільник~зв?яз) | (беспроводн~связ) | (бездрот~зв?яз) | (бесперебойн~связ) | (безперебійн~зв?яз) | j2me]| ems]| 3g]| gprs]| ggsn]| sgsn]| sms]| mms]| ems]| bluetooth]| mms]| tdma]| multipoint]| pcs]| cdma]| ofdm]| vpn]| wap]| umts]| gsm)&((моб~телефон)| (стільник~телефон)| (сотов~телефон)))! this.is

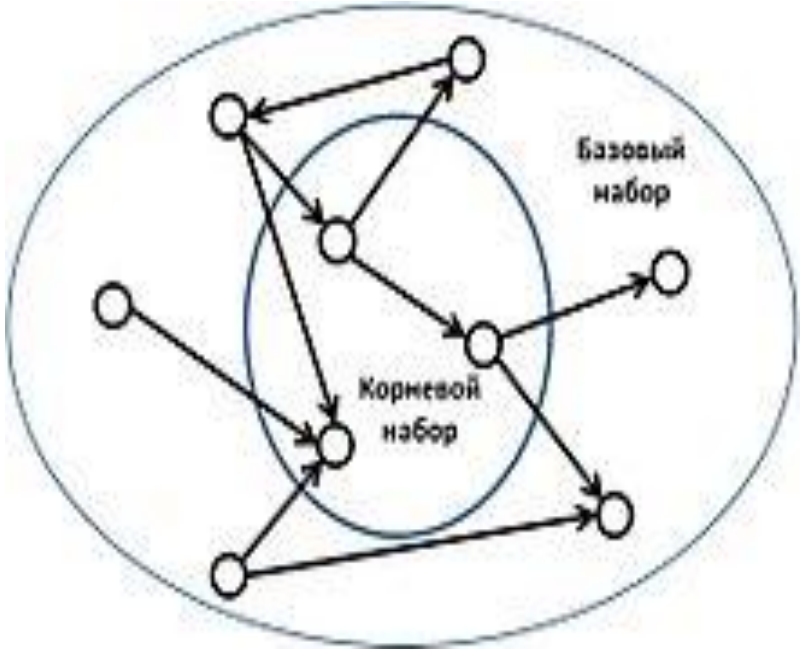


Сюжеты - результат построения семантических цепочек

Алгоритм HITS



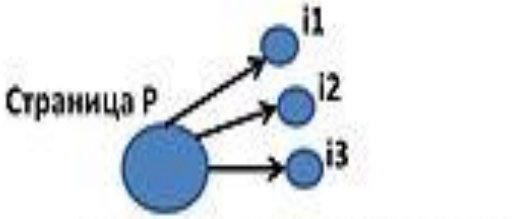
Плотно связанный набор авторитетных и хаб-документов



Расширение корневого множества релевантных страниц в базовом наборе

Начало ранжирования

$$\forall p,\ \mathbf{auth}(p) = 1 \text{ и } \mathbf{hub}(p) = 1$$



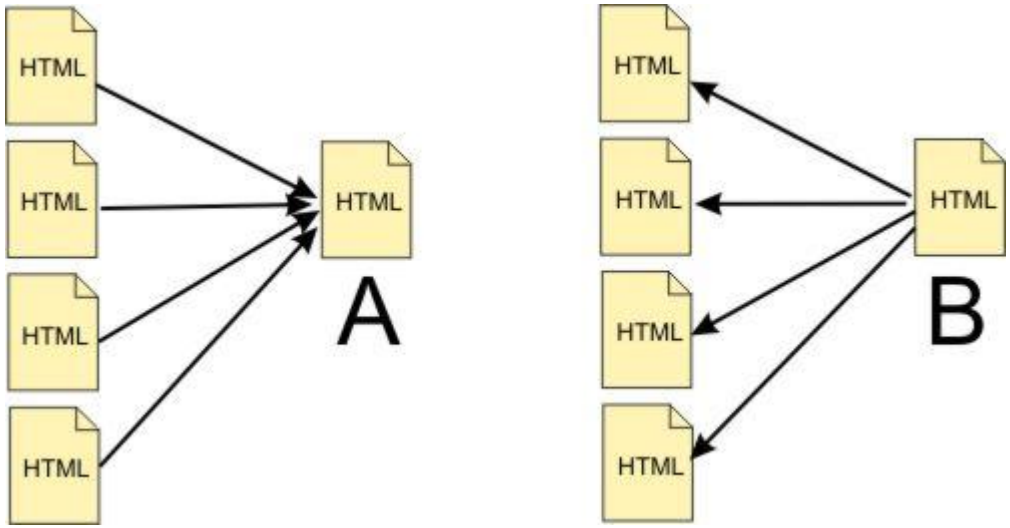
Правило обновления авторитетности

$$\mathbf{auth}(p) = \sum_{i=1}^{n} \mathbf{hub}(i)$$

Правило хаб-обновления

$$\mathbf{hub}(p) = \sum_{i=1}^{n} \mathbf{auth}(i)$$

Недостатки алгоритма HITS



Нормализация URL (примеры)

HTTP://www.Example.com/ → http://www.example.com/

http://www.example.com/a%c2%b1b → http://www.example.com/a%C2%B1b

http://www.example.com/%7Eusername/ → http://www.example.com/~username/

http://www.example.com:80/bar.html → http://www.example.com/bar.html

http://www.example.com/alice → http://www.example.com/alice/

http://www.example.com/../a/b/../c/./d.html → http://www.example.com/a/c/d.html

http://www.example.com/default.asp → http://www.example.com/
http://www.example.com/a/index.html → http://www.example.com/a/

http://www.example.com/bar.html#section1 → http://www.example.com/bar.html

http://208.77.188.166/ → http://www.example.com/

https://www.example.com/ → http://www.example.com/

http://www.example.com/foo//bar.html → http://www.example.com/foo/bar.html

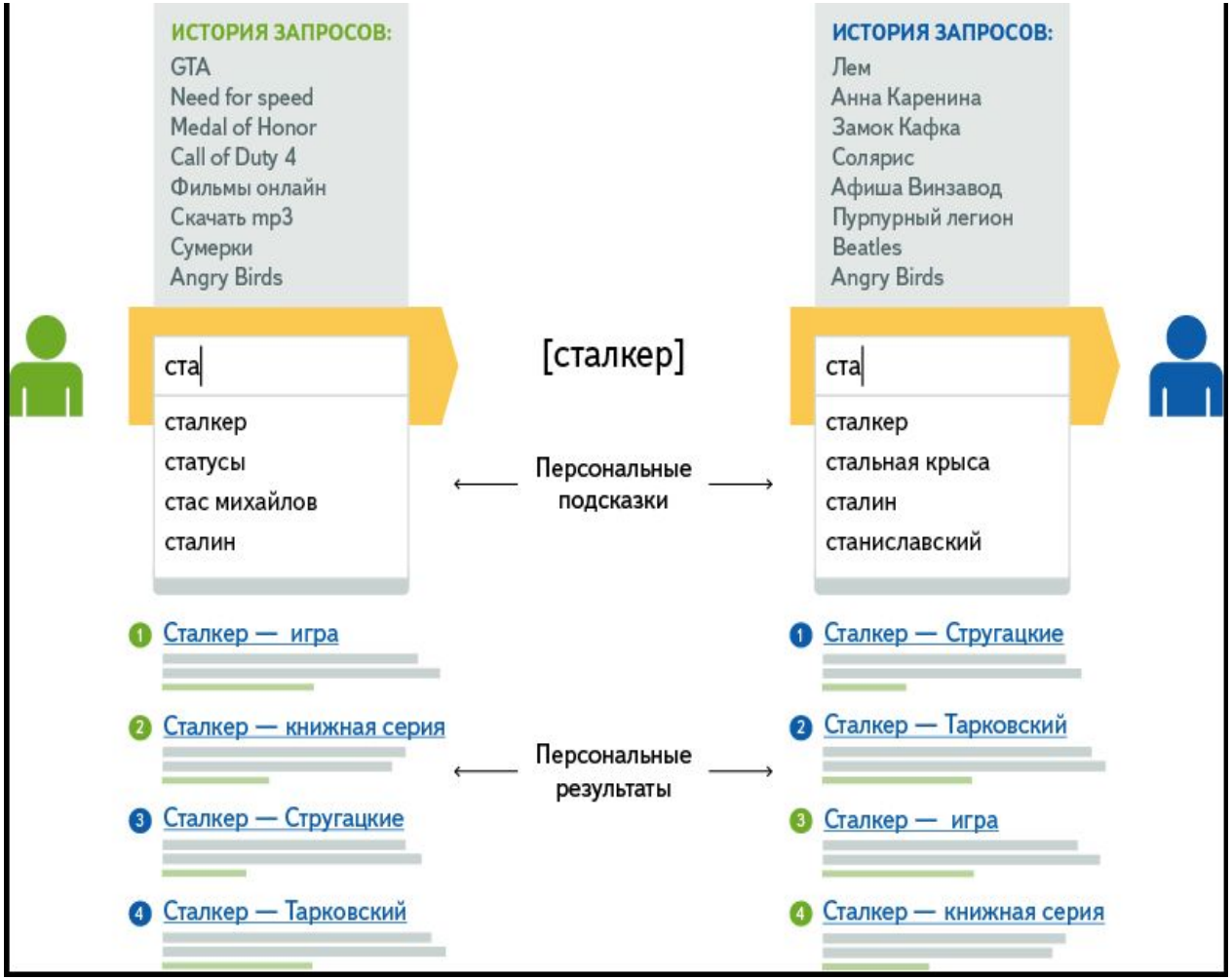
http://www.example.com/ → http://example.com/

http://www.example.com/display?id=123&fakefoo=fakebar → http://www.example.com /display?id=123

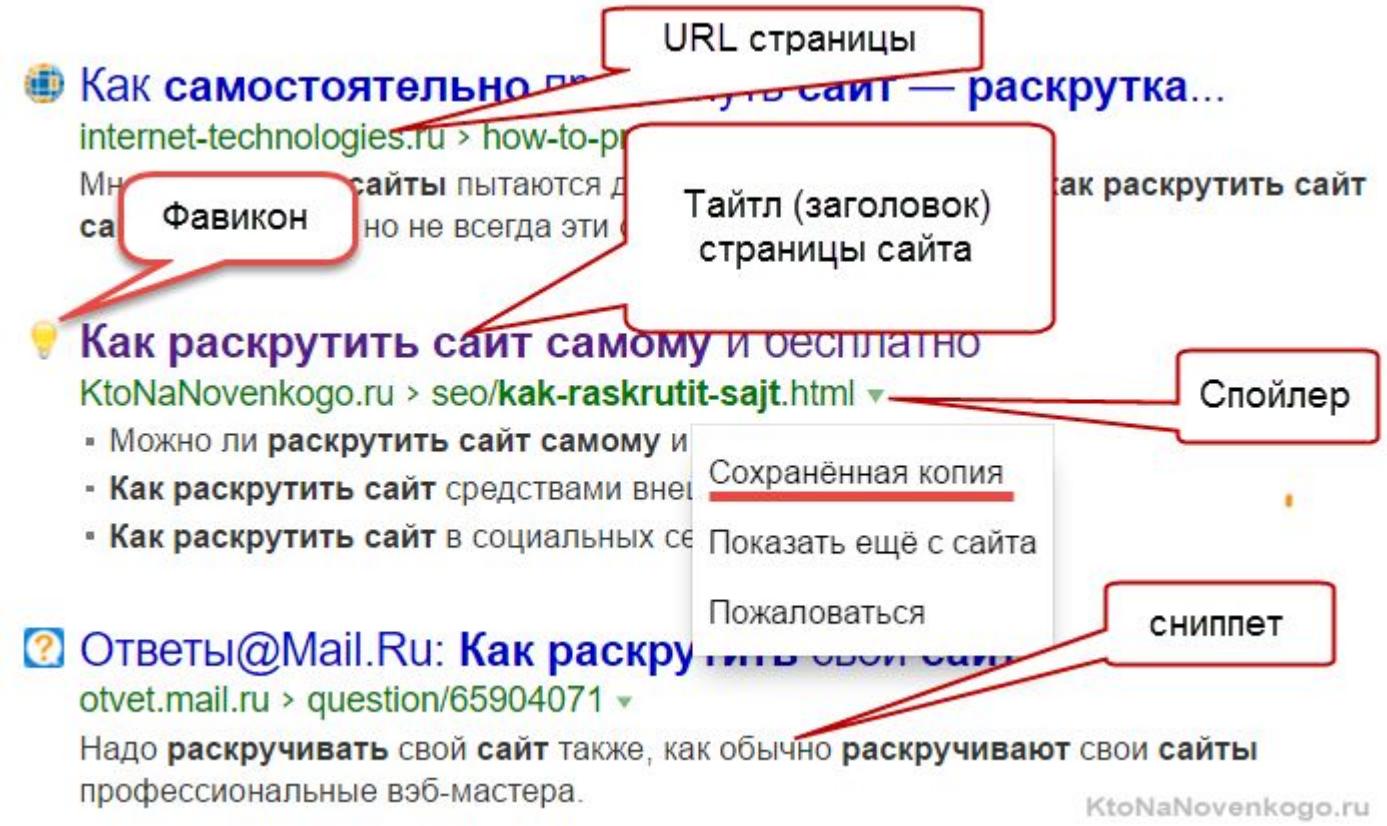
http://www.example.com/display?id=&sort=ascending → http://www.example.com/display

http://www.example.com/display? → http://www.example.com/display

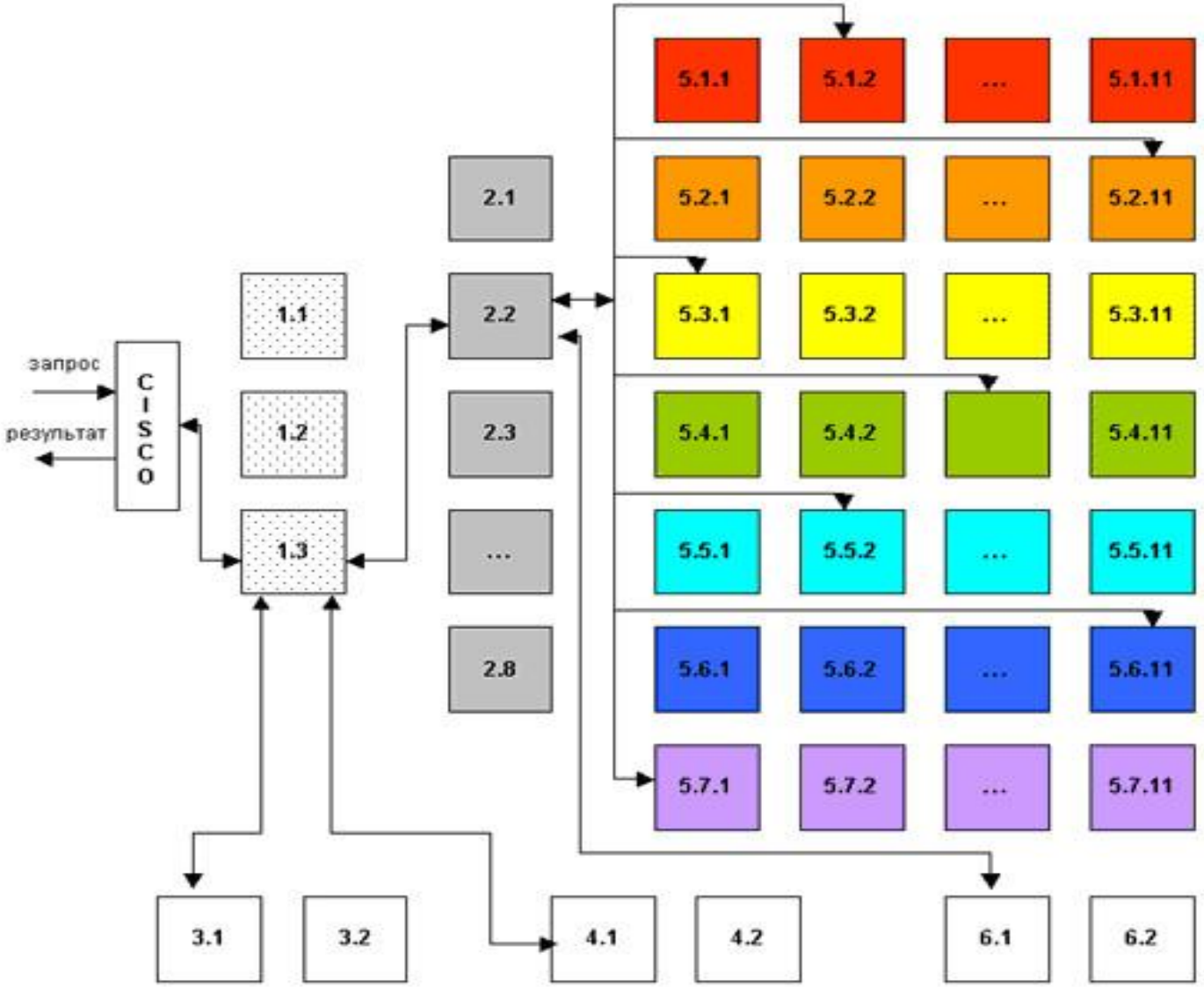
Пузырь фильтров



http://ktonanovenkogo.ru/seo/search/kak-rabotayut-poiskovye-sistemy-snippet-index.html



Обработка поискового запроса в системе «Рамблер»



1.1 – 1.3 – Frontend-сервера; 2.1 – 2.8 – Proxy-сервера; 3.1 – 3.2 – поиск по товарам; 4.1 – 4.2 – поиск по Top 100; 5.1.1 – 5.7.11 – Backend-сервера, содержащие основную индексную базу; 6.1 – 6.2 – Backend-сервера, содержащие быструю базу.

Самостоятельно

1. Алгоритм Беллмана-Форда -
2. Алгоритм Левита -
3. Алгоритм Ли
4. Двунаправленный поиск
5. Алгоритм Дейкстры -
6. Задача о кратчайшем пути
7. Информированный метод поиска
8. Лексикографический поиск в ширину -
9. Неинформированный метод поиска
10. Поиск в глубину -
11. Поиск в ширину
12. Поиск по первому наилучшему совпадению -
13. Поиск пути -
14. Интерполяционный поиск
15. Альфа-бета-отсечение -
16. Дихотомия
17. Задача поиска ближайшего соседа
18. Метод перебора
19. Троичный поиск