Вычислительный Центр Российской Академии Наук

«Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет»

Постановка задачи

- Разработать методы выявления структур сходства в клиентских средах
- Специфицировать методы для обработки логов поисковой машины
- Построить карты сходства ресурсов Интернет

Анализ Клиентских Сред

Задачи:

- **принятие решений**
- **маркетинг**
- **управление взаимоотношениями с клиентами**

Анализ Клиентских Сред

Схема:

- **Исходный протокол действий клиентов**
- **Выполнение цепочки процедур**
- **Решение конкретной задачи**

Исходные данные - лог поисковой системы

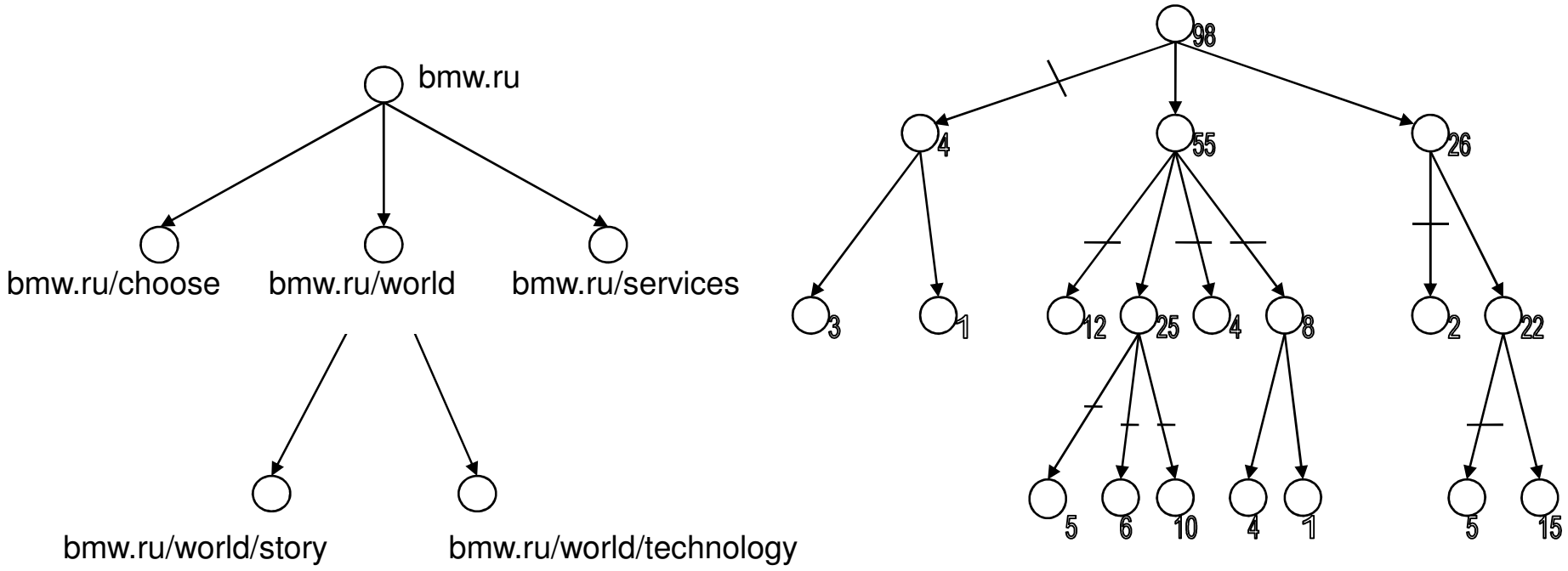
```
100002171080304956
    Валютное регулирование и валютный контроль в РФ 1110371715    281743  0
        http://tandem-forum.ru/seminars/info/seminar39.html    1110371732
        http://tandem-law.ru/seminars/info/seminar283.html    1110371757
        http://www.nns.ru/krizis/kabinet/2104/krizis8.html
        http://dit.perm.ru/articles/management/data/021217.htm  1110371805
        http://www.pomosch.com/article.php?sectionId=6&articleId=53
        http://www.e-ducate.ru/info/3523
        http://www.seminarium.ru/cgi-bin/viewer/events/data.cgi?eventID=1007519
        http://anons.webplan.ru/cgi-bin/shownew.cgi?numn=1518
        http://www.kadis.ru/daily/index.html?id=4405
        http://www.e-seminar.ru/root.phtml?PageID=168&EventID=486&CenterID=165  1110373262
    работа  1110527366    97008639      0
        http://www.job.ru/
        http://www.joblist.ru/
        http://rabota.mail.ru/
        http://www.zarplata.ru/ 1110527399
        http://www.ustroim.ru/
        http://www.rabota.ru/
        http://www.infocentre.net.ru/
        http://www.alljobs.ru/
        http://job.kulichki.net/
        http://www.webit.ru/
100002701110038640
                ...
```

- 14 606 пользователей
- 1 972 636 ресурсов (из них 129 600 были выбраны пользователями)
- Интервал времени: 1 неделя работы поисковой системы
- Объем лога 4Гб (время обработки около 30 секунд)

Общая схема обработки данных

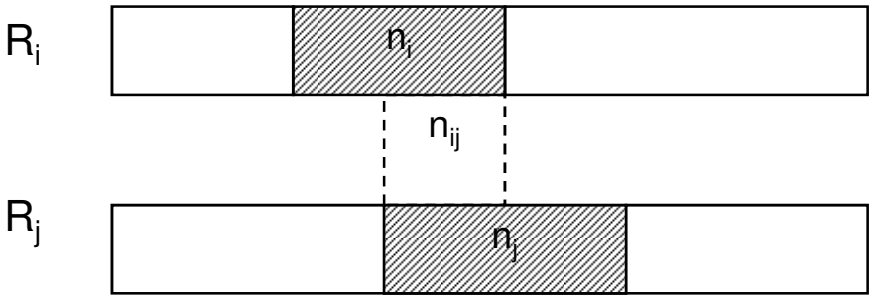
1. Формирование полных словарей пользователей и ресурсов
2. Редукция словарей
3. Формирование матрицы посещений
4. Формирование разреженной матрицы сходства ресурсов
5. Построение карты сходства всех ресурсов (карта Интернета)
6. Построение локальной карты сходства в окрестности заданного ресурса.

Формирование и редукция словарей пользователей и ресурсов

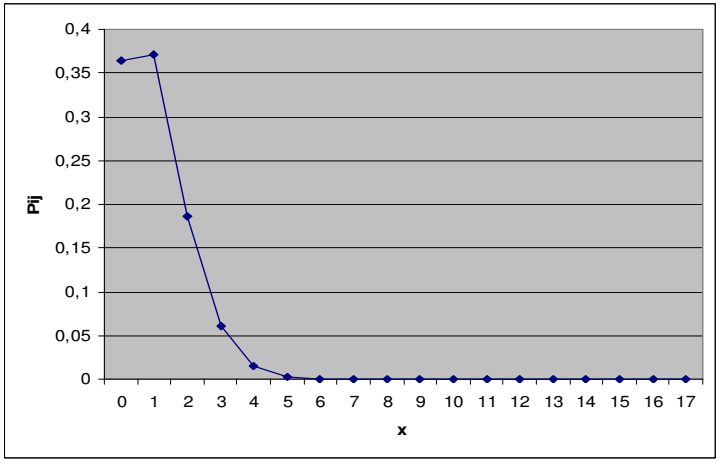


Пример построения и редукции словаря ресурсов с порогом 13 посещений

Вычисление оценок сходства ресурсов



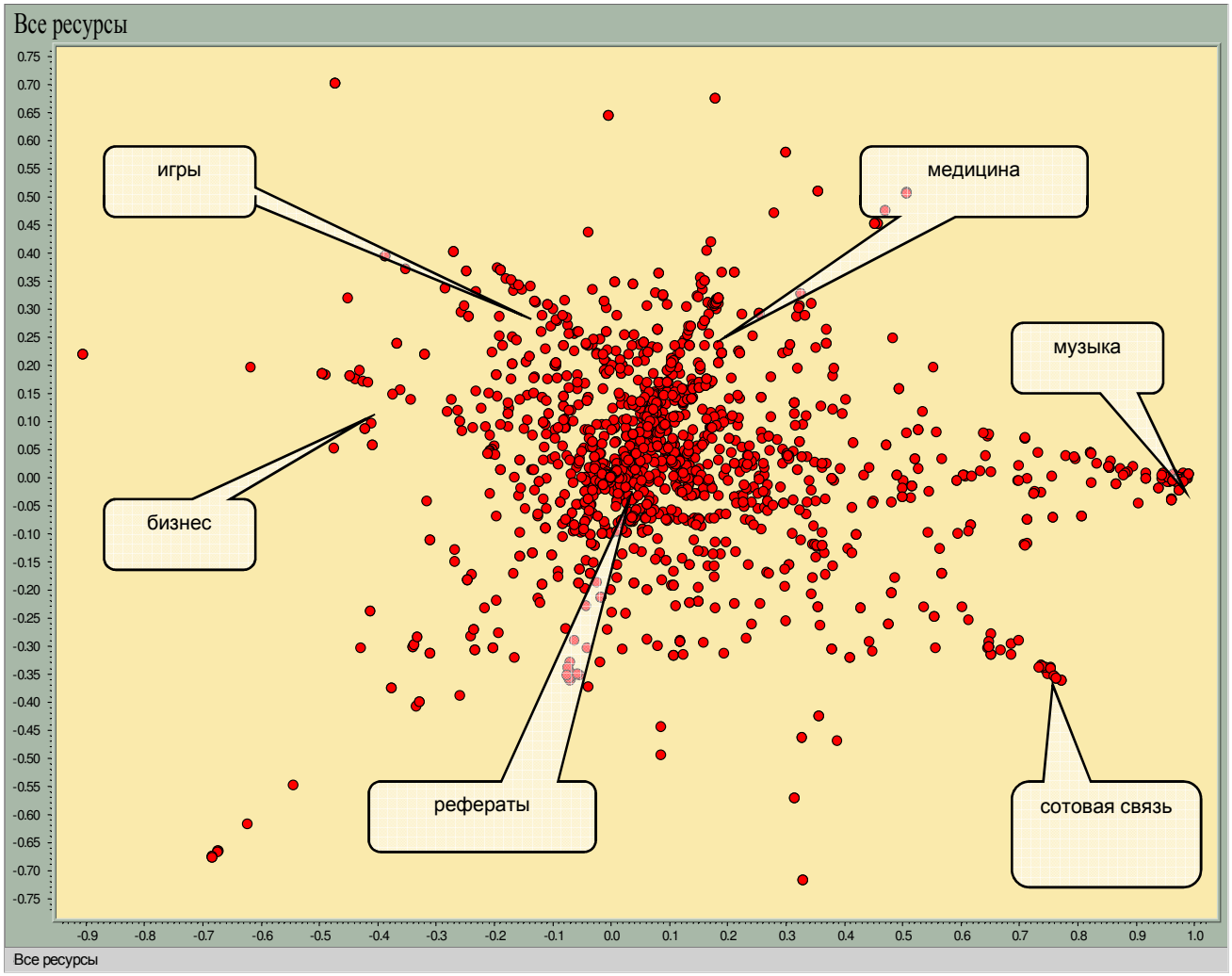
Гипергеометрическое распределение:



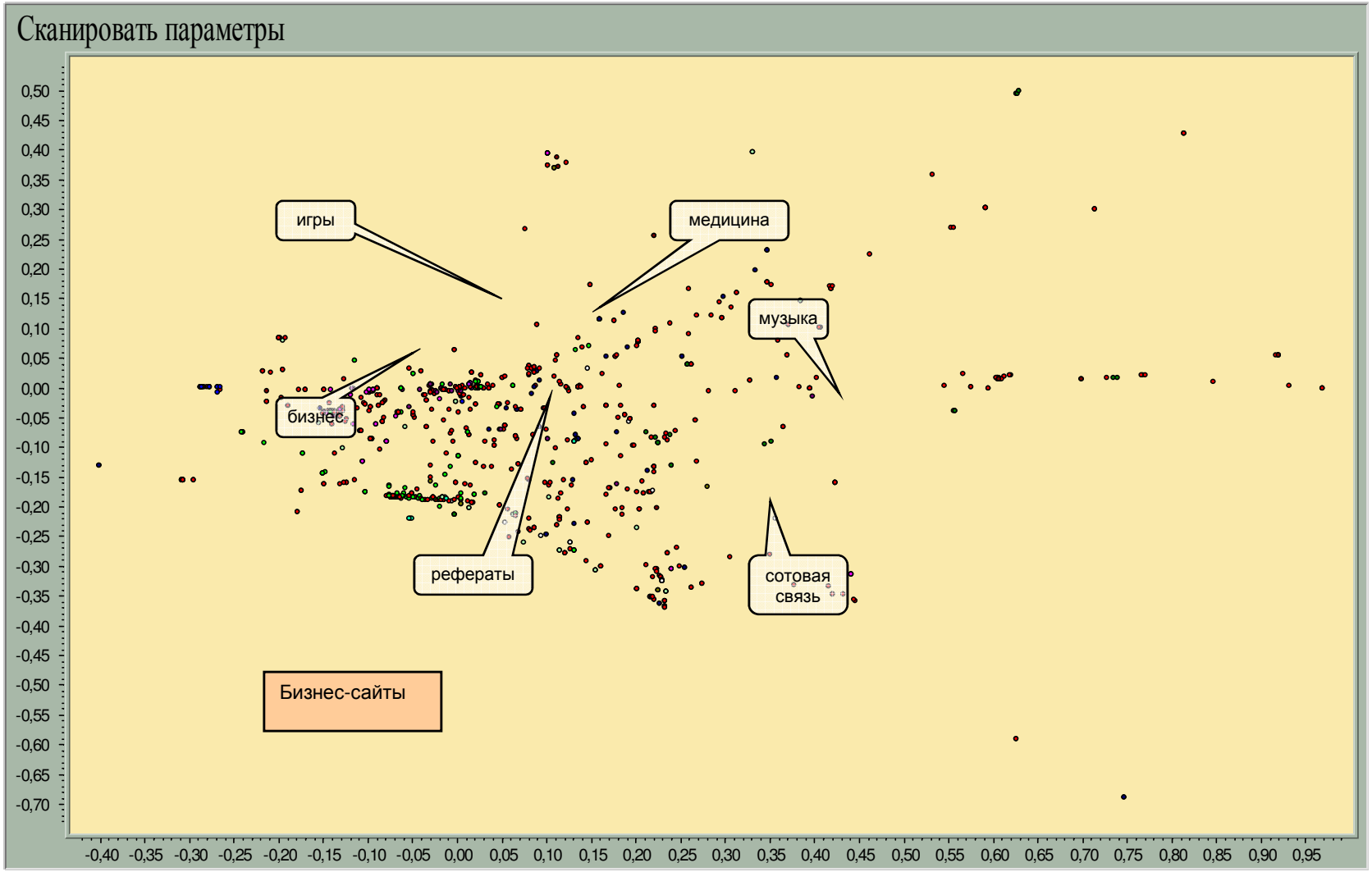
$$P_{ij} = P(n_{ij} = x) = \frac{C_{n_i}^{x} C_{U-n_i}^{n_j - x}}{C_U^{n_j}}$$

$$\rho(i, j) = \left(\frac{|\ln \alpha|}{|\ln P_{ij}|} \right)^3$$

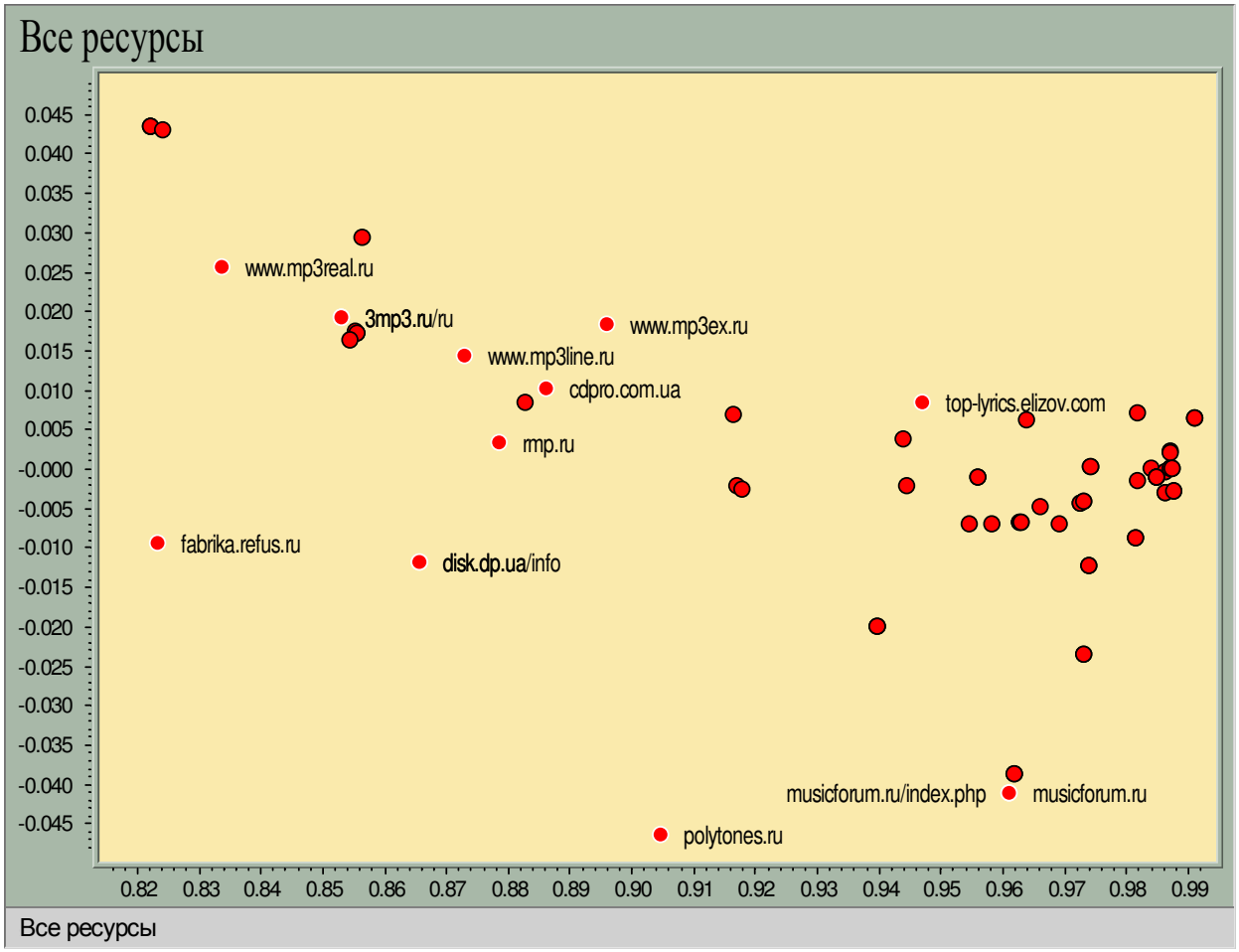
Карта сходства ресурсов



Карта сходства после оптимизации параметров

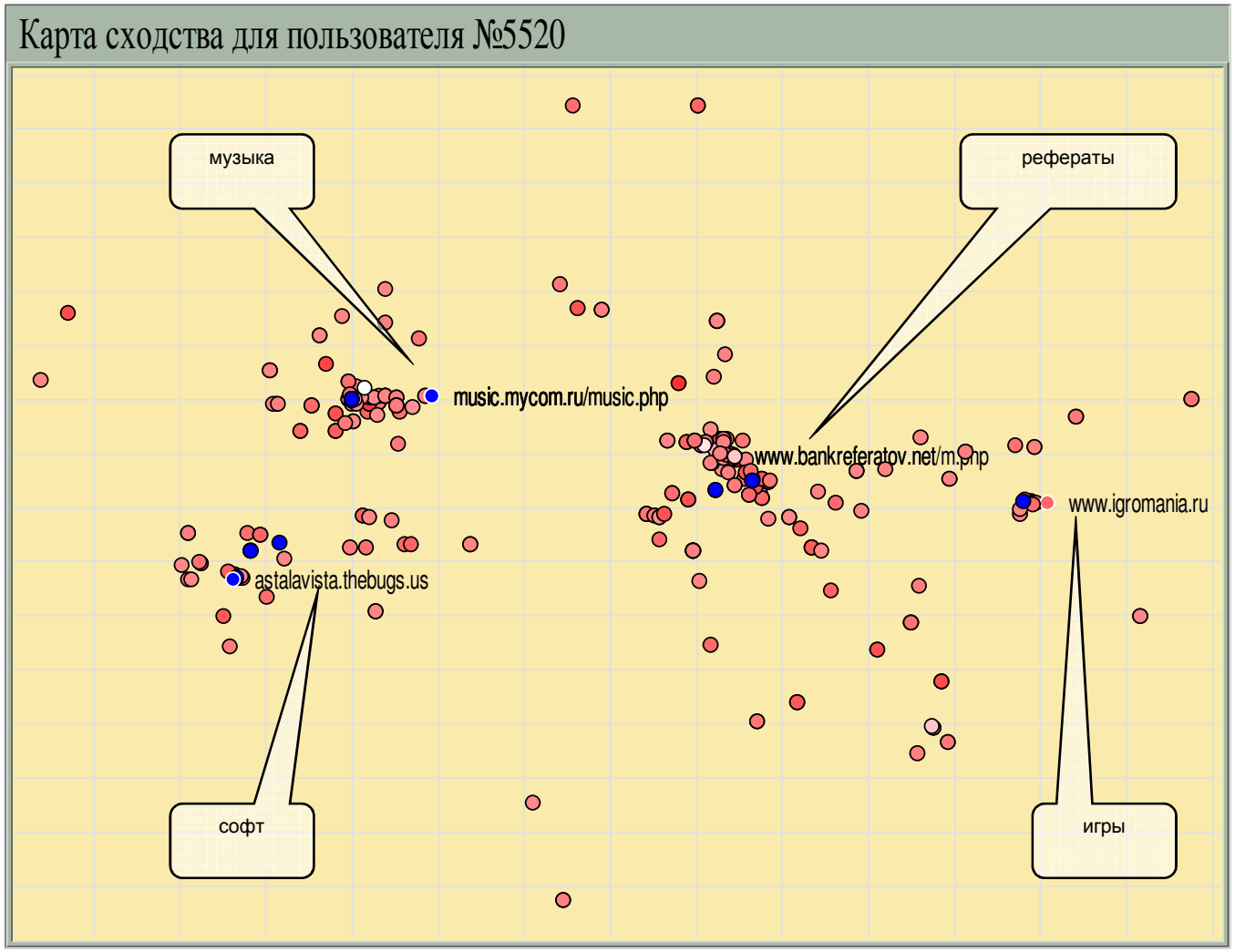


Фрагмент карты сходства



my3blka.narod.ru
mp3load.ru
mp3forum.ru
mp3rank.ru
top-lyrics.elizov.com
www.mp3ex.ru
www.mp3line.ru
www.oxid.ru
3mp3.ru
www.miditext.ru
www.rusmuz.ru
www.mp3search.ru
thesearch.ru
www.mp3.myfind.ru
www.russianmusic.de
www.mp3host.ru
www.mp3up.ru
mp3.volpi.ru
cdpro.com.ua
www.mp3zzz.ru
mp3-unique.liveinternet.ru
www.mp3collection.ru
www.mp3real.ru
rmp.ru
mp3.allabout.ru
www.filesale.ru
www.agharta.ru
moremusic.com.ru
ru.mp3s-download.com
meloman.tehnofil.ru
www.mp3tones.ru
www.muzzone.com
www.mp3zone.ru
www.zaycev.net
disk.dp.ua
www.uliss.ru
musicforum.ru

Персональная карта сходства



Основные результаты

- Разработаны и реализованы эффективные алгоритмы анализа логов поисковой машины
- Построена карта сходства ресурсов Интернет
- Предложена методика построения локальных карт сходства