НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Факультет математики

Евгений Ковалев и Иван Сухарев

Домашняя работа №4

Поиск закономерностей в данных

3 курс, майнор «Интеллектуальный анализ данных»,
группа ИАД-4

Москва, 2017 г.

1. Поиск частых множеств

Для решения этой задачи будем пользоваться SPMF. Для этого нужно сначала немного обработать данные и составить новый файл из 2000 строк (соответствующих компаниям), в которых будут расставлены числа, разделенные пробелами (номера использованных словосочетаний).

Код программы на Python:

```
import numpy as np
import pandas as pd
# считываем данные
data = pd.read_csv("a.txt", sep="\t")
"""
ищем все использования словосочетаний компаниями
(места, где в таблице стоит "1")
usages - массив, состоящий из
пар [номер компании, номер словосочетания]
"""
usages = np.argwhere(data.as_matrix())
# data_new - массив для обработанных данных
data_new = []
# company_number - номер обрабатываемой компании
company_number = 0
# company_temp - массив для номеров словосочетаний,
# использованных обрабатываемой компанией
company_temp = []
for index in usages:
    # обрабатывание одной компании
    if index[0] == company_number:
        company_temp.append(index[1])
    # переход на следующую компанию
    else:
        data_new.append(company_temp)
        company_temp = [index[1]]
        company_number += 1
# добавляем в массив для обработанных данных данные по последней компании
data_new.append(company_temp)
# вывод представляет из себя набор из 2000 строк (компании),
```

```
# состоящих из чисел, разделенных пробелами
# (номера использованных словосочетаний)
output = open("adv.txt", "w")
output.write("\n".join([" ".join(str(i) for i in data_new[j])
             for j in range(len(data_new))]))
output.close()
```

Далее было установлено, что для минимальной поддержки $minsupp = 35$ (35%) в пунктах получается 0, поэтому мы решали задачу для $minsupp = \frac{35}{2000} = 0.0175$ (1.75%) .

а) Количество частых множеств: 20910

б) Количество частых замкнутых множеств: 13812

в) Количество частых максимальных множеств: 4002

г) Найдем самые большие частые множества.

Код программы на Python:

In [2]:
```
# считываем результат (найденные частые множества)
with open('result.txt') as f:
    data = f.readlines()

# убираем значение поддержки
data = np.array(list(map(lambda x: x.split(' #SUP')[0], data)))

# вычисляем мощность каждого множества
data_str_lens = np.array(list(map(lambda x: len(x.split()), data)))

# выводим самые большие мощности
sorted(data_str_lens, reverse=True)[:10]
```

Out[2]:
```
[9, 9, 9, 9, 9, 9, 9, 9, 9, 8]
```

Итак, девять самых больших частых множеств содержат в себе 9 словосочетаний. Выведем эти группы словосочетаний.

Код программы на Python:

In [3]:
```
# самые большие множества
largest_frequent_sets = data[np.where(data_str_lens == 9)[0]]

# считываем исходный файл с данными
a = pd.read_csv('a.txt', sep='\t')

# во всех множествах меняем индексы словосочетаний на сами словосочетания
largest_frequent_sets = list(map(lambda x: a.columns[[int(y) for y in x.split()]],
                                 largest_frequent_sets))

# вывод
print('\n'.join([', '.join(x.tolist()) for x in largest_frequent_sets]))
```

```
casino gambling, casino game, casino game online, casino internet, casino line, casino net, casino online, gambling internet, g
ambling online
casino gambling, casino gambling online, casino game, casino game online, casino internet, casino line, casino net, casino onli
ne, gambling online
casino gambling, casino gambling online, casino game online, casino internet, casino line, casino net, casino online, gambling
 internet, gambling online
casino gambling, casino gambling online, casino game, casino game online, casino internet, casino line, casino online, gambling
internet, gambling online
affordable hosting web, cheap hosting site web, cheap hosting web, company hosting web, cost hosting low web, discount hosting
 web, hosting services web, hosting site web, hosting web
casino, casino gambling, casino gambling online, casino game, casino internet, casino online, gambling, gambling internet, gamb
ling online
casino, casino gambling, casino gambling online, casino game, casino game online, casino internet, casino online, gambling inte
rnet, gambling online
casino, casino gambling, casino gambling online, casino game, casino game online, casino internet, casino online, gambling, gam
bling online
casino, casino gambling, casino gambling online, casino game online, casino internet, casino online, gambling, gambling interne
t, gambling online
```

Видно, что пятое множество можно отнести к рынку веб-хостинга, все остальные — к рынку азартных игр и казино.

Эти же множества являются самыми большими частыми замкнутыми и самыми большими частыми максимальными.

2. Поиск ассоциативных правил

а) Количество ассоциативных правил: 10940

б) Количество замкнутых ассоциативных правил: 7098

в) 5 самых частых правил при минимальной достоверности $minconf = 0.8$:

1) based business home, business home opportunity ==> business home #SUP: 90 #CONF: 0.849 (90 компаний (занимающихся, вероятно, бизнесом на дому) используют словосочетания based business home, business home opportunity и business home; 84.9% компаний, использующих первые два словосочетания, используют третье тоже)

2) marketing online ==> internet marketing #SUP: 91 #CONF: 0.867 (91 компания (занимающаяся, вероятно, интернет-маркетингом) использует словосочетания marketing online и internet marketing; 86.7% компаний, использующих первое словосочетание, используют второе тоже)

3) based business home opportunity ==> based business home #SUP: 102 #CONF: 0.829 (102 компании (занимающихся, вероятно, бизнесом на дому) используют словосочетания based business home opportunity и based business home; 82.9% компаний, использующих первое словосочетание, используют второе тоже)

4) based business home opportunity ==> business home opportunity #SUP: 105 #CONF: 0.854 (105 компаний (занимающихся, вероятно, бизнесом на дому) используют словосочетания based business home opportunity и business home opportunity; 85.4% компаний, использующих первое словосочетание, используют второе тоже)

5) hosting site web ==> hosting web #SUP: 109 #CONF: 0.808 (109 компаний (занимающихся, вероятно, веб-хостингом) используют словосочетания hosting site web и hosting web; 80.8% компаний, использующих первое словосочетание, используют второе тоже)

3. Анализ посещаемости сайтов на основе решеток формальных понятий

а) Результаты после удаления объектов и признаков:

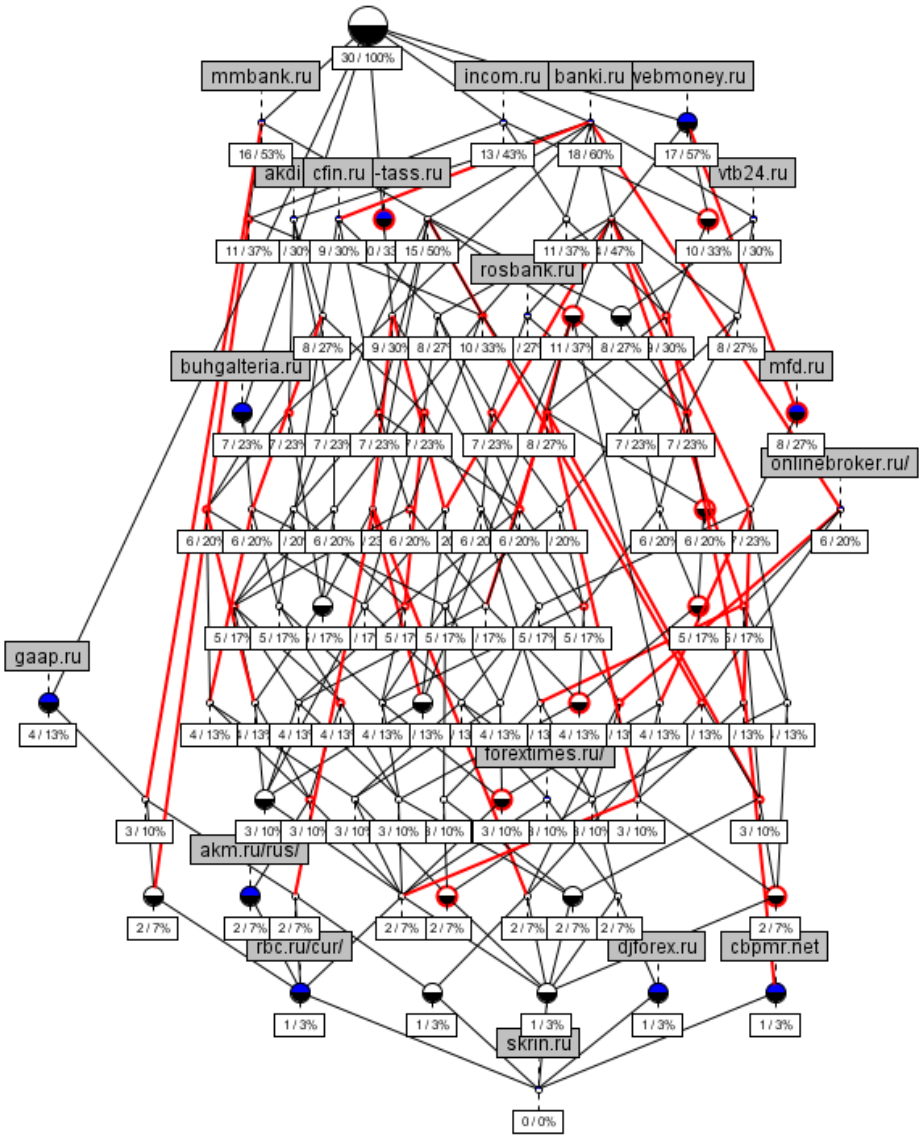
Финансы: 30 объектов, 24 признака, 101 понятие

Образование: 30 объектов, 19 признаков, 100 понятий

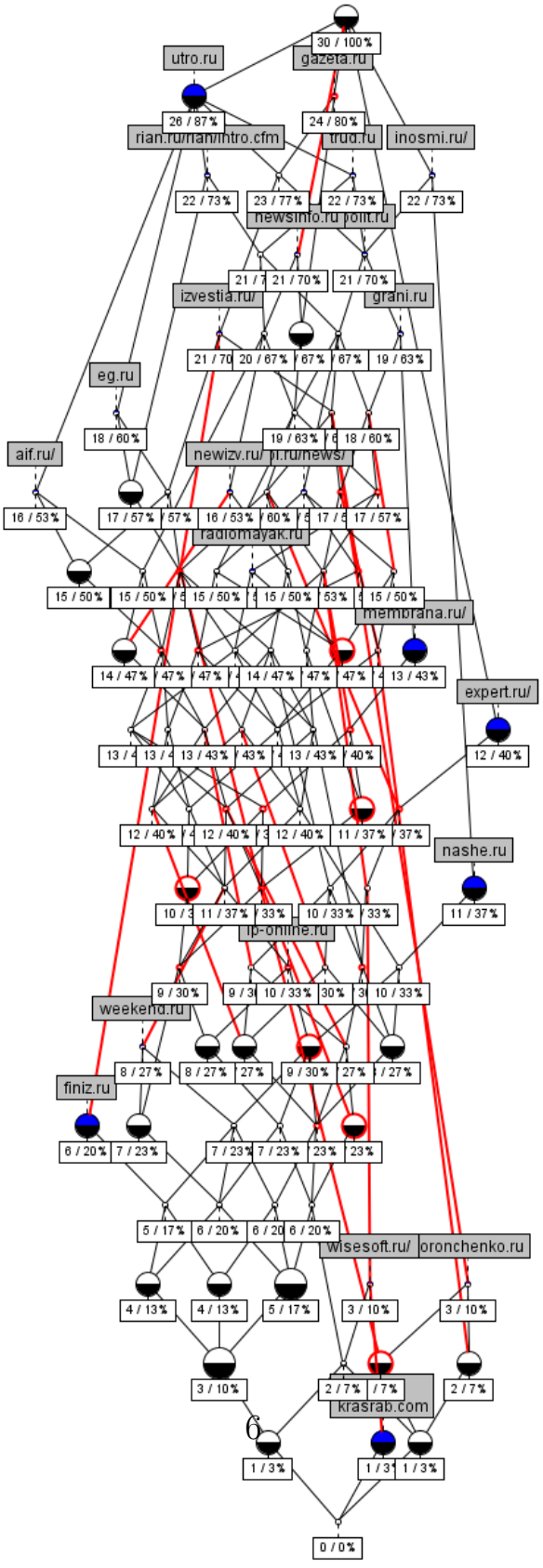
Новости: 29 объектов, 17 признаков, 100 понятий

б) Диаграммы решеток понятий:

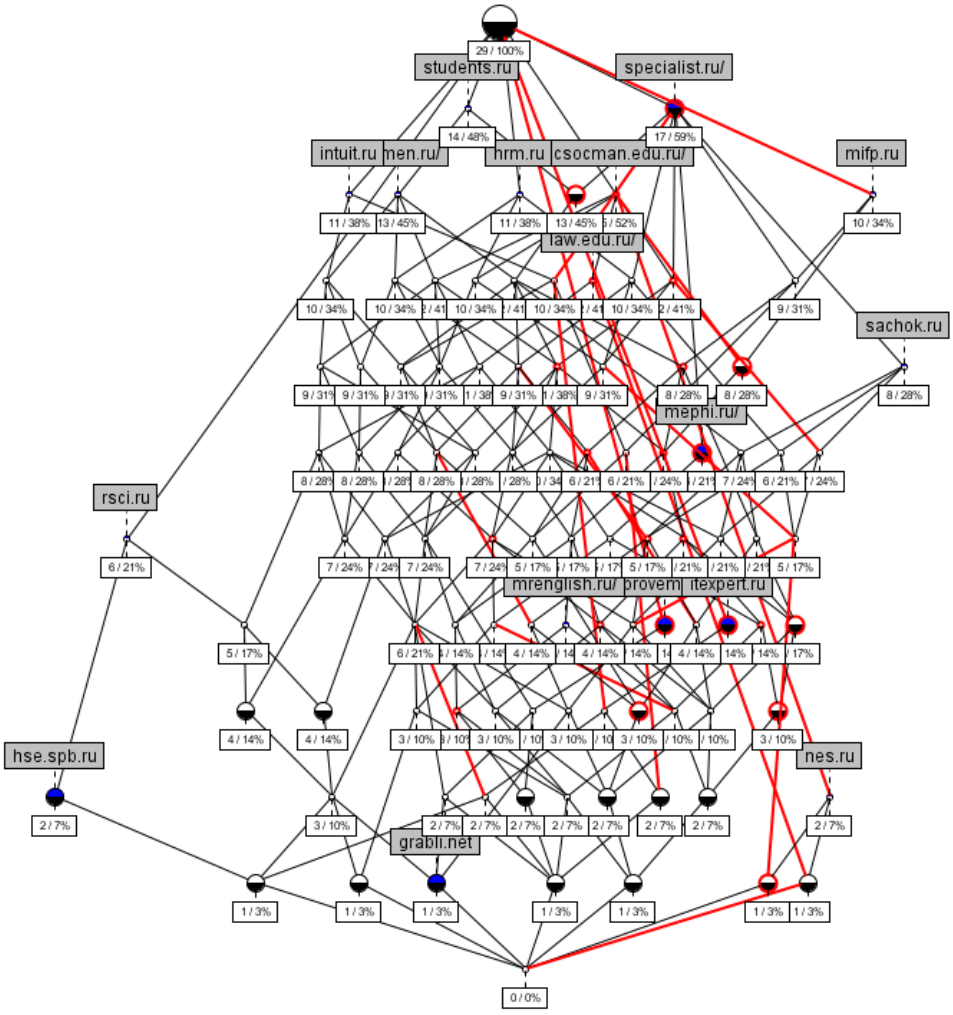
Финансы:


30 / 100%
mmbank.ru
incom.ru
banki.ru
vebmoney.ru
16 / 53%
13 / 43%
18 / 60%
17 / 57%
akdi
cfin.ru
-tass.ru
vtb24.ru
rosbank.ru
buhgalteria.ru
mfd.ru
onlinebroker.ru/
gaap.ru
forextimes.ru/
akm.ru/rus/
rbc.ru/cur/
djforex.ru
cbpmr.net
skrin.ru
0 / 0%

Образование:

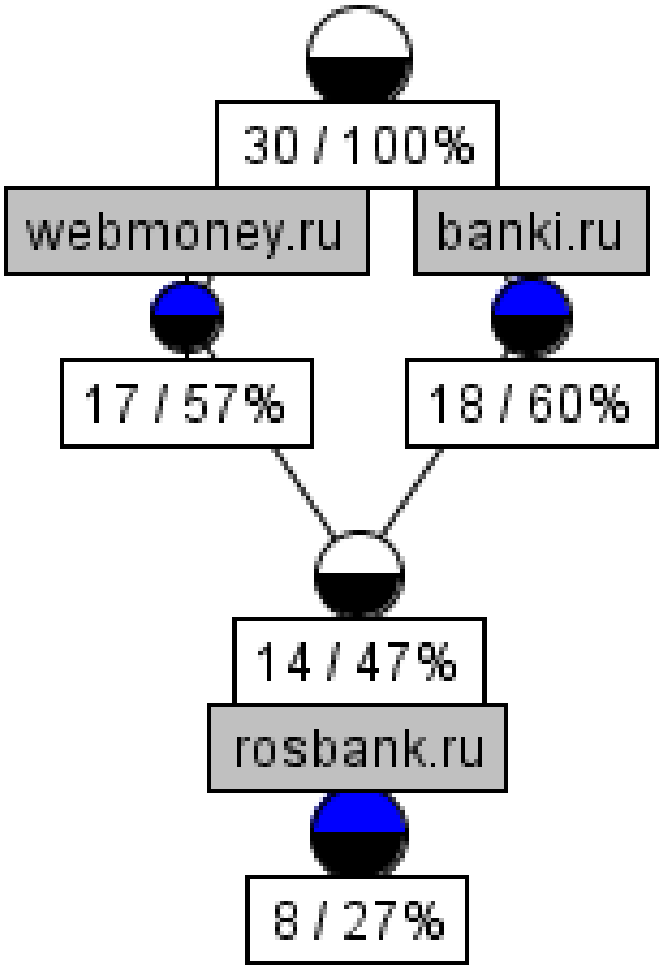


Новости:



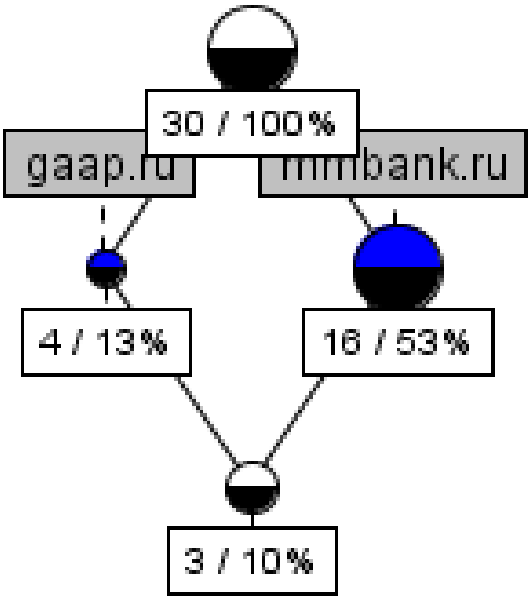
в) Финансы:

1) <8 объектов, {webmoney.ru, banki.ru, rosbank.ru}>



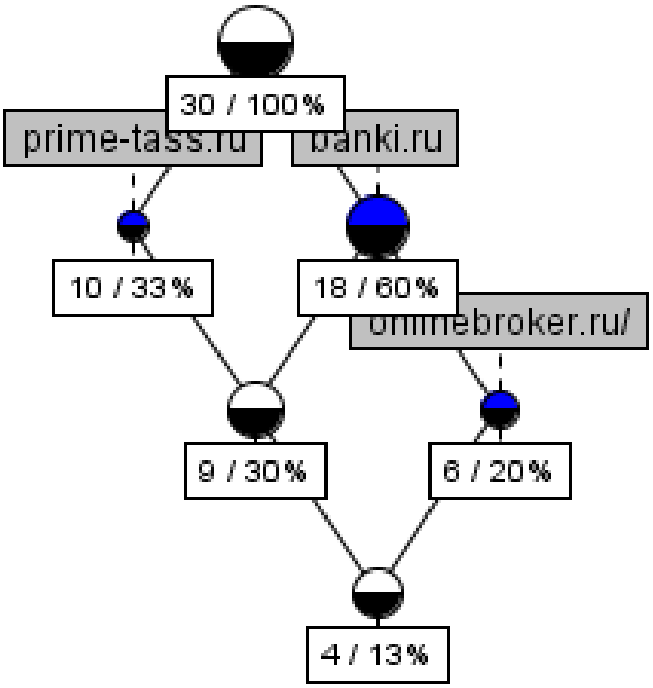
Около 27% рассматриваемых пользователей заходят на информационный портал banki.ru и на сайт системы расчетов webmoney.ru (возможно, у них там есть электронный кошелек), причем все они, вероятно, клиенты Росбанка, так как каждый из них заходит и на rosbank.ru тоже.

2) <3 объекта, {mmbank.ru, gaap.ru}>



Около 10% рассматриваемых пользователей заходят на сайт Банка Москвы mmbank.ru и сайт теории и практики управленческого учета gaap.ru; возможно, это студенты, изучающие близкие к этой области предметы.

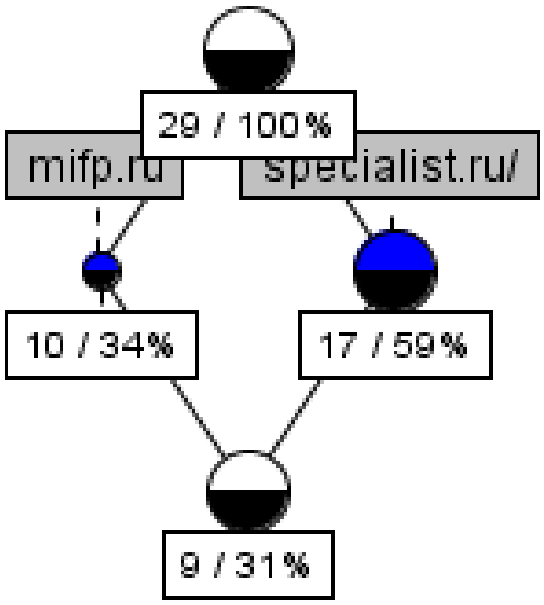
3) <4 объекта, {prime-tass.ru, onlinebroker.ru/, banki.ru}>



Около 13% рассматриваемых пользователей заходят на инвестиционный портал ВТБ24 onlinebroker.ru, на сайт агенства экономической информации prime-tass.ru (он же 1prime.ru) и banki.ru; возможно, это какие-то инвесторы или люди других профессий, которым нужна подобная экономическая информация.

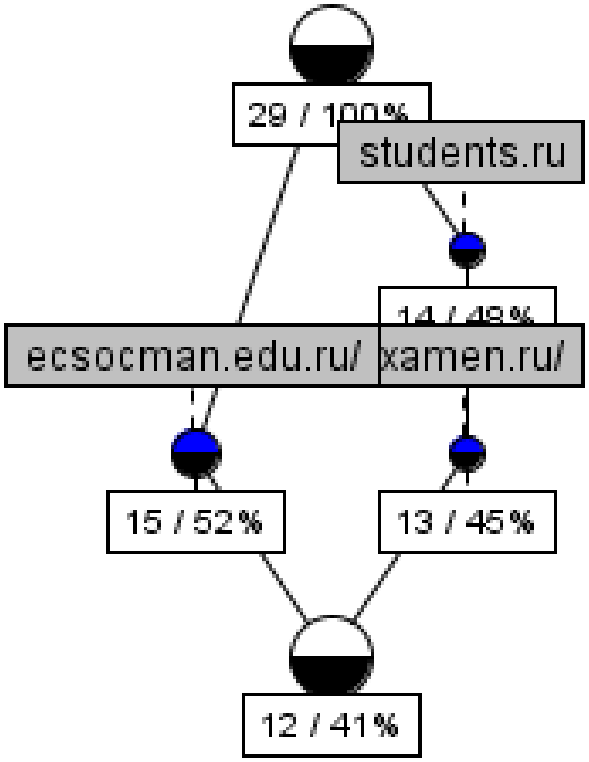
Образование:

1) <9 объектов, {specialist.ru/, mifp.ru}>



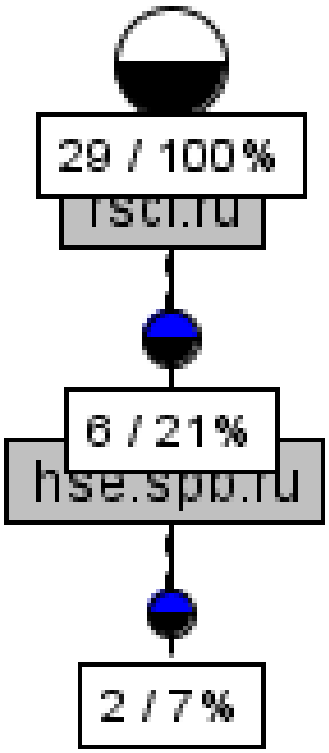
Около 31% рассматриваемых пользователей заходят на сайт, скорее всего, университета Синергия mifp.ru (ссылка устарела, теперь адрес synergy.ru) и сайт учебного центра «Специалист» при МГТУ им. Н.Э. Баумана specialist.ru/, возможно, это школьники, выбирающие себе какие-нибудь соответствующие курсы при университетах, либо абитуриенты, интересующиеся данными вузами.

2) <12 объектов, {ecsocman.edu.ru/, students.ru, examen.ru/}>



Около 41% рассматриваемых пользователей заходят на сайт федерального образовательного портала ЭСМ ecsocman.edu.ru/ и сайт с тестами по школьной программе, информацией об ОГЭ, ЕГЭ и т.д. examen.ru/, и вместе с ним каждый заходит также на сайт предположительно российского студенческого портала students.ru (по ссылке выдается явно не то, возможно, новый адрес — x-student.ru). Скорее всего, это школьники и абитуриенты, готовящиеся к экзаменам и поступлению в вузы.

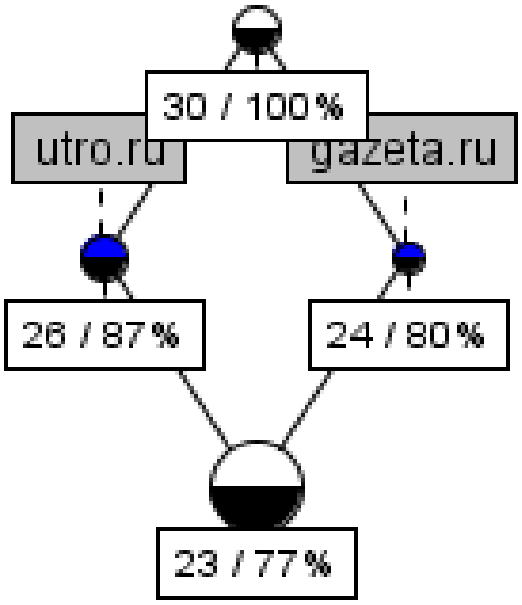
3) <2 объекта, {hse.spb.ru, rsci.ru}>



Около 7% рассматриваемых пользователей заходят на сайт питерского кампуса НИУ ВШЭ hse.spb.ru (новый адрес — spb.hse.ru), а также каждый из них заходит на сайт информационного интернет-канала NT-INFORM rsci.ru. Скорее всего, это студенты питерского кампуса НИУ ВШЭ, интересующиеся различными грантами и конкурсами и ищущие о них нужную информацию.

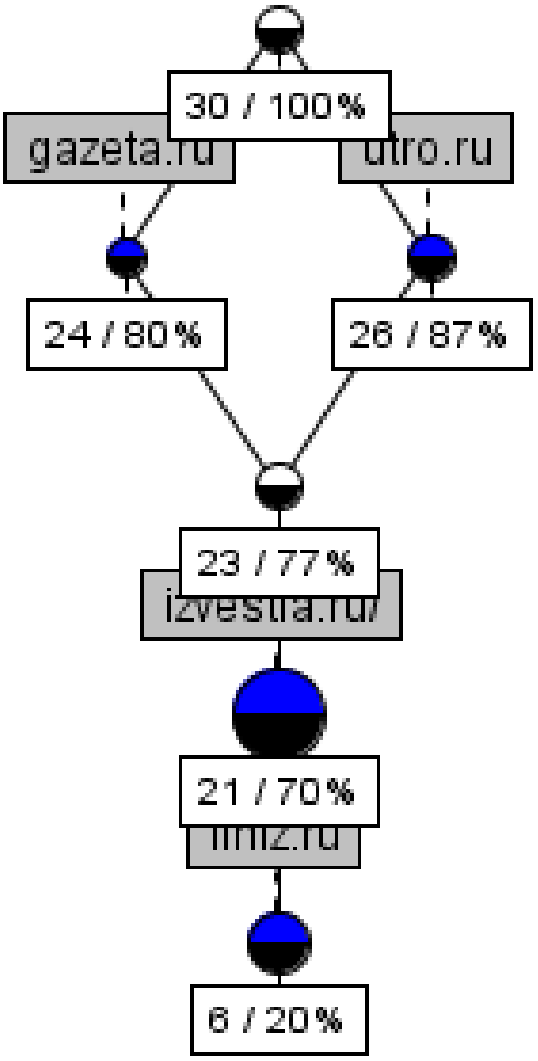
Новости:

1) <23 объекта, {gazeta.ru, utro.ru}>



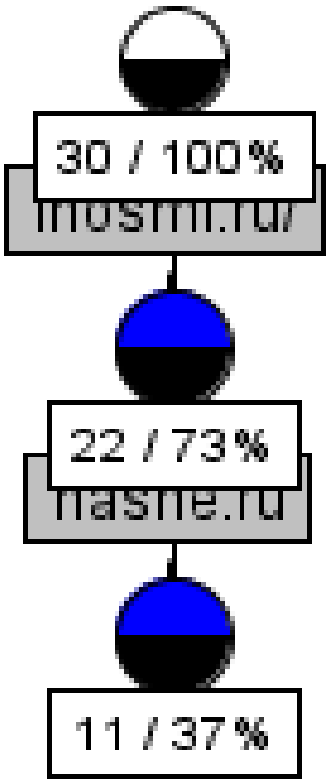
Около 77% рассматриваемых пользователей заходят на новостные сайты gazeta.ru и utro.ru — собственно, для того, чтобы прочитать последние новости.

2) <6 объектов, {gazeta.ru, utro.ru, izvestia.ru/, finiz.ru}>



Около 20% рассматриваемых пользователей заходят на новостные сайты finiz.ru, izvestia.ru/, gazeta.ru и utro.ru. Забавно, что каждый из посетителей сайта finiz.ru является также посетителем сайта izvestia.ru/ — на самом деле это одна и та же страница, и при переходе по первому адресу гостя перенаправляет на второй.

3) <11 объектов, {inosmi.ru/, nashe.ru}>

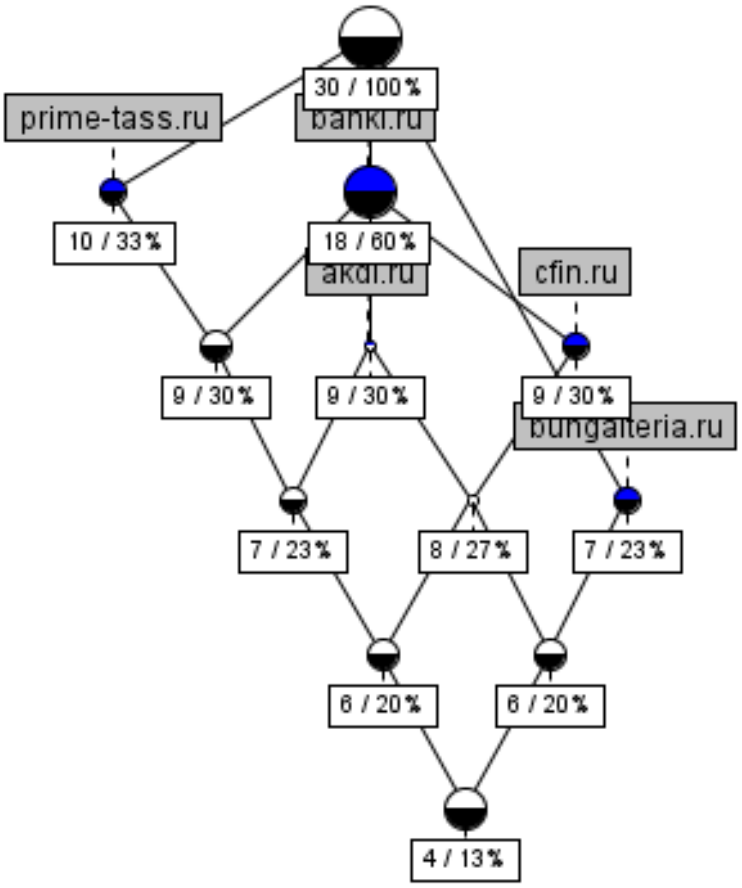


Около 37% рассматриваемых пользователей заходят на сайт радиостанции НАШЕ РАДИО nashe.ru, и каждый из них переходит также на новостной сайт inosmi.ru/. Честно говоря, не очень понятно, почему такое происходит; возможно, на сайте nashe.ru где-то есть мощная реклама данного новостного сайта, какой-нибудь баннер, через который идет перенаправление, потому что тематики этих двух сайтов не особо связаны друг с другом.

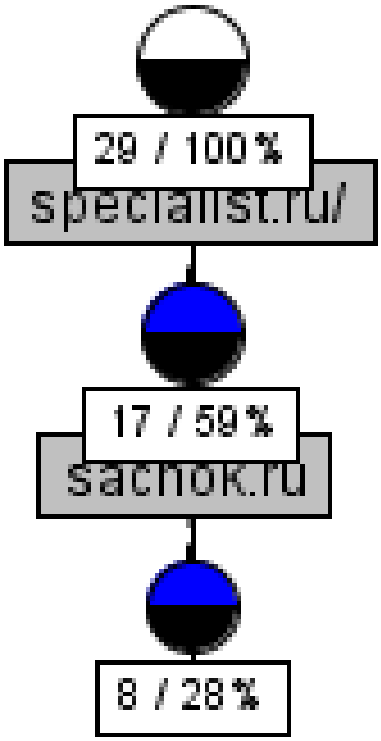
г) Импликации:

1) Финансы:

prime-tass.ru buhgalteria.ru ==> akdi.ru cfin.ru banki.ru ($support = 4$)



2) Образование:
sachok.ru ==> specialist.ru/ ($support = 8$)



3) Новости:
gazeta.ru inosmi.ru/ ==> trud.ru utro.ru polit.ru (*support* = 21)