# РАЗРАБОТКА СИСТЕМЫ ПОКАЗА МОБИЛЬНОЙ КОНТЕКСТНОЙ РЕКЛАМЫ

***Д.Н. Касымова,***
*программист-разработчик ООО «Мобил-2», Новосибирск,*
*Адрес: 630055, г. Новосибирск, ул. Мусы Джалиля, 11-819, Касымова Д.Н.,*
*Тел: 8-923-120-7864. E-mail: dilnara.kasymova@gmail.com*

***Статья посвящена развитию технологии использования сотовой связи для показа нацеленной мобильной рекламы. Рассмотрены алгоритмы подбора нацеленных рекламных сообщений абоненту сотового оператора. Проведен сравнительный анализ различных методов ускорения поиска по сходству для параметров, заданных сотовым оператором.***

**Ключевые слова:** Мобильная реклама. Поиск по сходству. Контекстная реклама. Индексы для поиска по сходству.

Мобильная реклама – это сравнительно новое явление на российском рекламном рынке. Использование данного канала коммуникаций в рекламных кампаниях началось несколько лет назад. Многие специалисты и эксперты рекламного рынка России считают, что у рекламы в мобильных телефонах есть хорошее будущее [Крапивинский 2008: 8]. Использование технологий мобильного маркетинга в рекламных кампаниях является эффективным способом продвижения товаров и услуг на рынок. Преимущества использования мобильного маркетинга очевидны: полностью автоматизированный двусторонний канал коммуникаций, работающий в режиме 24/7; возможность персональной коммуникации с целевой аудиторией, что позволяет формировать и поддерживать клуб лояльных к бренду потребителей; доступность для целевой аудитории, вне зависимости от местонахождения.

Обычно используются следующие способы доставки рекламы абонентам сотовой сети:

- SMS (*Short Message Service* – служба коротких сообщений) – система, позволяющая посылать и принимать текстовые сообщения при помощи сотового телефона. Как правило, абонент добровольно соглашается с рассылкой SMS-сообщений «рекламно-информационного характера» в обмен на получение бонусов (например, денежное вознаграждение за каждое сообщение, которое может использоваться для оплаты услуг сотовой связи);
- MMS (*Multimedia Message Service* – служба мультимедийных сообщений) – система, позволяющая посылать и принимать мультимедийные (изображения, мелодии, видео) сообщения при помощи сотового телефона. Аналогично SMS, абонент в обмен на согласие просматривать рекламу посредством MMS получает «информационно-развлекательный бонус». Например, в МТС абоненту высылается приглашение участвовать в акции, если абонент согласен, он отправляет в ответ на приглашение пустое MMS-сообщение. После этого ежедневно в течение определенного срока он будет бесплатно получать MMS-сообщение с погодой, анекдотом, гороскопом, кратким выпуском новостей и курсом валют. Верхнюю часть каждой страницы MMS-сообщения занимает баннер с рекламой товаров и услуг;
- USSD (*Unstructured Supplementary Service Data*) – стандартный сервис в сетях GSM (глобальный цифровой стандарт для мобильной сотовой связи), позволяющий организовать интерактивное взаимодействие между абонентом сети и сервисным приложением в режиме передачи

коротких сообщений. USSD-сервис в основном предназначен для обмена сообщениями между абонентом и дополнительными сервисами, в простейшем случае, службой автоинформатора расчётного счета, тогда как SMS в основном служит для обмена короткими сообщениями между абонентами. Этот канал так же может использоваться для размещения рекламы – в списке данного меню может быть пункт меню бренда, которое составляется с учётом задач рекламной компании бренда. Например, в меню могут быть внесены такие пункты, как условия проводимой брендом промо-акции, брендированный контент, купоны на скидку и т.д.;

- WAP (*Wireless Application Protocol* – протокол беспроводного доступа) – это средство получения доступа к ресурсам интернет посредством только мобильного телефона. По сути, это технический стандарт, описывающий способ, с помощью которого информация из интернета передается на дисплей мобильного телефона. Реклама на WAP-сайтах подчиняется тем же законам, что и обычная интернет-реклама.

Мобильная реклама нацелена на людей, которые с наибольшей вероятностью купят рекламируемый товар или воспользуются предлагаемой услугой, так называемую целевую аудиторию.

Нацеливание рекламы (таргетинг) бывает следующих видов:

- географический таргетинг – показ рекламы целевой аудитории, ограниченной некоторым географическим регионом, выбранным рекламодателем;
- социально-демографический таргетинг – по возрасту, полу, доходу, должности и т.д.;
- таргетинг по времени показа (утро или вечер, будни или выходные);
- таргетинг по интересам (контекстная реклама). Показ рекламы в соответствии с некоторым контекстом.

Как видно, нацеленная мобильная реклама – это развитие ставшей уже привычной для интернет пользователей нацеленной интернет рекламы. Считается, что история возникновения контекстной интернет-рекламы начинается с 1997 г., когда Билл Гросс, основатель молодой компании Idealab, придумал продавать ссылки, показываемые одновременно с результатами запросов. Сегодня услуги по контекстной рекламе предлагают тысячи различных сайтов.

Таким образом, на рынке мобильной рекламы происходит взаимодействие сотовых операторов и сервисов контекстной рекламы. У первых есть информация, позволяющая выбрать оптимальное решение по показу рекламы, у вторых – опыт.

Преимущество сотрудничества рекламодателя с сотовым оператором – возможность четкой сегментации целевой аудитории. Количество собираемой и хранимой сотовым оператором информации огромно (пол, возраст, прописка, тарифный план, среднемесячные расходы на связь, модель телефона, частота и продолжительность зарубежных поездок, уровень дисциплинированности при оплате счетов и т.д.).

Очевидно, сотовый оператор не может продавать информацию об абоненте рекламодателю. Вместо этого, оператор может предложить свои услуги по нацеливанию рекламы – в частности свою собственную систему показа контекстной рекламы, которая будет использовать известную об абоненте информацию для выдачи рекламного сообщения, соответствующего интересам абонента.

В нашей стране в настоящее время (начиная с 2007–2008 гг.) организуются стратегические партнёрства между операторами сотовой связи и агентствами мобильной рекламы. Например, у «большой тройки» (Российские лидеры сотовой связи – МТС, Билайн, Мегафон) существуют следующие партнёрства: Билайн+BrandMobile, Мегафон+CustomLine и МТС+NMM (IMHO VI).

Данная статья посвящена разработке системы показа мобильной контекстной рекламы («BannerEngine»), которая использует известную об абоненте информацию (собранную сотовым оператором) для выдачи рекламного сообщения, соответствующего интересам абонента.

**«BannerEngine»** – система показа мобильной контекстной рекламы, используемая сотовым оператором. «BannerEngine» реализует следующие основные функции:

- хранение и управление баннерами (рекламными сообщениями) в системе;
- хранение и управление сервисами (сервис – именованный информационный канал) в системе;
- хранение и управление регионами (регион – набор масок MSISDN абонентов, соответствующий географическому региону) в системе;
- выдачу баннеров по запросу от сторонних приложений с учётом ограничений, наложенных на выдачу баннеров;

- ✦ учёт количества показов баннеров;
- ✦ формирование статистики показов баннеров.

В рамках данной статьи интересующим нас ограничением, наложенным на выдачу баннеров, является т.н. ограничение *«рекламные группы»*. Баннер может быть привязан к определенным рекламным группам. Привязка производится путём назначения рекламодателем баннеру списка ключевых слов. Ключевое слово (КС) – строковое представление одной из характеристик абонента или одной из тем, интересных абоненту. Например, КС «MALE», «FEMALE» – характеристики абонента. КС «CARS», «FOOTBALL», «TRAVEL» – интересы абонента. Данная возможность позволяет привязывать баннеры к целевой аудитории, на которую они рассчитаны.

Система использует информацию о пользователе сотовой сети для показа ему нацеленного рекламного сообщения при включенном ограничении «рекламные группы».

Механизм выдачи баннера с учётом рекламных групп организован в виде «черного ящика», которому на вход подаются КС конкретного абонента (список его интересов), на выходе получается список подошедших баннеров по рекламным группам. Текущий критерий соответствия – КС баннера должны оказаться подмножеством КС абонента.

Если предположить, что 1 500 000 абонентов (для сотового оператора по региону) получают в день (в течение 12 часов) 4 SMS/MMS с рекламным сообщением – алгоритм выдачи баннера должен вернуть результат $(1\,500\,000 \cdot 4)/(12 \cdot 60 \cdot 60) = 139$ раз за секунду. Однако текущая реализация алгоритма выдачи подходящего баннера позволяет выдавать только 37 решений в секунду.

Таким образом, встает задача разработать оптимальный алгоритм выбора нужного баннера с минимальными временными затратами.

Главное требование к алгоритму – высокая скорость выдачи баннеров.

Решение зависит от того, как определяется соответствие баннера абоненту. В рамках данной статьи мы рассмотрим три способа определения соответствия:

1. Наиболее подходящими считаются баннеры, имеющие максимальное количество совпавших КС абонента и баннера. Например, для абонента [$КС_1$, $КС_2$] из баннеров [$КС_1$, $КС_2$, $КС_3$], [$КС_1$, $КС_2$] и [$КС_1$] будут выбраны [$КС_1$, $КС_2$, КС3] и [$КС_1$, $КС_2$] – по два совпавших КС.

2. Подошедшими считаются все баннеры, КС которых являются подмножеством КС абонента. Например, для абонента [KC1,KC2] из баннеров [KC1,KC2,KC3], [KC1,KC2] и [KC1] будут выбраны [KC1] и [KC1,KC2].

3. Подошедшими считаются все баннеры, КС которых в точности совпадают с КС абонента.

Для каждого способа определения соответствия мы проделаем следующее:

1. Рассмотрим адекватность. Т.е. насколько выбранные по такому критерию баннеры будут соответствовать интересам абонента.

2. Рассмотрим алгоритмы выбора баннеров согласно этому критерию и проанализируем границы применения этих решений.

3. Для критериев из п.п. 1, 2 – рассмотрим для выбранных алгоритмов возможность учёта весов КС абонента. Вес КС абонента может определяться, например, количеством откликов абонентом на баннеры, содержащие это КС. В случае если заданы веса КС абонента, соответствие по п.1 будем определять суммарным весом совпавших КС абонента и баннера. В случае соответствия по п.2 наиболее подходящими будем считать баннеры-подмножества с максимальным суммарным весом.

Исходя из полученных результатов по каждому критерию, мы определим наиболее компромиссный вариант между критерием соответствия, скоростью выбора и возможностью добавления учета весов КС (учёт весов КС позволит более качественно нацеливать баннеры).

**Экспериментальная часть**

Для генерации искусственных данных мы руководствовались следующими ограничениями:

- ✦ количество баннеров – от 100 000 до 300 000;
- ✦ количество уникальных КС – ~125;
- ✦ количество КС на абонента/баннер – от 1 до 50;
- ✦ желаемое время поиска – от 2 до 6 мс (6 мс соответствуют выдаче $1000/6 = 166$ результатов в секунду).

В наших искусственных данных КС равномерно распределены по всем объектам (т.е. каждое КС равновероятно).

Алгоритмы тестировались при следующих характеристиках:

- ✦ Intel Celeron CPU 2.40 GHz 0.98 GB of RAM;
- ✦ Java 1.6 в режиме –server -Xms512m -Xmx512m

### Решение задачи выбора баннера по максимальному количеству совпавших ключевых слов

Эту задачу мы рассмотрим наиболее подробно, так как из неё можно получить решение остальных трёх задач.

**1. Адекватность**

На первый взгляд это достаточно адекватный критерий. Более «точные» баннеры предпочитаются более «общим»: для абонента [$KC_1$, $KC_2$, $KC_3$] из баннеров [$KC_1$, $KC_2$, $KC_3$], [$KC_1$, $KC_2$] и [$KC_1$] будет выбран [$KC_1$, $KC_2$, $KC_3$]. Но рассмотрим следующий пример: абонент = {MALE, $KC_1$, ... $KC_N$}, а среди баннеров мы имеем $b$ = {FEMALE, $KC_1$, ... $KC_N$}. Несмотря на то, что в описании баннера присутствуют почти все интересы абонента, решение, полученное поиском максимального пересечения, не выглядит правильным. Так, например, для абонента {MALE, SPORT, SWIMMING} не исключена возможность выбрать баннер {FEMALE, SPORT, SWIMMING}, предлагающий женские купальники.

**2. Алгоритмы**

Задача выбора баннера поиском максимального пересечения множеств КС баннера и абонента является одной из задач поиска ближайшего соседа. Эта задача не имеет универсального решения. В каждом конкретном случае необходимо учитывать специфику данных и, исходя из этого, выбрать модель, которой будут представлены исходные данные, и определить функцию похожести или расстояния.

Для решения данной задачи наиболее подходит модель векторного пространства с функцией близости – скалярным произведением. Каждый объект (абонент, баннер) представляется в виде вектора размера N_KW, где N_KW – количество уникальных КС. В этом векторе выставляются K единиц следующим образом (K – количество КС объекта): в $i$-ой позиции выставляется единица, если $i$-ое слово встречается в объекте, иначе 0.

Следует отметить, что наша задача подходит под «проклятье размерности» [Koppen 2000: 4]. Это понятие обозначает явление, что все точные алгоритмы поиска ближайшего соседа эффективны по сравнению с «хорошим» последовательным поиском только для небольших размерностей (~ < 10) [Koppen 2000: 2]. В нашей же задаче объект (абонент) может описываться 50-ю признаками.

Были протестированы следующие алгоритмы:

- ✦ обратный индекс [Бартунов 2007: 7, Ilyinsky 2002: 3]. Обратный индекс –множество пар {$KC_i \leftrightarrow$ постинг-лист$_i$}, где постинг-лист$_i$ – список баннеров, содержащих $KC_i$. Для каждого КС абонента выбирается соответствующий постинг-лист, заводится счётчик для каждого баннера, и выполняется проход по всем листам с увеличением счётчика для каждого повстречавшегося баннера на единицу. Затем выбираются баннеры с максимальным значением счётчика;
- ✦ последовательный алгоритм в рамках векторной модели, базирующийся на битовых операциях. Векторы представляются в виде набора 32-битовых чисел, над которыми выполняется операция битовое «И». Заранее строится таблица соответствия «число ↔ количество выставленных битов». Очевидно, количество выставленных битов в результате операции битового «И» будет равно размеру пересечения. Таким образом, за одну операцию сравниваются 32 признака;
- ✦ метод разбиения пространства [Yianilos 1993: 6] с использованием BK-дерева [Burkhard 1973: 1] – структуры данных, основанной на неравенстве треугольника, адаптированной к высоким размерностям (не дает при поиске экспоненциальный рост времени от количества признаков). В качестве функции расстояния между двумя объектами (баннерами $b_1$, $b_2$) по $k$ КС использовалась $d$ ($b_1$, $b_2$) = $k$ – (количество совпавших КС у $b1$ и $b2$).

В *табл. 1* представлены результаты работы обратного индекса (II), последовательного алгоритма на битовых операциях (VM+SP(bit impl)) и BK-дерева.

Как видно из табл. 1, метод обратного индекса хорошо отражает проклятие размерности. Чем больше количество признаков, тем сильнее выигрывает последовательный алгоритм. Метод разбиения пространства демонстрирует линейную зависимость от количества признаков, однако показывает очень большое абсолютное значение времени поиска, что может быть связанно с недостаточно эффективной реализацией.

*Таблица 1*

**Время поиска максимального пересечения при использовании алгоритмов II, VM+SP (bit impl) и BK-tree в зависимости от количества КС на объект и баннеров. Количество уникальных КС равно 125**

| Кол-во КС на объект | Кол-во баннеров | II, ms | VM+SP(bit impl), ms | BK-tree, ms |
|---|---|---|---|---|
| до 10 | 100 000 | 2,18 | 6,22 | 23,08 |
| до 20 | 100 000 | 3,25 | 6,36 | 51,54 |
| до 30 | 100 000 | 5,12 | 6,35 | 78,15 |
| до 40 | 100 000 | 7,38 | 6,22 | 96,43 |
| до 50 | 100 000 | 9,70 | 6,09 | 120,49 |
| до 10 | 200 000 | 4,66 | 11,95 | 62,32 |
| до 20 | 200 000 | 11,60 | 12,05 | 117 |
| до 30 | 200 000 | 22,75 | 11,48 | 184,71 |
| до 40 | 200 000 | 31,96 | 11,63 | 240,23 |
| до 50 | 200 000 | 53,38 | 11,46 | 283,19 |
| до 10 | 300 000 | 8,44 | 18,45 | 93,87 |
| до 20 | 300 000 | 25,08 | 18,21 | 176,85 |
| до 30 | 300 000 | 51,69 | 16,82 | 259,22 |
| до 40 | 300 000 | 81,88 | 17,11 | 352,85 |
| до 50 | 300 000 | 123,14 | 17,08 | 442,46 |

В *табл. 2* результаты работы метода разбиения пространства оцениваются по количеству обойдённых при поиске узлов построенного дерева – т.е. по количеству рассмотренных баннеров.

*Таблица 2*

**Количество рассмотренных баннеров при поиске в BK-дереве в зависимости от количества КС на объект и баннеров. Количество уникальных КС равно 125**

| Кол-во КС на объект | Кол-во баннеров | Рассмотрено баннеров при поиске в BK-дереве |
|---|---|---|
| до 10 | 100000 | 32185 |
| до 20 | 100000 | 43083 |
| до 30 | 100000 | 37409 |
| до 40 | 100000 | 52829 |
| до 50 | 100000 | 43583 |
| до 10 | 200000 | 64384 |
| до 20 | 200000 | 89104 |
| до 30 | 200000 | 125885 |
| до 40 | 200000 | 61015 |
| до 50 | 200000 | 194430 |
| до 10 | 300000 | 192036 |
| до 20 | 300000 | 168947 |
| до 30 | 300000 | 237325 |
| до 40 | 300000 | 253814 |
| до 50 | 300000 | 184682 |

Алгоритм в рамках векторной модели с использованием битовых масок применим при небольшом количестве уникальных КС. В *табл. 1* приведены значения для количества уникальных КС = 125. Т.е. каждый объект представляется 125/32 + 1 = 4-мя 32-битовыми числами. При количестве КС до 32, каждый объект будет представлен одним числом. При увеличении количества уникальных КС скорость работы алгоритма будет ухудшаться.

Обратный индекс применим при условии, что есть большая база уникальных КС (не менее 100) и количество КС на баннер не превышает 10. В этом случае постинг-листы будут достаточно короткими, и их обработка займёт приемлемое время.

Метод разбиения пространства представляет тем больший интерес, чем сильнее данные (баннеры) разбиты на кластеры. Однако, чем меньше абонент будет иметь интересов, тем больше вероятность обойти все дерево полностью.

**3. Возможность добавления весов**

В последовательном «битовом» алгоритме и алгоритме обратного индекса есть возможность добавить работу с весами КС без потери производительности.

Для обратного индекса учёт весов означает увеличение счётчика не на единицу, а на вес КС абонента, для которого рассматривается текущий постинг-лист.

Для «битовой реализации» для учёта весов строится таблица по абоненту таким образом, чтобы по результату битового «И» можно было обратиться сразу за взвешенным значением текущего 32-битового кусочка.

При использовании метода разбиения пространства использование КС невозможно, т.к. метрическое дерево строится на основании заданной функции расстояния, которая точно в таком же виде должна использоваться и при поиске в дереве. Использование же весов КС абонента означает невозможность использовать единую функцию расстояния для всех баннеров/абонентов, т.к. каждый объект-абонент в этом случае определяет новую функцию расстояния.

## Решение задачи выбора баннера поиском «баннера-подмножества»

**1. Адекватность**

Данный критерий исключает пример, приведенный выше, в котором выбирается баннер с «лишними» КС. Выбранный баннер будет соответствовать интересам абонента и не предложит лишнего.

**2. Алгоритмы**

Решение данной задачи можно получить модификацией первой (проверять количество совпавших КС на равенство количеству КС баннера), однако задача на подмножества имеет следующие преимущества:

- ✦ если задан абонент с $k$ уникальными КС, то нет необходимости рассматривать баннеры с $\geq k + 1$ уникальными КС;
- ✦ если в баннере обнаружилось хотя бы одно «лишнее» КС, то нет смысла рассматривать этот баннер далее.

Для данного критерия мы сравнили три реализации:

1. Последовательный алгоритм, базирующийся на битовых операциях.
2. Последовательный алгоритм, базирующийся на простых числах [Clarke 1997: 2].
3. Последовательный алгоритм, базирующийся на сравнении сортированных идентификаторов КС.

Алгоритм, базирующийся на битовых операциях, был получен модификацией «битового» алгоритма, вычисляющего максимальное пересечение КС абонента и баннера. Баннеры и абоненты так же представляются в виде набора 32-битовых чисел, над которыми выполняется операция битовое «И». Если $a[i]$, $b[i]$ – 32битовые кусочки абонента и баннера, и $a[i]$ «И» $b[i] != b[i]$, – то сразу ясно, что баннер не подходит абоненту по данному критерию, прекращаем его рассмотрение и переходим к следующему.

Алгоритм, базирующийся на простых числах, заключается в следующем: каждому уникальному КС ставится в соответствие простое число. И каждый баннер представляется в виде произведения своих простых чисел. Когда на вход поступает абонент, он точно так же представляется в виде произведения. Если «баннер делит абонента» без остатка, то баннер – подмножество абонента.

Алгоритм, базирующийся на сравнении сортированных идентификаторов КС, требует, чтобы идентификаторы (уникальные номера) КС баннеров/абонента были отсортированы. Для каждого идентификатора КС баннера бинарным поиском ищется такой же идентификатор среди идентификаторов КС абонента. Если поиск дал отрицательный результат, сразу переходим к следующему баннеру.

В *табл. 3* представлены результаты для всех трех реализаций: сравнение сортированных идентификаторов КС (Sorted IDs), с использованием простых чисел (Prime based), с использованием битовых масок (Bit mask), с использованием битовых масок + сортировка баннеров по количеству уникальных КС баннера (Bit mask (sorted)).

*Таблица 3*

**Время поиска «баннера-подмножества» при использовании алгоритмов Sorted IDs, Prime Number, Bit mask и Bit mask (sorted) в зависимости от количества КС на абонента/баннер и баннеров. Количество уникальных КС равно 125**

| Кол-во КС на объект | Кол-во баннеров | Sorted IDs, ms | Prime Number, ms | Bit mask, ms | Bit mask (sorted), ms |
|---|---|---|---|---|---|
| до 10 | 100000 | 17,54 | 10,52 | 2,32 | 1,55 |
| до 20 | 100000 | 22,35 | 9,85 | 2,18 | 1,45 |
| до 30 | 100000 | 23,03 | 11,03 | 2,26 | 1,51 |
| до 40 | 100000 | 23,95 | 11,77 | 2,24 | 1,49 |
| до 50 | 100000 | 24,23 | 12,47 | 2,32 | 1,55 |
| до 10 | 200000 | 37,91 | 25,47 | 4,64 | 3,09 |
| до 20 | 200000 | 41,88 | 22,77 | 4,77 | 3,18 |
| до 30 | 200000 | 45,72 | 27,88 | 3,87 | 2,58 |
| до 40 | 200000 | 48,87 | 25,51 | 4,21 | 2,81 |
| до 50 | 200000 | 2,07 | 24,12 | 3,83 | 2,55 |
| до 10 | 300000 | 54,15 | 23,97 | 8,24 | 5,49 |
| до 20 | 300000 | 57,84 | 36,58 | 6,38 | 4,25 |
| до 30 | 300000 | 66,48 | 41,09 | 5,77 | 3,85 |
| до 40 | 300000 | 64,79 | 34,25 | 7,17 | 4,78 |
| до 50 | 300000 | 76,75 | 33,64 | 5,65 | 3,76 |

Метод с использованием битовых масок показал большую эффективность по сравнению с остальными, поэтому именно его мы улучшили сортировкой баннеров по количеству уникальных КС, что дало улучшение еще в 1,5 раза.

Как и в предыдущем случае, выбранный алгоритм зависит от количества уникальных КС. Но если для поиска максимального пересечения необходимо было обойти все 32-битовые кусочки, то в данном случае мы прекращаем рассмотрение текущего баннера, как только не выполнилось равенство ($a[i]$ & $b[i] = b[i]$). Благодаря этому в сочетании с сортировкой баннеров по количеству уникальных КС становится приемлемым использование довольно большого количества уникальных КС.

**3. Возможность добавления весов**

Метод битовых масок модифицируем для учета весов КС абонента точно так же как и в случае поиска максимального пересечения.

### Решение задачи выбора баннера по точному совпадению ключевых слов

**1. Адекватность**

Выбранный баннер 100% отражает интересы абонента, однако данный критерий очень сильно ограничивает набор подходящих баннеров и есть достаточно большая вероятность вернуть пустой результат для абонента.

**2. Алгоритмы**

Здесь применимы все решения, описанные выше, но также на этот случай есть специфичное очень «быстрое решение» – trie-дерево [Shang 1995: 5].

*Таблица 4*

**Время поиска точного совпадения в trie-дереве в зависимости от количества КС на абонента/баннер и баннеров. Количество уникальных КС равно 125**

| Кол-во КС на объект | Кол-во баннеров | Trie-дерево, ms |
|---|---|---|
| до 10 | 100000 | 0,01 |
| до 20 | 100000 | 0,01 |
| до 30 | 100000 | 0,02 |
| до 40 | 100000 | 0,04 |
| до 50 | 100000 | 0,17 |
| до 10 | 200000 | 0,01 |
| до 20 | 200000 | 0,16 |
| до 30 | 200000 | 0,03 |
| до 40 | 200000 | 0,03 |
| до 50 | 200000 | 0,11 |
| до 10 | 300000 | 0,01 |
| до 20 | 300000 | 0,03 |
| до 30 | 300000 | 0,01 |
| до 40 | 300000 | 0,04 |
| до 50 | 300000 | 0,12 |

Как видно из таблицы, поиск в таком дереве происходит за доли миллисекунды.

### Выводы

Были проанализированы критерии соответствия баннера интересам абонента и методы выбора баннеров по этим критериям. При текущих параметрах системы (критерий соответствия «баннер-подмножество», 125 КС, 100 000–300 000 баннеров, до 25 КС на абонента/баннер) найдено удовлетворительное решение, позволяющее выдавать 1000/~2.62 = = ~384 решения в секунду и учитывать на лету веса КС абонента. Были даны рекомендации по использованию других методов при изменении соотношения параметров и критерия соответствия. ■

**Литература.**


1. Burkhard W. A., Keller R. M. Some approaches to best-match file searching // Communication of the ACM.-1973.-V. 16. – P. 230–236.
2. Clarke I. Testing Subsets Using Prime Numbers // Ian Clarke's Homepage/Subsets [Электронный ресурс]. – Электрон. дан. – 1997. – Режим доступа: http://www.sanity.uklinux.net/subsets.html.
3. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of web documents with the use of inverted index // Proceedings of the eleventh Int. World Wide Web Conference (WWW). – 2002.
4. Koppen M. The curse of dimensionality // Proceedings of the fifth Online Conference on Soft Computing in Industrial Applications (WSC5). – 2000.
5. Shang H., Merrettal T. H. Tries for Approximate String Matching // IEEE Transactions on Knowledge and Data Engineering. – 1996. – V. 8.- № 4. – P. 540–547.
6. Yianilos Peter N. Data structures and algorithms for nearest neighbor search in general metric spaces // Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms. – 1993. – 25–27 Jan. – P. 311–321.
7. Бартунов О. С., Сигаев Ф. Г. Специализированные типы данных для цифровых библиотек // 9-ая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL). – 2007.
8. Крапивинский А. Кому рекламу на «сотовый»? // «Медиа-Онлайн»/Аналитика/Медиабизнес/ [Электронный ресурс]. – Электрон. дан. – 2008. – 18 авг. – Режим доступа: http://www.media-online.ru/index.php3?id=290255.