Санкт-Петербургский государственный университет

Математико-механический факультет

Кафедра системного программирования

Лапин Сергей Константинович

Анализ и прогнозирование пользовательской активности

Выпускная работа бакалавра

Допущен к защите

Зав. кафедрой:

д.ф.-м.н., проф. А.Н. Терехов

Научный руководитель:

к.ф.-м.н., доцент Д.Ю. Бугайченко

Рецензент:

к.ф.-м.н., доцент Н.Г. Графеева

Санкт-Петербург

2012

Saint-Petersburg State University

Faculty of Mathematics and Mechanics

System engineering department

Analysis and prediction of users activity

Graduate paper

Lapin Sergey

gr. 461

Scientific advisor	……	D.Y. Bugaychenko
	/ signature /	
Reviewer	……	N.G.Grafeeva
	/ signature /	
“Admitted to proof”	……	A.N. Terehov
Head of the chair,	/ signature /	

Saint-Petersburg

2011

Содержание

Оглавление

Введение 4

Цель исследования 5

Постановка задачи 5

Обзор существующих методов выбора оптимального времени публикации. 6

Обзор существующих методов для анализа и прогнозирования временных рядов 8

Реализация поставленной задачи 16

Общая архитектура системы 16

Особенности реализованной системы 17

Сбор данных 17

Подготовка данных 19

Разделение ряда 22

Предсказание 23

Пост обработка 24

Нахождение оптимального времени публикации 24

Валидация осуществлённого предсказания 26

Полученные Результаты 28

Список литературы 29

Введение

В настоящий момент наиболее популярным инструментом общения в интернете является социальные сети, такие как ВКонтакте и Facebook. В основном информация массового характера в них распространяется по принципу подписки или «френдования». Каждый пользователь является и поставщиком и потребителем информации. Обмен информацией в широких кругах происходит благодаря механизму новостной ленты.

Мы предполагаем, что какой-либо контент, находящийся в топе новостной ленты, будет вероятнее всего замечен пользователем.

Для того чтобы попасть в топ нам необходимо опубликовать информацию приблизительно перед началом сессии пользователя. В иных случаях мы либо проиграем более свежей информации, либо информация вообще не будет доставлена.

В сущности, нет достоверной информации, позволяющей гарантированно утверждать о наличии активности пользователя в то или иное время. Тем не менее, поведение человека до некоторой степени предсказуемо. Существует суточная и недельная сезонность в активности пользователя, которая зависит от его образа жизни. Таким образом, пользовательскую активность можно представить композицией трех составляющих: сезонной, шумовой и трендом. На основе периодической составляющей и тренда можно предсказать будущую активность.

Можно повысить вероятность распространения информации за счет предсказания активности целевой группы пользователей.

В рамках данной исследовательской работы ставится задача об эффективном распространении информации. Более конкретно нас интересует оптимальное время публикации контента. Также мы абстрагируемся от сущности контента, нас интересует только время публикации.

Цель исследования

В рамках данной квалификационной работы ставятся следующие цели исследования:

- Нахождение оптимального времени публикации
- Выделение сезонности в поведении пользователей
- Осуществление предсказания

Постановка задачи

Задачей работы является разработка подхода к построению системы анализа, предсказания и кластеризации активности пользователей на основе данных о прошлой активности. Реализация системы использующей это решение, анализ эффективности и её тестирование.

Задача разделяется на следующие подзадачи:

- Собрать необходимые для анализа данные
- Подготовить данные для анализа данных
- Произвести анализ и смоделировать предсказание

Обзор существующих методов выбора оптимального времени публикации.

В области проблемы оптимизации времени публикации существует множество научных работ рассматривающих стратегии публикации основанных на теории коммуникации и социологии. (ZOLLMAN)

Для анализа пользовательской активности широко используется сервис Google Analitics с помощью которого, можно получить статистику посещаемости сайта. Сервис позволяет определить следующие значения показателей с различными временными интервалами:

- Количество уникальных посетителей
- Среднюю длительность просмотра страницы
- Количество просмотров страницы
- Среднюю длительность пребывания на сайте

При этом пользуясь Google Analitics нет возможности осуществить предсказание по целевой группе пользователей.

Особенно хорошо изучен в смысле нахождения оптимального времени публикации twitter. Для него есть широкий инструментарий решающий данную задачу [(HOW TO Post Your Tweets At Optimal Times) (HOW TO Post Your Tweets At Optimal Times).

Периодически публикуется статистика использования большинства популярных сервисов, таких как Facebook (When is the best time to post on Facebook? Here’s the Answer…)StackOverflow (The Best Time to Ask a Stack Overflow Question?), Tumblr (Best Time To Post On Facebook, Twitter, Tumblr)

Тем не менее, осуществлять публикацию в «прайм тайм» использования сервиса не эффективно, если у нас есть выбранная целевая аудитория. В случае если предсказание для групп укажет отличное от «прайм тайм» время конкуренция между постами будет ниже, а вероятность того, что пост будет замечен выше. Кроме того существенен анализ определённой активности, а не простой факт присутствия пользователя.

Например, если мы хотим получить качественный ответ на форуме StackOverflow в определённой области, то наша целевая группа пользователей - люди указавшие в интересах тему нашего вопроса, а оптимальное время публикации вопроса – время спрогнозированное на основе времени их предыдущих ответов.

Таким образом, если мы располагаем временным рядом активности определённого пользователя можно, проанализировав его осуществить индивидуальное предсказание на различных интервалах времени. Используя такие предсказания, улучшим оптимальное время публикации, для выбранных групп пользователей.

Обзор существующих методов для анализа и прогнозирования временных рядов

В области осуществления анализа и предсказаний над временными рядами разработано множество алгоритмов и моделей.

Выделение тренда

Один из простейших способов выделения тренда временных рядов - модель скользящего среднего.

Простое скользящее среднее

Или (simple moving average) численно равно среднему арифметическому значений исходной функции за установленный период и вычисляется по формуле:

$$\mathrm{SMA_t} = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-1} = \frac{p_t+p_{t-1}+\cdots+p_{t-i}+\cdots+p_{t-n}+p_{t-n+1}}{n}$$

Взвешенные скользящие средние

Является модифицированной версией *SMA*. Их применяют при построении скользящей средней, когда некоторые значение исходной функции целесообразно сделать более значимыми. Например, если предполагается, что внутри интервала сглаживания имеет место нелинейная тенденция, или в случае временных рядов, последние — более актуальные данные могут быть весомее предыдущих.

$$\mathrm{WMA_t} = \frac{2}{n*(n+1)}\sum_{i=0}^{n-1}(n-i)p_{t-1} = \frac{n*p_t+(n-1)p_{t-1}+\cdots+(n-i)p_{t-i}+\cdots+2*p_{t-n}+1*p_{t-n+1}}{n+(n-1)+\cdots+(n-i)+\cdots+2+1},$$

где $\mathrm{WMA_t}$— значение взвешенного скользящего среднего в точке t, n — количество значений исходной функции для расчёта скользящего среднего, p_{t-i} — значение исходной функции в момент времени, отдалённый от текущего на i интервалов.

Экспоненциальное скользящее среднее

Применяют в случае, если важность данных убывает экпонентенциально со временем**.** Это разновидность взвешенной скользящей средней, веса которой убывают экпонентенциально и никогда не равны нулю. Определяется следующей формулой:

$$\mathrm{EMA_t} = \alpha * p_t + (1 - \alpha) * \mathrm{EMA_{t-1}},$$

где $\mathrm{EMA_t}$ — значение экспоненциального скользящего среднего в точке t (последнее значение, в случае временного ряда)

Выделение периодичности (метод Фурье – спектральный анализ)

Классический метод разложения функции на периодики – спектральный анализ Фурье.

Цель спектрального анализа - разложить ряд на функции синусов и косинусов различных частот, для определения тех, появление которых особенно существенно и значимо. Один из возможных способов сделать это - решить задачу линейной множественной регрессии, где зависимая переменная - наблюдаемый временной ряд, а независимые переменные или регрессоры: функции синусов всех возможных (дискретных) частот. Такая модель линейной множественной регрессии может быть записана как

$$x_t = a_0 + \sum_{k=1}^{q} (a_k \cdot \cos(\lambda_k \cdot t) + b_k \sin(\lambda_k \cdot t)), \text{ где}$$

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)dx$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx)\, dx$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx)\, dx$$

Получить коэффициенты ряда можно с помощью быстрого преобразования Фурье (БПФ).

Коэффициенты ряда Фурье “быстро” убывают:

- Степенное убывание коэффициентов Фурье присуще функциям класса $C^{(k)}$

- Экспоненциальное — аналитическим функциям.

Частоты с достаточно малыми коэффициентами можно отбросить как шум.

Декомпозиция

Уже на основе моделей скользящих средних и спектрального анализа можно осуществить декомпозицию ряда, представив его в следующем виде:

$$F(t) = trend(t) + season(t) + noize(t)$$

Тем не менее, такой подход не даст нам убедительные результаты, существуют методы разделяющие ряд более эффективно.

Метод «Гусеница»

«Гусеница» (Д. Л. Данилов) (Н.Э., Метод «Гусеница»-SSA: анализ временных рядов) или Singular spectrum analysis, метод основанный на преобразовании одномерного временного ряда в многомерный ряд и последующего применения к полученному многомерному временному ряду метода главных компонент. Это один из способов понижения размерности, состоящий в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных. Вдоль первой оси нового базиса дисперсия максимальна, вторая ось максимизирует дисперсию при условии ортогональности первой оси, и так далее, последняя ось имеет минимальную дисперсию из всех возможных. Такое преобразование позволяет понижать информацию путем отбрасывания координат, соответствующих направлениям с минимальной дисперсией.

Предполагается, что если нам надо отказаться от одного из базисных векторов, то лучше, если это будет тот вектор, вдоль которого набор входных данных меняется менее значительно.

Описание алгоритма

Шаг 1. (Развертка одномерного ряда в многомерный)

Выберем некоторое число $M < N$, называемое «длиной гусеницы» или окном, и представим первые M значений последовательности f в качестве первой строки матрицы X. В качестве второй строки матрицы берем значения последовательности с x_2 по x_{M+1}.

Последней строкой с номером $k = N - M + 1$ будут последние M элементов последовательности. Эту матрицу, элементы которой равны $x_{ij} = x_i + j - 1$, можно рассматривать как M - мерную выборку объема k или M - мерный временной ряд, которому соответствует M - мерная траектория (ломаная в M - мерном пространстве из $k - 1$ звена.

Далее по обычной схеме (за исключением стандартизации признаков) проводится анализ главных компонент (АГК).

Шаг 2. (Анализ главных компонент: сингулярное разложение выборочной ковариационной матрицы)

Сначала вычисляется матрица $V = \frac{X^T X}{k}$

Следующий шаг, как обычно в АГК, состоит в вычислении собственных чисел и собственных векторов матрицы V, т.е. разложение ее $V = P\Lambda P^T$, где Λ - диагональная матрица, на диагонали которой стоят упорядоченные по убыванию собственные числа, а P - ортогональная матрица собственных векторов матрицы V.

Матрицы Λ и P совместно имеют множество интерпретаций, основанных на АГК. В частности, матрицу P можно рассматривать как матрицу перехода к главным компонентам $XP = Y = (y_1, y_2, \dots, y_M)$

Шаг 3. (Отбор главных компонент)

В силу свойств матрицы P мы можем представить матрицу ряда X как $X = Y\ P^T X$. Таким образом, мы получаем разложение матрицы ряда по ортогональным составляющим (главным компонентам).

В то же время преобразование $y_j\ =\ \boldsymbol{X}\ p_j$ является линейным преобразованием исходного процесса с помощью дискретного оператора свертки, т.е.

$$y_{jj}\ [l]\ =\ \sum_{q=1}^{M} x_{lq}\ p_{jq} = \sum_{q=1}^{M} x_{l+q-1}\ p_{jq}$$

Таким образом, процедура "Гусеница" порождает набор линейных фильтров, настроенных на составляющие исходного процесса. При этом собственные векторы матрицы $\boldsymbol{V}$ выступают в роли переходных функций соответствующих фильтров.

Среди главных компонент можно выделить

1. относящиеся к тренду (медленно меняющиеся) ,
2. периодические,
3. шумовые

К тренду относится первая компонента. Шумовые компоненты - те, которые вносят вклад ниже определённого порового значения. Остальные компоненты периодическими или сезонными.

Шаг 4 . (Восстановление одномерного ряда)

Следующим ключевым элементом метода "Гусеница" является процедура восстановления. Эта процедура основана на разложении $\mathrm{X}\ =\ \mathrm{Y}\ \mathrm{P}^{\mathrm{T}}$. Будем говорить, что восстановление проводится по данному набору главным компонентам, если при применении формулы восстановления $\mathrm{X}\ =\ \mathrm{Y}^{*}\ \mathrm{P}^{\mathrm{T}}$ матрица Y^{*} получена из матрицы Y обнулением всех не входящих в набор главных компонент. Таким образом, мы можем получить интересующее нас приближение матрицы ряда или интерпретируемую часть этой матрицы.

Подбор входных параметров «Гусеницы»

Выбор компонент

Можно использовать два критерия - либо выбирать одно и тоже число компонент (K) для всех прогнозов, либо выбирать число компонент таким образом, чтобы суммарный вклад в общий процесс был не меньше определённого процента (например - 99.99%). Недостаток первого заключается в том, что на разных интервалах можно получать разное разложение по числу значимых компонент и при таком способе можно ухудшить качество аппроксимации и как следствие - качество прогноза. Второй критерий в этом смысле является более надёжным, выбора компонент по данному критерию следует построить полное разложение ряда «гусеницей» получив собственные числа компонент и взять нужное количество компонент исходя из того что, собственное число компоненты равняется её вкладу в разложенном ряде.

Выбор размера окна

В ходе данной работы не получилось автоматизировать выбор окна гусеницы, экспериментально подобрана длина в 1/3 от длины анализируемого ряда.

Предсказание с помощью метода «Гусеница»

Числовой ряд $(f_i)_{i=1}^{N+1}$ называется продолжением ряда $(f_i)_{i=1}^{N}$, если порождаемая им при гусеничной обработке выборка лежит в той же гиперплоскости, что и у исходного ряда.

Пусть у нас есть некоторый набор выбранных главных компонент $i_1, i_2 \dots i_r$Определим

$$w = \begin{pmatrix} v_\sigma^{(i_1)} & v_\sigma^{(i_2)} & \cdots & v_\sigma^{(i_r)} \\ v_{2\sigma}^{(i_1)} & v_{2\sigma}^{(i_2)} & \cdots & v_{2\sigma}^{(i_r)} \\ \vdots & \vdots & \ddots & \vdots \\ v_\tau^{(i_1)} & v_\tau^{(i_2)} & \cdots & v_\tau^{(i_r)} \end{pmatrix} \text{ и } V_* = \begin{pmatrix} v_1^{(i_1)} & v_1^{(i_2)} & \cdots & v_1^{(i_r)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{\sigma-1}^{(i_1)} & v_{\sigma-1}^{(i_2)} & \cdots & v_{\sigma-1}^{(i_r)} \\ v_{\sigma+1}^{(i_1)} & v_{\sigma+1}^{(i_2)} & \cdots & v_{\sigma+1}^{(i_r)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{2\sigma-1}^{(i_1)} & v_{2\sigma-1}^{(i_2)} & \cdots & v_{2\sigma-1}^{(i_r)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{\tau-1}^{(i_1)} & v_{\tau-1}^{(i_2)} & \cdots & v_{\tau-1}^{(i_r)} \end{pmatrix}$$

Также положим

$$Q = \left(f_{N-\sigma+2}^{(1)}, \dots, f_N^{(1)}, f_{N-\sigma+2}^{(2)}, \dots, f_N^{(2)}, \dots, f_{N-\sigma+2}^{(s)}, \dots, f_N^{(s)}\right)^T$$

Тогда прогнозируемые значения системы в точке $N+1$вычисляются по формуле:

$$f_{N+1} = w(V_*^T V_*)^{-1} V_*^T Q$$

Обзор применяемых алгоритмов и технологий

В ходе исследовательской работы применяются следующие алгоритмы:

- Гусеница как базовый алгоритм предсказания
- Интерполяция сплайнами для подготовки ряда

Используемые языки:

Система пишется на двух языках: C# и R и использует их преимущества.

Доступ к функциям языка R в языке C# осуществляется с помощью библиотек R.Net и R-(D)COM.

Сравнение языков C# и R.

	C#	**R**
Достоинства	Широкий инструментарий для промышленной разработки ПО	• Идеально подходит для статистической обработки данных - широкий инструментарий для подобных задач • В частности есть открытый код гусеницы
Недостатки	Язык не предназначен для статистического анализа данных	• Скриптовый • Слабый инструментарий для промышленной разработки

Сравнение библиотек R.Net и R-(D)COM.

	R.Net	***R-(D)COM***
Достоинства	• Легкая интеграция • Удобный интерфейс	• Стабильность • Библиотека дорабатывается и поддерживается
Недостатки	• Прекращает работу в неожиданный момент. • Разработка библиотеки приостановлена	Сложность при установке

Реализация поставленной задачи

Стоит договориться об определениях временных рядов. Временные ряды различаются автором статьи на два типа по формату хранения данных:

Название	IrregularSeries	RegularSeries
Расстояние между отметками	Между отметками временные интервалы произвольной длины	Между отметками временные интервалы постоянной длины
Значение отметки	Наличие отметки гарантирует активность в указанный момент времени	Отметка хранит "вероятность" наличия активности в указанный момент времени
Занимаемая память	Значительно более экономный формат	Избыточный по памяти формат

Общая архитектура системы

Доступ к алгоритму гусеница и визуализации данных реализованных на языке R осуществляется с помощью технологии R.Net или R(D)Com.

Для инкапсуляции всех команд языка R используется шаблон проектирования «фасад». Это позволяет использовать C# как основной язык для разработки. Такая архитектура дает преимущества в гибкости при разработке системы.

В процессе реализации данного механизма исследователь столкнулся с рядом трудностей связанных использованием вышеупомянутых библиотек:

- R.net падает на циклах и вызывает исключения при специфической последовательности команд
- Библиотека RCom сложна при установке и не удобна в использовании

Для разрешения этих проблем автор использует шаблон проектирования «заместитель», реализация которого отлавливает исключения связанные с недоработкой библиотеки R.Net, инициирует новое соединение и дублирует запрос последней команды.

Особенности реализованной системы

- Разделение на модули позволяет обеспечить стабильность и расширяемость всей системы
- Ряды занимают много оперативной памяти. Для корректной работы промежуточные результаты сохраняются в дисковое пространство. Система в течение своей работы порождает много файлов
- Взаимодействие с языком R с помощью библиотеки R.Net в некоторых случаях нарушается, для передачи данных так же используется файловое пространство

Сбор данных

Для качественного предсказания необходим анализ работы системы на достаточно количестве данных. Автор рассматривает два способа сбора данных: импорт, из какой либо существующей системы с предварительной их очисткой от персональной информации пользователей и моделирования потоков активности пользователей.

Собраны данные из следующих источников:

Skype

Была написана программа очищения логом Skype и собрана статистика использования сервиса у порядка 20 пользователей. Данные представляют собой иррегулярный ряд с, где временная метка соответствует времени входящего или исходящего сообщения. Целесообразно отфильтровать данные, оставив только метки времени исходящих сообщений.

StackOverflow

Удобной статистикой использования располагает сервис StackOverflow.

Собраны 200 временных рядов активности пользователей с наиболее высокой репутацией за годичный интервал [18]. Ряды содержат информацию о времени публикации ответов на вопросы других пользователей StackOverflow[19], а так же комментарии к другим постам.

При написании парсера были преодолены сложности связанные с ограничением количества постов пользователя выдаваемых на страницу – максимальное количество записей 20. Синхронными запросами информация об одном пользователе закачивалась в среднем за 40 минут. Такая низкая скорость прежде всего связанна с блокировкой IP на сервере сервиса.

Проблему блокировки IP сайтом не удалось разрешить полностью, однако после реализации асинхронных запросов среднее время закачки уменьшилось до 10 минут на пользователя.

Механизм скачивания данных с StackOverflow

Создаются n асинхронных запросов для загрузки n страниц. Текст скачиваемой страницы хранится в контейнере, в котором содержится также информация:

- Номер обрабатывающего потока в массиве ожидания,
- Bool флаг isLoaded

Поток, либо успешно загружает страницу, тогда isLoaded присваивается true, либо срабатывает WebExeption(Сервер не доступен, страница не успела загрузиться), тогда isLoaded присваивается false.

Затем поток завершает свою работу. Как только все потоки завершили свою работу загружаются следующие n страниц.

После того как все страницы обработаны происходит восстановление потерянных страниц:

1) Находятся номера потерянных страниц

2) Страницы отправляются на загрузку

3) Шаги 1 и 2 повторяются до загрузки всех выбранных страниц.

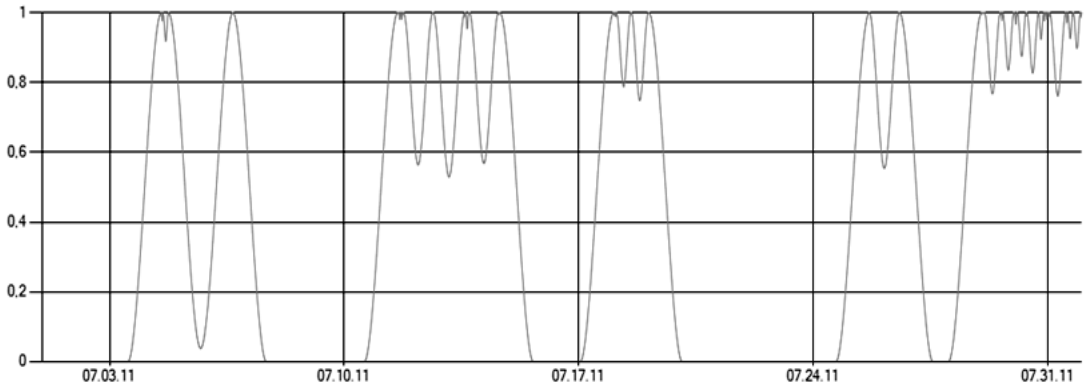
Подготовка данных

Для дальнейшего анализа нам необходимо получить TimeSeries из TimeRecords. Для этого сконструируем сессии активности по тому принципу, что любые записи из TimeRecords будут находиться в одной сессии, если расстояние между ними не превышает определённого значения(в дальнейшем коэффициент угасания или FadeCoef). «Физический» смысл у такого преобразования следующий: из набора временных точек получаем, набор сессий/«несессий».

Если временная точка принадлежит определённой сессии, то в пределах заданной нами временной точности есть другая временная отметка об активности пользователя. Сессии образуют точки отличные от нуля, «несессии» соответственно образуют нулевые точки.

Можно так же представить это другим образом: множество дискретных значений TimeRecords преобразуется в непрерывную функцию. Каждая точка, которой показывает вероятность активности пользователя.

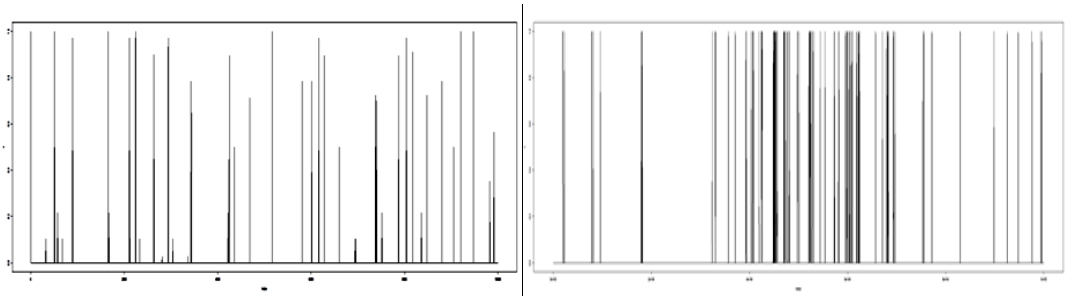
Визуализируем полученный ряд. Для этого построим интерполяционный полином в точках перегиба. Используем для этого сплайны Эрмита обозначив значения производных по всех точках перегиба равными нулю.



На рисунке график активности пользователя Skype в рамках одного месяца

Обработав ряд таким образом, мы теряем информацию связанную с интенсивностью активности пользователя в единицу времени. Близко лежащие точки сливаются в одну и не вносят существенного вклада в использованном представлении данных.

Представление содержит информацию исключительно о взаимном расположении сессий во времени. Это важно для осуществления дальнейшего предсказания, так как ищется сезонность именно во взаиморасположении сессий/«не сессий».

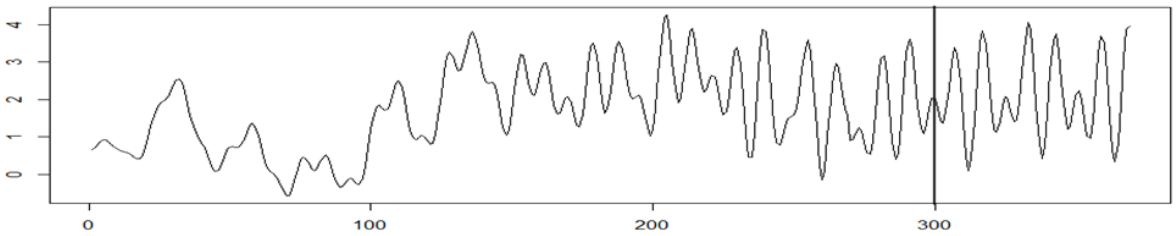


Ряд активности пользователя Skype и Stack Overflow за год со сформированными сессиями (коэффициент угасания = 10 сек)

Предсказания малой точности

В случае если необходимо осуществить предсказание в масштабе недели или месяца можно в качестве анализируемого ряда разумно взять скользящее суммирование от ряда, смоделированного с относительно высоким коэффициентом угасания. Такой ряд хорошо прогнозируется классической «Гусеницей» из-за малого количества стабильных участков.

Реализованное предсказание скользящего суммирования:



Интервал активности 2 года

Проблема стабильных участков

Исходный ряд сессий проблематично предсказывать гуссеницей так как в рядах часто встречаются участки стабильных значений, когда активность пользователя не меняется в течении времени большего, чем интервал измерений. Применив метод «Гусеница»-SSA к таким рядам, прогнозов со стабильными участками получено не было. Как правило, ряд прогноза незначительно колебался в местах, где значения, исходя из вида прогнозируемого ряда, должны быть постоянны. Имея стабильные участки в исходных данных, разумно ожидать их и в результате, чего при использовании классического метода не наблюдается.

Склейка ряда

Был рассмотрен метод «склейки» временного ряда обходящий проблему стабильных участков. (А.А)

Описание алгоритма «Склейки»

1. $F = (f_0, \ldots, f_n)$ - исходный ряд.

2. $G = (g_0, \ldots, g_m)$ - склейка стабильных участков F,

$g_{i-1} \neq g_i \neq g_i + 1 \; \forall i \in 1 \ldots m - 1.$

Каждый элемент G представляет собой стабильный участок в F.

3. $L = (l_0, \ldots, l_m)$ - длины стабильных участков G. $\sum_{i=0}^{m} l_i = n = |F|$.

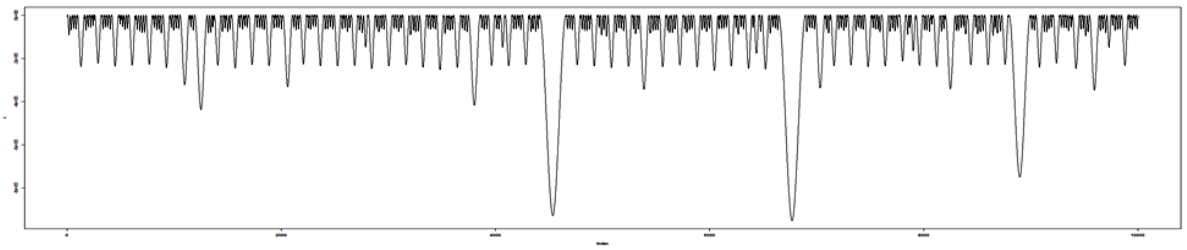
4. G' - прогноз G, L' - прогноз L, полученные методом «Гусеница»-SSA.

Тогда, прогноз исходного ряда F, находится как G'_i, повторяющееся L_i раз. Находя прогноз стабильных участков в новом временном ряду, метод обеспечивает их наличие и в результирующем графе.

Однако «склеивая» ряды мы теряем информацию о нелинейной сезонности во взаиморасположении сессий и к тому же сильно укорачиваем ряд. В некоторых случаях ряд укорачиваться фатально для возможного предсказания.

Ряд взвешенный по длине сессий

Интуитивно понятно, что последовательности сессий разной длины будут образовывать разные периодические составляющие ряда. Можно использовать длительные промежутки сессий/«несессий» для нового представления, в рядах которого максимальные абсолютные значения – равны длине исходного промежутка. Остальные значения ряда получены интерполяцией сплайнами Эрмита.

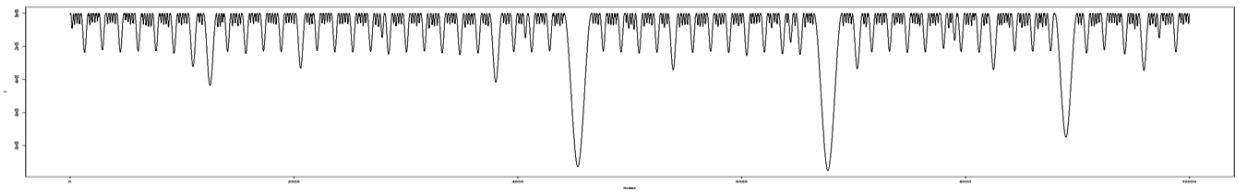


Взвешенный ряд по длине сессий/«несессий» пользователя Skype

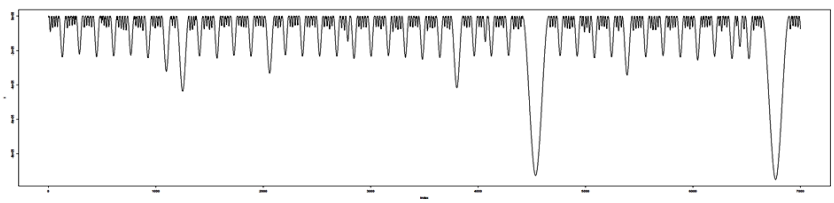
Такой ряд так же как и после склейки не будет содержать стабильных участков и на нём удобно использовать метод классической «Гуссеницы»-SSA.

Разделение ряда

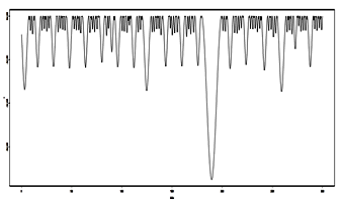
Для оценки качества предсказания проводится разделение анализируемого ряда на две части. По первой осуществляется предсказание, а вторая сравнивается с полученным предсказанием.



Взвешенный ряд по длине «сессий-несессий» за год



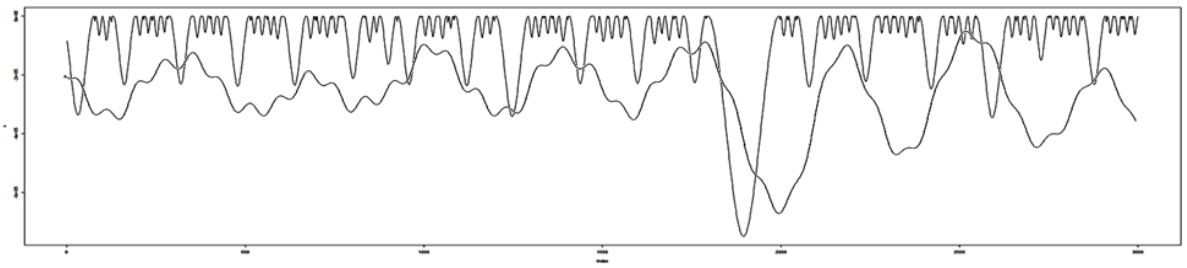
Первый ряд для предсказания



Второй ряд для проверки (один квартал)

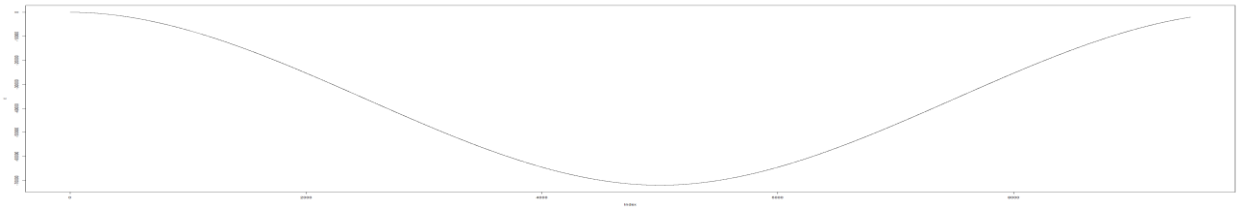
Предсказание

По первой части ряда осуществим предсказание. И сравним с реальными значениями.



Реальное продолжение ряда

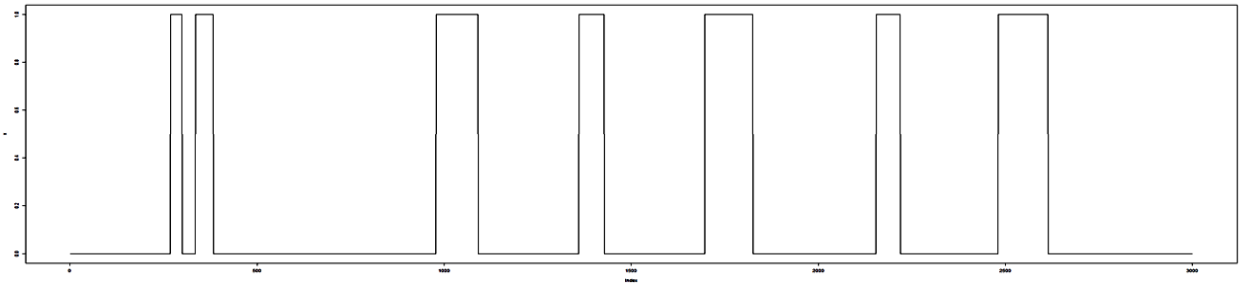
Следует также учесть то, что нам часто будут попадаться пустые участки. Например, при прогнозе малых интервалов времени, так как за основу предсказания берется пропорциональное количество точек и велика вероятность, что на небольших интервалах пользователь не совершал никакой активности. В таком случае мы получим следующее представление:



На таких значениях гусеница дает ложный положительный прогноз, так как тренд на последней половине анализируемых значений растёт. Таким рядам соответствует тривиальный прогноз - отсутствие активности пользователя.

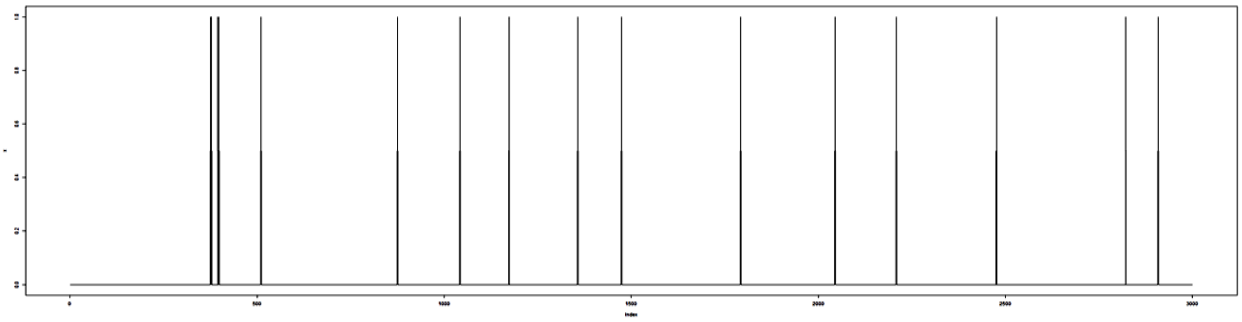
Пост обработка

Любой взвешенный ряд по длине сессий можно интерпретировать в бинарном виде, значение точки в котором будет означать, принадлежит ли точка сессии либо «не сессии». Интерпретировав, таким образом, получившееся предсказание мы получим предсказанное расположение сессий.



Предсказанное расположение сессий в интервале квартала

Сравним полученные результаты с реальным продолжением ряда.



Реальное расположение сессий

Видно, что на длительном интервале качество предсказания ухудшается, однако первая группа сессий распознана с достаточной точностью

Нахождение оптимального времени публикации

Оптимальное время прогноза для целевой группы пользователей высчитывается суммированием предсказанных рядов бинарного вида.

Краткий алгоритм проведённого анализа

1) Из IrregularSeries по заданному коэффициенту угасания с помощью сплайнов Эрмита получить RegularSeries сессий.

2) Из RegularSeries получить взвешенный по длине сессий/«несессий» RegularSeries

a. В случае, если полученный RegularSeries тривиальный – не меняет знака, предсказание делать тривиальным того же знака.

3) Из взвешенных RegularSeries осуществить предсказание «Гусеницы»

4) Полученный ряд предсказания бинаризовать.

5) Суммировать получившиеся бинаризованные ряды пользователей из целевой группы.

6) Найти точки максимумов.

Валидация осуществлённого предсказания

Ниже приведены параметры полученного предсказания для 30 пользователей Skype на временных интервалах. Длина 2 части рядов указана в таблице в часах.

Часы	1	2	3	5	8	12	24	48
Активных пользователей в выбранный момент	0	2	12	12	17	18	18	17
Максимум активных пользователей на интервале	1	4	18	15	19	22	25	27
10 минут после начала сессии (предсказание)	0	1	12	18	22	22	21	24
10 минут после начала сессии (максимум)	0	2	17	22	24	27	30	30
КПД нахождения в выбранном интервале	0	0.5	0.6	0.8	0.9	0.9	0.9	0.65
КПД нахождения пользователей спустя 10 минут от предполагаемой сессии	NA	0.5	0.7	0.8	0.9	0.8	0.7	0.8

Результаты апробированы на статистике комментирования 30 пользователей StackOverflow. Было взято 10 случайных интервалов и произведена валидация предсказания, количество пользователей усреднено:

Часы	1	2	3	5	8	12	24	48
Активных пользователей в выбранный момент	1	2	10	12	13	13	16	13
Максимум активных пользователей на интервале	3	4	13	14	15	15	18	20
10 минут после начала сессии (предсказание)	1	1	11	13	13	10	10	16
10 минут после начала сессии (максимум)	1	2	14	15	15	12	13	20
КПД нахождения в выбранном интервале	0.3	0.5	0.8	0.9	0.9	0.9	0.9	0.65
КПД нахождения пользователей спустя 10 минут от предполагаемой сессии	1	0.5	0.8	0.9	0.9	0.8	0.7	0.8

Из полученных статистик видно, что оптимальный интервал для поиска времени публикации - 5-10 часов.

Полученные Результаты

- Разработана система импорта и моделирования данных
- Предложены и применены методы подготовки данных
- Произведена оценка качества предсказания
- Предложены оптимальные для публикации временные интервалы
- Система разработана на языках R и C#

Список литературы

3 Twitter Tools To Determine The Best Time To Tweet. Получено из http://www.shoutmeloud.com/3-twitter-tools-to-determine-the-best-time-to-tweet.html

Automatic Time Series Forecasting: The forecast Package for R.. Получено из http://www.jstatsoft.org/v27/i03/paper

Graphics with R R Development Core Group. Получено из http://csg.sph.umich.edu/docs/R/graphics-1.pdf

HOW TO Post Your Tweets At Optimal Times. Получено из http://blog.bufferapp.com/how-to-post-your-tweets-at-optimal-times

R пакет алгоритма «Гусеницы».. Получено из http://cran.r-project.org/web/packages/Rssa/Rssa.pdf

R пакет для анализа временных рядов.. Получено из http://cran.r-project.org/web/views/TimeSeries.html

The Best Time to Ask a Stack Overflow Question?. Получено из http://blog.stackoverflow.com/2009/01/the-best-time-to-ask-a-stack-overflow-question/

When is the best time to post on Facebook? Here's the Answer….. Получено из http://www.danieldecker.net/when-is-the-best-time-to-post-on-facebook-heres-the-answer/

ZOLLMAN, K. J.. *OPTIMAL PUBLISHING STRATEGIES*. Получено из http://www.andrew.cmu.edu/user/kzollman/research/Papers/Zollman-OptimalPublishing.pdf

А.А, Дзюба. *Сбор рваных временных рядов в централизованное хранилище.*

Алгоритмы и интерполяция сплайнами. Получено из http://alglib.sources.ru/interpolation/spline3

Грешилов А. А., С. В. *Математические методы построения прогнозов.*

Д. Л. Данилов, А. А. *Главные компоненты временных рядов: метод «Гусеница».*

Н.Э., Г. . *Метод «Гусеница»-SSA: прогноз временных рядов.*

Н.Э., Г.. *Метод «Гусеница»-SSA: анализ временных рядов* .

Прогнозирование параметров вращения земли с помощью сингулярного спектрального анализа Горшков В.Л., Миллер Н.О.. Получено из http://www.gao.spb.ru/english/as/publ/gorshkov_miller_izv219.pdf