УДК 330.101.541

НЕГАТИВНО ОКРАШЕННЫЕ СЛОВА В ПОИСКОВЫХ ЗАПРОСАХ ИНТЕРНЕТА КАК ИНДИКАТОР УРОВНЯ СРЕДНЕДУШЕВЫХ ДОХОДОВ НАСЕЛЕНИЯ В ФЕДЕРАЛЬНЫХ ОКРУГАХ РФ

А.В. Болдырева[1,2], М.А. Александров[2,3], Д.В. Суркова[1]

[1] *Московский физико-технический институт (государственный университет),*

[2] *Российская академия народного хозяйства и государственной службы при Президенте РФ,*

[3] *Автономный Университет Барселоны*

anna.boldyreva@phystech.edu, malexandrov@mail.ru, darina10@mail.ru

У роботі досліджується залежність частоти використання негативно забарвлених слів та рівня достатку населення. Використані часові ряди пошукових запитів Яндекса і алгоритми Методу Групового Урахування Аргументів (МГУА). Ці алгоритми включені в пакет GMDH Shell для прогнозу доходу населення.

Ключові слова: Інтернет, пошукові запити, добробут, негативні слова

In the paper the dependence of frequency of words having negative sentiment intensity and level of population welfare is studied. We use time series of Yandex search queries and algorithms of Group Method of Data Handling (GMDH). These algorithms are included to the software package GMDH Shell to forecast incomes of population.

Keywords: Internet, search queries, , income, negative sentiment

В работе исследуется зависимость частоты использования негативно окрашенных слов и уровня достатка населения. Использованы временные ряды поисковых запросов Яндекса и алгоритмы Метода Группового Учета Аргументов (МГУА). Эти алгоритмы включены в пакет GMDH Shell для прогноза дохода населения.

Ключевые слова: Интернет, поисковые запросы, благосостояние, негативные слова

Вступление

В настоящей работе исследуется степень удовлетворенности (уровень счастья) населения регионов России методом «от противного» — через взаимосвязь среднедушевых доходов населения и интенсивности употребления негативных слов в поисковых запросах в Интернет по федеральным округам (ФО) России. Исходная гипотеза: чем меньше негативных слов используется в лексике людей, тем выше их материальная обеспеченность. Такой подход может быть оправдан тем, что довольные люди в меньшей степени выражают свою радость, чем расстроенные. В работе также изучается возможность прогнозирования экономического состояния регионов с использованием динамики поисковых запросов.

Принципиальные отличия данной работы от других исследований в этой области:

- впервые были получены прогнозные модели для определения экономического положения в регионах с помощью поисковых запросов;

- впервые использован статистический сервис Яндекс для оценки экономического положения регионов;
- впервые исследуется зависимость между поисковыми запросами и экономическим состоянием регионов Российской Федерации;

1. Терминология и сокращения

В исследовании используется следующая терминология:

- Поисковый запрос — запрос, исходящий от пользователя сети Интернет для получения информации в поисковой системе;
- Поисковая система — специализированные сервисы для поиска информации;
- Дескриптор — часть слова, слово или словосочетание, служащие для формулировки запроса при поиске информации в поисковой системе;
- Индикаторы — экономические, социальные, демографические показатели, публикуемые на официальных или авторитетных сайтах;
- Топовая выборка дескрипторов — выборка динамик поисковых запросов, наиболее высоко коррелирующих с заданными индикаторами;
- Барометры — динамика средних значений в «топовой выборке» по временной шкале.

В исследовании применяются сокращения:

- НО-дескрипторы (NS, negative sentiment) — эмоционально негативно-окрашенные слова, используемые пользователями поисковой машины.
- МГУА — метод группового учета аргументов;
- GMDH Shell (GS) — пакет программного обеспечения, использующий алгоритмы МГУА;
- ФО — федеральный округ;
- PCI (per capita income) — среднедушевые доходы населения;
- μSQ (μ search queries) — среднее значение динамики поисковых запросов;
- abs — абсолютные значения поисковых запросов;
- rel — относительные значения поисковых запросов;
- Pop — популяция, количество населения в регионе;
- DI (development of the Internet by region) — развитие сети Интернет в регионе;
- LW (living wage) — прожиточный минимум;
- SL (standard of living) — уровень жизни;
- norm — нормализованные значения.

2. Состояние вопроса

Только на одном портале PLoS One [1] выложено более 50 тысяч работ, посвященных анализу социальных медиа. С помощью поисковых запросов

Google анализируются уровни безработицы [2], прогнозируются эпидемии гриппа [3]. Используя метод случайного блуждания, ученые из Финляндии смогли выявить самые спорные и противоречивые темы, волнующие пользователей Твиттера [4].

Исследования поисковых запросов позволяют улучшать традиционные регрессионные модели. В своей работе «Forecasting private consumption: survey-based indicators vs. Google trends» [5] исследователи сравнивали прогнозирование уровня частного потребления — авторегрессионную модель, и ту же модель с добавлением информации из Google Trends. С помощью приложения Google Trends и Insights for Search аналитики скачали набор категорий и экспертно отобрали подходящие потреблению запросы. Исследование доказывает, что прогнозные модели, выполненные с учетом информации Гугл Трендов, превосходят обычные регрессионные модели.

Исследователи утверждают, что с помощью социальных медиа можно также выявлять и анализировать экономические проблемы в разных странах и регионах.

Проведя исследование более 19 миллионов сообщений пользователей социальной сети Твиттер из 340 регионов Испании, ученые из Института инженерии знаний и Университета Карлоса III в Мадриде, Алехандро Льоренте с коллегами [6], выявили связи между содержанием сообщений и экономическим состоянием регионов. Похожее исследование провели учёные-лингвисты из Пенсильвании совместно с IT-специалистами Microsoft Research [7]. Исследование проводилось так же с использованием площадки Твиттер. Было проанализировано более 10 миллионов сообщений от 5000 пользователей. Отметив, что существует устойчивое мнение о том, что хорошее образование и высокий интеллект определяют более высокие доходы, ученые пришли к парадоксальному выводу — высокие доходы в регионе коррелируют с использованием агрессивной лексики и слов, связанных со страхом.

В 2010 году Ингмар Вебер и Карлос Кастильо скачали большую выборку персонализированных поисковых запросов с «Yahoo!» и проанализировали соотношение длины поисковых запросов у разных типов людей (в том числе в зависимости от дохода) [8]. Анализировалось, каким образом люди с разным доходом ищут одну и ту же страницу. Например, страницу http://www.braces.org в среднем ищут по запросу «braces», а богатые — по навигационному запросу «braces.org».

Шарада Гоел, Джейк М. Хоффман и М. Ирмак Сайрер опубликовали обширное исследование поведения людей в Интернете. Анализ пользовательской активности людей позволил ученым доказать, что у историй браузеров имеется сильная связь с доходами домашних хозяйств [9].

В то же время, не все индикаторы подаются прогнозированию с помощью динамики поисковых запросов. Например, цены на полезные ископаемые [10] или цены на аренду коммерческой недвижимости почти невозможно спрогнозировать с помощью динамики поисковых запросов. Тогда как индексы

потребительских цен, уровни рождаемости или безработицы, цены на вторичное жилье — прогнозируются с высокой долей точности. В работе [11] было сделано предположение, что качественные прогнозные модели строятся там, где есть высокая активность пользователей, а формирование процесса или индикатора не регламентировано. Поиск показателей, которые поддаются подобным исследованиям — отдельное направление в этой области.

3. Индикаторы (зависимые переменные)

В нашей работе использовались данные по всем федеральным округам РФ, кроме ФО Крым. Это связано с отсутствием данных прошлых периодов об экономическом положении в этом регионе.

В качестве индикаторов были взяты данные среднедушевых денежных доходов с сайта Госкомстата [12,13,14]. Для анализа каждого ФО отбирались данные, соответствующие динамике среднедушевых доходов этого конкретного региона.

В России очень высокое территориальное расслоение экономического состояния, как по уровню цен, так и уровню доходов населения. Для Республики Калмыкия сумма 25 тысяч рублей в месяц — очень высокая зарплата, для Магаданской области это крайне мало. Поэтому мы дополнительно скорректировали показатели среднедушевых доходов с учетом региональных коэффициентов. Таким образом, в исследовании применяются 3 индикатора:

- Среднедушевые доходы в ФО;
- Среднедушевые показатели с учетом стоимостей жизни в регионах. Для каждого региона индекс рассчитывался по ежеквартальным данным Госкомстата Российской Федерации [12,13];
- Среднедушевые показатели с учетом коэффициента уровня жизни в регионе [14]. Рейтинг социально-экономического положения регионов вычисляется ежегодно аналитиками Рейтингового агентства «РИА Рейтинг и рассчитывается на основе агрегирования 15 групп показателей, характеризующих экономическую, социальную и бюджетную сферы. Учитываются, например, индекс промышленного производства в регионе, динамика инвестиций, оборот розничной торговли [15] и т.д.

4. Барометры (независимые переменные)

Выборка негативно окрашенных слов (далее НО-дескрипторы) была собрана согласно рекомендациям экспертов в области анализа тональности высказываний с использованием материалов, представленных в [16].

Пул содержит одинаковый набор из 575 НО-дескрипторов для всех ФО.

Диапазон, используемый в исследовании: сентябрь 2013-го — август 2015-го года, определяется временными рамками сервиса Яндекса [17].

В барометрах использовались дескрипторы, имеющие максимальную корреляцию с индикаторами исследуемого региона.

Процесс подготовки данных для прогнозирования заключается в сборе файлов с расширением .html со статистического сервиса Яндекса. Далее файлы расшифровываются, создаются динамические ряды дескрипторов и таблицы временных рядов в корреляции с заданными индикаторами и одновременным лагированием в 1, 2 и 3 месяца. Такие подготовительные операции позволяют сократить базы от сотен тысяч временных рядов дескрипторов до «топовой выборки» из десятков самых значимых в корреляции с заданным индикатором.

Для создания «барометров» из полученной топовой выборки временные ряды нормируются, группируются по взаимной корреляции и вычисляются их средние значения. В результате топовая выборка сокращается до 8 барометров. 4 барометра содержат средние значения временных рядов дескрипторов, дающих высокую корреляцию с индикатором — без лагирования, с лагом в 1 месяц, в 2 месяца и в 3 месяца. Другие 4 барометра содержат средние значения временных рядов дескрипторов, которые дают высокую антикорреляцию с индикатором — без лагирования, и с лагом в 1-2-3 месяца.

Создание барометров необходимо, прежде всего, для обработки их в прогнозной программе GMDH Shell с целью, для сокращения сроков обработки данных. Подобные барометры также создавала в своих исследованиях компания Google [3]. Таким образом, мы получаем страховку от случайностей, которые могут быть следствием неучтенных информационных и общественных процессов. Если какой-либо из 10-40 дескрипторов, входящих в барометр, на следующий месяц даст нестандартную динамику, то остальные динамики в барометре эту погрешность нивелируют.

5. Анализ

Была исследована вся база НО-дескрипторов по всем ФО РФ, кроме федерального округа Крым. Это связано с отсутствием данных прошлых периодов об экономическом положении в этом регионе. Таким образом, были исследованы данные округов: Центральный, Северо-Западный, Южный, Северо-Кавказский, Уральский, Сибирский, Дальневосточный и Приволжский. Для предварительного анализа было взято среднее ежемесячное значение всех запросов, приходящихся на регион, начиная с октября 2013-го года в абсолютном выражении. Был построен график (Рис. 1) соответствия ежемесячного среднедушевого дохода (PCI) усредненным нормированным значениям динамики НО-дескрипторов в абсолютном выражении (μSQ,NS,abs,norm) по каждому месяцу и по каждому ФО, всего 176 точек.

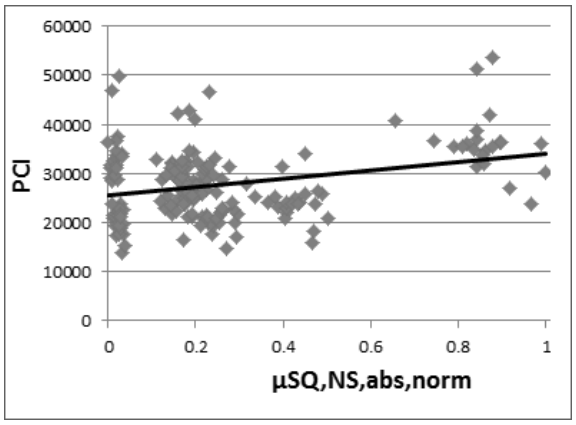

Рис. 1. Соотношение среднедушевых доходов (PCI) и абсолютных значений поисковых запросов (µSQ)

На основе этого графика мы можем предположить, что с увеличением доходов населения растет и частота использования НО-дескрипторов. На графике так же видно, что данные сгруппированы в 3 кластера. Анализ кластеров показал, что левая группа содержит в основном значения поисковых запросов Дальневосточного и Северо-Кавказского ФО. Крайне правая группа содержит запросы Центрального Федерального округа. Эти группы федеральных субъектов различаются, в том числе, количеством населения и уровнем распространения Интернета. Поэтому было принято решение ранжировать значения НО-дескрипторов по регионам с учетом популяции (Pop) в регионе и развитием сети Интернет (DI — development of the Internet by region) — µSQ,NS,abs,Pop,DI,norm.

На итоговом графике (Рис. 2) мы наблюдаем, что распределение стало более равномерным, и тренд всей динамики по-прежнему направлен в сторону увеличения доходов с ростом использования НО-дескрипторов в ФО.

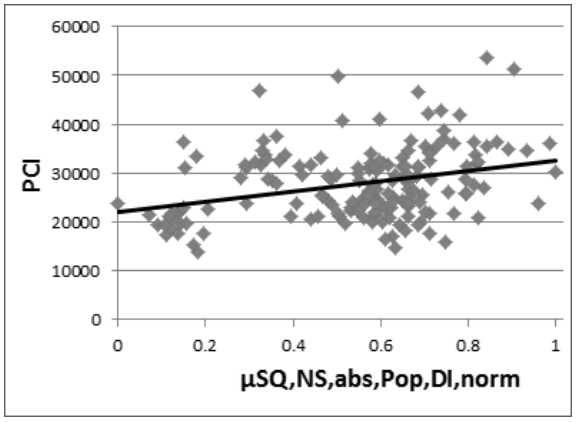

Рис. 2. Соотношение PCI и абсолютными значениями µSQ, с учетом Pop и DI

Чтобы учесть неоднородность доходов и расходов в регионах России, показатели доходов были ранжированы с учетом прожиточного минимума (living wage — LW — PCI/LW) в регионе, а также с учетом рейтинга социального развития регионов (standard of living — SL — PCI*SL/100).

Итоговые графики (Рис. 3) показывают разное рассеивание значений, но при этом тренд не меняется. В отличие от графика на рисунке 1, крайне правые значения принадлежат регионам Западного Кавказа и Дальнего Востока. Точки, расположенные слева, принадлежат в основном регионам ЦФО, СЗФО и Уралу.

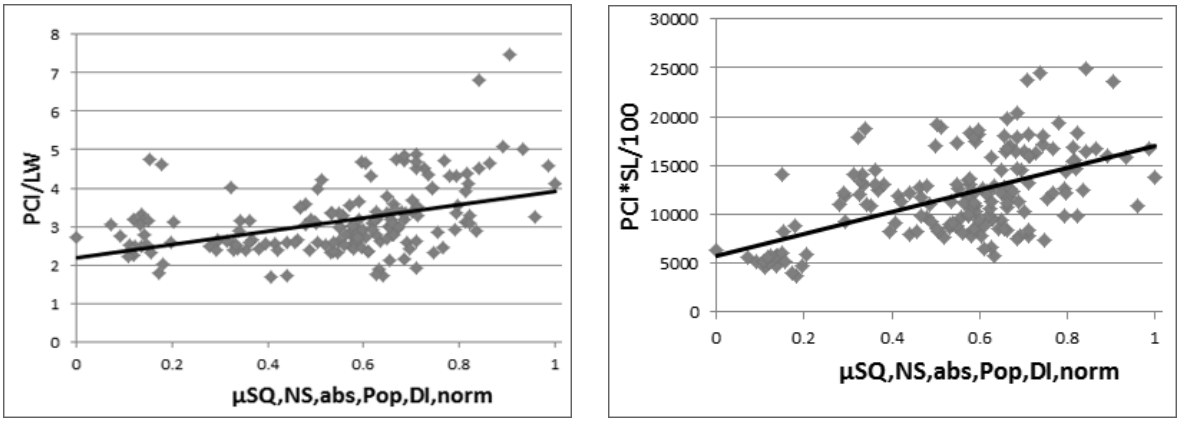

Рис. 3. Соотношение PCI/LW, PCI/ SL и абсолютными значениями µSQ, с учетом Pop и DI

Можно с большой уверенностью предположить, что жители обеспеченных регионов действительно больше употребляют НО-дескрипторов, чем жители малообеспеченных ФО. Но, возможно, это связано с тем, что более обеспеченные граждане в целом больше используют Интернет.

Статистический сервис Яндекса предоставляет данные поисковых запросов не только в абсолютном, но также и в относительном выражении. При этом абсолютное значение данного конкретного запроса Яндекс ранжирует по общему количеству всех поисковых запросов, сделанных в данном регионе.

Было взято среднее ежемесячное значение всех запросов, приходящихся на регион, начиная с октября 2013-го года, в относительном выражении. Затем были получены графики соответствия ежемесячного PCI и усредненного значения динамики НО-дескрипторов также в относительном выражении (µSQ,NS,rel,norm). Всего мы имели 176 точек. Данные по доходам населения были ранжированы аналогичным образом: и с учетом прожиточного минимума в регионе, и с учетом рейтинга социального развития регионов.

На графиках (Рис. 4) мы видим, что направление тренда изменилось по сравнению с графиками (Рис.3), где использовались абсолютные значения поисковых запросов.

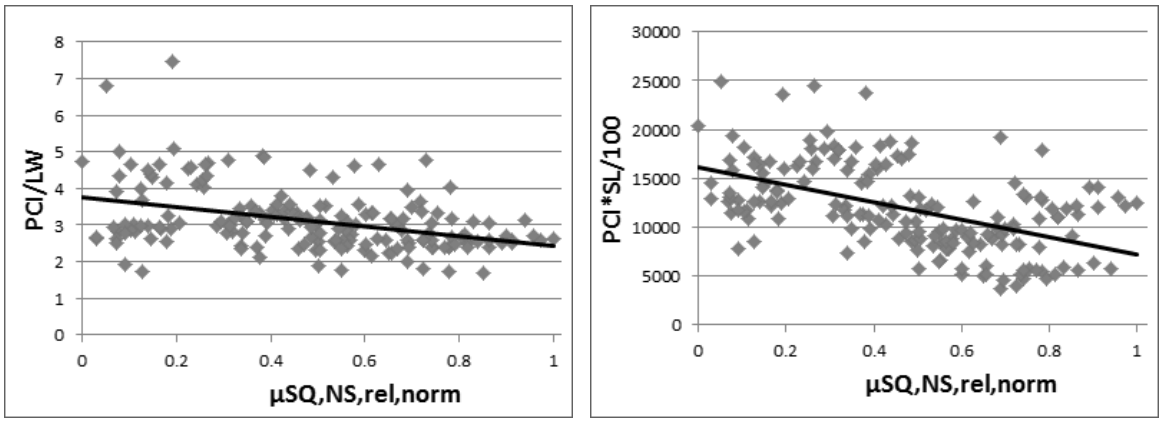

Рис. 4. Соотношение PCI/LW, PCI/ SL
и динамики относительных значений μSQ

Теперь можно предположить, что с увеличением среднедушевого дохода люди начинают меньше употреблять негативно окрашенные слова. Но такой вывод противоречит исследованию ученых группы Дэниела Преотьюк-Пьетро [7], которая вывела прямо пропорциональную зависимость использования эмоционально агрессивных слов и слов, связанных с проявлением страха — с уровнем PCI. Чтобы разрешить это противоречие, возможно, необходимо ранжировать нашу базу всех НО-дескрипторов по 2-м группам слов, обозначающих, соответственно, агрессию и страх.

После разделения базы на указанные группы мы повторили все исследования отдельно для дескрипторов агрессии и страха для абсолютных (Рис. 5,6) и для относительных (Рис. 7,8) значений μSQ. Был отмечен интересный факт, что линия тренда дескрипторов страха изменила направление (Рис. 6), по сравнению с другими графиками, которые также содержат абсолютные значения поисковых запросов (Рис. 1,2,3,5). С увеличением доходов использование в поисковых запросах слов, связанных со страхом (как в абсолютном, так и в относительном выражении) падает. Также, несмотря на все процедуры ранжирования, графики абсолютных значений дескрипторов страха демонстрируют устойчивое распределение точек по региональным кластерам. Возможно, это свидетельствует о культурных особенностях регионов. Высокий уровень использования дескрипторов страха показывает Дальний Восток. Кластер посередине — Северный Кавказ. Слева на графиках, ранжированных по дескрипторам страха — кластеры Центрального и Приволжского округов.

На графиках с относительными значениями дескрипторов Дальний Восток так же показывает высокий уровень использования дескрипторов, при этом Центральный федеральный округ показывает крайне низкий уровень их использования (Рис. 7,8).

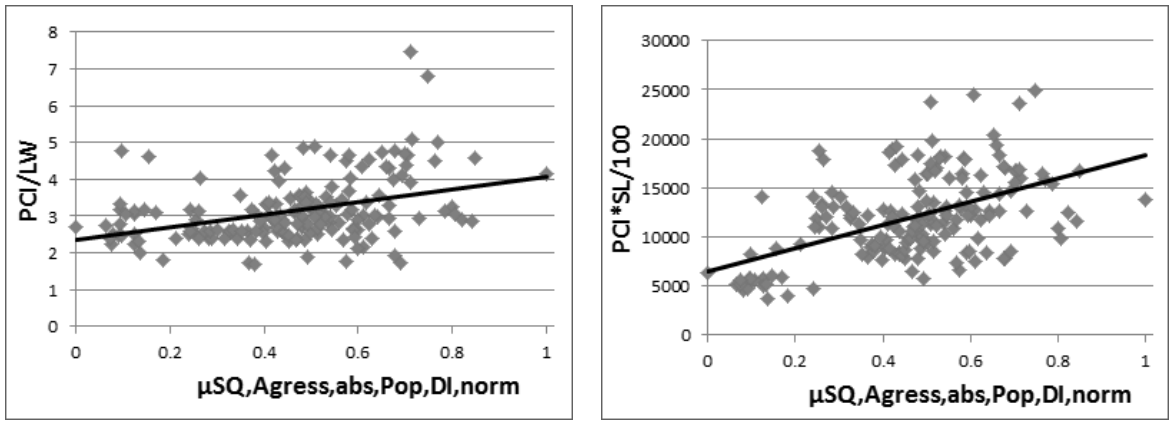

Рис. 5. Соотношение PCI/LW и PCI*SL/100 и динамики абсолютных значений μSQ агрессии (Agress), с учетом Pop и DI

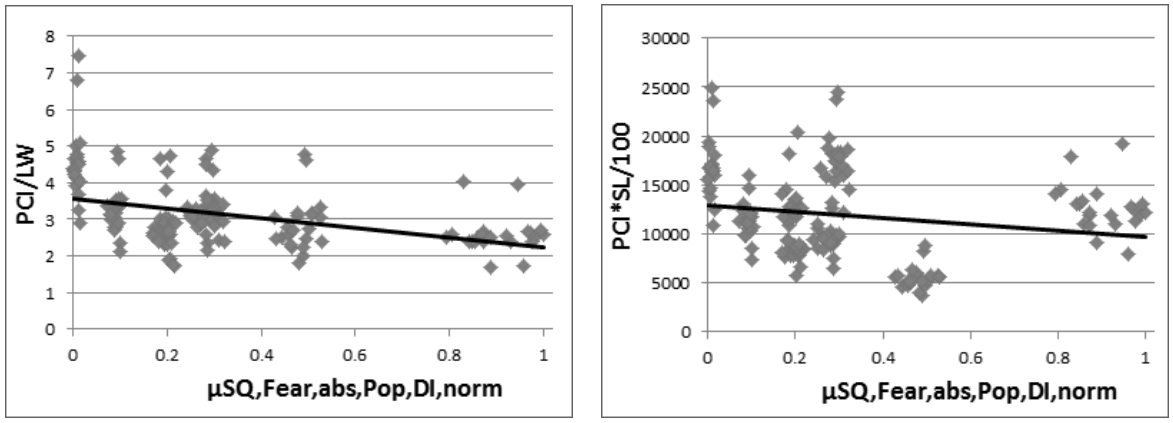

Рис. 6. Соотношения PCI/LW и PCI*SL/100 и динамики абсолютных значений μSQ страха (Fear), с учетом Pop и DI

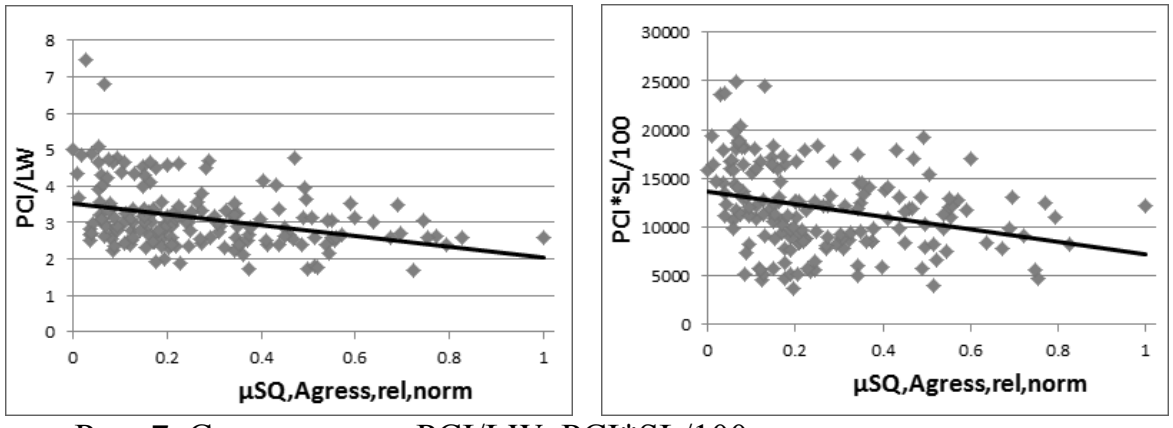

Рис. 7. Соотношение PCI/LW, PCI*SL/100 и динамики относительных значений μSQ агрессии (Agress)

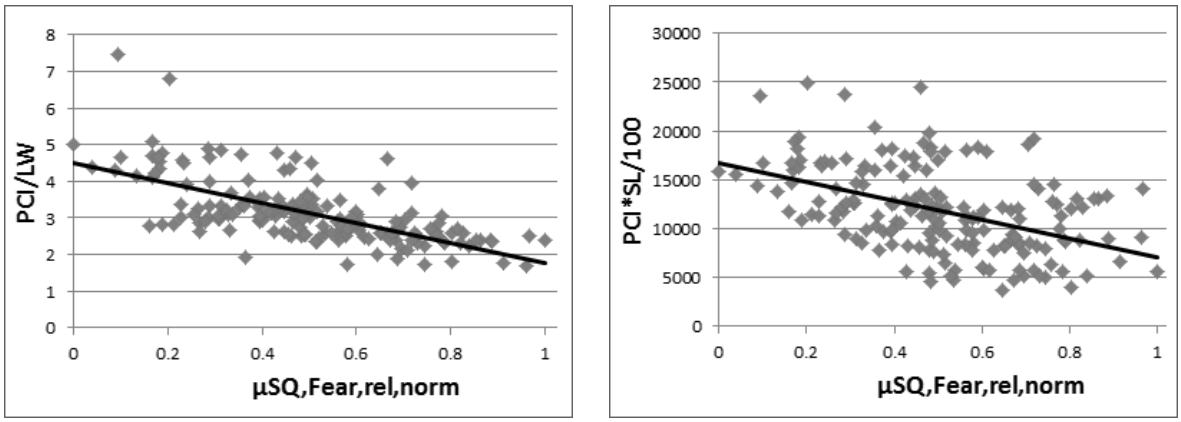

Рис. 8. Соотношение PCI/LW, PCI*SL/100 и динамики относительных значений μSQ страха (Fear)

Все графики относительных значений показывают обратно пропорциональную зависимость дескрипторов агрессии и страха от PCI.

Интересно посмотреть (Рис. 9), как распределяются средние значения относительных HO-дескрипторов после нормирования общей выборки по регионам.

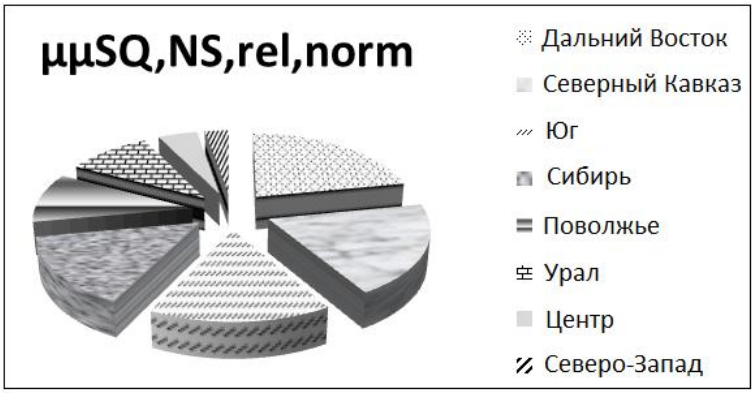

Рис. 9. Распределение средних значений относительных HO-дескрипторов после общей нормировки

Лидеры: Дальний Восток, Северный Кавказ, Южный федеральный округ и Сибирь. Данный материал – это хороший повод региональным властям задуматься над сложившейся ситуацией.

6. Инструменты моделирования

6.1 МГУА

Для моделирования был применен МГУА — Метод Группового Учета Аргументов. Метод был предложен академиком А. Ивахненко в 70-е годы прошлого столетия и получил развитие в 80-е и 90-е годы в его работах, а также работах его учеников [18,19,20]. Новые алгоритмы МГУА были предложены В.

Степашко в его работе [21]. Можно отметить две принципиальные характеристики МГУА:

a. метод может применяться при отсутствии или почти отсутствии априорной информации о структуре и значениях параметров модели;
b. метод может применяться в условиях крайне ограниченного набора экспериментальных данных, когда число наблюдений может быть даже меньше числа параметров модели.

Существует несколько базовых алгоритмов МГУА и десятки их модификаций. В нашей работе мы используем нейроподобные алгоритмы МГУА, поскольку именно они обеспечили наиболее точные результаты на наших данных по сравнению с другими алгоритмами.

6.2 GMDH Shell

В качестве инструмента моделирования использовался пакет GS, который упоминали выше. GS предлагает различные процедуры обработки:

1. Преобразование данных. Здесь могут применяться различные элементарные функции, такие как, например, логарифмы, квадратные корни, и т.п. Могут вводиться лаги, дополнительные переменные времени, а также фиктивные переменные.
2. Контроль качества модели. Здесь могут быть заданы различные соотношения размеров обучающей и контрольной выборки, k-кратная перекрестная проверка и т.п.
3. Выбор внешнего критерия. Например, может использоваться среднеквадратичная ошибка прогноза, или корреляционная оценка расхождения модельных и наблюденных данных и т.п.
4. Ограничения на сложности модели. Пользователь может задать: максимальное число членов в уравнениях модели в комбинаторном алгоритме, максимальное количество слоев в нейроподобном алгоритме.

Существенное преимущество GS — автоматическое тестирование различных вариантов решения. При этом лучшие решения предлагаются пользователю.

7. Моделирование

7.1 Параметры

В качестве индикатора мы рассмотрели средний доход, деленный на прожиточный минимум.

Для нейронного алгоритма мы использовали квадратичные полиномы. Были приняты ограничения: начальная ширина слоя - 5, максимально допустимое число слоев - 6. Все наблюдения были перемешаны случайным образом. Модель проверялась перекрестно на основе двух подвыборок.

Последний месяц был исключен из выборки для последующего контроля качества прогноза.

В качестве внутреннего критерия мы использовали метод наименьших квадратов. Внешний критерий определялся среднеквадратичной ошибкой прогноза. Для сравнения моделей между собой мы использовали среднюю абсолютную процентную ошибку (MAPE).

Точность прогноза на июль 2015 года мы оценивали по процентному отклонению от реального значения P.

7.2 Прогнозные модели

Всего было построено 8 моделей. Примеры моделей с наилучшими результатами для наиболее отличающихся округов представлены ниже. На Рис. 10 показан график модели для Дальневосточного ФО.

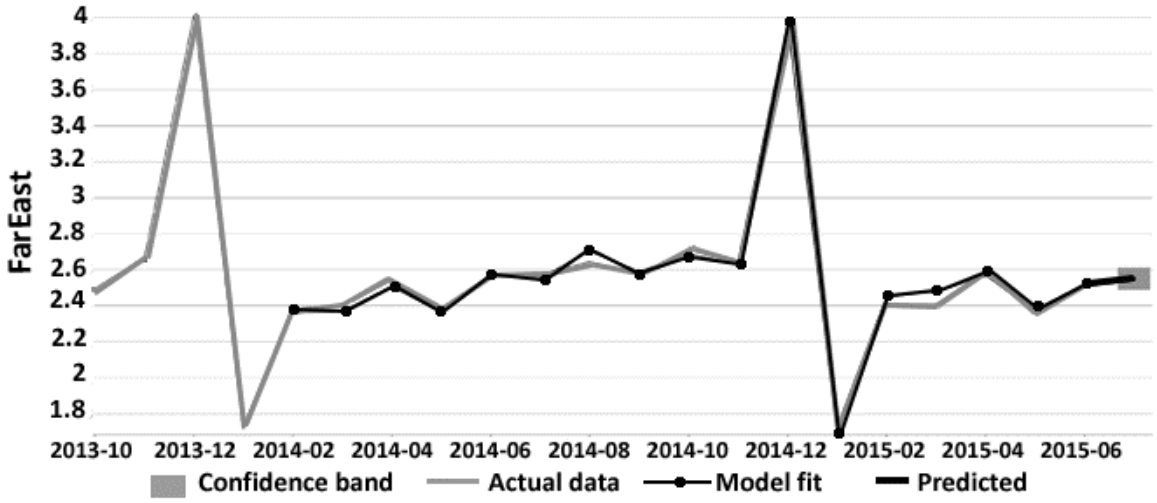

Рис. 10. График реальных и прогнозных значений для Дальневосточного ФО. Нейроподобный алгоритм. MAPE = 1.0%, P = –1.6%

Система уравнений, представляющая данную модель:

$$Y1\ t\ = 0.005 - Offset3mNeg\ \ t-3\ \ * \mathrm{N}\,2 * 0.033 + N2 * 1.019;$$

$$N2\ \ t\ \ = 0.016 - Offset3mPos\ \ t-3\ \ * N3 * 0.041 + N3 * 1.006;$$

$$N3\ \ t\ \ = -0.063 - Offset2mNeg\ \ t-2\ \ * N4 * 0.097 + N4 * 1.056;$$

$$N4\ \ t\ \ = -0.068 + Offset1mNeg\ \ t-1\ \ * N6 * 0.076 + N6 * 1.000;$$

$$N6\ \ t\ \ = 4.032 + Offset2mPos\ \ t-2\ \ * (N9 * 2.414 - 5.903) - N9 * 0.636;$$

$$N9\ \ t\ \ = 2.978 + Offset2mPos\ \ t-2\ \ * (Offset2mPos\ \ t-2\ \ * 7.151 - 4.160);$$

где $Y1$ — индикатор;

Значения в квадратных скобках — месяцы;

$Offset1mPos$, $Offset2mNeg$, $Offset3mPos$ — барометры, которые наиболее высоко положительно коррелируют с индикатором со сдвигами, соответственно, в 1-2-3 месяца.

$Offset1mNeg$, $Offset2mNeg$, $Offset3mNeg$ — барометры, которые наиболее высоко отрицательно коррелируют с индикатором со сдвигом, соответственно, в 1-2-3 месяца.

Примеры НО-дескрипторов, которые вошли в барометр для Дальневосточного ФО, представлены в таблице 1.

Таблица 1.

Примеры НО-дескрипторов, вошедших в барометры для Дальневосточного ФО

Барометр	НО-дескрипторы
$Offset1mPos$	урод, подлец, редкий, выродок
$Offset1mNeg$	грубо, зануда
$Offset2mPos$	яд, блуд, дрянной, предательство
$Offset2mNeg$	беда, высокомерный, самоубийство, депрессия, трудно
$Offset3mPos$	жлоб, отчаяние, обида, тревога
$Offset3mNeg$	придурок, крик, недовольно, слабак

На Рис. 11 показаны результаты для Северо-Западного ФО.

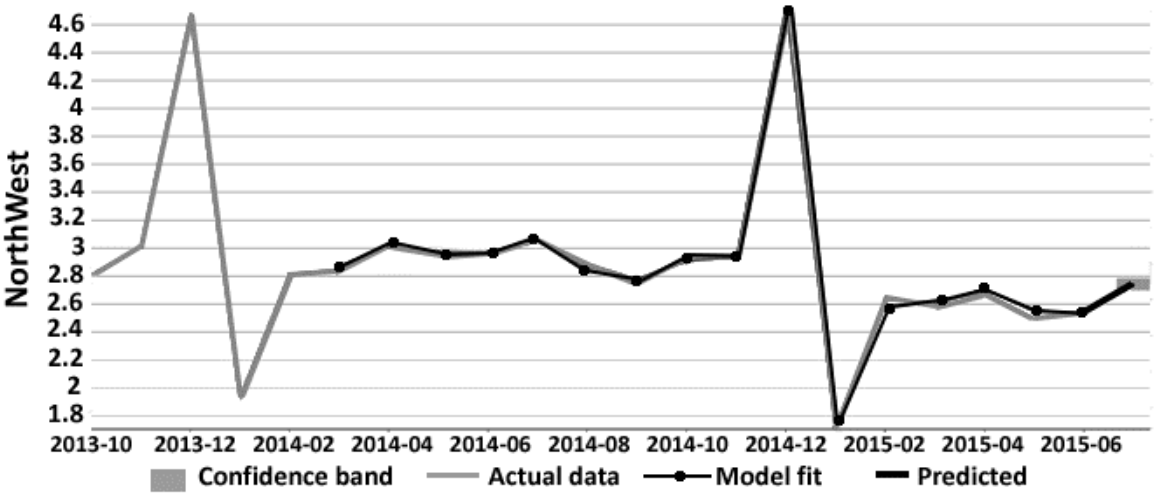

Рис. 11. График реальных и прогнозных значений для Северо-Западного ФО. Нейроподобный алгоритм. MAPE = 0.5%, P = –4.4%

На рис. 12 показаны результаты для Сибирского ФО.

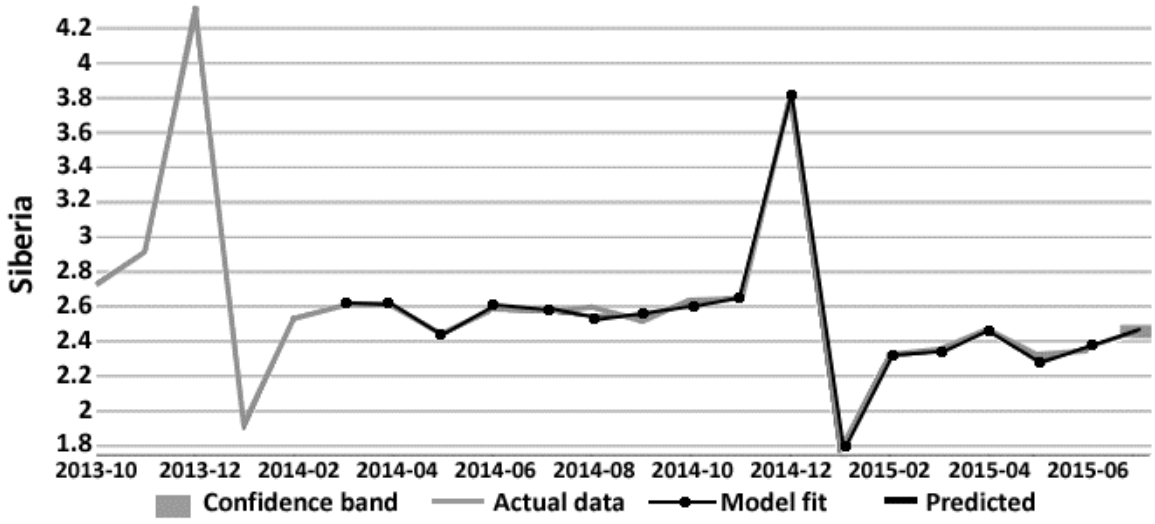

Рис. 12. График реальных и прогнозных значений для Сибирского ФО. Нейроподобный алгоритм. MAPE = 0.5%, P = –2.5%

Мы получили MAPE: 0.4%-4.9% и абсолютное процентное отклонение от реального значения прогноза на июль 2015 года P: 1.7%-10.9%. Таблица 2 содержит результаты для каждой модели и индикатора.

Нейроподобный алгоритм показал хорошие результаты для всех округов, кроме Южного ФО. Это туристический регион, и можно предположить несколько объяснений такого расхождения:

- поисковые запросы, отправляемые приезжими, искажают статистику;
- официальные публикуемые данные по среднедушевым доходам не отражают реальные стандарты жизни в этом регионе.

Таблица 2.

MAPE и ошибки будущих прогнозов

ФО	MAPE	P
Центральный федеральный округ	4,9%	5,7%
Дальневосточный федеральный округ	1,0%	-1,6%
Северо-Кавказский федеральный округ	0,5%	0,3%
Северо-Западный федеральный округ	0,5%	-4,4%
Сибирский федеральный округ	0,5%	-2,5%
Южный федеральный округ	4,4%	10,9%
Уральский федеральный округ	2,8%	-1,7%
Приволжский федеральный округ	0,4%	6,0%

Заключение. По результатам исследования мы можем сделать предположение, что при росте среднего уровня доходов на душу населения недовольство жителей российских регионов уменьшается, что отражается в относительно менее частом использовании негативно окрашенных слов в поисковых запросах. Кроме того, по относительному уровню использования НО-дескрипторов можно выявить регионы, где жители испытывают большее недовольство. Исследование показало, что жители Дальнего Востока, Северного Кавказа, Южного федерального округа и Сибири используют в относительном выражении на порядок больше негативно-окрашенных слов в поисковых запросах, чем жители Центральных и Западных регионов России. Такие выводы могут быть полезны руководителям Федеральных округов, а также политикам, социологам, экономистам.

Результаты моделирования показывают возможность получить очень высокое качество прогноза среднедушевого дохода населения на основе исследования динамики НО-дескрипторов. А именно, были получены ошибки прогнозирования 0.3%-6%. Следует отметить, что предложенный нами метод прогноза не подходит для регионов с флуктуирующим населением и непостоянными доходами. К таким регионам относится, в частности, Южный федеральный округ.

В будущих работах предполагается:

- использовать авторегрессию для улучшения качества прогноза наряду с регрессией;
- использовать динамику «упоминаний в сети» вместо динамики «поисковых запросов», что позволит учесть большее количество мнений.

Благодарности

Авторы выражают признательность д.т.н. проф. В.С. Степашко за интерес к работе и ценные замечания, к.т.н. А.А. Кошулько, главному конструктору пакета GMSH Shell, за многочисленные консультации.

Литература

1. Multidisciplinary Open Access journal PLoS One, Web: https://www.plos.org/
2. Pavlicek J. et al. Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. – PloS One, 2015, Web: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0127084.
3. Ginsberg G. et al. Detecting influenza epidemics using search engine query data. – Nature, vol.457, 2009. - P. 1012-1014.
4. Garimella K. et al. Quantifying Controversy in Social Media. – ArXiv, 2015, Web: http://arxiv.org/abs/1507.05224.
5. Vosen S. Forecasting private consumption: survey-based indicators vs. Google trends. – Journal of Forecasting, vol. 30, issue 6, 2011. - P. 565-578.
6. Llorente A. et al. Social Media Fingerprints of Unemployment. – PloS One, 2015, Web: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128692.

7. Preoțiuc-Pietro D. et al. Studying User Income through Language, Behaviour and Affect in Social Media. – PloS One, 2015, Web: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0138717.
8. Weber I. et al. The demographics of web search. – Proc. of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010. - P. 523-530.
9. Goel S. et al. Who Does What on the Web: A Large-scale Study of Browsing Behavior. – Proc. of ICWSM, 2012. – 8 p.
10. Болдырева А. Построение прогнозных моделей экономической конъюнктуры и преступлений экономической направленности по интенсивности запросов в поисковой системе Интернет. – Дипломная работа, РАНХГС, 2015. – 109 с.
11. Болдырева А. Применение метода МГУА на основе интенсивности поисковых запросов в сети Интернет для прогноза рынка недвижимости. – Сборник трудов «Задачи современной информатики», 2015. – С. 46-51.
12. Уровень жизни и доходы населения, Госкомстат РФ: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/level/.
13. Прожиточный минимум по регионам РФ в 2015 и в 2016 году, Госкомстат РФ: http://potrebkor.ru/prozhitochnyi-minimum.html.
14. Федеральная служба государственной статистики, Информация для ведения мониторинга социально-экономического положения субъектов Российской Федерации: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/doc_1246601078438.
15. Рейтинг социально-экономического положения субъектов РФ: http://vid1.rian.ru/ig/ratings/rating_regions_2014.pdf.
16. Alexandrov M. et al. Models for opinion classification of blogs taken from Perivian Facebook. – Proc. of 4-th Intern. Conf. on Inductive Modeling, 2013, - P. 241-246.
17. Статистический сервис поисковой машины Яндекс: https://wordstat.yandex.com/.
18. Ивахненко А. Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наукова думка, 1981. – 296 с.
19. Ивахненко А., Степашко В. Помехоустойчивость моделирования: монография, – Киев: Наукова думка, 1985. - 216 с.
20. Ivakhnenko A. et al. Inductive learning algorithms for complex systems modeling, – NY: CRC Press, 1994. – 373 c.
21. Stepashko V. Ideas of academician A. Ivakhnenko in Inductive Modeling field from historical perspective. – Proc. of 4th Intern. Conf. on Induct. Modeling (ICIM-2013), 2013. – P. 31-37.