УДК 51-77

Прогнозирование социальных, демографических и макроэкономических показателей с помощью исследования поисковых запросов сети Интернет. Мифы и реальность

А.В. Болдырева[1,2]

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

[2]Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации

Абстракт

Поисковые запросы – живое отражение происходящих в обществе событий и явлений. Зарубежные ученые, аналитики, экономисты, трейдеры ежегодно публикуют сотни исследований в поисках связи между этими, вроде бы, хаотичными данными, и экономическими, социальными, демографическими показателями, которые регулярно публикуют государственные органы на соответствующих порталах. В России таких исследований практически не проводится. В докладе анализируются причины такого невнимания российских аналитиков к этой сфере, перечисляются основные мифы и предубеждения, сложившиеся вокруг подобных исследований. Показаны основные сервисы, методы и инструменты, используемые современными аналитиками для исследования динамики поисковых запросов.

Терминология

В исследовании используется терминология:

Поисковый запрос – запрос, исходящий от пользователя сети Интернет с целью 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 – дескрипторы с высоким значением коэффициента корреляции относительно выбранного индикатора;

«Барометр» – динамика средних значений «топовой выборки» по временной шкале;

«Тематические базы» – базы динамик поисковых запросов, сгруппированные по специализированным терминам: база юридических, экономических терминов, база названий брендов и т.д.;

«Технические базы» – базы динамик поисковых запросов, сгруппированные по количеству букв в дескрипторе.

Введение

Многочисленные зарубежные исследования доказали, что благодаря анализу пользовательской активности поисковых машин интернета можно прогнозировать экономическую конъюнктуру, эпидемии, направления трендов в туристической отрасли, показатели в криминальной, социальной и демографической областях [1-7]. Сервис Google Scholar по запросу «Search Queries» выдает 55 тысяч исследований, по сдвоенному запросу «Search Queries» and «Economy» – около 5 тысяч. Российских исследований среди них нет. Единственная работа написана в 2011-м году кандидатом экономических наук М. Столбовым: «Статистика поиска в Google как индикатор финансовой конъюнктуры» [8]. В докладе анализируются причины такого игнорирования российских аналитиков этой области исследований, показаны примеры работ и применяемые методы.

I. Мнения относительно сферы исследований поисковых запросов

1. Бесполезность исследований

Существует стойкое убеждение, что данные поисковых запросов хаотичны и не подходят для задач прогнозирования. Это не соответствует действительности. К примеру, в 2013-м году английский профессор Тобиас Прайз и его коллеги [9] отобрали 98 терминов, которые, по их мнению, могли прямо или косвенно влиять на биржевые индексы, и соотнесли динамику поисковых запросов по этим терминам с динамикой индекса DJIA (индекс Доу Джонса). Исследование показало, что рост поисковых запросов, имеющих отношение к финансовому миру, таких как «долг», «рынок», «акции» и т.д. ведет к падению рынка. Результатом работы стала интересная методика – если частота «финансовых» дескрипторов падает, нужно скупать акции и занимать «длинную позицию», если частота таких запросов растет – стратегия биржевой игры должна заключаться в «короткой продаже». В своем интервью порталу «Бизнес-инсайдер» ученый рассказал, что за период с 2004 по 2011-й год, пользуясь такой стратегией, можно было получить 326% прибыли, в то время как обычная традиционная игра принесла бы всего лишь 16%.

2. «Поисковые запросы» – это не серьезная область исследований

Часто встречается отношение к поисковым запросам как к излишней простой и недостойной серьезного аналитика области исследований. Предыдущий пример о

получении прибыли на бирже благодаря поисковым запросам является тому опровержением.

Чаще всего данные поисковых дескрипторов используются для задач прогнозирования. Поиск необходимых дескрипторов сопряжен с трудностями технического и аналитического плана. И тогда недостаточно просто зайти на сервис Google Correlate, ввести понятие интересующего объекта исследования или временной ряд его динамики, и получить список коррелируемых дескрипторов, который можно использовать для прогнозирования показателей, например, следующего месяца. На короткий период времени сервис, как правило, не выдает ни одного коррелятора. К примеру, на дескриптор «ищу квартиру в Москве», сервис Google Correlate не находит корреляторов даже и на длинные временные диапазоны, хотя другими способами для своей статьи мы их нашли около тысячи. Там, где сервис выдает данные, почти всегда это будут или ложные корреляции, или неприменимые на коротких горизонтах прогнозирования, которые наиболее интересны для исследования. В частности, на запрос «оружие» мы получим 90 корреляторов, но это будут незначительные дескрипторы «пвп», «скачать игру бесплатно», «Хемингуэй» и т.д. Понятно, что речь идет об играх и книгах, и такие запросы не имеют никакого отношения к подпольному рынку оружия. Сдвоенный запрос «купить оружие» сервис оставит без внимания на любых диапазонах, зато сервис Google Trends покажет динамику таких запросов (рис. 1, 2).

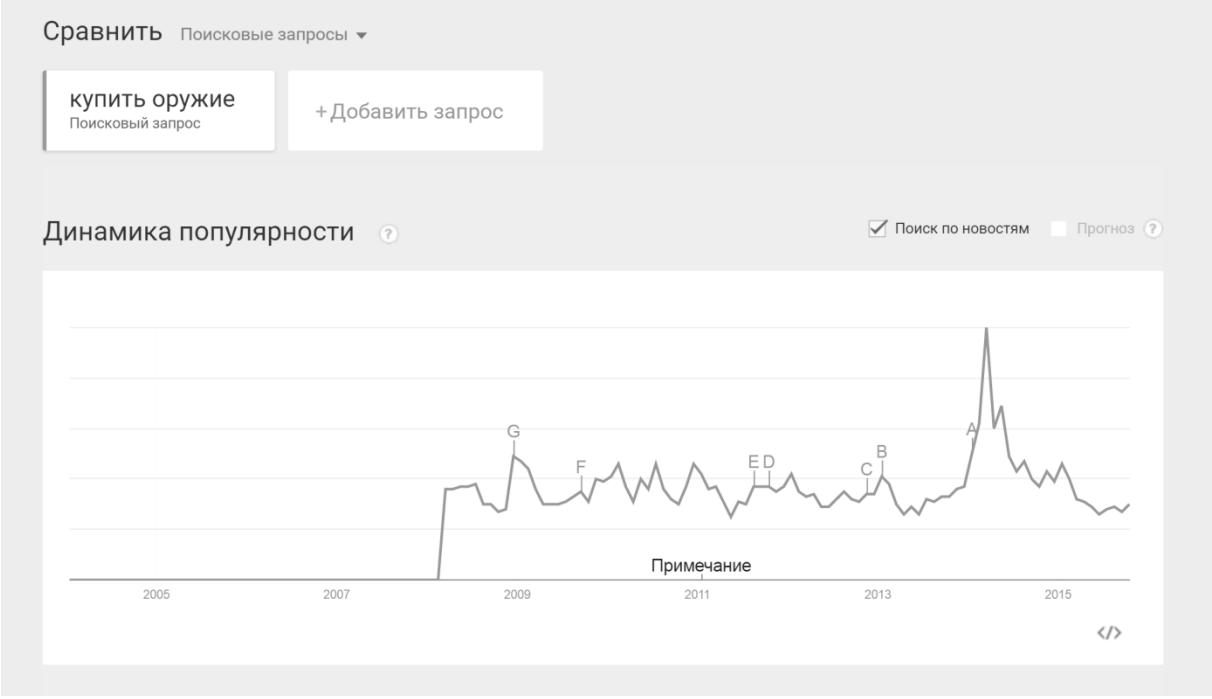

Рис. 1. Динамика поискового запроса «купить оружие» на территории Украины, максимум=100

Рис. 2. Карта поискового запроса «купить оружие»

Несложно заметить, что статистика этого запроса связана с актуальными событиями в обществе. Можно предположить, что должны быть и другие корреляторы.

Но, даже получив другими способами действительные подходящие базы дескрипторов, мы просто получим результат, действующий на исследуемом отрезке времени. Найти набор поисковых запросов, который давал бы приемлемые и устойчивые прогнозные результаты на будущие периоды, достаточно сложно. Для этого существует множество методик, которые регулярно анализируются, улучшаются, комбинируется аналитиками. Кроме того, не до конца решена проблема наличия информационного шума. В целом иногда на поиски устойчивой прогнозной модели уходит несколько месяцев сбора динамик, опытов и экспериментов.

3. Не нужны другие исследования

Существует убеждение, противоречащее вышеупомянутым. Оно состоит в том, что любой процесс, динамика которого отражена в каких-то официальных индикаторах, можно спрогнозировать с помощью поисковых запросов, и, соответственно, одной работы на эту тему достаточно. Это не так. Не все индикаторы поддаются прогнозированию с помощью поисковых запросов. Например, цены на полезные ископаемые или цены на аренду коммерческой недвижимости почти невозможно спрогнозировать описываемыми методами, а индексы потребительских цен, уровни рождаемости или безработицы прогнозируются с высокой долей точности. Качественные прогнозные модели строятся там, где есть высокая активность пользователей, а формирование индикатора не регламентировано. Существуют показатели, к которым аналитики не могут подобрать

подходящую методику, но причины этого пока не ясны. Например, не прогнозируются продажи легковых автомобилей, хотя динамика продаж грузовых автомобилей прогнозированию с помощью поисковых запросов вполне поддается [11]. Поиск показателей, которые поддаются подобным исследованиям – отдельное направление в этой области.

4. Неточные данные по индикаторам

Часто встречается убеждение, что данные некоторых индикаторов корректируются, и поэтому не подходят для прогнозирования. Это частично верно, но не отменяет получаемые результаты исследований. Как правило, сама корректировка этих индикаторов происходит однотипно из месяца в месяц, и подчиняется своим правилам. Но, действительно, при построении прогнозных моделей, можно предположить, где официальные индикаторы подвергались корректировке, особенно, если сравнивать данные по регионам.

II. Инструменты и методы

1. Сервисы, используемые аналитиками для исследования динамик поисковых запросов

Для сбора данных, прежде всего, используются статистический сервис Яндекса[1], расширенный поиск Яндекса[2], сервис Google Тренды[3], Google Correlate[4], Google AdWords[5] и сервис анализа рынков[6].

Сервисы предоставляют данные с разной периодичностью и с разными временными лагоми, но все они, как правило, недружелюбны к аналитикам и к автоматическим программам сбора информации. К примеру, на статистическом сервисе Яндекса, чтобы получить необходимые цифры, требуется 3 раза ввести капчу[7].

[1] Статистический сервис поисковых запросов Яндекса http://wordstat.yandex.com/
[2] Расширенный поиск Яндекса http://yandex.ru/search/advanced
[3] Web-приложение корпорации Google, показывающее, как часто определенный термин ищут по отношению к общему объему поисковых запросов в различных регионах мира www.google.ru/trends/
[4] Сервис, позволяющий отслеживать зависимости между исследуемыми дескрипторами и остальными поисковыми запросами или произвольной функцией, описывающей какую-то тенденцию, заданную в форме данные www.google.com/trends/correlate/
[5] Сервис контекстной рекламы компании Google www.google.com/adwords/
[6] Сервис исследования рынков http://translate.google.com/globalmarketfinder/g/index.html?locale=ru
[7] Капча (CAPTCHA) — от англ. Completely Automated Public Turing test to tell Computers and Humans Apart — компьютерный тест, используемый для определения, является ли пользователь системы человеком или автоматической программой

2. Методы исследования поисковых запросов

2.1 Анализ количества запросов в абсолютном выражении

Несмотря на кажущуюся простоту метода, он очень эффективен для сравнения полученных данных, например, по регионам.

В первых трех месяцах 2015 года в семи регионах Российской Федерации был зафиксирован повышенный рост смертности, на 5.2%[8]. Для прояснения ситуации была создана комиссия Росздравнадзора и Федерального фонда обязательного медицинского страхования.

Как можно использовать динамику поисковых запросов для оперативного мониторинга ситуации?

Были отсортированы имеющиеся базы динамик по стандартному коэффициенту корреляции Пирсона с динамикой «Число зарегистрированных умерших» по данным Росстата и выбраны все положительно коррелирующие дескрипторы. Дополнительно была создана так называемая «медицинская база». В нее вошли специфические медицинские термины, алфавитный перечень симптомов, алфавитный перечень лекарств, находящийся в открытом доступе, а также дескрипторы, относящиеся к области взаимодействия жителей с медицинскими учреждениями, такие как «телефон дежурного врача», «больничный», «вызов скорой» и т.д. Далее была собрана динамика запросов по этим семи регионам по всем выбранным дескрипторам. Затем были отсортированы полученные данные по проценту превышения количества запросов в сравнении с общероссийскими показателями. Так, например, в Ленинградской области в январе 2015 г. было выявлено превышение запросов по дескрипторам:

[8] Статья на электронном портале «Свободная пресса»: «Скворцова рассказала о повышенном росте смертности в семи регионах РФ» http://svpressa.ru/society/news/123271/

– потеря вкусовых ощущений – 208%;
– дежурный врач – 205%;
– полный пульс – 191%;
– кашель с желтой мокротой – 173%;
– онемение шеи – 170%;
– нечувствительность – 170%;
– ночная потливость – 169%;
– стерильные бинты – 167%;
– абстинент – 153%;
– дежурная больница – 155%;
– вертиго – 154%;
– горький вкус во рту – 152%;
– телефон аптеки – 146%;
– приемный покой – 146%;
– маниакальная фаза – 145%;
– нафтизин – 143%;
– кровотечение из ушей – 138%;
– вызвать врача 138%;
– эфералган – 137%;
– лекарства купить – 136%;
– свистящее дыхание – 134%.

По остальным базам значительное превышение поисковых запросов по Ленинградской области зафиксировано в дескрипторах:

– водоем – 386% превышения;
– рост-падение – 176%;
– самоизоляция – 145%;
– сальдо отрицательное – 137%;
– дефляция – 136%.

Можно заметить, что дескрипторы, процент которых превышает общероссийские показатели, касаются в целом экономической ситуации в стране, также выделяется дескриптор «водоем», который оказался специфическим именно для Ленинградской области.

Мониторинг и анализ динамики поисковых запросов может дать важный материал для выяснения причин аномальных явлений в социальной и демографической сферах.

2.2 Исследование тематических баз по процентному отношению поисковых запросов с корреляцией выше 0.7 с различными индикаторами

В процессе исследований был обнаружен интересный эффект, ранее не отмеченный в работах других исследователей. Он появляется только при использования в исследованиях разных тематических баз, в технических базах такого эффекта не наблюдается. Были собраны динамики временных рядов дескрипторов 7-ми тематических баз. Динамики поисковых запросов были ранжированы по значениям коэффициента корреляции Пирсона с 15-ю индикаторами[9]. Была создана выборка всех поисковых запросов с корреляцией более 0.7 по 15-ти разным индикаторам и исследовано их количество в процентном

[9] Данные индикаторов взяты на сайте Госкомстата http://www.gks.ru/

отношении в различных базах. По результатам было сделано много интересных выводов и наблюдений. Например, на инфографике (рис. 3) можно наглядно наблюдать, какие индикаторы более значимы в базе поисковых запросов «Бренды и товары» (расшифровка индикаторов в «прил. 1»).

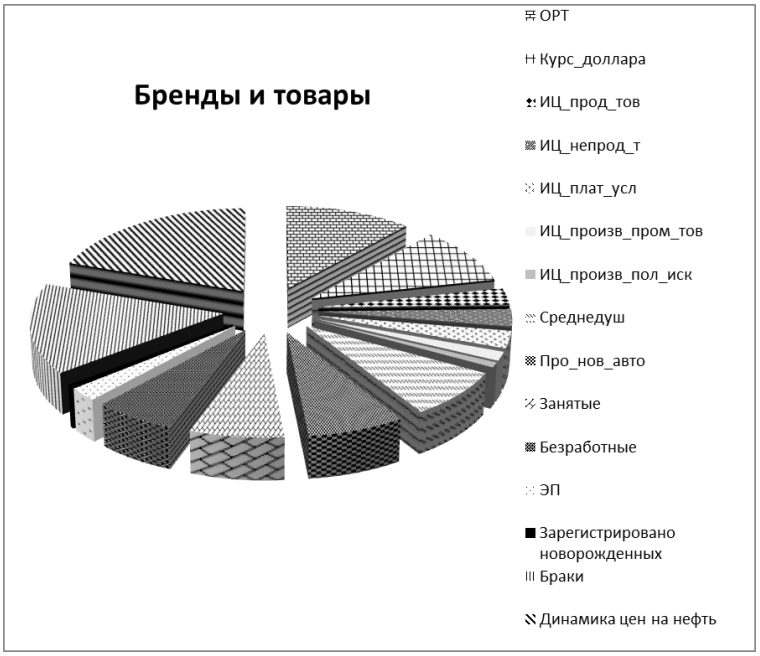

Рис. 3. База «Бренды и товары» и процентное соотношение в ней количества поисковых запросов, с корреляцией выше 0.7 по разным индикаторам

Так, например, можно отметить, что в базе «Бренды и товары» больше представлено поисковых запросов с корреляцией выше 0.7 по индикатору «зарегистрировано браков» (15.7%), чем по индикатору «зарегистрировано новорожденных» (всего лишь 0.4%). Можно предположить, что молодожены больше интересуются покупками, чем молодые родители. Мы привели этот пример, как самый очевидный, но подтверждающийся методикой исследования. Проводя подобные исследования, можно выявлять, какие слои населения интересуются какими именно брендовыми товарами. Подобная информация была бы интересна тем, чья деятельность связана с торговлей.

2.3 Прогнозирование

Основная область исследования поисковых запросов – это, все же, прогнозирование. Существует очень много методов построения моделей, например, с разными видами препроцессинга переменных, с включением авторегрессии, с учетом сезонности и календарных эффектов[10] и т.д. Как правило, западные исследователи в своих работах

[10] Календарные эффекты – это отклонения, связанные с определенными предсказуемыми календарными событиями, такими как праздничные дни, количество рабочих дней за месяц, високосность года и т.п. http://www.bseu.by/russian/faculty5/stat/docs/4/EconometricsBook3.pdf

используют стандартный регрессионный анализ. Метод Группового Учета Аргументов [12], реализованный в системе GMDH Shell, дает результаты намного более точные.

Для получения прогнозных моделей, как правило, применяется лагирование поисковых запросов. Формируются «барометры» динамических рядов с лагами в 1, 2, 3 и более дней/недель/месяцев. Анализируется поисковая активность, предшествующая исследуемому периоду. В 2011-м году Хьюньянг Чой и Хэл Вэриан [11] в своей статье «Прогнозирование с помощью Google Trends» проанализировали взаимосвязь поисковых запросов и динамики продажи автомобилей и запчастей. Использование временного лага в 4-5 недель раскрыло значительное соответствие динамики продаж и динамики поисковых запросов. Авторы наглядно показали, что данные поисковых запросов Google по автомобилям положительно коррелируют с продажами спустя несколько недель (рис. 4).

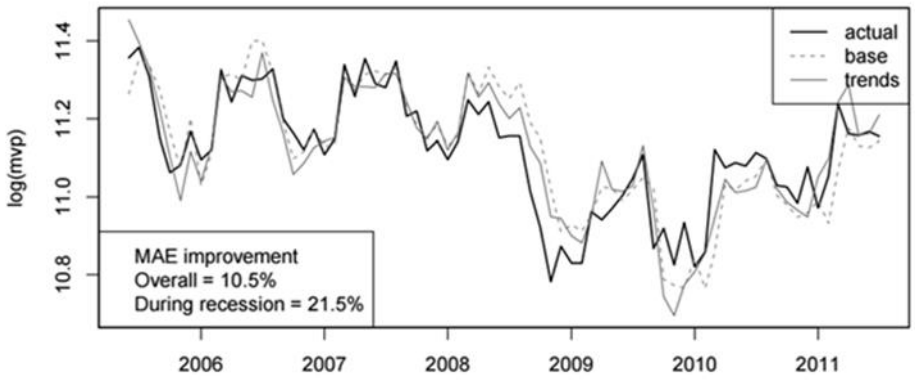

Рис. 4. Поисковые запросы (черная линия) и объемы продаж автомобилей и запчастей спустя несколько недель (серая линия)

В своем бакалаврском дипломе [13] я описываю построение 52 прогнозных моделей: по 4 модели для 13 индикаторов, с использованием разных алгоритмов прогнозирования (комбинаторного и нейронного) [14], с применением «барометров» и без, с разными формами препроцессинга переменных. На графике (рис. 5) приведен пример построения прогнозной модели без барометров для индикатора «Курс доллара к рублю, в среднем за период».

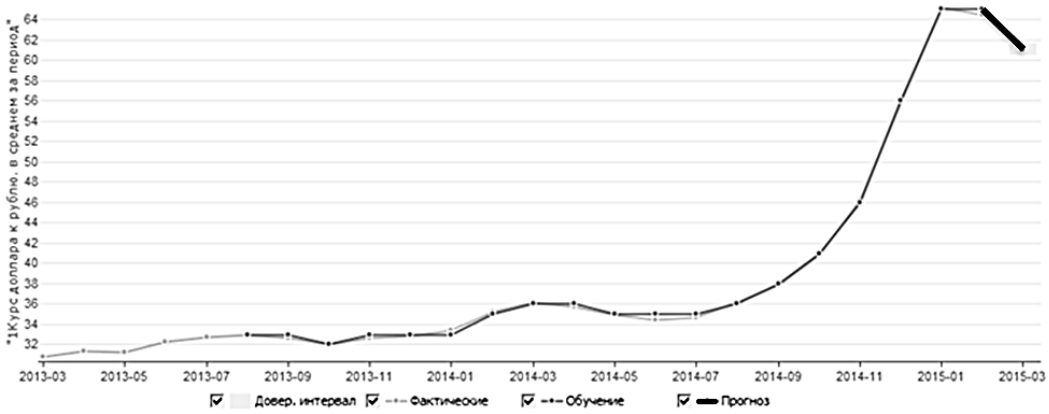

Рис. 5. Прогнозная модель. Индикатор – «Курс $», алгоритм – нейросеть, без корней. MAPE=0.64%. Прогноз на март 2015: 61 рубля/$. Реальное значение: 60 рублей/$

Система уравнений, характеризующая данную модель:

$$Y1[t] = 0.169144 - цена\ лицензии[t-2]*0.000448188 + N2*1.02275$$
$$N2[t] = -0.370442 + новый\ выпуск[t-2]*2.43664e-05 + N3*0.99092$$
$$N3[t] = -2.18959 + недомогание[t-4]*0.00161548 + N4*0.989241$$
$$N4[t] = -0.0541967 + N6*0.558507 + N8*0.442858$$
$$N8[t] = -2.61279 + паф[t-4]*0.00184835 + N9*0.964146$$
$$N9[t] = 16.7458 + фул[t-2]*0.00110923 + иов[t-3]*0.00193864$$
$$N6[t] = -0.109233 + N11*0.509554 + N12*0.493197$$
$$N12[t] = 1.17085 + игги[t-3]*0.00351227 + условия[t-1]*1.62631e-05$$
$$N11[t] = 25.2778 + выг[t-4]*0.00498369 + папироска[t-2]*0.00878033$$

В трех из четырех прогнозных моделей, оценивающих курс доллара к рублю, с высокой точностью было предсказано неожиданное падение доллара в марте 2015.

Выводы

Исследование поисковых запросов с целью прогнозирования или анализа экономических, социальных, демографических показателей представляется очень перспективным направлением для ученых. Такие исследования могут быть использованы:

- Для мониторинга экономического состояния в регионах в режиме реального времени. При этом появляется возможность обойти сложности получения данных, например, из-за финансовых ограничений;
- Для параллельного контроля достоверности официальной информации. Это позволяет выявить искажения, которые могут вносить официальные органы, например, в данные, касающиеся миграции, безработицы, доходов населения;
- Для возможности мониторинга ситуации в случае полного или частичного отсутствия признанной официальной методологии;

- Для прогнозирования экономических, демографических, социальных показателей в кризисные периоды. В настоящее время качественный анализ этих показателей проводится с использованием авторегрессии, которая неэффективна в кризисные периоды;
- Для прогноза динамики различных процессов в других странах, не опираясь на текущие данные и данные прогнозов, публикуемые официальными органами в открытом доступе с большим опозданием;
- Для оценки сведений, которые могут получить исследователи других стран о России, если они будут иметь доступ к запросам отечественных поисковых машин;
- Для создания соответствующих бизнес-приложений, для отслеживания падения или наращивания спроса на определенные виды товаров;
- Для сбора и анализа оперативной информации для использования ее на рынке акций.

Литература

1. *Nikolaos Askitas, Klaus F. Zimmermann* Google econometrics and unemployment forecasting. – Applied Economics Quarterly. – 2009. – V. 55. – № 2. – P. 107-120.
2. *Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen [et al.]* Web search queries can predict stock market volumes [Электронный ресурс]. – PLoS One. – 2012. – Режим доступа:
   URL: http:// journals.plos.org/plosone/article?id=10.1371/journal.pone.0040014
3. *Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, Nikolaos Aletras* Studying User Income through Language, Behaviour and Affect in Social Media [Электронный ресурс]. – PLoS One. – 2015. – Режим доступа:
   URL: http:// journals.plos.org/plosone/article?id=10.1371/journal.pone.0138717
4. *Michael Ettredge, John Gerdes, Gilbert Karuga* Using web-based search data to predict macroeconomic statistics. – Commun. ACM. – 2005. – T. 48. – № 11. – C. 87-92.
5. *Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski [et al.]* Detecting influenza epidemics using search engine query data. – Nature. – 2009. – T. 457. – № 7232. – C. 1012-1014.
6. *Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, Duncan J. Watts* Predicting consumer behavior with Web search // Proc. Natl. Acad. Sci. USA. – 2010. – T. 107. – № 41. – C. 17486-17490.

7. *Ladislav Kristoufek* Can Google Trends search queries contribute to risk diversification? – Sci. Rep. – 2013. – Т. 3. – № 2713.
8. *Столбов М.И.* Статистика поиска в Google как индикатор финансовой конъюнктуры. – Вопросы экономики. – 2011. – № 11. – С. 79-93.
9. *Chester Curme, Tobias Preis, H. Eugene Stanley, Helen S. Moat* Quantifying the semantics of search behavior before stock market moves // Proc. Natl. Acad. Sci. USA. – 2013. – Т. 111. – № 32. – С. 11600-11605.
10. *Hyunyoung Choi, Hal Varian* Predicting the Present with Google Trends. – The Economic Record. – 2012. – Т. 88. – Issue Supplement № 1. – С. 2-9.
11. *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем – Киев: Наукова думка, 1981. – 296 с.
12. *Болдырева А.В.* Построение прогнозных моделей экономической конъюнктуры и преступлений экономической направленности по интенсивности запросов в поисковой системе Интернет. – 2015. – ВКР. – РАНХиГС. – 109 с.
13. *Степашко В.С.* Метод критических дисперсий как аналитический аппарат теории индуктивного моделирования. – Проблемы управления и информатики. – 2008. – № 2. – С. 8-26.

Приложения

Приложение 1

Расшифровка индикаторов:

ОРТ – оборот розничной торговли (млн.руб);
Курс доллара – средневзвешенный курс доллара к рублю, в среднем за месяц;
ИЦ_прод_тов – индекс потребительских цен на продовольственные товары;
ИЦ_непрод_т – индекс потребительских цен на непродовольственные товары;
ИЦ_плат_усл – индекс потребительских цен на платные услуги населению;
ИЦ_произв_пром_тов – индекс цен производителей промышленных товаров;
ИЦ_произв_пол_иск – индекс цен производителей полезных ископаемых;
Среднедуш – среднедушевые денежные доходы населения (тыс. руб);
Про_нов_авто – продажи новых легковых и легких коммерческих автомобилей в России (штук);
Занятые – численность занятых в экономике (тыс. чел.);

Безработные – численность безработных, зарегистрированных в государственных учреждениях службы занятости населения (тыс. чел.);

ЭП – зарегистрировано преступлений экономической направленности;

Зарегистрировано новорожденных – число зарегистрированных новорожденных (тыс.чел);

Браки – число зарегистрированных браков (тыс. чел);

Нефть – динамика цен на Нефть Brent (ICE.Brent), USD/баррель;