УДК 681.513.8

ПРОГНОЗИРОВАНИЕ ЭПИДЕМИОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ ГРИППА НА ОСНОВЕ ЗАПРОСОВ ПОИСКОВОЙ СИСТЕМЫ

Е.А. Настенко[1], Е.К. Носовец[1], О.К. Белошицкая[1], Ан.В. Павлов[2], А.Д. Соколов[1], Ал.В. Павлов[1]

[1]Национальный технический университет Украины Киевский политенический институт им. Игоря. Сикорского

[2]Международный научно-учебный центр информационных технологий и систем НАНУ и МОНУ

nastenko@inbox.ru,e.nosovets@ya.ru,ksenia,bil@ukr.net,andriypavlove@gmail.com, sokolov.alexey1994@gmail.com,cheshirelk@gmail.com

В роботі розглянуто можливість підвищення ефективності прогнозування медико-біологічних процесів за рахунок використання даних пошукових запитів. Для моделювання використаний релаксаційний алгоритм методу групового урахування аргументів. Отримано моделі прогнозу епідеміологічних показників грипу на основі фактичного звернення пацієнтів до лікувального закладу і з доповненням даних запитів пошукової системи Yandex. У другому випадку модель показала істотне поліпшення властивостей прогнозування.

Ключові слова: прогнозування, пошукові запити, ресурс Yandex, грип, епідеміологічні показники, релаксаційний алгоритм, метод групового урахування аргументів (МГУА).

The paper considers the possibility of increasing the forecasting medical and biological processes effectiveness through the use of search query data. For modeling used relaxation algorithm of group method of data handling. Get model forecast influenza epidemiological indicators on the basis of the actual treatment of patients in the hospital and with the addition of data the Yandex search engine queries. In the second case, the model showed a significant improvement in predictive properties.

Keywords: forecast, search queries, resource Yandex, flu, epidemiological indicators, relaxation algorithm. group method of data handling GMDH.

В работе рассмотрена возможность повышения эффективности прогнозирования медико-биологических процессов за счет использования данных поисковых запросов. Для моделирования использован релаксационный алгоритм метода группового учета аргументов. Получены модели прогноза эпидемиологических показателей гриппа на основе фактического обращения пациентов в лечебное учреждение и с дополнением данных запросов поисковой системы Yandex. Во втором случае модель показала существенное улучшение прогнозирующих свойств.

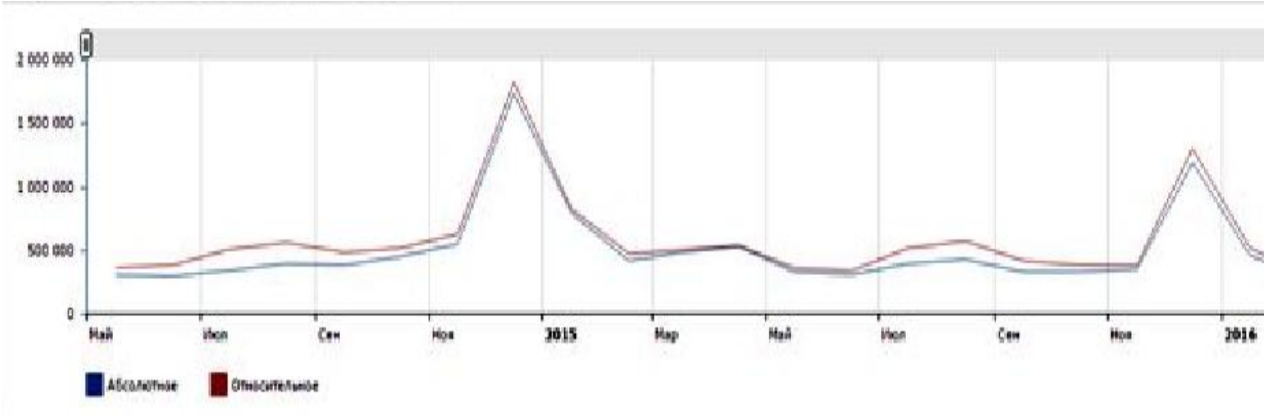
Ключевые слова: прогнозирование, поисковые запросы, ресурс Yandex, грипп, эпидемиологические показатели, релаксационный алгоритм, метод группового учета аргументов (МГУА)

Введение

Статистические данные сформированные по пользовательским запросам поисковых систем характеризуют одну из форм реакции интернет сообщества на происходящие в реальности процессы, события, явления. Представляет оп-

ределенный интерес исследовать, в какой степени указанные данные могут способствовать повышению эффективности прогнозирующих систем для предсказания экономических, социологических, медицинских показателей, которые достоверно характеризуют процессы в экономических, социальных, биологических объектах. Ряд современных исследований показали возможность открытия поисковыми системами доступа к информационным ресурсам, предоставляющим возможность построить более точные прогнозирующие модели, а в некоторых случаях появляется возможность дополнительного многостороннего анализа для определения новых свойств и структурных связей событий, процессов и явлений.

В 2009 году опубликованы работы по прогнозу показателей микро- и макро- состояний экономики по динамике запросов в Интернете [1]. Исследователи применили для прогнозирования данные поисковых сервисов предполагая наличие существенной корреляции между текущим уровнем деловой активности и динамикой соответствующих запросов пользователей поисковиков. Результаты моделирования подтвердили предположение авторов на примере прогнозов динамики продаж автомобилей Ford, Chevrolet и Toyota на рынке США и туристического потока в Гонконге. Полученные результаты моделирования с включением переменных, учитывающих колебания запросов в Google превосходили по критериям оценки качества прогноза во всех вариантах примененных выборочных данных. Впоследствии эффективность применения данных поисковых запросов для мониторинга и прогнозирования деловой активности была подтверждена в десятках опубликованных работ. В качестве примера приведем работу Столбова [2]. Он исследовал применения сервисов Google для оценки текущего финансового состояния рынков и показал, что при удачном формировании выборки дескрипторов есть возможность получать более точный результат прогноза чем при применении только опросных методов. Также интересна работа [3], где примененные дескрипторы позволяют определить структурные особенности экономической преступности. Наглядным примером связи динамики праздничной активности населения и соответствующего поискового запроса является отражение приготовления украинцев к Новому Году (скачки перед январем 2015 и 2016 гг.) в динамике запросов слова "салат".



Таким образом есть все основания предполагать полезность применения данных поисковых запросов для повышения точности прогнозирующих моделей и в других областях. Ниже рассмотрено применение статистики поисковых запросов для прогнозирования обращения населения в лечебные заведения по поводу заболевания остро-респираторной вирусной инфекцией и гриппом.

Постановка задачи

Целью настоящей работы является исследование информативности данных поисковых запросов с точки зрения повышения точности прогнозов динамических процессов. Рассматривается один из наиболее часто встречающихся в природе заболеваний человека эпидемиологический процесс - процесс заболевания (суммарный) остро-респираторной вирусной инфекцией и гриппом, наблюдаемый в некотором ограниченном регионе. В данном конкретном случае мы не выделяем для прогнозирования определенный штамм гриппа, а пользуемся значением зарегистрированного суммарного обрашения в поликлинику в одном из районов города Киева по поводу ОРВИ и гриппа. График значений средних за месяц соответствующего ряда приведен на рис.1.

Рис.1. График обращений в поликлинику по поводу гриппа и ОРВИ

Значения на графике отображают числа первичных обращений населения в поликлинику на протяжении 18 месяцев в 2014-2015 гг.

Требуется построить прогнозирующую модель с наилучшими показателями на экзаменационном подмножестве. В качестве такого подмножества выделены 2 последние точки выборки - значение ряда в августе и сентябре 2015г.

Потребителем нашего прогноза может быть администрация соответствующего лечебного заведения, использующего его для планирования графика работы участковых и дежурных терапевтов и опережающей закупки лечебных препаратов в стационар.

Кроме того в работе необходимо получить ответ о возможности улучшения прогнозирующих свойств модели за счет дополнительных переменных - статистики реализации поисковых запросов на интересующую нас тему. Данные моделирования должны быть использованы для аргументации эффективности применения статистики запросов поисковых систем для прогнозирования динамических процессов.

Основная часть.

Расчет модели прогноза по значениям исходного ряда. Ввиду достаточно малого количества значений имеющегося ряда в качестве инструмента создания прогнозирующей модели выберем алгоритмы метода группового учета аргументов МГУА. Аргументация целесообразности и преимущества применения данного подхода при малых объемах выборки обоснованы в многочисленных работах [4,5]. Модели прогноза ниже рассчитаны программной реализацией алгоритма РИА - релаксационного итерационного алгоритма МГУА [7]. В качестве аргументов используем значения ряда с четырьмя запаздываниями. Ниже приведены модели, графики исходного ряда, прогноза, и среднее абсолютное отклонение MAD на рабочей выборке и на экзамене, нормированные на коридор вариации всей выборки - Mean Relative Derivation;

$$MRD_{раб} = 100 / n_{раб} \frac{\sum_{i=1}^{n_{раб}} |y_i - \hat{y}_i|}{\frac{2}{n}\sum_{1=1}^{n} |y_i - \bar{y}|} \qquad MRD_{экз} = 100 / n_{экз} \frac{\sum_{i=1}^{n_{экз}} |y_i - \hat{y}_i|}{\frac{2}{n}\sum_{1=1}^{n} |y_i - \bar{y}|}$$

Здесь $n, n_{раб}, n_{экз}$ - количество точек всей выборки, рабочей выборки и экзамена соответственно, $y_i, \hat{y}_i$ $\bar{y}$ - значения исходного ряда, модели и среднее ряда соответственно.

Модель 1, полученная РИА МГУА на четырех запаздываниях:

$$y(t+1) = 177.23596 + 3.5875\,E001 * y(t)/(y(t-1))^{-2/3} - 2.3444\,E003 * y(t-2)/y^2(t-2) -$$
$$- 7.5672\,E000 * (y(t))^{4/3}/y(t-3) - 3.1234\,E002 * y(t-2)/(y(t-1)y^{-1/3}(t-3)) - 2.8897\,E003 *$$
$$* y(t)/(y(t-1) * y(t-2))$$

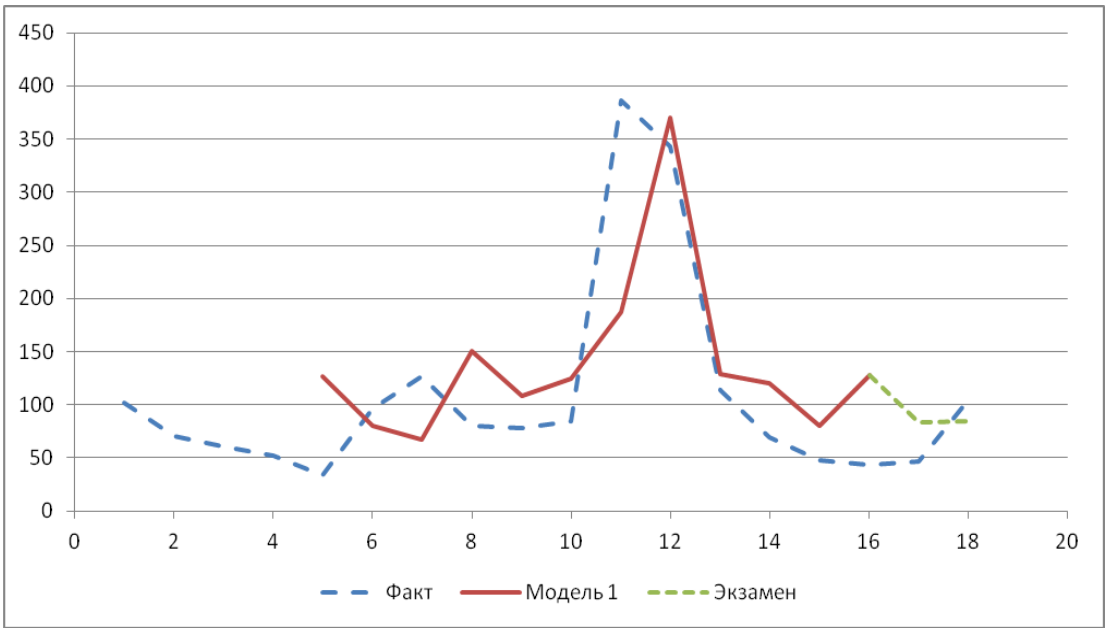


Рис.2 Графики исходного ряда и модели 1 РИА на рабочих точках и экзамене.

$MRD_{раб\ РИА_1} = 7{,}51\%$ $MRD_{экз\ РИА_1} = 23{,}93\%$

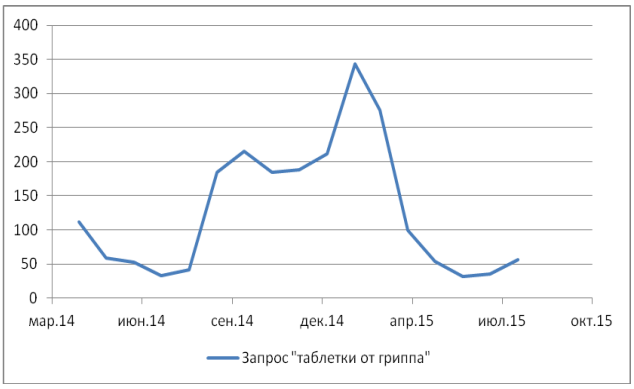
Расчет модели прогноза с привлечением статистики запросов поисковой системы Яндекс.

Данный материал служит цели привлечь внимание специалистов к источнику информации, использование которого способно повысить эффективность прогнозирующих систем.

В нашем случае воспользуемся возможностями инструмента ***Yandex Wordstat*** позволяющего получить доступ к статистике запросов по ключевым словам. Для удобства представления данных предложен программный продукт для обработки запросов, куда импортируется статистика ресурса Yandex.

В сервисе программы выбор региона Украины, где учтены поисковые запросы, выбор разбиения по месяцам или неделям, рассчитываются характеристики связи рядов.

Ниже представлены графики рядов статистики запросов по ключевым словам "таблетки от гриппа", "грипп", "ОРВИ", полученные для месячного периода усредненния данных по городу Киеву за время с апреля 2014г. по август 2015г.



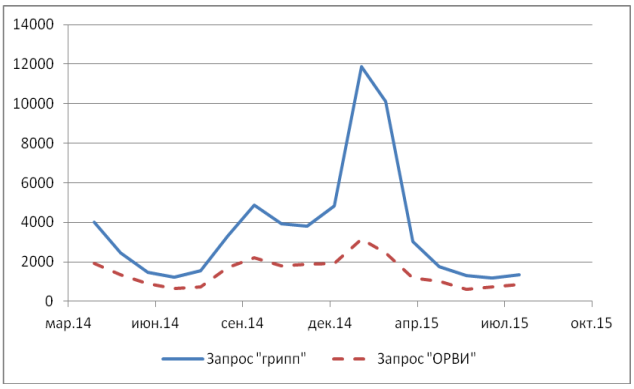


Рис.3 График запросов "таблетки от гриппа" Рис.4 Графики запросов "грипп" и "ОРВИ"

Далее при моделировании будут использованы статистики именно данных запросов так как их корреляция с исходным рядом оказалась наиболее высокой. Обозначения, принятые в модели: x_2 - переменная статистики запроса "грипп", x_3 - переменная статистики запроса "ОРВИ" x_4 - переменная статистики запроса "таблетки от гриппа".

Модель2, полученная РИА МГУА на четырех запаздываниях:

$$y(t+1) = 32.49466 + 2.8912\,E000 * x_4(t) * y^{-1/3}(t-3) / y^{-1/3}(t-1) - 4.5879\,E001 * x_3(t) *$$
$$* y^{-1/3}(t-2) / x_3(t-2) + 7.2297\,E009\,/(x_3(t-2) * x_2(t-1)y(t)) - 1.5572\,E000 * x_3(t-2) /$$
$$/(x_4^{-1/3}(t) * y(t-1)) - 1.7566\,E-006 * y^3(t)$$

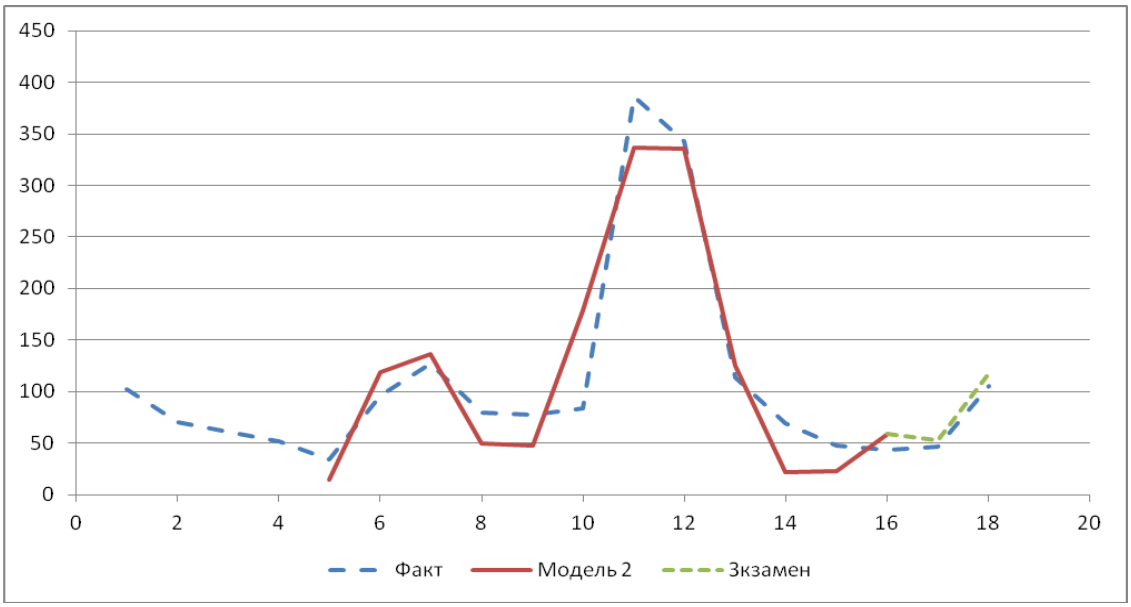


Рис.5 Графики исходного ряда и модели2 РИА на рабочих точках и экзамене.

$MRD_{раб\ РИА1} = 2{,}94\%$ $MRD_{экз\ РИА1} = 8{,}02\%$

Выводы

Получены модели прогноза обращения населения в лечебное учреждение по поводу заболеваний ОРВИ и гриппа с хорошим качеством на экзаменационной выборке. Показано существенное улучшение прогнозирующих свойств модели при учете статистики запросов поискового ресурса.

Литература

1.Choi, H., Varian, H.: Predicting the Present with Google Trends. // Working paper, Google Inc., 2009

2. Столбов, М.И.: Статистика поиска в Google как индикатор финансовой конъюнктуры // Журнал "Вопросы экономики" 2011, №11

3. Болдырева А. Построение прогнозных моделей экономической конъюнктуры и преступлений экономической направленности по интенсивности запросов в поисковой системе Интернет. // Выпускная квалификационная работа, Москва, 2015

4. Ивахненко А, Степашко В. Помехоустойчивость моделирования. // Киев: Наук. думка. 1985

5. Степашко В. Индуктивное моделирование в исторической перспективе. // Труды 4-й межд. конф. по индуктивному моделированию (ICIM-2013), НАНУ, Чешский технический университет, Киев, 2013, сс.31-37

6. Интернет-ресурс *https://www.gmdhshell.com/*

7. Павлов.В.А. Синтез нечутливих до зсуву нелінійних моделей [Текст] / В.А. Павлов, О.В. Павлов // Комп'ютерні системи та мережі.[Зб. наук. праць] - Л.: Видавництво Національного університету "Львівська політехніка"/ Вісник Національній університет "Львівська політехніка", 2009. - №658. - С. 111-115.