Миллионы публикаций, выкладываемых пользователями в сеть ежедневно, невозможно обработать вручную для определения и исследования общественного мнения. Данный факт выдвигает на первый план потребность в методах автоматизации анализа текстовой информации, позволяющих за короткое время обработать огромные объемы интернет-данных и понять позицию пользователей исходя из содержимого сообщений. Таким образом, необходимость развития методологического аппарата для решения комплекса задач автоматизации мониторинга общественного мнения является актуальной проблемой в сфере анализа данных.

Для получения эмоциональной окраски сообщения необходимо разрешить проблемы алгоритмического извлечения аспектов мнения и определения полярности (положительная, отрицательная или нейтральная) мнения текстового сообщения. Извлечение аспектов направлено на идентификацию целевых объектов, относительно которых и сформировано какое-либо мнение. Наиболее распространенные для этого механизмы – это метод обучения без учителя и статистический метод (выделение n-грамм), для которых не требуются размеченные тренировочные данные.

Задача определения полярности сообщения сводится к анализу её эмоциональной окраски по заранее определенным критериям со своей системой баллов. Наиболее часто процесс определения полярности происходит в два этапа: проверка на объективность/субъективность и непосредственно определение окраски исходного сообщения. Объективное предложение выражает фактическую информацию об объекте и не рассматривается с точки зрения определения полярности. Субъективное же, напротив, чаще всего содержит мнение автора об объекте, что и является основным источником анализа. Существует несколько наиболее распространенных методов оценки полярности сообщения: алгоритмы обучения без учителя, метод опорных векторов (SVM), наивный байесовский классификатор, словарный метод и метод максимальной энтропии.

Из вышесказанного видно, что на данный момент существует большое число методов оценки эмоциональной окраски сообщений, что свидетельствует о том, что «серебряной пули» в данной области еще не найдено. Каждый приведённый метод может выигрывать у других в одной области применения и проигрывать в других. Постоянное совершенствование этих методов свидетельствует о том, что универсального варианта не существует.

Для улучшения качества оценки сообщений, в рамках своей научной работы в этой области, планируется использование комбинированного метода, состоящего из SVM и наивного байесовского классификатора. Объединение результатов работы отдельных методов должно увеличить точность распознавания, что будет проверено путем тестирования работы каждого алгоритма отдельно. Оба приведенных выше метода относятся к алгоритмам обучения с учителем, что требует заранее подготовленную размеченную обучающую выборку. Для извлечения аспектов из текста используется статистический метод извлечения n-грамм, содержащих только прилагательные и существительные обработанные C-value фильтром.

Список использованных источников:

1. Pang B, Lee L. Thumbs up? Sentiment Classification using Machine Learning Techniques. - Ithaca, 2002. – 353-362 p.
2. Bing Liu. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing. – Chicago, 2010. – 217 p.

МОДЕЛИ ПРОГНОЗИРОВАНИЯ И ОЦЕНКА ЗАВИСИМОСТЕЙ ФИНАНСОВЫХ ВРЕМЕННЫХ РЯДОВ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Савостин А.А.

Фролов И.И. – к.т.н.

Финансовые ряды постоянно находятся в поле зрения нашего внимания. К ним относятся новостные сообщения о значениях индекса фондового рынка, процентных ставках, ценах на электроэнергию и т.д. Существуют две основные цели исследования финансовых временных рядов [1]. Во-первых, важно понять, как ведут себя цены. Завтрашняя цена не определена, поэтому она описывается с помощью распределения вероятностей. Это означает, что статистические методы являются обычным способом исследования цен. Вторая цель – использовать знания о поведении цен для снижения риска либо принятия правильных решений.

Целью данной статьи является анализ и прогнозирование временных рядов. В статье поставлено две цели – первая по существующим данным определить степень влияния одних рядов на другие. Вторая – смоделировать изменение интересующего ряда с целью построения прогнозов.

В качестве анализируемых исходных факторов выбраны данные финансового рынка России, такие как товары, акции, облигации, цены на сырую нефть и др. В качестве наблюдаемых показателей – экономические показатели Беларуси.

Анализируется влияние исходных факторов на экономические показатели Беларуси (инфляция, ценовые индексы, курс рубля и др.). Одним из индексов цен, характеризующих уровень инфляции, является

индекс потребительских цен (Consumer Price Index, CPI).

Инфляция – процесс обесценивания денег, приводящий к повышению цен на большинство категорий продукции, не обусловленному улучшением ее качества. При инфляции расчете учитываются затраты на приобретение определенных товаров и услуг, формирующих так называемую потребительскую корзину благ. В ее состав включаются важнейшие статьи расходов: продукты питания, жилье, одежда, транспортные издержки, расходы на медицинские и образовательные услуги [2]. В качестве источника статистических данных для беларуси используются данные, предоставляемые национальным статистическим комитетом РБ.

Для оценки зависимости между двумя величинами рассчитывается коэффициент корреляции. На основании оценки корреляции отобраны факторы, наиболее сильно влияющие на оцениваемые параметры ($|r_{xy}| > 0.6$). Был выбран коэффициент корреляции Пирсона, который для двух выборок$x^m = (x_1, .., x_m)$ и $y^m = (y_1, .., y_m)$ рассчитываемый по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x, y)}{s_x^2 s_y^2},$$

где $\bar{x}$ и $\bar{y}$ – средние значения выборок, s_x^2 и s_y^2 – выборочные дисперсии,$r_{xy} \in [-1,1]$.

Корреляции рассчитываются на разные периоды: 1, 2, 3, 5, 10, 20 лет. Для прогнозирования на периоде 1-3 года используется регрессионная линейная модель [3], дающая приемлемые результаты. Для больших периодов даже визуально видно, что корреляция на большие промежутки времени невысокая (рис. 1), следовательно, для прогнозирования требуются более сложные модели с большим количеством параметров.

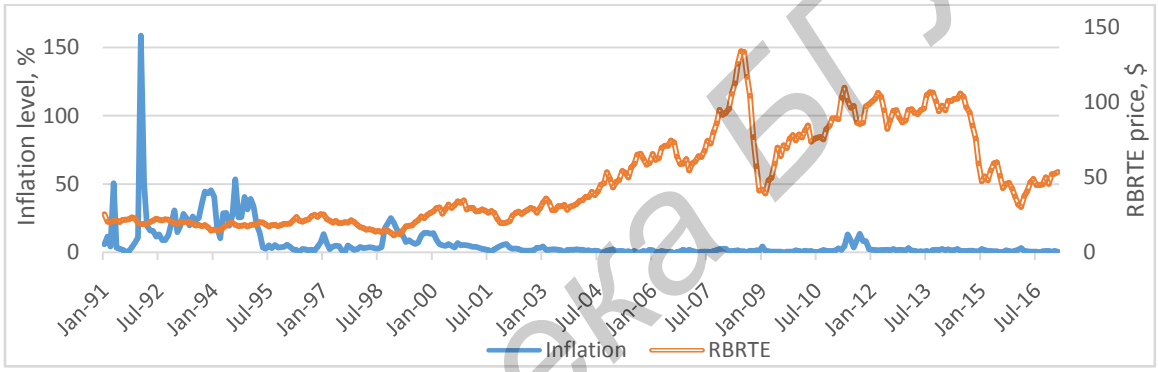

Рис. 1 Графики инфляции в Беларуси и цены на нефть с 1991 года.

Списокиспользованныхисточников:

1. Aas, K., Statistical modelling of financial time series: An introduction / K. Aas, Xeni K. Dimakos. NorwegianComputingCenter, Oslo, 2004.
2. Национальный статистический комитет Республики Беларусь. Официальная статистика. - Режим доступа: http://www.belstat.gov.by/ofitsialnaya-statistika/. - Дата доступа: 02.27.2017
3. Freedman D. Statistical Models: Theory and Practice / D. Freedman. CambridgeUniversityPress, NY, USA. 2009.

РАЗРАБОТКА ЧЕРЕЗ ТЕСТИРОВАНИЕ НА ПРИМЕРЕ НЕБОЛЬШИХ ПРИЛОЖЕНИЙ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Стрельцов Г. Ю.
Искра Н. А. – м.т.н., старший преподаватель

С каждым днём в сфере разработки программного обеспечения появляется всё больше и больше различного рода задач, концепций, методологий и требований. Одними из ключевых факторов, которые влияют на успех и качество разрабатываемого программного продукта, можно с уверенностью назвать выбор в пользу того или иного архитектурного решения, объём тестирования и скорость разработки. И чем больше и сложнее задача стоит перед разработчиками, тем больше внимания им приходится обращать на данные факторы. Цель данной исследовательской работы заключается в том, чтобы найти баланс между выбором архитектуры, объёмами тестирования и скоростью разработки в различных ситуациях, чаще всего встречающихся в разработке небольших приложений. Всё это позволит добиться большего качества и меньшей себестоимости программного продукта. Особый акцент сделан на тестирование. В данной работе оно