УДК 330.47

АЛГОРИТМ ИМПОРТА И ПЕРВИЧНОЙ ОБРАБОТКИ ДАННЫХ ПРИ АНАЛИЗЕ ЭКОНОМИЧЕСКОЙ БЕЗОПАСНОСТИ РЕГИОНОВ РОССИИ

Митяков Е.С., Митяков С.Н.

ФГБОУ ВПО «Нижегородский государственный технический университет им. Р.Е. Алексеева», Нижний Новгород, Россия (603950, ГСП-41, Н.Новгород, ул. Минина, д. 24), e-mail:nntu@nntu.nnov.ru

В статье предложен алгоритм импорта и первичной обработки данных при анализе экономической безопасности регионов России. На первом этапе алгоритма производится поиск исходной информации. Показано, что исходные данные, как правило, относятся к разряду слабоструктурированных. Приведен обзор типов и основных алгоритмов обработки слабоструктурированных данных. В ходе первичной обработки данных выявлены характерные особенности, на основе которых для обработки выбран метод последовательного перебора. На втором этапе производилось преобразование исходных данных в индикаторы экономической безопасности с учетом их размерностей. На третьем этапе осуществляется обработка и визуализация преобразованной информации с целью последующего анализа и прогнозирования. Для автоматизации анализа первичной информации разработана информационная система, состоящая из модуля импорта, модуля обработки, модуля редактирования данных, модуля визуализации.

Ключевые слова: экономическая безопасность, слабоструктурированные данные, импорт информации, обработка информации, информационная система.

ALGORITHM OF IMPORT AND PRIMARY DATA PROCESSING IN THE ANALYSIS OF ECONOMIC SECURITY OF REGIONS OF RUSSIA

Mityakov E.S., Mityakov S.N.

Nizhny Novgorod State Technical University n.a. R.E. Alekseev, Nizhny Novgorod, Russia, (603950, GSP- 41, Nizhniy Novgorod, Minin str., 24), e-mail: nntu@nntu.nnov.ru

In this paper developed an algorithm of import and primary processing of data for analysis of economic security of regions of Russia. The first stage is the search for primary information.Shown that the initial data, as a rule, classified as semi-structured. The paper gives an overview of the main types and processing algorithms with semi-structured data. During the initial processing of the data revealed their features. On their basis for the processing method used by one.At the second stage transformation of basic data to indicators of economic security taking into account their dimensions was made. At the third stage processing and visualization of the transformed information for the purpose of their subsequent analysis and forecasting was made. For automation analysis of primary information was developed information system, consisting of the importer, the processing module, module of data editing, visualization module.

Keywords: economic security, semi-structured data, import data, information processing, information system.

Анализ и обработка информации – одна из самых динамично развивающихся и актуальных областей. Это связано в первую очередь с развитием интернет технологий и, как следствие, с увеличением потоков различных данных. В официальных таблицах Росстата содержится более 300 показателей. Однако далеко не все из них полезны при анализе экономической безопасности хозяйствующих субъектов. В этой связи возникает актуальная задача поиска и первичной обработки необходимой информации для формирования статистических данных о динамике индикаторов экономической безопасности хозяйствующих субъектов на федеральном и региональном уровнях. Решение данной задачи способствует эффективному мониторингу экономической безопасности.

Основой мониторинга экономической безопасности является междисциплинарный подход, основанный на информационных моделях обработки информации, математических ме-

тодах анализа и прогнозирования экономических процессов. Анализ динамики индикаторов дает представление о состоянии систем экономической безопасности, позволяют субъектам хозяйствования определить как уязвимые сферы своей деятельности, так и сильные стороны. Одной из важнейших стадий мониторинга выступает сбор и обработки первичной информации. Данная стадия состоит из ряда этапов.

На *первом этапе* производится поиск информации, источником которой в большинстве случаев является официальный сайт Росстата. Говоря об извлечении данных из глобальной сети, можно выделить следующие их уровни:

- исходные данные – необработанные массивы данных, получаемые в результате наблюдения за некой динамической системой или объектом и отображающие его состояние в конкретные моменты времени;
- информация – обработанные данные, которые несут в себе некую информационную ценность для пользователя; сырые данные, представленные в более компактном виде;
- знания – представляют наибольшую ценность для пользователя, отображают скрытые взаимосвязи межу объектами, которые не являются общедоступными.

Исходные данные, которые удается найти, как правило, относятся к разряду слабоструктурированных. Это – нерегулярные, несогласованные данные, которые не имеют постоянной, четко определенной структуры, то есть их структура, тип и состав могут динамически изменяться.

Проблеме анализа слабоструктурированных данных посвящены многочисленные работы отечественных и зарубежных ученых. Так, М. Когаловский определяет слабоструктурированные данные как «данные, которые не имеют регулярной структуры, обладают динамичной (по отношению к экземплярам описываемых ею данных) схемой». Схема часто не представляется явным образом, поскольку данные являются самоописываемыми, и метаданные содержатся непосредственно в самих данных. Схема, если она задана явно, не является для таких данных предписывающей – один и тот же атрибут в разных экземплярах данных может иметь значения разных типов. Характерным примером слабоструктурированных данных являются гипертекстовые данные Web» [2]. Томас Коноллии, Каролин Бегг определяют слабоструктурированные данные как «данные, обладающие определенной структурой, но эта структура может оказаться непостоянной, недостаточно изученной или неполной» [6].

Значительный вклад в развитие направления анализа слабоструктурированных данных внес В.И. Левенштейн [7], который ввел понятие расстояния Левенштейна (функция Левенштейна, в теории информатики и компьютерной лингвистики является мерой разницы двух последовательностей символов (строк) относительно минимального количества операций вставки, удаления и замены, необходимых для перевода одной последовательности в дру-

гую). Практическим применением дистанции Левенштейна является определение сходства последовательностей символов. Это широко используется при обработке гибридных данных, поиске дубликатов, проверке текстовых данных на различного рода ошибки.

Проблематика обработки слабоструктурированных данных поднимается в работах Питера Бунемана. В статье [8] он описывает не только алгоритмы обработки, но и модели представления полуструктурированных массивов информации в базе данных.

Проблема обработки слабоструктурированных данных и поиска в них, несмотря на нарастающую актуальность и увеличение потоков данных, освещена и изучена в достаточной степени. Рассмотрим основные алгоритмы, которые можно использовать для решения подобных задач.

1. Современные методы обработки массивов слабоструктурированной информации в информационных, коммуникационных и управляющих системах на основе теории фильтров Калмана и Пугачева развиты в трудах И.Н. Синицына [5].

2. Альтернативным методом обработки массивов слабоструктурированных данных является использование генетических алгоритмов, которые представляют собой адаптивные методы поиска и используют прямую аналогию с «механизмом выживания».

3. В ряде случаев целесообразно использовать нейросетевое моделирование, которое предлагает в качестве распознаваемого образа различные по своей природе объекты: символы текста, изображения, образцы звуков.

4. Следующим классом алгоритмов, которые могут использоваться при обработке подобной информации, являются алгоритмы нечеткого поиска. Они имеют высокую ценность при выявлении плагиата, поиске и фильтрации спама, архивировании документов, но в нашем случае такие алгоритмы малоэффективны.

5. Еще одним методом для обработки слабоструктурированной информации является метод последовательного перебора. Простейшее решение задачи перебора состоит в последовательном сравнении, начиная с t(1) и p(1), символов и слов T и P до тех пор, пока не будет обнаружено равенство или неравенство сравниваемых символов. В последнем случае следует вернуться к началу сравнения и, сдвинувшись на один символ по тексту (теперь это будет t(2)), и повторить процедуру. Применение данных алгоритмов обусловлено их высокой эффективностью. При последовательном переборе строки считываются последовательно и сравниваются непосредственно с поисковым образцом. Для сравнения строк используются битовые алгоритмы. Однако данный алгоритм работает достаточно медленно.

6. Сигнатурные алгоритмы [1]. Их основой является буквенное сэмплирование. При этом используется хеш-функция, которая однозначно определяет преобразование строки в

целое число. Метод хеширования по сигнатуре позволяет осуществлять поиск с высокой скоростью, отличается простотой реализации.

7. Алгоритм шинглов[4] позволяет определять меру схожести документов в численном виде. Основная идея заключается в разбиении всего текста на некие равные части. Перед разбиением происходит очистка (канонизация) текста от предлогов, союзов, дополнительной текстовой разметки. Этот алгоритм достаточно эффективен по скорости.

8. Метод расширения выборки довольно часто применяют при проверке орфографии. Суть метода заключается в сведении поиска по сходству к точному поиску. Для этого формируется множество всех «ошибочных» слов, которые получаются из поискового образца в результате одной-двух операций редактирования: вставки, замены, удалении и транспозиции, после чего построенные термины ищутся в словаре (на точное соответствие).

В ходе первичной обработки данных выявлены следующие особенности:

- перманентные изменения названий регионов и состава Федеральных округов;
- изменение структуры статистических документов (например, введение Общероссийского классификатора видов экономической деятельности);
- модификация структур статистических таблиц;
- изменение структуры цен;
- отсутствие данных в ряде периодов и др.

Для первичной обработки данных при анализе экономической безопасности регионов России был выбран метод последовательного перебора. Он достаточно прост для программной реализации, что при решении данной задачи не отражается на его эффективности.

На *втором этапе* производилось преобразование исходных данных в индикаторы экономической безопасности с учетом их размерностей. Далеко не все импортируемые показатели являются индикаторами экономической безопасности. Многие индикаторы получаются из исходных путем различных расчетов и преобразований. При этом создаются трехмерные массивы, последующая обработка которых приводит к формированию новых знаний. Первый индекс в таких массивах – номер региона, второй – номер индикатора экономической безопасности, третий – время. В настоящее время в системе Росстата по ряду индикаторов доступны данные с 1995 года. Периодичность их измерения – один год. Кроме основной системы, для оперативных прогнозов возможно использовать систему краткосрочных индикаторов, измеряемых с периодичностью 1 месяц (они доступны с 1999 г.).

На *третьем этапе* производится обработка и визуализация преобразованной информации с целью их последующего анализа и прогнозирования.

Для более эффективного мониторинга представляется целесообразным автоматизировать предложенный алгоритм. В работе [3] обоснована необходимость разработки информа-

ционной системы мониторинга индикаторов экономической безопасности регионов России. К функциям данной системы можно отнести:

- импорт и первичную обработку индикаторов экономической безопасности;
- хранение и предоставление доступа к данным: редактирование, добавление, удаление;
- построение линейных и лепестковых диаграмм для анализа динамики индикаторов.

Информационная система должна иметь достаточно удобную навигационную среду для выбора необходимой опции: построение диаграмм, которые позволяют легко и наглядно просматривать динамику развития определенного региона или же тенденцию поведения заданного индикатора; импортирование, первичная обработка (в том числе и приведение индикаторов к безразмерному виду), редактирование данных, настройки.

В основном первичные данные импортируются из официальных источников (сайты Росстата, Минфина, Центрального Банка и т.д.) и обычно представлены в двух форматах: заархивированные документы и HTML-страницы.

При импорте данных в виде архивов программе потребуется сначала скачать архив, потом его распаковать и только затем извлекать данные из файла для последующей обработки. Еще один способ заключается в извлечении архива без помощи информационной системы, что достаточно нерационально, требует дополнительной памяти для хранения архива, занимает больше времени и представляет некоторые неудобства для пользователя. Импорт данных в виде HTML-страниц имеет ряд преимуществ. Во-первых, извлечение данных требует меньшего количества действий и, как следствие, меньших затрат ресурсов. Во-вторых, структура электронной версии состоит из одной таблицы, в то время как файл содержит несколько таблиц со значениями индикаторов, что осложняет извлечение данных. Однако данный вариант имеет и ряд недостатков. К ним можно отнести:

1. Необходимость подключения к сети Интернет, что представляется возможным не во всех случаях.

2. Данные, представленные в виде HTML-страниц, имеют нерегулярную структуру, то есть являются слабоструктурированными, что значительно усложняет задачу извлечения.

3. Существуют ряд проблем доступа к данным сайта. Причиной могут быть различные работы, проводимые администрацией сайта, проблемы с сетью и многое другое. В этом случае импорт информации невозможен.

Перейдем к рассмотрению структуры самих страниц. Первый столбец каждой таблицы – это регионы и округа Российской Федерации. Самая верхняя строка – это года, за которые представлены значения индикаторов по данным субъектом страны. Вторая верхняя строка – значения индикатора в целом по России.

Примеры таблиц изображены на рис. 1.

	2000	2005	2006	2007	2008	2009	2010	2011
Валовой региональный продукт по субъектам Российской Федерации (валовая добавленная стоимость в текущих основных ценах) - всего	**39532,3**	**125658,7**	**157233,0**	**195819,0**	**237552,2**	**224163,3**	**263828,6**	**316626,6**
Центральный федеральный округ	**48205,0**	**164887,9**	**208806,5**	**267272,1**	**331472,2**	**297793,0**	**350204,2**	**420102,4**
Белгородская область	27969,5	95911,2	118211,4	156225,1	208548,1	199046,1	260015,6	333502,0
Брянская область	17413,5	49923,4	62187,8	78518,8	96885,4	98014,5	114777,6	141682,8
Владимирская область	21073,3	58261,0	76184,8	99682,5	119941,8	127815,1	155494,2	178491,9
Воронежская область	20365,1	56534,5	70492,7	94849,5	122591,1	129112,5	148432,6	191652,4
Ивановская область	14240,0	40039,1	50271,5	68865,7	80708,5	81286,7	103280,0	120349,8
Калужская область	22438,0	69192,2	84317,4	109790,3	147929,5	152611,6	186347,8	232255,6
Костромская область	21984,7	63304,4	78226,9	95687,2	119071,5	116856,2	146536,9	167845,2
Курская область	23677,7	72995,3	88949,4	111348,4	146276,4	141833,5	171322,1	207690,8
Липецкая область	39050,9	121376,2	150197,1	176534,6	219135,8	192165,2	211610,6	244560,6
Московская область	26687,7	104738,3	137092,1	188565,3	237595,8	217339,7	259421,5	313635,7
Орловская область	25168,4	64180,4	79341,5	95387,1	120531,4	113848,8	134533,8	167149,9

Рис. 1.Фрагмент таблицы «ВВП на душу населения
(в процентах к предыдущему году)» электронной версии

Для анализа первичной информации разработана информационная система, состоящая из следующих модулей (рис. 2): модуль импорта, модуль обработки, модуль редактирования данных, модуль визуализации.

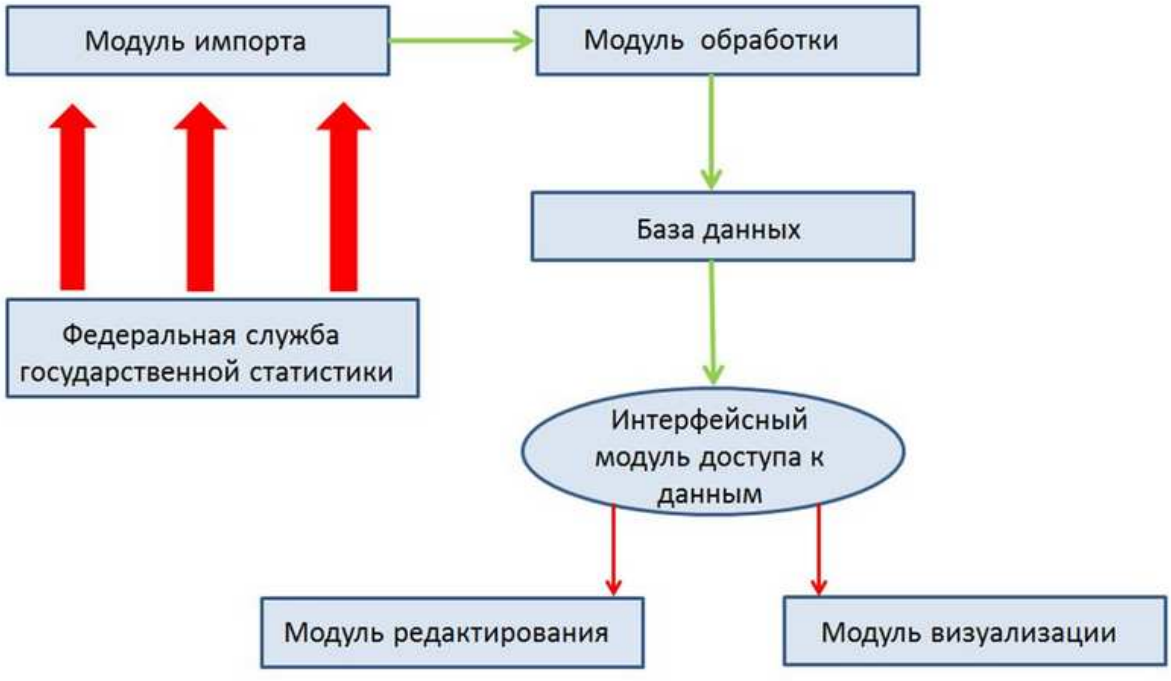

Рис. 2. Структурная схема системы импорта и обработки данных

Основной функцией этих модулей является взаимодействие с базой данных. Модули импорта и первичной обработки наполняют базу данных. Модуль редактирования данных позволяет редактировать, удалять, добавлять индикаторы, регионы, проекции экономического развития, ссылки для импорта, пороговые значения. Модуль визуализации позволяет представлять информацию в виде графиков и диаграмм.

Стрелками на рисунке изображены потоки информации, где красным цветом выделены потоки входных и выходных данных, зеленым – движение данных внутри системы.

В заключение отметим, что количество данных постоянно возрастает, и поэтому проблема обработки слабоструктурированной информации актуализируется каждый год с новой силой точно так же, как понятие экономической безопасности в условиях новых вызовов и угроз.

Работа выполнена при финансовой поддержке РГНФ в проведении научных исследований «Методологические основы анализа экономической безопасности региона (на примере Нижегородской области)», проект №14-02-00093.

Список литературы

1. Бойцов Л.М. Использование хеширования по сигнатуре для поиска по сходству. Прикладная математика и информатика. – М.: Изд-во факультета ВМиК, МГУ, 2000. – № 7.
2. Когаловский М.Р. Абстракции и модели в системах баз данных // СУБД. – 1998. – № 4-5. – С. 73-81.
3. Митяков Е.С. Структура информационной системы экономической безопасности регионов России // Труды НГТУ им. Р.Е. Алексеева. – 2014. – № 1 (102). – С. 268-273.
4. Родненко В. Python: Алгоритм Шинглов – поиск нечетких дубликатов текста // URL: http://www.codeisart.ru/python-shingles-algorithm (дата обращения: 11.02.2015).
5. Синицын И.Н. Фильтры Калмана и Пугачева // Рос. акад. науки, Ин-т проблем информатики. – М.: Логос, 2007. – 772 с.
6. Томас Коннолли, Каролин Бегг. Базы данных. Проектирование, реализация и сопровождение. Теория и практика. 3-е изд. – М: Вильямс, 2003. – 1440 с.
7. Delsarte P.; Levenshtein, V. I. (1998), Association schemes and coding theory, IEEE Transactions in Information Theory 44 (6): 2477–2504.
8. Peter Buneman. Semistructured data. In *Proc.* ACM Symposium on Principles of Database Systems, pages 117-121, Tucson, AZ., 1997. Abstract of invited tutorial.

Рецензенты:

Дмитриев М.Н., д.э.н., профессор, зав. кафедрой «Экономика, финансы и статистика» Нижегородского государственного архитектурно-строительного университета, г. Нижний Новгород;
Рузанов А.И., д.т.н., профессор кафедры информационных технологий и инструментальных методов в экономике Нижегородского государственного университета им. Н.И. Лобачевского, г. Нижний Новгород.