

ОБРАБОТКА ГРУППЫ НЕЧЕТКИХ ВРЕМЕННЫХ РЯДОВ С ПРИМЕНЕНИЕМ ТЕМПОРАЛЬНОЙ ГРАММАТИКИ

С.П. Ивченко, Л.И. Сучкова

Алтайский государственный технический университет им. И.И. Ползунова
г. Барнаул

Статья посвящена разработке алгоритмов и реализации методики выявления закономерностей в группе временных рядов с помощью грамматики специального вида, а также разработке средства формализации представления результатов анализа.

Ключевые слова: анализ данных, темпоральная грамматика, нечеткий ряд.

Эволюция технических средств записи и хранения данных вызвала взрывной рост объемов собираемых данных в различных предметных областях. Сырые данные содержат в себе много избыточной информации, которая может быть использована для совершенствования способов хранения и обработки данных. Тем не менее извлечение информации весьма затруднительно, объемы данных настолько внушительны, что человеку просто невозможно проанализировать их самостоятельно. В связи с этим активно разрабатываются методы автоматизированного поиска закономерностей в больших массивах. В данной статье рассматривается метод анализа данных, называемый универсальной временной грамматикой [1]. Достоинством метода является возможность обработки групп временных рядов [2] как распространенной формы сырых данных, тяжело поддающейся автоматизированному анализу.

Вышеизложенное послужило основанием для проведения исследований, основной целью которых является:

- разработка алгоритмов и программной реализации методики выявления темпоральных закономерностей в группе рядов с помощью грамматики специального вида;
- разработка грамматики, способной в полной мере описать результаты, найденные в процессе анализа временных рядов.
- применение нечеткой логики для реализации возможности работы с нечеткими данными.

Для формализации данных за основу взята универсальная темпоральная грамматика, описывающая взаимосвязи нескольких простых для понимания уровней временной абстракции [3]. Особенностью данной грамматики является то, что она работает не с

численными данными, а с семиотическими, то есть с символьными. Анализируемые входные данные представлены в виде матрицы чисел, каждый столбец соответствует одному входному временному ряду, каждая строка соответствует данным измерений, совершенным в один и тот же момент времени. Для формирования матрицы и преобразования данных в символьную форму с помощью специального менеджера рядов производится их загрузка. Скриншот диалога загрузки и преобразования рядов показан на рисунке 1.

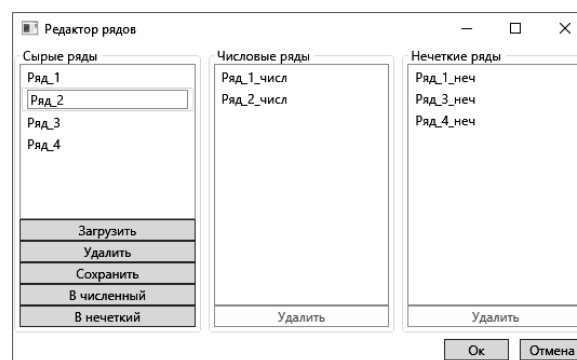


Рисунок 1 – Диалог загрузки
и преобразования временных рядов

Программа различает три вида рядов: «сырые», численные и нечеткие ряды. Диалог загрузки рядов содержит три списка, каждый из которых соответствует своему типу ряда. Первичным является сырой ряд, после загрузки он может быть преобразован либо в численный ряд, либо в нечеткий нажатием на соответствующую кнопку.

Для работы с нечеткими рядами разработан редактор лингвистических переменных (РЛП). Скриншот РЛП показан на рисунке 2.

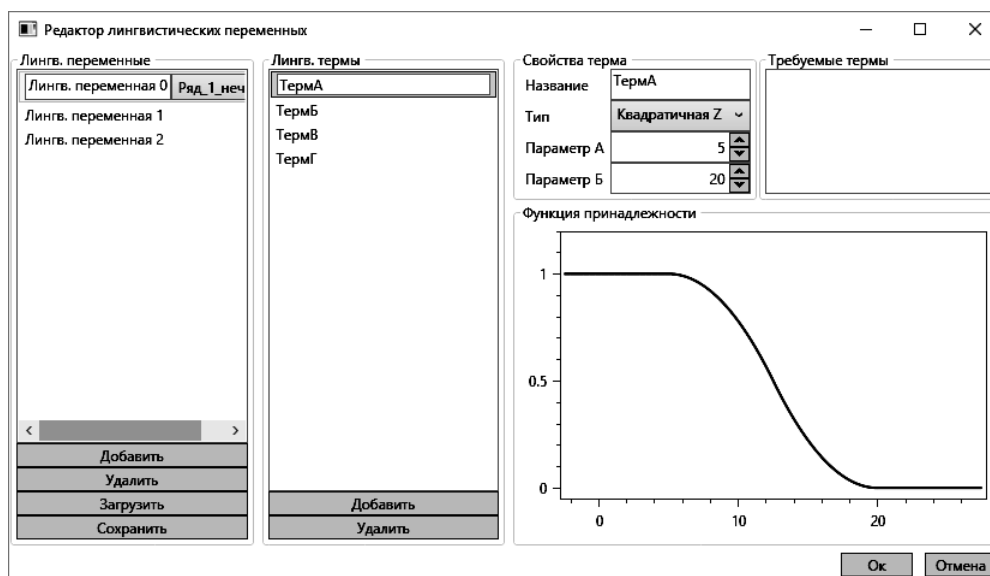


Рисунок 2 – Редактор лингвистических переменных

РЛП предназначен для создания лингвистических переменных, термы которых будут использоваться далее в нечетких выражениях. Каждая лингвистическая переменная привязана к нечеткому ряду и имеет несколько лингвистических термов. Каждый лингвистический терм характеризуется функцией принадлежности. В программе доступно 14 видов функций принадлежности. Список лингвистических переменных в диалоге находится слева. Список термов расположен посередине окна. Справа расположена область редактирования параметров лингвистического терма.

Основой для формирования сложных темпоральных конструкций являются примитивные шаблоны, поэтому для их формирования задаются правила вычисления. Данные правила определяют, как именно загруженные ряды будут преобразованы в ряды примитивных шаблонов, с которыми далее будет работать основной алгоритм. Предусмотрены численные и нечеткие правила. Численные правила определяют, как преобразовывать численные ряды в ряды примитивных шаблонов, а нечеткие правила определяют, как преобразовывать нечеткие ряды. На рисунке 3 показан пример скриншота редактора правил.

В данном диалоге есть три списка. Первый список содержит наборы правил. Каждый набор правил будет использован для создания одного ряда примитивных шаблонов. Наборы численных правил помечены синим кружком, наборы нечетких помечены зеле-

ным. Список в центре отображает правила из набора, выбранного в первом списке. Список справа показывает компоненты выбранного правила.

Каждый набор правил, таким образом, соответствует одному ряду примитивных шаблонов, а каждое правило в наборе соответствует одному примитивному шаблону.

Численные ряды преобразовываются в ряды примитивных шаблонов следующим образом. Составляется набор численных правил. Каждое из правил в наборе содержит численный интервал, заданный пользователем. При преобразовании каждое число в ряду заменяется на примитивный шаблон, соответствующий правилу, в интервал которого входит текущее число.

Нечеткие ряды преобразовываются с помощью наборов нечетких правил. Каждое правило содержит нечеткое выражение, в котором используются лингвистические переменные, определенные на предыдущем этапе. В нечетких выражениях допустимо использование операций И, ИЛИ, и НЕ. Выражения состоят из термов лингвистических переменных соединённых нечеткими операциями. Например для лингвистических переменных «Температура» и «Высота»: можно составить выражение: «Не Холодно И Низко». При преобразовании вычисляются все нечеткие выражения, включенные в набор правил для каждого набора отсчетов в исходных нечетких рядах. Под набором отсчетов подразумеваются данные измерений, сделанные в один момент времени и нахо-

ОБРАБОТКА ГРУППЫ НЕЧЕТКИХ ВРЕМЕННЫХ РЯДОВ С ПРИМЕНЕНИЕМ ТЕМПОРАЛЬНОЙ ГРАММАТИКИ

дящиеся в разных нечетких рядах. При вычислении нечетких выражений для каждого термина в выражении берется замер из нечеткого ряда, привязанного к лингвистической переменной термина. Для замера вычисляется его степень принадлежности с использованием функции принадлежности текущего лингвистического термина. После того как для каждого термина в выражении вычислена степень

принадлежности соответствующего замера, к результатам применяются нечеткие операции И, ИЛИ и НЕ. Результат вычисления выражения - это степень принадлежности набора отсчетов к нечеткому выражению. В результирующий ряд добавляется примитивный шаблон, соответствующий правилу, для которого была вычислена наибольшая величина принадлежности набора отсчетов.

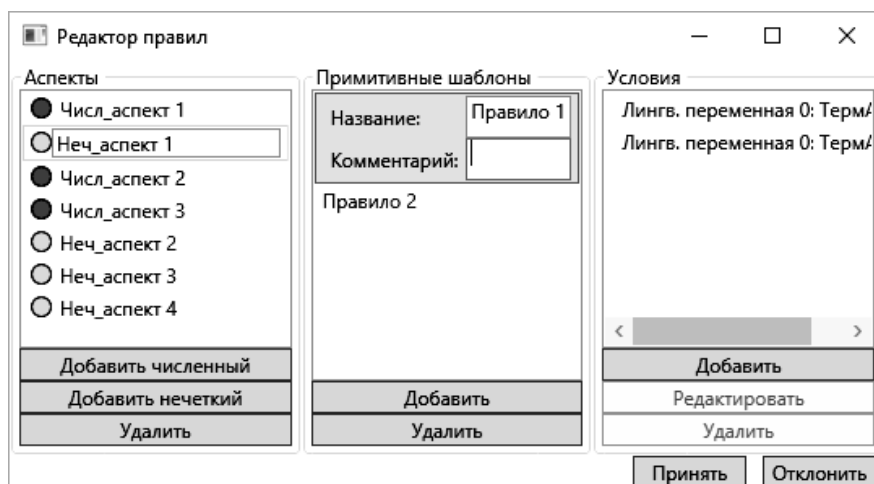


Рисунок 3 – Редактор правил

После того как закончен этап предварительной обработки и сформированы ряды примитивных шаблонов, осуществляется основной этап анализа. На рисунке 4 схематично изображен процесс поиска временной информации с применением темпоральной грамматики.

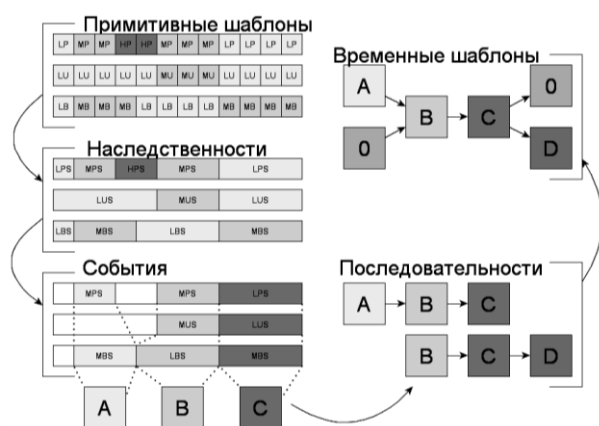


Рисунок 4 – Схема анализа временных рядов

Сначала осуществляется нахождение наследственностей, отражающих временной концепт длительности. Следующим шагом алгоритма является поиск событий. Событие – это совпадение, или частичное совпадение

начала и конца нескольких наследственностей в разных временных рядах. Частичное совпадение означает что концы, или начала интервалов не должны совпадать точно, и допустима некая погрешность. События отражают временной концепт совпадения.

После выявления событий происходит поиск последовательностей - нескольких событий, идущих друг за другом по времени. При поиске последовательностей тоже допускается погрешность, то есть конец одного события не обязательно должен быть началом следующего события.

Далее осуществляется поиск временных шаблонов. В большинстве случаев многие найденные последовательности будут похожи друг на друга, и могут иметь всего лишь небольшие различия. Подобные последовательности объединяются во временной шаблон. Степень сходства временного шаблона определяется с помощью специального алгоритма основанного на алгоритме Левенштейна [4]. Временной шаблон является объединением нескольких похожих последовательностей и отражает временной концепт вариативности.

Описанная выше методика анализа временных рядов была реализована программно. Программа написана на языке C#[5] с ис-

пользованием технологии WPF. Создана специализированная грамма для лингвистического описания найденных темпоральных закономерностей в процессе анализа. Для лингвистического описания найденных темпоральных закономерностей была разработана специализированная грамматика. Для парсинга был создан лексический анализатор [6], реализованы средства синтаксического [7] и семантического анализа [8].

Разработанное программное обеспечение предназначено для анализа данных с помощью темпоральной грамматики. Основное окно программы содержит основные элементы управления, а также поля, служащие для вывода информации.

В программе присутствует два способа обработки данных, комплексный и последовательный. В последовательном режиме программа обрабатывает данные постепенно, позволяя производить промежуточную корректировку данных перед поиском конструкции более высокого уровня. Данный способ предназначен для обработки очень больших объемов данных, где удаление лишних конструкций на нижних уровнях грамматики значительно ускорит обработку данных. Комплексный метод автоматически производит поиск конструкций на всех уровнях, и выводит результат на экран. После того как конечный результат был получен, можно убрать лишние конструкции прямо в окне с конечным результатом. Затем программа удалит лишние конструкции, а также конструкции вышестоящих уровней, содержащие удаленные конструкции.

Вывод результатов анализа производится через текстовое поле, занимающее большую часть главного окна. В нижней части окна расположена строка состояния, информирующая пользователя о текущем выполняемом действии, и успешности результатов анализа данных. Все результаты анализа выводятся в текстовое поле в главном окне программы, при желании их можно отредактировать, удалив не интересующие пользователя данные.

По результатам работы можно сделать следующие **выводы**.

Разработан ряд алгоритмов, реализующих все этапы поиска закономерностей в

данных на основе темпоральной грамматики, включая предварительную обработку исходных данных. Для представления результатов работы алгоритмов была разработана специализированная грамматика, позволяющая сформировать итог поиска в виде понятного для человека текста. На основе разработанных алгоритмов создано программное обеспечение, предназначенное для анализа групп временных рядов. Алгоритм расширен с помощью применения нечеткой логики на этапе предварительной обработки.

СПИСОК ЛИТЕРАТУРЫ

1. Mörchen F. Unsupervised pattern mining from symbolic temporal data [Электронный ресурс] / F. Mörchen. Режим доступа: <http://www.mybytes.de/papers/moerchen07unsupervised.pdf>.
2. Дюк В. Data Mining. Учебный курс [Текст] / В. Дюк, А. Самойленко. – СПб.: Питер, 2001. – 368.: ил. + 1 эл.опт. диск (CD ROM).
3. Mörchen F. Mining hierarchical temporal patterns in multivariate time series [Электронный ресурс] / F. Mörchen, A. Ultsch. Режим доступа: <http://www.unimarburg.de/fb12/datenbionik/pdf/pubs/2004/moerchen04mining>.
4. Карахтанов Д.С. Программная реализация алгоритма Левенштейна для устранения опечаток в записях баз данных [Текст] / Д.С. Карахтанов // Молодой ученый. – 2010. – №8. – С. 158-162.
5. Фримен А. LINQ. Язык интегрированных запросов в C# 2010 для профессионалов [Текст] / А. Фримен, Д. Раттц. – М.: Вильямс, 2011. – 656 с.
6. Карпов Ю.Г. Теория автоматов [Текст] / Ю.Г. Карпов. – СПб.: БХВ-Петербург, 2003. – 208 с.
7. Малявко А.А. Формальные языки и компиляторы: учебник НГТУ [Текст] / А.А. Малявко. – Новосибирск: Изд-во НГТУ, 2013. – 431 с.
8. Вылиток А.А. Металингвистические формулы и синтаксические диаграммы [Текст] А.А. Вылиток. – М.: МАКСПресс, 2012. – 24 с.

Ивченко Сергей Павлович – магистрант кафедры информатики вычислительной техники и информационной безопасности, тел.: 8-983-606-23-45, e-mail: antyrox@gmail.com;
Сучкова Лариса Иннокентьевна – д.т.н., профессор кафедры информатики вычислительной техники и информационной безопасности АлтГТУ им. И.И.Ползунова, тел.: 8-(3852) 29-07-86, e-mail: lara8370@yandex.ru.