УДК 004

Разработка алгоритмов и системы анализа неструктурированной и слабоструктурированной информации

Гнатюк Евгений Сергеевич
Волжский политехнический институт (филиал) Волгоградского государственного технического университета
Студент

Рыбанов Александр Александрович
Волжский политехнический институт (филиал) Волгоградского государственного технического университета
Кандидат технических наук, доцент, заведующий кафедрой «Информатика и технология программирования»

Аннотация
В статье рассмотрены методы анализа неструктурированной и слабоструктурированной информации, а также представлен программный модуль, в котором разработан алгоритм структуризации слабоструктурированной информации.
Ключевые слова: неструктурированная и слабоструктурированная информация, методы анализа неструктурированной и слабоструктурированной информации, алгоритм структуризации неструктурированной и слабоструктурированной информации.

Development of algorithms and systems for analysis of unstructured and weakly structured information

Gnatyuk Evgeny Sergeevich
Polytechnic Institute of Volzhsky branch of the Volgograd State Technical
Student

Rybanov Aleksandr Aleksandrovich
Polytechnic Institute of Volzhsky branch of the Volgograd State Technical
Ph.D., Associate Professor, Head of the Department «Computer technology and programming»

Abstract
The article methods of the analysis of unstructured and poorly structured information are considered and the program module in which the algorithm of structuring of the poorly structured information is developed.
Keywords: unstructured and weakly structured information, methods of analysis of unstructured and weakly structured information, algorithm for structuring unstructured and weakly structured information.

На сегодняшний день существует большое количество самых разнообразных систем, основной целью которых является увеличение уровня интеллектуальной информационной поддержки современного специалиста, занимающегося принятием различных решений. Данные системы относятся к категории интеллектуального анализа данных и дают возможность обнаружить скрытые закономерности. Также в категорию данных систем включаются программные продукты области имитационного моделирования, экспертные системы, в основе которых лежат знания и опыт экспертов с сфере принятия решений каких-либо конкретных предметных областей, а также множество прочих систем, которые помогают пользователю с выбором наиболее лучшего варианта из совокупности предлагаемых ему [2, с.201].

В роли самых сложных для анализа, а также принятия решений, выступают слабоструктурированные и неструктурированные предметные области. Данные предметные области широко используются медицине, биологии, экономике, задачах прогнозирования и других сферах. Необходимо отметить, что перед анализом информации и нахождением оптимальных вариантов решений, должен осуществляться процесс формализации модели, а именно должны выявляться факторы, их взаимосвязь и влияние друг на друга.

Рассматривая слабоструктурированные и неструктурированные области знаний можно сказать, что вышеуказанный процесс должен осуществляться непосредственно с участием экспертов, а для сверхсложных областей необходима помощь компьютерной поддержки [4, с.93].

Для того чтобы соответствовать современным тенденциям, системам анализа слабоструктурированной и неструктурированной информации необходимо иметь внутри себя совокупность разнообразных методов анализа, а также оценки и подготовки решений. Вместе с этим такие системы должны обладать продвинутым интерфейсом для пользователей, в роли которых выступают эксперты. Немаловажным является наличие в данных системах модулей, позволяющих визуализировать процесс принятия решений целиком, а также его конечные результаты.

Системы, архитектуры которых подходят под указанные выше требования, сегодня очень динамично развиваются, поскольку использование таких систем дает большие преимущества в скорости и качестве исследования сложных предметных областей. Исходя из этого можно сделать вывод о том, что темы исследований, связанных с алгоритмами анализа неструктурированной и слабоструктурированной информации, являются востребованными и актуальными.

Под неструктурированной информацией понимают данные, которые представлены в совершенно произвольной форме и включают в себя различные тексты, графические материалы, изображения, звуки и т.д. Такая форма представления данных на сегодняшний день широко используется в социальных сетях, поисковых системах и других источниках [3, с.172].

К слабоструктурированной информации относятся данные, определенные общими правилами и форматами.

Перечислим наиболее используемые задачи в области анализа неструктурированных и слабоструктурированных данных:

- классификация;
- кластеризация;
- построение семантических сетей;
- извлечение фактов, понятий (feature extraction);
- извлечение мнений;
- аннотирование, суммаризация (summarization);
- ответ на запрос (question answering);
- тематическое индексирование (thematic indexing);
- поиск по ключевым словам (keyword searching);
- создание таксономий и тезаурусов.

Для того чтобы найти полезные знания в слабоструктурированной или неструктурированной информации наиболее часто используется процесс под названием Knowledge Discovery in Databases (KDD).

Данный процесс состоит из вопросов касающихся подготовки данных, извлечения информативных признаков, очистки данных, применения методов Data Mining (DM), постобработки данных и интерпретации полученных результатов. В роли основы всего этого процесса выступают методы DM, при помощи которых и происходит процесс обнаружения знаний. Примерами данных знаний могут являться правила, наиболее часто встречающиеся шаблоны (наборы ассоциативных правил), а также итоги классификации (нейронные сети) и кластеризации данных и т.д. [5, с.264].

Рассмотренная выше методика не привязана к конкретной предметной области и по сути является набором простых атомарных операций, составляя комбинации из которых, появляется возможность получить требуемое решение. Общий алгоритм методики извлечения знаний представлен на рисунке 1.

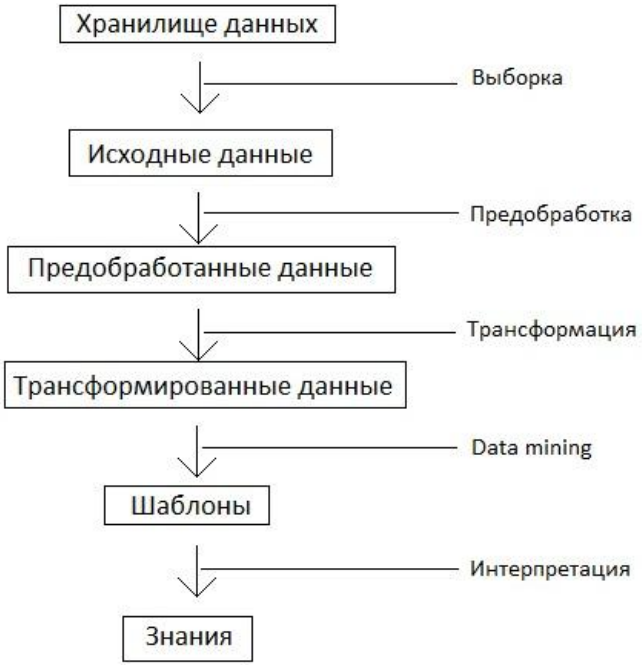

Рисунок 1 – Алгоритм методики извлечения знаний

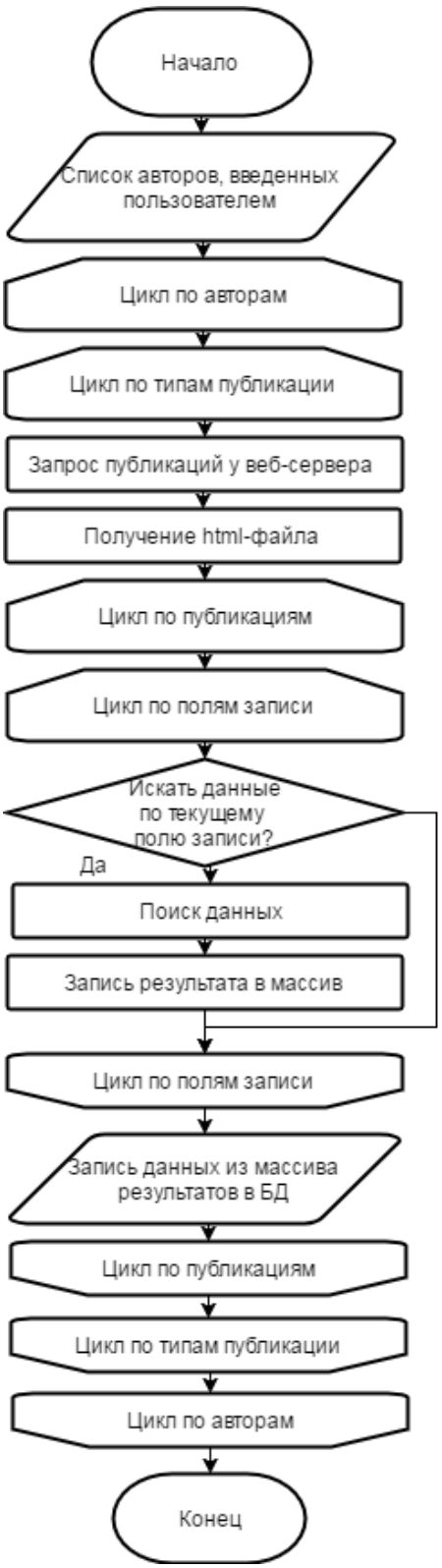

Рисунок 2 – Блок-схема разработанного алгоритма

Разработанный алгоритм работает следующим образом: при вводе пользователем списка авторов требуемых для парсинга, отправляется запрос на сервис vstu.ru в связке: первый введенный автор и первый тип публикации (монография). Все типы публикаций изначально введены в программный модуль. Поэтому парсинг происходит последовательно по каждому типу публикации. При получении html-файла, по заданному паттерну выделяется

публикация. Исполняется цикл по публикациям, в который вложен цикл по полям записи, т.е. начинает процесс парсинга. Поля проверяются на их существование и на наличие конца публикации. При обнаружении необходимого поля функция заносит данные в массив. По окончанию парсинга одного типа публикации отправляется новый запрос с тем же автором, но с другим типом публикации (учебник). Повторяет проверка на существования и наличие конца публикации, при обнаружении результат заносится в массив. После обработки по всем типам публикаций данные из массива записываются в собственную базу данных. Далее отправляется запрос со следующим автором и снова с первым типом публикации (монография). Далее происходит тот же самый процесс поиска и извлечения информации. При повторном требовании запроса к программному модулю алгоритм не будет заполнять базу данных дублями, так как производится проверка на существование записей в базе данных. Так же дубля не будет, если встретятся два разных автора, у которых общая публикация, при прохождении проверки на запись, данные не запишутся в базу данных.

Все публикации разные так же могут отличаться их структура, например, может отсутствовать город или год издания, а может отсутствовать город, а год издания будет или наоборот. В подобных моментах, чтобы не нарушать логику парсера, производятся проверки.

Блок-схема алгоритма обработки проверки на наличие сущности представлена на рисунке 3.

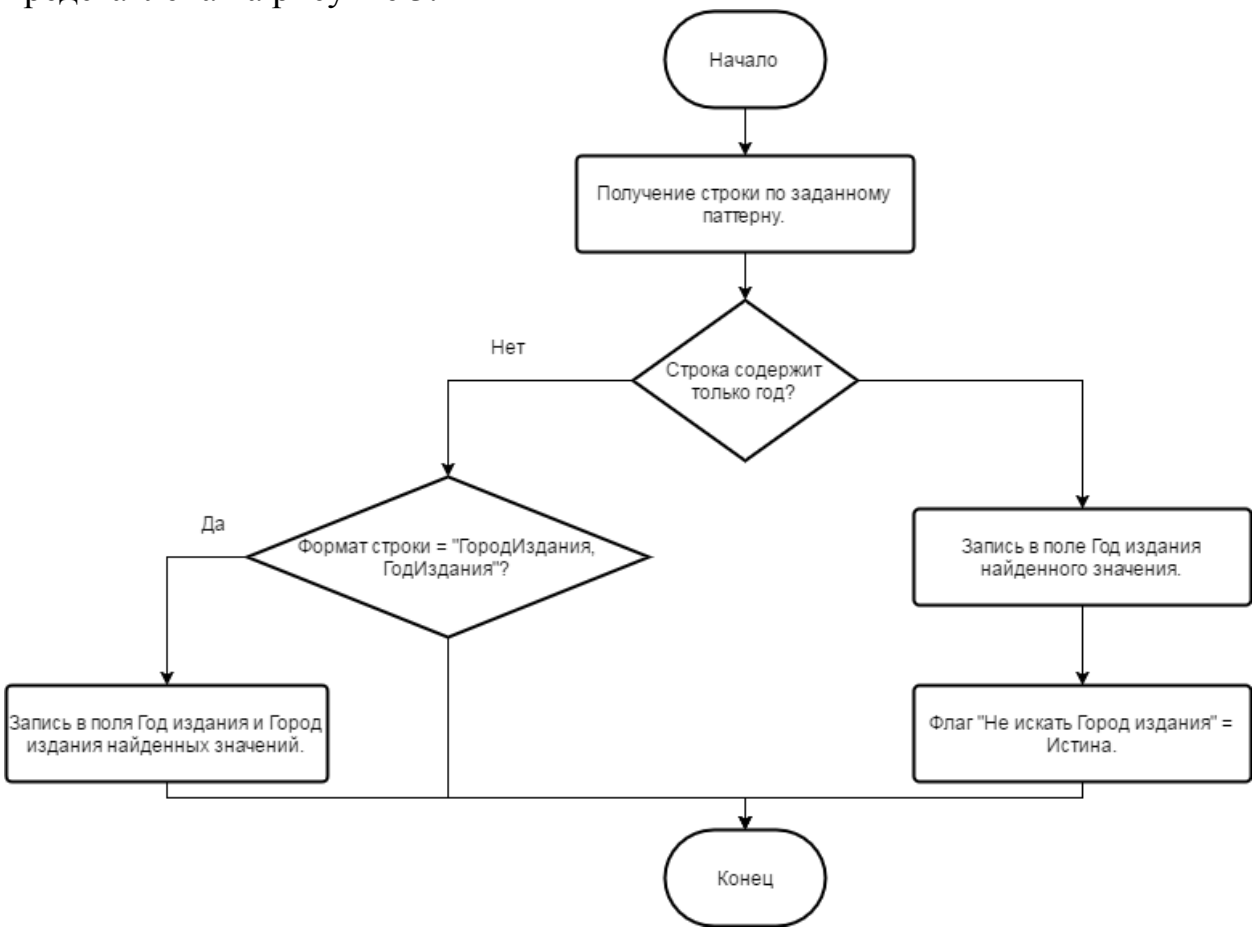

Рисунок 3 – Алгоритм обработки проверки на наличие сущности

Модуль для структуризации неструктурированного и слабоструктурированного списка научных публикаций сотрудников ВолгГТУ позволяет выполнить автоматизацию следующих действий:

- отправлять сразу нескольких сотрудников на обработку информации;
- просмотреть количество публикаций всех доступных типов
- выгрузка информации в собственную базу данных в структурированном виде;
- запрос информации с баз данных;
- расширенный функционал поиска по публикациям;

После запроса данных об авторе (авторах) на вход модуля поступает html файл с результатом сервиса. Далее модуль обрабатывает информацию и структурирует данные по категориям. Структуризация информация выполняется за счет парсинга. Последовательный синтаксический анализ информации, размещённой в результате получения html-файла осуществляется за счет регулярных выражений.

В базе данных модуля выделены 14 объектов для критерий структуризации информации и 2 служебных объекта (таблица 1).

Таблица 1 – Описание полей БД модуля

Наименование поля	Тип	Описание
id	int	Идентификатор публикации
defaultString	varchar(1024)	Исходная строка с публикацией
mainAuthor	varchar(512)	Главный автор
name	varchar(512)	Название публикации
authors	varchar(512)	Авторы
titlePublication	varchar(512)	Заголовок публикации
executiveEditor	varchar(512)	Редактор
placePublication	varchar(512)	Место издания
cityPublication	varchar(512)	Город издания
yearPublication	varchar(512)	Год издания
issueNumber	varchar(512)	Номер выпуска издания
pages	varchar(512)	Страницы
accessMode	varchar(512)	Режим доступа
advancedInfo	varchar(512)	Дополнительная информация
buf	varchar(512)	Остаток после парсинга
typePublication	varchar(512)	Тип публикации

Структурированные данные выгружаются в личную базу данных в структурированном виде. Данные в базе предоставляются в виде таблицы, в которой можно фильтровать данные по полю заголовка.

Просмотр таблицы базы данных реализован отдельно. Осуществляется запрос к базам данных с выводом информации и пользователю

предоставляется таблица структурированного вида в конечной форме, где присутствует возможность фильтровать данные по полям заголовок таблицы.

Для сравнения на рисунке 4 изображен результат обработки запроса сервисом сайта vstu.ru

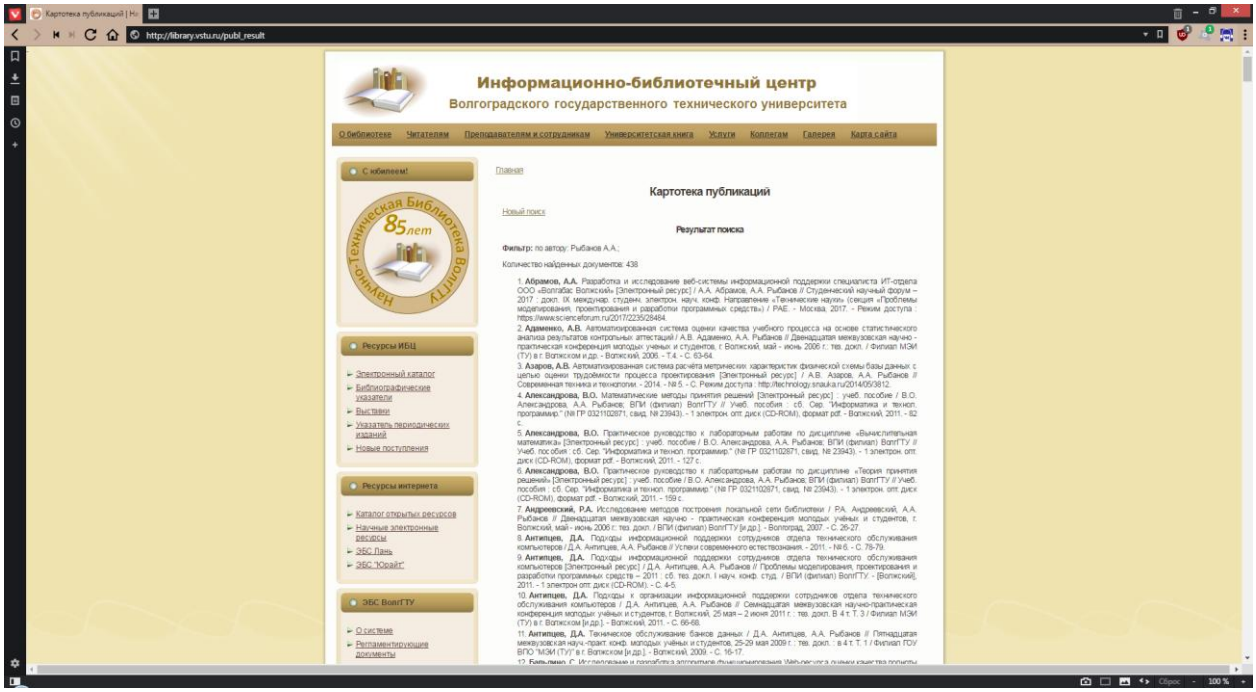

Рисунок 4 – Результат запроса сервиса сайта vstu.ru

На рисунке 5 изображен результат обработки разработанного программного модуля.

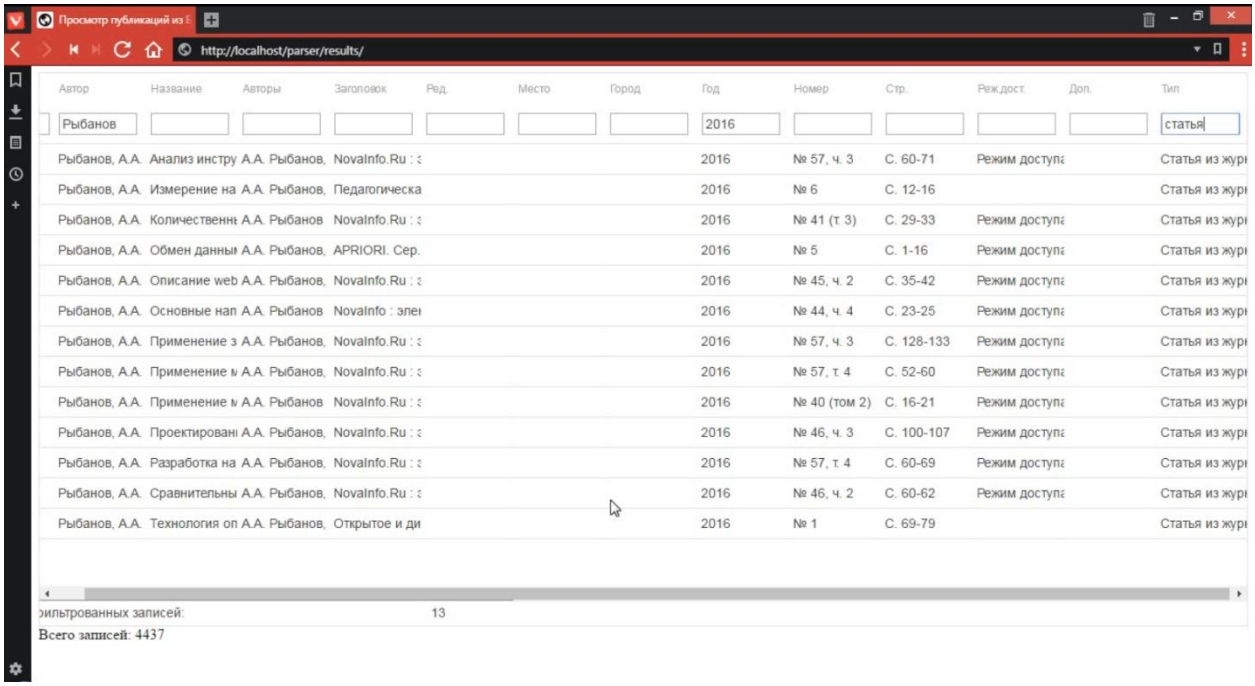

Рисунок 5 – Результат обработки разработанного программного модуля

Итогом работы стала реализация программного модуля по структуризации информации по итогам обработки сервиса базы данных «Публикации сотрудников ВолгГТУ».

Библиографический список

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа данных: Учебное пособие. М.: ИД ГУ ВШЭ, 2012.
2. Наследов А.Д. IMB SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. СПб.: Питер, 2013.
3. Романко В.К. Статистический анализ данных: Учебное пособие. М.: БИНОМ. ЛЗ, 2013.
4. Уткин Л.В., Шубинский И.Б. Нетрадиционные методы оценки надежности информационных систем. СПб.: Любавич, 2013.
5. Чесноков С.В. Детерминационный данных. М.: КД Либроком, 2013.