

TECHNOLOGY OF DATA CLEANSE FOR THEIR INTEGRATION PROCESSES

Viacheslav V. Paramonov^(1, 3), Alexey O. Shigarov^(1, 2), Gennagy M. Ruzhnikov^(1, 3)

⁽¹⁾ Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia

⁽²⁾ Irkutsk State University, Irkutsk, Russia

⁽³⁾ Irkutsk Scientific Center SB RAS, Irkutsk, Russia

The paper deals with approaches to heterogeneous data cleanse in processes of their preparation to integration. This process allows identifying and fixing different inconsistencies and errors in data to improve data quality. It provides effective versatile usage of data. The paper considered approaches that include a combination of methods fuzzy string comparison, phonetic algorithms, and the usage of the regular expression technique. The development of effective methods of data cleanse allows to improve the quality information obtained from various sources.

Keywords: integration, textual data, data cleanse, analysis.

ТЕХНОЛОГИЯ ОЧИСТКИ ДАННЫХ ДЛЯ ИХ ИНТЕГРАЦИИ

Парамонов В.В.^(1,3), Шигаров А.О.^(1, 2), Ружников Г.М.^(1, 3)

⁽¹⁾ Институт динамики систем и теории управления им. В.М. Матросова СО РАН, г. Иркутск

⁽²⁾ Иркутский государственный университет, г. Иркутск

⁽³⁾ Иркутский научный центр СО РАН, г. Иркутск

В работе рассматривается комплексный подход для очистки гетерогенных текстовых данных на стадии их подготовке к интеграции. Это позволяет выявить несоответствия и, в автоматическом режиме, исправить ошибки, связанные с опечатками, некорректностью форматов, повтором в данных. Цель – улучшение качества данных и обеспечение возможности их эффективного совместного использования. В работе рассматриваются методов очистки данных, включающих комбинацию нечеткого сравнения строк, фонетических алгоритмов и применение аппарата регулярных выражений. Развитие методов эффективной очистки текстовых данных позволяет повысить качество интегрируемой информации, полученной из различных источников. Интеграция же, в свою очередь, предоставляет возможность проводить более глубокий, разноаспектный, анализ информации и тем самым повысить качество научных исследований.

Ключевые слова: интеграция, текстовые данные, очистка, анализ.

Введение. Научным сообществом накоплен большой объем различных тематических данных. Некоторые из них доступны для совместного использования. Однако исследователям смежных областей, как правило, доступны уже агрегированные данные, а не исходные, что снижает возможности их анализа. Использование первичных данных предоставляет возможность проводить исследования с учетом различных

аспектов, на более полном материале. В свою очередь это может привести к открытию новых взаимосвязей и закономерностей. Таким образом, интеграция разнородных первичных данных является значимым процессом. Однако, учитывая, что первичные данные собираются, как правило, с учетом требований предъявляемых их владельцами, то они могут существенно отличаться по схемам организации и качеству. Под качеством данных понимается характеристика, показывающая степень пригодности данных к анализу.

Данные, содержащие множество различных характеристик позволяют более полно представить описание предметную область. Однако при этом требуется обеспечить высокое качество интегрируемых (исходных) данных. В противном случае, анализ набора данных будет затруднён.

Таким образом, для обеспечения эффективности взаимодействия исследователей, в том числе смежных областей, актуальна разработка инструментальных средств, обеспечивающие возможность повышения качества и последующей генерализации пользовательских данных, в частности, сбор, загрузку и приведение пользовательской информации к единообразному виду. Примерами таких средств выступают различные программное обеспечение, поддерживающее ETL-процесс [1].

В научных, государственных и коммерческих учреждениях существует множество разнообразных открытых тематических данных, представленных в виде таблиц, в форматах PDF (Portable Document Format), Excel, CSV (Comma Separated Values). В подобных форматах структура данных может быть выражена не строго и не очевидно. Для извлечения данных из подобных слабоструктурированных источников использованы методы, приведенные в работе [2]. Для проведения интеграции, результатом которой будет более полный, качественный набор данных требуется решить ряд задач, одна из которых, рассматриваемая в работе – очистка данных [3].

Процедура очистки данных позволяет выявить и, в ряде случаев, исправить ошибки, а также несоответствия в сырых (пользовательских) данных. Целью очистки является повышение качества данных.

Выделяют несколько проблем, связанных с необходимостью очистки данных перед проведением процедур консолидации [1]. Это, в частности, обеспечение корректности форматов и представлений данных; уникальность первичных ключей в таблицах базы данных; полнота и целостность данных; полнота связей; соответствие некоторым аналитическим ограничениям и т.д.

В данной работе рассматриваются вопросы очистки данных, связанные с корректностью форматов представления данных, целостностью данных, полнотой связей.

Представление данных. Данные, используемые для интеграции, как правило, представлены в виде таблиц, но в документах различных форматов и, соответственно, имеют отличную структуру. Связано это с тем, что источниками являются не только выполненные в строгом, определенном формате документы, сформированные, например, подразделениями службы Государственной статистики, различными министерствами и ведомствами, но и предоставленные заинтересованными в обмене, накоплении, обработке информации организациями и частными лицами. Иными словами, данные доступные для интеграции могут собираться по принципу краудсорсинга. Вследствие этого, интегрируемая информация представляются в форматах и структурах удобных именно

владельцам данных. При сборе данных, в таком случае, как правило, не рассматриваются вопросы о возможности дальнейшего их использования другими исследователями. Таким образом, даже однотипные данные, относящиеся к одной предметной области, но собранные различными исследователями (заинтересованными субъектами) могут быть организованы совершенно по-разному. К тому же в собранных данных могут содержаться ошибки, возникшие в результате внесения данных или связанные с некорректной схемой их организации.

В связи с этим представляет интерес создание инструментальных средств, позволяющие проводить очистку тематических данных перед слиянием в интегрирующую базу данных.

«Грязные» данные. Для внесения новых данных исследователи используют разнообразное программное обеспечение. Это могут быть и специальные программные средства, поддерживающие высокое качество данных – например, посредством использования масок ввода, специализированных словарей, подстановок значений из фиксированного набора (исключающих некорректное написание), а также текстовые файлы, электронные таблицы – в которых качество данных может быть существенно ниже. Понижение качества вводимых данных связано с отсутствием жесткого контроля. Это приводит к возникновению опечаток, дублирования, некорректного отображения форматов данных и т.п.

Рассмотрим некоторые из характерных ошибок, возникающих при вводе данных и влияющих на их качество, а также предлагаемые методы их идентификации и устранения.

Идентификация типов данных. Данные, экстрагируемые из документов, в частности из электронных таблиц, могут быть представлены в виде строк, однако, фактически, не являются строковыми типами. Например, данные, представлены числом, но тип данных ячейки, задан как «текст». Однако, возможно идентифицировать и преобразовать основные типы (целое число, число с плавающей точкой, дата, строка и пр.) данных, что позволит повысить их качество. Для этого проводится анализа массива данных с целью идентификации их типа. Корректная идентификация типов данных позволяет повысить аккуратность, точность объединения данных из различных источников и тем самым расширить возможности последующего анализа.

На текущий момент, разработанное в рамках исследования, инструментальное программное обеспечение позволяет идентифицировать такие типы данных как числа, даты, булевы данные и строки.

Для идентификации типов данных использован аппарат регулярных выражений. Регулярное выражение – представляет собой некий шаблон, с которым сопоставляются подстроки. Таким образом, посредством разработки шаблонов, учитывая допустимые (часто используемые) варианты представления данных, их возможно идентифицировать и привести к эталонному виду. Шаблоны представления данных разрабатывались на основе проведенного анализа документов, предоставленных для интеграции. В соответствии с шаблоном осуществляется автоматическое приведение данных к требуемому виду. Пример подобного приведения представлен в таблице 1.

Таблица 1. Пример преобразования данных

Данные пользователя	Представление в интегрирующей БД
01.01.11	2011-01-01
01.01.2011	
1-1-11	
2011/01/1	
234.3	234,3
234,3	

Сформированные шаблоны могут быть повторно использованы для идентификации, преобразования других данных, но получаемых из этого же источника. Также при идентификации типов данных возможно их преобразование – например из дюймов в сантиметры. Такая операция может проводиться в случае, если из спецификации данных известны исходная и целевая единицы измерения.

Следующие методы очистки данных связаны с обработкой именно текстовой информации и приведения ее к некоторому эталонному виду.

Нечеткое сравнение строк. Текстовые научные данные кроме числовых, булевых значений, дат могут иметь также строковые выражения. Такие выражения могут носить как описательный характер объекта исследования, так и его содержать некоторое значение, сопоставленное со значением классификатора – название растения, животного, местности, вида исследования и т.п. Ввиду того, что пользовательские данные могут содержать ошибки, то необходимо проверить возможность автоматической трансформации введенного слова так, чтоб оно однозначно соответствовало значению классификатора или же сообщить о невозможности установления соответствия.

Одним из способов установления похожести строки с эталонным образцом является нечеткое сравнение строк. Данный алгоритм позволяет оценить похожесть строк, т.е. минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую. В настоящее время существует множество методов нечеткого сравнения строк, отличающихся по быстродействию, например, алгоритмы Левенштейна и Укконена. При этом ни один из методов не обладает языковой зависимостью и может быть использован для обработки текстов на различных языках. Однако для некоторых слов может понадобиться одинаковое число модификаций строки. Например, слово с ошибкой *мук* с одинаковой похожестью может соответствовать названию растений: *лук*, *бук*, *дуб*.

Для исправления таких ошибок предлагается применять фонетические алгоритмы.

Фонетические алгоритмы. Основное назначение фонетических алгоритмов определение оценки похожести слов на основе их фонетического сходства. Все существующие на сегодняшний день фонетические алгоритмы используют кодирование слов в зависимости от особенностей произношения текста.

Применение фонетических кодов позволяет повысить качество сравнения слов при неправильном или различном написании. В основном фонетические алгоритмы используются для сопоставления фамилий. Нам представляется также достаточно эффективным использование фонетических алгоритмов для сопоставления пользовательских строк с эталонными значениями из различного рода классификаторов.

Предполагается комплексно применять фонетические алгоритмы в совокупности с алгоритмами нечеткого сравнения строк для повышения качества выявления и автоматического устранения орфографических ошибок исходных (сырых) данных.

Практически все фонетические алгоритмы, ориентированы на использование фонетических правил английского языка. Реализованы некоторые модификации алгоритмов, например, для французского, испанского языков [4]. Существуют и другие адаптации фонетических алгоритмов для языков, отличных от английского. Как правило, в этом случае используется транслитерация, а в качестве алгоритма используются вариации Soundex [4, 5]. Транслитерация в большинстве случаев не позволяет учесть особенности фонетики искомого языка. Для обработки строк на русском языке, с учетом фонетических особенностей предлагается применять разработанный авторами алгоритм Polyphon [6].

При обработке текстов на русском языке, фонетические алгоритмы целесообразно использовать для сопоставления слов, представленных в виде имен существительных, в именном падеже. Таким образом, этот метод эффективно использовать для идентификации значений, которые можно сопоставить с классификаторами. Связано это с тем, что русский язык является морфологически богатым и изменение словоформы может существенно отразиться на фонетическом представлении слова.

Практическая апробация.

Представленный в работе комплексный подход позволил обработать и загрузить в информационно-аналитическую систему (<http://tbd.icc.ru/>) слабоструктурированные данные об активности иксодовых клещей на территории Иркутской области и Республики Бурятия. В частности методы очистки использовались при обработке дат укусов, результатов исследования и геокодирования. Это позволило сформировать тематическую карту укусов. Инструментальные средства для комплексной очистки данных также использованы для интеграции разнородной информации об опасных геологических процессах в байкало-монгольском регионе, а также для обработки данных по социально-экономическому развитию Иркутской области.

Заключение. В работе рассмотрены возможности по применению комплексного подхода к очистке данных, в том числе для подготовки их к интеграции. Представленные методы использованы для интеграции разнородных данных в разрешении различных научных задач.

Работа выполнена при частичной финансовой поддержке РФФИ (грант № 17-47-380007_p_a, 17-57-44006_Монг_a, 18-07-00758_a).

Результаты получены при использовании сетевой инфраструктуры ЦКП «Интегрированная информационно-вычислительная сеть Иркутского научно-образовательного комплекса» (<http://net.icc.ru>).

ЛИТЕРАТУРА

- [1] Loshin D. Master Data Management. Morgan Kaufmann Publisher. - 2009. – 273 p.
- [2] Shigarov A., Paramonov V., Belykh P., Bondarev A. Rule-Based Canonicalization of Arbitrary Tables in Spreadsheets // Communications in Computer and Information Science. Springer.– 2016.– Vol. 639, pp. 78-91.

- [3] *Maletic J.I., Marcus A.* Data Cleansing. In: Maimon O., Rokach L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.
- [4] *Alotaibi Y., Meftah A.* Review of distinctive phonetic features and the Arabic share in related modern research // Turkish Journal of Electrical Engineering & Computer Sciences 2013, Vol. 21 Issue 5, pp.1426-1439
- [5] *Howida A. Shedeed A.* New Intelligent Methodology for Computer based Assessment of Short Answer Question based on a new Enhanced Soundex phonetic Algorithm for Arabic Language // International Journal of Computer Applications (0975 – 8887) Volume 34–No. 10, November 2011. pp. 40-47.
- [6] *Paramonov V., Shigarov A., Ruzhnikov G. Belykh, P.* Polyphon: An Algorithm for Phonetic String Matching in Russian Language. // Communications in Computer and Information Science. — 2016. — V. 639, — pp. 568 – 579.