Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Кафедра информационно-аналитических систем

Калина Алексей Игоревич

Оценка качества слабоструктурированных данных при сопоставлении независимых источников

Выпускная квалификационная работа

Научный руководитель:
д. ф.-м. н., профессор Новиков Б. А.

Рецензент:
архитектор ПО ЗАО "Диджитал Дизайн" Котов А. В.

Санкт-Петербург
2018

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Chair of Analytical Information Systems

Kalina Alexey

Quality assessment of semi-structured data by independent sources matching

Graduation Project

Scientific supervisor:
professor Boris Novikov

Reviewer:
software architect at Digital Design Alexander Kotov

Saint-Petersburg
2018

Оглавление

Введение

В наши дни объемы данных увеличиваются более чем в два раза каждые два года [9]. Количество ошибок и несоответствий с реальным миром растет вместе с объемами данных. Качество данных является обширной и все более актуальной темой в современном мире. Разные авторы определяют термин *качество данных* по-разному. Одни из них утверждают, что это степень пригодности данных для конкретного использования [1][11]. Другие делают упор на том, что это понятие многомерное и складывается из точности, полноты и других критериев [21][20][23]. Оценка качества данных – первый и очень важный шаг в трудоемком процессе, который называется *Улучшение качества данных*.

В течение нескольких последних десятилетий были разработаны различные методы оценки качества данных [22][18]. Большинство из них относятся к реляционной модели данных и основываются на анализе отдельных значений без использования других таблиц. Исключением является метод кросс-доменного анализа, который позволяет обрабатывать избыточность и несогласованность данных в нескольких таблицах [6]. В этой работе предлагается метод оценки качества, основанный на сопоставлении нескольких источников. Этот подход позволяет определить качество экземпляра данных в контексте различных критериев и с применением нескольких метрик для оценки.

Цель работы заключается в разработке нового подхода к оценке качества данных. Многообразие форматов и моделей представления данных существенно усложняет эту задачу. Изучение способов обработки неструктурированных и слабоструктурированных данных продолжается, хотя еще не так давно основное внимание исследователей привлекали данные, представленные в реляционной модели и имеющие четкую структуру. Подход, который описывается в этой работе, учитывает текущие тенденции и предоставляет возможность оценивать качество слабоструктурированных данных. Под этим термином мы понимаем возможное отсутствие схемы данных с фиксированными типами

данных, иерархическую структуру атрибутов и возможный пропуск атрибутов. В экспериментах при оценке использовались наборы данных, представленные в слабоструктурированной форме. Они имеют различную структуру, модель и формат представления.

В разделе *Качество данных* излагаются основные теоретические сведения, необходимые для описания работы метода. Сначала мы расширим понятие жизненного цикла качества данных на случай использования нескольких источников данных. Оно показывает, какое место в процессе улучшения качества данных занимает оценка. После этого мы определим критерии и метрики, с помощью которых можно оценивать качество данных.

Далее описан процесс сопоставления источников, заимствованный из процессов интеграции данных. Сопоставление источников является ключевой особенностью предлагаемого метода оценки качества данных. Интеграция данных может быть одним из возможных способов улучшения качества данных [8]. Она представляет собой комплекс задач, возникающих и в научных исследованиях (использование данных из разных биоинформационных репозиториев), и в коммерческой сфере (при объединении баз данных схожих фирм). Суть интеграции данных в слиянии записей, соответствующих одному и тому же объекту реального мира, из различных источников.

В разделе *Метод* описывается подход к оценке качества данных. В разделе *Эксперименты* описаны результаты применения метода для оценки качества коллекции о книжных публикациях и трех коллекций с информацией о футболистах. Цель экспериментов состоит в оценке стабильности результатов применения метода. В *Заключении* подводится итог работы и озвучиваются задачи для будущих работ.

Отметим, что эта работа была представлена на конференции SEIM'18 [12], прошла этап пост-рецензирования и будет опубликована в RSCI.

1. Качество данных

1.1. Жизненный цикл

В нашей работе оценка качества данных рассматривается как часть процесса улучшения качества данных. Для описания этого процесса модифицируем понятие *жизненного цикла качества данных*, которое было введено авторами методологии Luzzu для оценки связанных данных [5]. Использование метода оценки качества на основе сопоставления источников требует дополнительного этапа в начале цикла. В итоге жизненный цикл включает в себя 6 этапов:

1. *Выбор стратегии сопоставления источников.* При оценке качества данных методом сопоставления источников добавляется первый этап. Он включает в себя идентификацию и анализ взаимодействия нескольких источников.

2. *Определение метрик.* Для каждой предметной области и конкретного приложения может иметь место свое понимание термина качества данных. Поэтому определение критериев, которые имеют более весомое значение, является первым шагом в этом процессе. Далее определяются метрики, по которым будет проводиться оценка качества.

3. *Оценка.* На этом этапе набор данных оценивается по метрикам качества, определенным на предыдущем этапе жизненного цикла. В результате этого этапа появляется информация о наличии ошибок различного рода. По окончании этапа информация об ошибках структурируется в отчетах и передается на следующую стадию жизненного цикла.

4. *Очистка данных.* Для обеспечения высокого качества и постоянного его улучшения требуется произвести очистку для набора данных. Очисткой данных называется процесс исправления ошибок в коллекции. Некоторые операции очистки, например – по восста-

новлению пропущенных данных, могут быть произведены автоматически. В общем случае этот процесс требует вмешательства человека.

5. *Хранение, каталогизация и архивирование.* На этом этапе наборы данных, возможно очищенные, хранятся и архивируются вместе с их метаданными качества. Коллекции публикуются и каталогизируются на основе различных критериев качества.

6. *Анализ.* Анализ метаданных позволит определить, какие наборы данных требуют новой итерации жизненного цикла качества.

Эта работа сфокусирована на первых трех этапах жизненного цикла качества данных.

1.2. Критерии и метрики

Оценка качества данных включает в себя измерение *критериев* качества. Критерии – свойства данных, которые могут быть измерены или оценены в соответствии со стандартами и использованы для определения качества данных [23]. В разных областях ключевыми могут являться разные характеристики. Например, анализ статистических данных обычно требует значительного и репрезентативного количества данных для проведения анализа. В этом случае важно способствовать полноте, допуская несогласованность. И наоборот, при публикации таблицы результатов экзаменов студентов важнее иметь проверенные на непротиворечивость результаты, чем полные.

Критерии могут рассматриваться как некоторые характеристики набора данных. Сами по себе критерии не предоставляют никаких количественных мер, но могут быть измерены с помощью *метрик оценки качества* [1]. Эти метрики разрабатываются с учетом конкретного приложения и позволяют оценивать качество данных по конкретным критериям. При этом одому критерию качества может соответствовать несколько метрик. В данной ситуации термин *метрика* рассматривается не в общеизвестном математическом смысле.

Существуют различные варианты разделения критериев качества на группы. К примеру, Кристиан Бизер определяет 3 категории в соответствии с типом оцениваемых данных: *Информация о контенте*, *Информация о контексте* и *Рейтинги* самих данных и их поставщиков [2]. Другой вариант разделения представлен в статье Завери, в которой рассматриваются критерии связанных данных: *Доступность*, *Внутренние*, *Контекстные* и *Репрезентативные* [17]. В этом разделе мы рассмотрим критерии группы Информации о контенте или же в другой нотации – Внутренней группы. В различных источниках в эту группу входят: синтаксическая и семантическая точность, полнота, согласованность и другие. Разберем некоторые из них.

В общем случае можно определить *точность* как степень достоверности данных, то есть как меру совпадения значения в наборе данных со значением объекта реального мира [15]. При этом точность бывает двух видов: *синтаксическая* и *семантическая*.

1. *Синтаксическая точность*. Этот критерий качества определяет совпадение значения атрибута со значением, принадлежащим предметной области, соответствующей этому атрибуту [1]. В роли метрик синтаксической точности могут выступать различные функции расстояния. Например, *расстояние Левенштейна*, которое определяется как минимальное число вставок, перестановок и удалений символов для преобразования одной стоки в другую [24]. Другой пример метрики – использование синтаксических правил, таких как проверка на допустимость символов или на соответствие значения шаблону [17][23].

2. *Семантическая точность*. В свою очередь семантическая точность определяется как соответствие заданного значения реальному [14]. Для вычисления этого критерия можно использовать статистические методы, например, поиск выбросов [2]. Другой способ определения семантической точности значения – поиск данных об одном объекте в различных источниках и определение корректности путем сравнения [1].

3. *Полнота.* Полнота набора данных – это степень, в которой в наборе данных присутствуют все релевантные для него данные. Метрикой может выступать отношение количества непустых значений коллекции к мощности всего набора данных [23].

Отметим, что рассмотренные критерии отличаются от одноименных понятий, используемых в теории информационного поиска.

1.3. Сопоставление источников

Интеграция данных включает в себя объединение данных из различных источников и предоставление данных пользователям в унифицированном виде [7][3]. Она состоит из трех этапов:

1. *Сопоставление схем данных.* Задача на этом этапе состоит в том, что из схем представления данных нескольких независимых источников необходимо составить одну и предоставить отображение каждой из схем на результирующую.

2. *Связывание записей.* Данный этап также называют проблемой идентификации объекта. Задача заключается в определении того, соответствуют ли две сущности из разных наборов данных одному объектом реального мира.

3. *Слияние данных.* При слиянии данных происходит разрешение конфликтов. Конфликтами называются ситуации, когда один и тот же реальный объект имеет в нескольких источниках различные значения атрибутов результирующей схемы.

Процесс интеграции данных существенно пересекается с подходом, предлагаемым в этой работе. В ходе метода, описываемого в следующем разделе, также решаются задачи сопоставления схем данных и идентификации объекта. Тем не менее, конечным результатом интеграции данных является новый набор данных, полученный из нескольких независимых источников. В то время как предлагаемый подход позволяет

получить оценку качества существующего набора данных и впоследствии применять различные техники для улучшения общего качества данных.

2. Метод

2.1. Стратегия идентификации объекта

Метод предполагает, что для поставленной задачи оценки качества идентифицированы источники данных, с помощью которых она будет решаться. В качестве источников могут выступать несколько наборов данных с пересекающимися множествами объектов реального мира. На этом этапе необходимо решить типичную задачу для сценариев интеграции данных – задачу идентификации объекта. Также в других источниках ее называют проблемой слияния/очистки [10]. Суть задачи заключается в выборе алгоритма для определения сущностей из различных источников, соответствующих одному объекту реального мира.

Существуют разные способы решения этой проблемы, и метод позволяет использовать любой из них для достижения конечной цели – результатов оценки качества данных. Рассмотрим два подхода к решению задачи идентификации объекта.

Простейшим способом сопоставления объектов из разных наборов данных является определение уникальных идентификаторов (ключей), имеющихся у сущностей в коллекциях. Тогда объекты с одинаковыми идентификаторами можно считать одним реальным объектом. Этот способ меньше всего подвержен ошибкам, но не во всех случаях имеющиеся данные обладают уникальным идентификатором.

Использование предикатов соответствия [13] – другой способ для решения этой задачи. К каждому атрибуту во всех наборах данных применяются атомарные метрики схожести. Они вычисляют насколько два значения близки друг другу. Среди таких функций выделяют расстояние Левенштейна [24], метрику Джаро-Уинклера, фонетические функции и другие [19]. Далее к результатам этих вычислений применяют более общие функции, уровня записи. Как правило, на этом этапе применяются статистические и вероятностные алгоритмы. На последнем шаге метода производится анализ контекста сущностей, включающий внешние ключи с ассоциативными и структурными ссылками.

2.2. Определение критериев и метрик

Рассмотрим критерии и метрики качества, использовавшиеся в эксперименте с книжными публикациями. Все измерения проводились на уровне атрибутов данных. Оценка применялась к двум типам данных - строковым значениям и элементам, хранящим дату.

Так как оценка в контексте описываемого метода проводится с использованием значений соответствующих объектов из разных источников, метрики качества основываются на сравнении содержимого атрибутов этих объектов. При работе со строковыми значениями оцениваемое значение принимается за неполное, если оно является подстрокой соответствующего ему значения из другого источника. Оценка полноты дат заключается в проверке на наличие значения в необходимом поле. Если необходимый атрибут имеет какое-либо содержимое, то он в достаточной мере представляет реальный мир. Правильность же этого значения нужно проверять другими критериями. Кроме того, важно не забывать, что при оценке слабоструктурированных данных необходимое поле может отсутствовать вовсе. Это также пример неполноты.

Синтаксическая точность основывается на принадлежности значения тому множеству всех вариантов, которые соответствуют исследуемому атрибуту. Поэтому оценка для синтаксической точности проводится без использования объекта другого источника. В случае со строковыми атрибутами всевозможные значения определяются естественным языком. Для проверки на синтаксическую точность таких данных можно использовать орфографические анализаторы или другие инструменты, оценивающие правильность написанного текста. При обработке атрибутов-дат важно учитывать определение формата представления полей этого типа в системе. Например, в определенных реализациях даты могут содержать только год, в других – только полноценные записи, включающие день или в разных системах могут использоваться разные разделители для чисел. Все эти требования определяются политиками, разработанными до оценки качества. В качестве метрики синтаксической точности можно использовать предикат, про-

веряющий соответствие значения регулярному выражению.

При использовании метода сопоставления источников для оценки качества могут возникнуть ситуации, когда оба значения соответствующих объектов синтаксически точны, но тем не менее различны. Например, такое течение событий может произойти в случае, когда одно из значений неточно семантически. Это означает, что данные значения соответствуют разным объектам реального мира. Для проверки этой гипотезы в работе использовались возможности современных поисковых движков. Алгоритм заключается в том, что к поисковой системе отправляются два запроса, содержащие строковые значения первого и второго объекта. Далее проводится анализ выдачи на превышение порогового значения количества совпавших результатов из двух запросов. В случае такого совпадения делается вывод о соответствии двух сущностей одному реальному объекту и как следствие семантическая точность строковых значений. Для дат этот критерий можно интерпретировать как проверку на теоретическую реальность такого значения. То есть, к примеру, дата публикации книги не может быть указана позже текущего года или до даты издания первого варианта. Такие правила также предварительно определяются политиками оценки качества.

2.3. Оценка качества данных

На этом этапе проводится основная вычислительная работа метода. Псевдокод алгоритма оценки качества данных приведен в блоке Algorithm 1. Ход алгоритма:

1. Для объекта, подвергающегося оценке, проводится поиск соответствующего ему объекта в другом источнике. Эта процедура выполняется в соответствии со стратегией идентификации объекта, определенной на первом шаге метода. В результате получаем пару объектов, являющихся одной сущностью реального мира.

2. Теперь необходимо провести оценку качества по интересующим критериям с использованием выбранных метрик и найденного объ-

екта из другого источника. Существуют разные подходы для применения метрик качества к данным. Метод не ограничивает разработчика в выборе таких подходов. Для примера, при проведении экспериментов для применения метрик качества использовалась идея конвейера. Вычисление оценки по разным критериям проводилось последовательно, и от результатов оценки текущей характеристики зависит то, какой критерий будет оцениваться следующим. Рассмотрим подробнее конвейер для эксперимента с книжными публикациями.

(a) Первоначально необходимо сравнить значения атрибутов соответствующих объектов. Для тех полей, которые содержат идентичные данные, оценка не требуется. Мы пользуемся предположением о том, что эквивалентные значения означают правильность содержимого этого атрибута в обоих источниках данных.

(b) К несовпадающим атрибутам объекта применяются метрики для оценки полноты. В случае неполноты атрибута, дальнейшая оценка для него не требуется, так как не отражающее в достаточной мере значение реального мира нельзя проверить на точность. Если атрибут обладает свойством полноты, переходим к следующему шагу конвейера.

(c) На этом этапе конвейера необходимо определить какой анализ выполнять: синтаксической или семантической точности. Для этого используется предположение о том, что похожие, но не идентичные значения означают, что данные могут иметь синтаксическую неточность. В свою очередь сильно отличающиеся данные могут означать, что в них есть семантическая ошибка. Для сравнения значений можно использовать разные метрики сравнения, в зависимости от предметной области. При проведении экспериментов для сравнения данных применялось расстояние Левенштейна.

(d) Для близких значений проводится оценка синтаксической точ-

ности, а для далеких - семантической. В случае, если при вычислении не было обнаружено ошибок этих типов, то качество данных исследуемого атрибута определяется как неизвестное и требует ручного анализа.

(e) Алгоритм конвейера применяется ко всем объектам набора данных.

Algorithm 1 Алгоритм оценки качества данных

```
1: procedure DATA QUALITY ASSESSMENT
2:     reports ← Array()
3:     for book ∈ books do
4:         book2 ← IdentificationObject(book)
5:         report ← Initialize()
6:         for attr ∈ book.Attributes do
7:             if Equals(attr, book2) then
8:                 report.Add(attr, accurate)
9:             else if IsNotComplete(attr, book2) then
10:                report.Add(attr, incomplete)
11:            else if Leven(attr, book2) < thresh then
12:                if IsNotSyntAccur(attr, book2) then
13:                    report.Add(attr, syntacticInaccurate)
14:                end if
15:            else if IsNotSemAccur(attr, book2) then
16:                report.Add(attr, semanticInaccurate)
17:            else
18:                report.Add(attr, unknown)
19:            end if
20:        end for
21:        reports.Add(report)
22:    end for
23:    return reports
24: end procedure
```

2.4. Демонстрация результатов

На заключительном этапе проводится демонстрация полученных результатов. После проведения процедуры оценки качества данных, необ-

ходимо представить результаты в том виде, который позволит экспертам понять, как можно повысить качество этих данных. Как правило, реализация этого шага метода требует меньших трудозатрат нежели непосредственная оценка, однако, некачественное проведение этого этапа может перечеркнуть всю проделанную ранее работу. Рассмотрим два подхода к демонстрации результатов, использовавшихся при экспериментах:

1. Один из способов заключается в предоставлении общей картины качества данных по набору данных. Это достигается путем вычисления метрик по полученным результатам для разных критериев качества. Например, для того чтобы определить общую полноту набора данных, можно рассчитать отношение количества атрибутов, в которых полнота достигается, к неполным атрибутам. Получив общую картину качества данных, принимается решение о необходимости улучшения текущего состояния набора данных и происходит выбор подходящих способов для этого.

2. В противоположность первому подходу можно предоставлять результаты оценки качества в виде отчетов по каждому содержащему ошибки объекту. Такой отчет предоставляет информацию об атрибуте, в котором есть недочет, то есть его значение и критерий качества, который не удовлетворяет определенным политикам. Результаты в такой форме можно автоматически отправлять экспертам для ручного исправления и повышения качества набора данных.

3. Эксперименты

Было проведено два эксперимента в разных предметных областях. В первом эксперименте используются два набора данных, асимметричных по размеру. Так как в большей коллекции подавляющая часть записей не имеет пересечения с другой коллекцией, оценка качества проводилось только для меньшей из них. Во втором эксперименте оценивается качество трех наборов данных.

3.1. Книжные публикации

Для проведения первого эксперимента использовались два набора данных с информацией о книжных публикациях. Коллекции были взяты из двух независимых источников: сообщество “Book-Crossing” [4] и портал “Open Library” [16]. Наборы данных представлены в слабоструктурированном виде, в форматах csv и json соответственно. Первая коллекция насчитывает порядка 270 тыс. записей, в то время как вторая – около 25 млн. Для оценки качества использовался набор данных с информацией о публикациях, взятый из сообщества по буккроссингу.

Listing 1: Запись из набора данных Open Library

```
{
  "title": "To Kill a Mockingbird",
  "publishers": [
    "The Audio Partners"
  ],
  "authors": [
    {
     "key": "/authors/OL498120A"
    }
  ],
  "isbn_10": [
    "157270036X"
  ],
  "publish_date": "May 1997"
}
```

Рис. 1: Отчеты качества данных

Поле	Значение	Дефект качества
ISBN	451525221	
Автор	Nathaniel Hawthorne	
Название	Scarlet Letter *The Scarlet Letter (Signet Classics)*	Неполнота
Год издания	1993 *1-Aug-59*	Неизвестно
Издательство	Signet Book *Signet Classics*	Семантическая неточность

Поле	Значение	Дефект качества
ISBN	342311360X	
Автор	Gabriel Garcia Marquez	
Название	Die Liebe in Den Zelten *Die liebe in den Zeiten der Cholera*	Семантическая неточность
Год издания	0 *1991*	Неполнота
Издательство	Deutscher Taschenbuch Verlag (DTV) *DTV*	

Поле	Значение	Дефект качества
ISBN	743486226	
Автор	Dan Brown	
Название	Angels & Demons *Angels & demons*	Синтаксическая неточность
Год издания	2003 *2003*	
Издательство	Atria *Atria Books*	Неполнота

Поле	Значение	Дефект качества
ISBN	067102535X	
Автор	Ann Rule	
Название	Last Dance, Last Chance (Ann Rule's Crime Files) *Last dance, last chance and other true cases*	Семантическая неточность
Год издания	2002 *2003*	Неизвестно
Издательство	Pocket *Pocket Books*	Неполнота

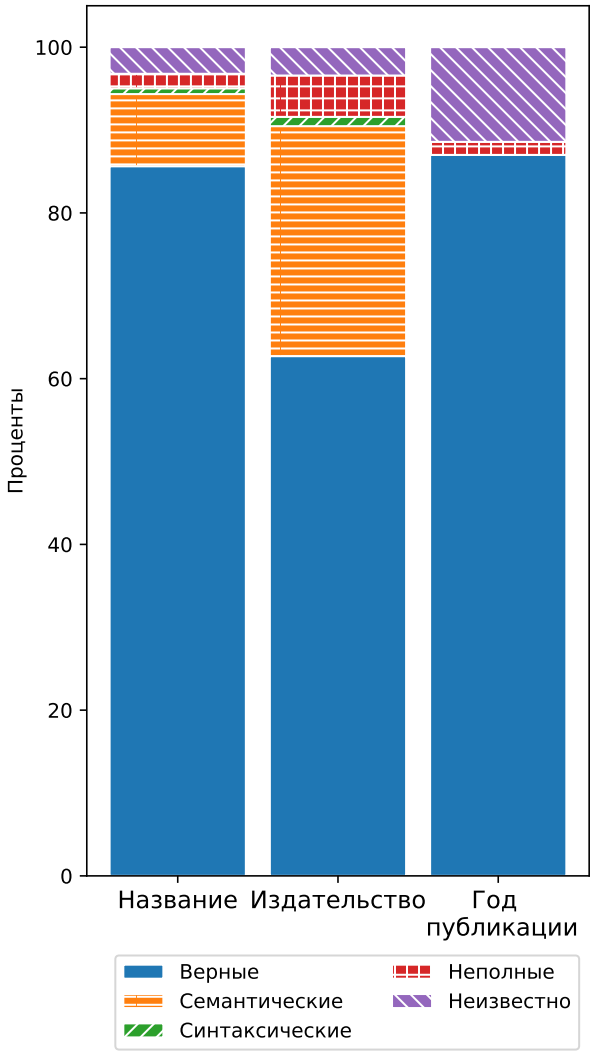
Listing 2: Запись из набора данных Book-Crossing

```
''157270036X'';''To Kill a Mockingbird'';''Harper Lee'';''1997'';''Audio Partners''
```

Оценка качества данных производилась по отношению к следующим атрибутам: название книги, дата публикации и издательство. Проблема идентификации объекта для данной задачи решалась сопоставлением в наборах данных ключевого поля ISBN10 (международный стандартный книжный номер). Этот атрибут является уникальным идентификатором любой книжной публикации. На листингах 1 и 2 представлены примеры записей из рассматриваемых наборов данных, описывающие однин и тот же объект реального мира (часть полей не показана).

Оценка качества данных проводилась по критериям полноты, синтаксической и семантической точности в соответствии с метриками, описанными в предыдущем разделе. Последний этап метода заключается в демонстрации полученных результатов. Два разных подхода к представлению качества данных проиллюстрированы на рисунках 1 и 2.

Рис. 2: Статистика качества по набору данных ”Book-Crossing”



Для определения стабильности результатов применения метода, набор данных “Book-Crossing” был разбит на десять равномощных независимых частей, и для каждой из них была произведена оценка трех атрибутов. Поле Publication Year оценивалось по трем критериям, однако синтаксический анализ не выявил ошибок, поэтому таблица результатов не содержит соответствующего им поля. Результаты эксперимента приведены в таблицах 1, 2, 3.

Таблица 1: Качество атрибута Title, %

	1	2	3	4	5	6	7	8	9	10
Верные	86,83	87,26	86,94	87,04	87,15	85,89	86,74	86,49	86,84	87,54
Синтаксические ошибки	0,59	0,61	0,55	0,51	0,65	0,68	0,71	0,67	0,72	0,68
Семантические ошибки	7,02	6,97	8,13	8,18	8,27	8,38	8,16	8,25	8,53	7,66
Неполные	1,96	1,78	1,50	1,60	1,35	1,69	1,55	1,65	1,38	1,57
Неизвестно	3,60	3,38	2,89	2,66	2,58	3,36	2,84	2,94	2,54	2,54

Таблица 2: Качество атрибута Publisher, %

	1	2	3	4	5	6	7	8	9	10
Верные	61,62	62,07	63,43	63,37	63,83	63,29	62,63	62,34	61,94	62,34
Синтаксические ошибки	0,88	1,00	0,89	0,51	1,14	1,15	1,49	1,61	1,42	1,10
Семантические ошибки	26,93	27,98	26,55	27,11	27,66	27,87	28,17	28,28	29,10	27,71
Неполные	7,38	5,89	5,81	5,02	4,22	4,04	3,95	4,15	4,05	5,41
Неизвестно	3,19	3,06	3,32	3,48	3,15	3,65	3,76	3,61	3,50	3,43

Таблица 3: Качество атрибута Publication Year, %

	1	2	3	4	5	6	7	8	9	10
Верные	43,51	41,82	40,32	40,12	45,91	40,78	38,83	41,28	40,02	41,95
Семантические ошибки	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,01
Неполные	1,43	1,83	1,62	1,64	1,70	1,43	1,64	1,54	1,49	1,39
Неизвестно	55,05	56,35	58,06	58,23	52,38	57,79	59,52	57,18	58,48	56,65

По результатам проведенного исследования видно, что большая часть оцененных записей являются верными. Также есть большая группа объектов с неизвестным критерием качества для года публикаций. Это обуславливается тем, что для несовпадающих значений, но удовлетворяющих условиям формата и предшествования текущему году, дополнительный анализ не проводится.

Для каждого критерия качества атрибутов были рассчитаны среднеквадратические отклонения. Результаты представлены в таблице 4. Максимальное значение среднеквадратического отклонения достигает 1,38%. Малый разброс результатов говорит о том, что результаты применения предложенного метода оценки качества стабильны.

Таблица 4: Среднеквадратичные отклонения, %

	Название	Издательство	Год публикации
Верные	0,43	0,71	1,38
Синтаксические ошибки	0,07	0,31	0
Семантические ошибки	0,52	0,7	0
Неполные	0,17	1,07	0,13
Неизвестно	0,36	0,21	1,37

Полный набор данных портала “Open Library”, насчитывающий более 25 млн записей, был проиндексирован в поисковом движке *ElasticSearch*. Данная процедура позволила быстро искать публикации по их *ISBN*. Модуль оценки качества данных для обоих экспериментов был написан на *Java*. В роли функции расстояния использовалось расстояние Левенштейна, реализованное в библиотеке *info.debatty*. Для проверки орфографии применялась библиотека *org.languagetool*. С помощью *Yandex XML API* были загружены все необходимые для семантической оценки результаты поисковых запросов.

3.2. Футболисты

Второй эксперимент заключался в оценке качества данных трех коллекций с информацией о футболистах. Для этого были собраны данные со спортивных сайтов ”championat.com”, ”bombardir.ru” и ”eurofootball.ru”. Результирующие наборы данных содержат 3808, 3377 и 3495 записей, соответственно. Коллекции создавались с единой схемой данных, поэтому проблемы сопоставления схем данных не возникло. Пример записи из набора данных ”championat.com” представлен в листинге 3.

Listing 3: Запись из набора данных ”championat.com”

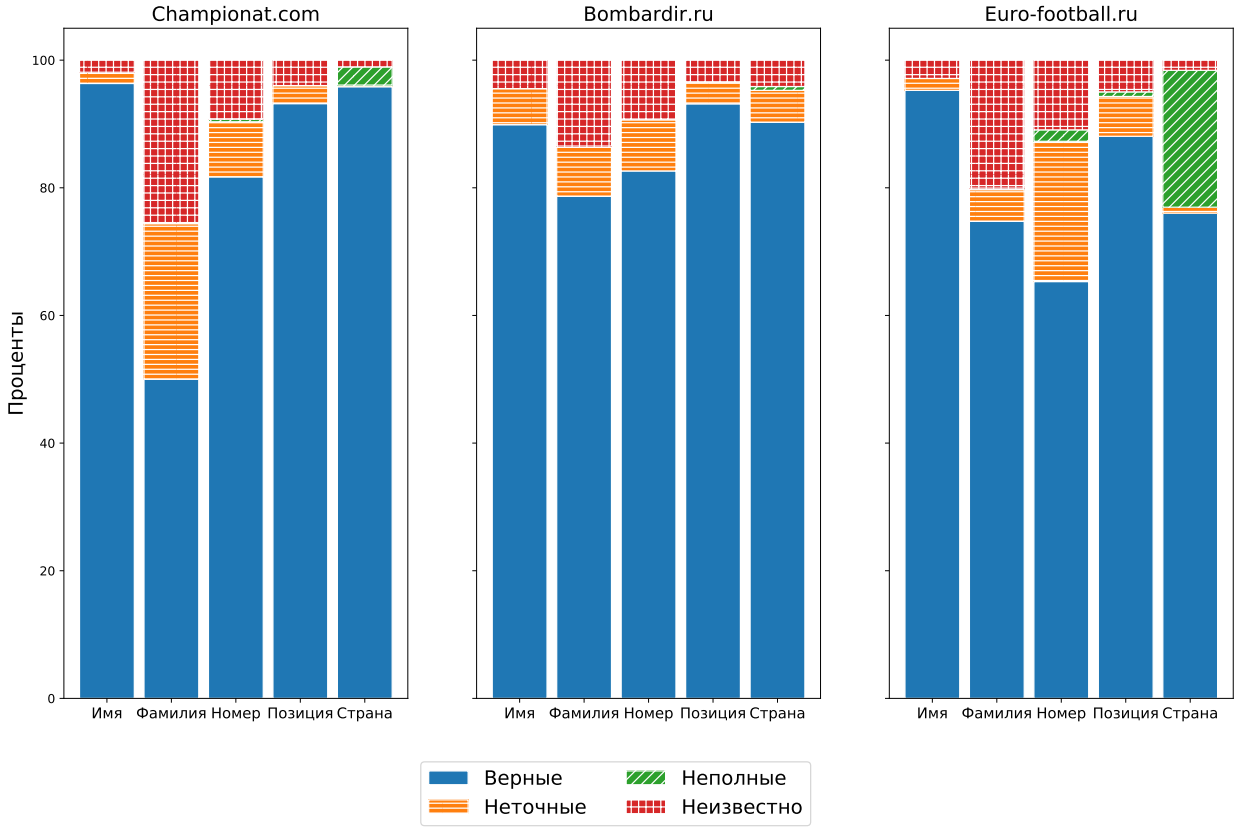
```
{
    "FirstName": "Рафаэл",
    "LastName": "Толой",
    "Nationality": [
        "Бразилия",
        "Италия"
    ],
    "Position": "защитник",
    "BirthDate": "1990-10-10",
    "Height": 185,
    "Weight": 75,
    "Number": 3
}
```

Оценка качества проводилась по пяти атрибутам: Имя, Фамилия,

Позиция, Номер, Гражданство. Проблема идентификации объекта решалась с помощью расстояния Левенштейна по двум атрибутам: Имя и Фамилия. Если вычисленное расстояние оказалось меньше порогового значения, то записи признавались соответствующими одному объекту.

Для оценки использовались два критерия качества: точность и полнота. В этом эксперименте точность не разделялась на синтаксическую и семантическую, так как решение об ошибке принималось только на основе различия значений. Если в двух записях из трех, соответствующих одному объекту, значения атрибута совпадали, то они признавались точными, а третье — неточным. Для оценки полноты производилась проверка на отсутствие значения у атрибута. Кроме того, для поля Гражданство неполнота имела место, если список не содержал все необходимые элементы.

Рис. 3: Результаты оценки качества для эксперимента с футболистами



Результаты оценки качества трех наборов данных представлены на рисунке 3. Высокий процент неточных значений атрибута Фамилия связан с различным представлением фамилий иностранных футболистов

в источниках данных. Для атрибута Гражданство набора данных Euro-Football.ru 21% процент от всех значений – неполные. Это обуславливается ограничением источника данных, заключающимся в том, что каждому футболисту на сайте соответствует только одно гражданство.

Заключение

В ходе работы были изучены различные критерии и метрики оценки качества данных. Также были рассмотрены существующие способы решения задачи оценки качества данных. В результате был разработан новый метод, основанный на сопоставлении независимых источников данных. Этот подход достаточно гибок и может использоваться в различных предметных областях и задачах.

В качестве примера применения метода было проведено два эксперимента. В первом из них использовались две коллекции с информацией о книжных публикациях, для одной из которых была проведена оценка качества. Во втором эксперименте качество данных было оценено для трех наборов данных о футболистах. Описанный метод показал стабильные результаты на этих наборах данных.

Целью будущих работ является использование разработанного метода для оценки качества других данных. Под этим понимается, как использование других атрибутов с новыми типами данных, так и оценка других коллекций.

Список литературы

[1] Batini C., Scannapieca M. Data quality. // Springer-Verlag, Berlin, Germany. — 2006. — P. 19–31.

[2] Bizer Christian, Cyganiak Richard. Quality-driven information filtering using the WIQA policy framework // Web Semantics: Science, Services and Agents on the World Wide Web. — 2009. — Vol. 7, no. 1. — P. 1–10.

[3] Bleiholder Jens, Naumann Felix. Data fusion // ACM Computing Surveys (CSUR). — 2009. — Vol. 41, no. 1.

[4] Book-Crossing Dataset. — URL : http://www2.informatik.uni-freiburg.de/ cziegler/BX/.

[5] Debattista Jeremy, Auer SÖren, Lange Christoph. Luzzu—A Methodology and Framework for Linked Data Quality Assessment // Journal of Data and Information Quality (JDIQ). — 2016. — Vol. 8, no. 1. — P. 4.

[6] Discover dependencies from data—a review / Jixue Liu, Jiuyong Li, Chengfei Liu, Yongfeng Chen // IEEE transactions on knowledge and data engineering. — 2012. — Vol. 24, no. 2. — P. 251–264.

[7] Dong Xin Luna, Srivastava Divesh. Big data integration // Data Engineering (ICDE), 2013 IEEE 29th International Conference on / IEEE. — 2013. — P. 1245–1248.

[8] Endler Gregor. Data quality and integration in collaborative environments // Proceedings of the on SIGMOD/PODS 2012 PhD Symposium / ACM. — 2012. — P. 21–26.

[9] Gantz John, Reinsel David. Extracting value from chaos // IDC iview. — 2011. — Vol. 1142, no. 2011. — P. 1–12.

[10] Hernández Mauricio A, Stolfo Salvatore J. Real-world data is dirty: Data cleansing and the merge/purge problem // Data mining and knowledge discovery. — 1998. — Vol. 2, no. 1. — P. 9–37.

[11] Herzog Thomas N., Scheuren Fritz J., Winkler William E. What is Data Quality and Why Should We Care? // Data Quality and Record Linkage Techniques. — New York, NY : Springer New York, 2007. — P. 7–15. — ISBN: 978-0-387-69505-1. — URL: https://doi.org/10.1007/0-387-69505-2_2.

[12] Kalina A., Novikov B. Quality assessment of semi-structured data by independent sources matching. — Accepted at SEIM, 2018. — URL: http://seim-conf.org/en/about/accepted-papers/.

[13] Koudas Nick, Sarawagi Sunita, Srivastava Divesh. Record linkage: similarity measures and algorithms // Proceedings of the 2006 ACM SIGMOD international conference on Management of data / ACM. — 2006. — P. 802–803.

[14] Lei Yuangui, Uren Victoria, Motta Enrico. A framework for evaluating semantic metadata // Proceedings of the 4th international conference on Knowledge capture / ACM. — 2007. — P. 135–142.

[15] Nelson R Ryan, Todd Peter A, Wixom Barbara H. Antecedents of information and system quality: an empirical examination within the context of data warehousing // Journal of management information systems. — 2005. — Vol. 21, no. 4. — P. 199–235.

[16] Open Library. — URL : https://openlibrary.org/.

[17] Quality assessment for linked data: A survey / Amrapali Zaveri, Anisa Rula, Andrea Maurino et al. // Semantic Web. — 2016. — Vol. 7, no. 1. — P. 63–93.

[18] Sæbø Hans Viggo. Quality Assessment and Improvement Methods in Statistics–What works? // Statistika. — 2014. — Vol. 94, no. 4. — P. 5–14.

[19] Sumathi V. P. Kousalya K. Kalaiselvi R. A Comparative study on Syntax Matching Algorithms in Semantic Web // WSEAS TRANSACTIONS on COMPUTERS. — 2016.

[20] Veregin Howard. Data quality parameters // Geographical information systems. — 1999. — Vol. 1. — P. 177–189.

[21] Wang Richard Y, Kon Henry B, Madnick Stuart E. Data quality requirements analysis and modeling // Data Engineering, 1993. Proceedings. Ninth International Conference on / IEEE. — 1993. — P. 670–677.

[22] Woodall Philip, Oberhofer Martin, Borek Alexander. A classification of data quality assessment and improvement methods // International Journal of Information Quality 16. — 2014. — Vol. 3, no. 4. — P. 298–321.

[23] The six primary dimensions for data quality assessment : Rep. / Technical report, DAMA UK Working Group ; Executor: Nicola Askham, Denise Cook, Martin Doyle et al. : 2013.

[24] Левенштейн Владимир Иосифович. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук / Российская академия наук. — Vol. 163. — 1965. — P. 845–848.