УДК 004.9

ФАЛЬТИНА Е.В., студентка гр. ПИм-221 (КузГТУ)
Научный руководитель БУЙНАЯ Е.В., к.э.н, доцент (КузГТУ)
г. Кемерово

РЕШЕНИЕ ПРОБЛЕМЫ ДУБЛЕЙ В СПРАВОЧНИКАХ БАЗЫ ДАННЫХ «1С: БУХГАЛТЕРИЯ» С ПОМОЩЬЮ МАТЕМАТИЧЕСКОГО АЛГОРИТМА РАСЧЕТА КОСИНУСНОЙ ДИСТАНЦИИ

В любой без исключения организации при работе в программах обязательно происходит заполнение справочников. Справочники — это прикладные объекты конфигурации, позволяющие хранить в информационной базе данные, имеющие одинаковую структуру и списочный характер [1]. При этом если в организации отсутствует жесткое регулирование такого пополнения, то в ней часто возникает проблема дублей. При этом в базе также можно встретить перестановки слов и их сокращения, при этом самые разнообразные. Из-за всего этого список приходит в неустойчивое состояние, что приводит к производным сложностям учета: пересортице, отрицательным остаткам, избыточной инвентаризации.

Как следствие, очень важно систематизировать данные и исправить все ошибки, чтобы не потерять прибыль из-за проведения операций с неверными или ошибочными данными. В связи с этим весьма актуально решение проблем хаоса в справочниках при занесении первичных данных.

В ходе работы со справочниками были выявлены следующие проблемы:

- перестановка слов в наименовании («гвозди плоские» и «плоские гвозди»);
- излишняя информация (исходит из желания точнее описать сущность предмета);
- сокращения наименований (например, «пл. гвозди»);
- копирование лишних символов при переносе информации из других систем (например, перенос строки);
- числовые ошибки (например, вместо 250 гр. пишут 25 гр.);
- орфографические ошибки в словах (например, «плоские гвозьди»).

Если не решать данные проблемы, то в конечном итоге придётся столкнуться со следующими последствиями:

- сложность прочтения данных, введенных другим работником;
- наличие нескольких карточек с похожими названиями, а значит, и непонимание того, какую именно карточку использовать при поступлении/реализации товара;
- пересортица в базе по нескольким карточкам, при которой также неясно, какую карточку выбрать при списании;
- сложность при синхронизации с другими учетными системами.

Для решения данных проблем существует множество методов. Все их можно разделить на административные и инженерные. К административным методам относятся:

- ввод единых классификаторов;
- разработка и внедрение инструкций по ведению справочников;
- увольнение всех сотрудников и ввод информации самостоятельно (что долго и мешает другим делам);
- использование бумажных справочников (это неудобно и не технологично).

К инженерным методам относятся:

- ввод через сканеры (покупка доп. оборудования);
- загрузка через файлы обмена, почту;
- ввод с помощью математических методов и расчетов.

Нами предполагается разработка типа обработки, решающего выявленные проблемы на основании математических алгоритмов. В качестве исходных данных для разрабатываемой программы будут использованы записи справочников номенклатуры и контрагентов конфигурации «1С: Бухгалтерия»; их структура представлена в таблицах 1 и 2. Основная логика программы будет содержаться в Python. Также будет настроена интеграция между конфигурацией «1С: Бухгалтерия» и приложением на Python, так как в системе 1С необходимо будет вызывать функции из Python, а в основной программе — использовать записи из базы данных. Для этого будут использоваться встроенные библиотеки Python.

Таблица 1. Структура данных справочника "Номенклатура"

Реквизит	**Тип**
Наименование	Строка (250)
Вид номенклатуры	СправочникСсылка.ВидНоменклатуры
Единица измерения	СправочникСсылка.ЕдиницаИзмерения
Артикул	Число (250)
Цена реализации	Число (250)
Номенклатурная группа	СправочникСсылка.ГруппаНоменклатуры

Таблица 2. Структура данных справочника "Контрагенты"

Реквизит	**Тип**
Наименование	Строка (250)
Полное наименование	Строка (250)
Вид контрагента	ПеречислениеСсылка.ВидКонтрагента
ИНН	Число (12)
КПП	Число (9)
Страна регистрации	Строка (250)

В результате работы приложения в качестве выходных данных в базе будут храниться вектора слов (двоичные числа), словосочетаний и предложений, а также число, называемое косинусной дистанцией.

Для применения математических методов с входными данными необходимо поработать, т.е. сделать предобработку. Она заключается в следующем:

- необходимо удалить все междометия («в», «на» и т.д.);
- нужно удалить точки, запятые, двоеточия, т.е. все спецсимволы («*», «,» и т.д.);
- привести существительные к «начальной» форме (т.е. перевести в именительный падеж);
- привести все слова к нижнему регистру;
- произвести обратную транслитерацию слов (т.е. перевести все слова на русский).

После такой работы с данными необходимо выделить в них признаки. Признаки — это некоторые свойства класса-объекта, его характерные черты, уникально выделяющие его среди других классов-объектов. Признаками слов являются биплеты или триплеты. Биплет — это обычное разбиение слова на последовательности по две буквы с шагом смещения в один символ; триплет — всё то же самое, но здесь последовательность имеет по три буквы.

Однако для того, чтобы применять математические методы, необходимо будет также закодировать слова — то есть, в этом случае, представить их в виде чисел. Для этого удобно использовать гистограммы, так как для программистов гистограмма – это вектор. По своей сути гистограмма – графическое представление данных, содержащее информацию о том, сколько и каких событий случилось за определенный интервал. На ней по горизонтали откладывается номер биплета, а по вертикали — количество таких биплетов в предложении/слове.

Для такого рода работы первоначально нужно будет составить словарь гистограммы. Для этого можно вывести все возможные комбинации биплетов русского языка. Так, если исключить из алфавита ненужные нам экзотические буквы «Ъ» и «Ь», останется 31 буква. Всего получится 31 × 31 комбинаций — 961. Это значительное количество, которое можно попробовать уменьшить.

С целью уменьшения количества комбинаций можно использовать словарь и найти на основе его текста частоту появления каждого биплета. Так, если частота очень большая, то этот биплет будет отнесен к групповому биплету, т.е. биплету, который очень часто встречается в тексте. Частоты посчитаем так: количество раз появления биплета во всем тексте поделим на общее количество биплетов в тексте.

Появление в предложении биплета, частота появления которого очень высока, приводит к малому уменьшению информационной энтропии сообщения. Информационная энтропия — это мера непредсказуемости появления какого-

либо символа первичного алфавита. В свою очередь, первичный алфавит — это набор букв или, в данном случае, биплетов, которыми мы можем оперировать.

Следовательно, мы можем попробовать сгруппировать биплеты с большой частотой появления в один общий сводный биплет. На гистограмме такие биплеты будут «занимать» ровно одну позицию по координате X. В итоге у нас получится N биплетов — это и будет словарь.

Благодаря всему вышеперечисленному мы, имея некоторое предложение, можем посчитать количество биплетов в нем. В массиве длиной N, где каждая позиция соответствует отдельному биплету из словаря, проставим количество биплетов в предложении на позициях биплетов в гистограмме. В итоге мы получим гистограмму предложения. Эта гистограмма зависит не от порядка слов и биплетов, а только от конечного набора последних в предложении. На рисунке 1 представлен пример; числа, которые можно видеть на схеме, — это количество появления биплетов в проверяемом предложении.

Рисунок 1. Слово, представленное в виде вектора

После преобразования данных в нужную нам форму можно наконец применить математические методы. Наиболее удобно в данном случае будет использовать косинусную дистанцию между векторами.

Косинусное подобие — метрика, используемая для определения сходства между двумя векторами. Она вычисляется как косинус угла между двумя векторами и может принимать значения от −1 до 1. Значение 1 означает полное сходство, а значение −1 — полное несходство.

Косинусная дистанция (расстояние) получается, когда мы из единицы вычитаем косинусное подобие. В этом случае ситуация противоположна приведённой выше: чем меньше показатель, тем лучше.

$$1 - \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||} = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Рисунок 2. Формула расчета косинусной дистанции (где A и B — векторы слов, $||A||$ и $||B||$ — их длина, а n — количество элементов массива, соответствующих вектору)

По формуле, показанной на рисунке 2, рассчитываем расстояние между двумя словами (предложениями), представленными в виде векторов. Чем меньше это расстояние, тем одно слово (предложение) ближе ко второму по смыслу.

В пользовательском режиме программы 1С форма реализованной обработки будет выглядеть так, как представлено на рисунке 3, то есть под полем ввода наименования будет выпадающий список со всеми схожими значениями из базы, а справа от них – рассчитанная косинусная дистанция.

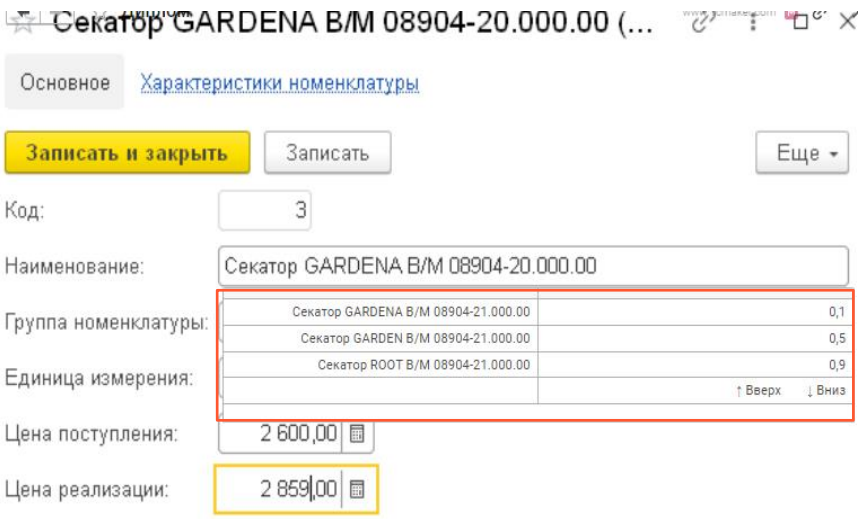

Рисунок 3. Макет приложения

В ходе воплощения проекта нами также был реализован контрольный пример на маленькой выборке, необходимый для выявления недочетов на начальном этапе разработки. Для этого были взяты следующие данные: вводимая строка – «Гвозди, в коробке.», строка из базы – «Белые болты в коробке». Сначала происходила предобработка введенной строки — так, как это было описано выше (см. рис. 4).

```
Гвозди, в коробке.
Переведенное слово:
 ['Гвозди,', 'в', 'коробке.']
Строка без междометий:
 Гвозди, коробке.
Строка без спецсимволов:
 Гвозди коробке
Строка в именительном падеже:
 гвоздь коробка
Строка в нижнем регистре:
 гвоздь коробка
Итоговая строка:
 гвоздь коробка
```

Рисунок 4. Предобработка введенной строки

Далее производилось построение гистограмм по данным, после чего все полученные результаты были выведены в виде отчета в соответствии с рисунком 5. Слева в отчете представлена гистограмма введенной строки (на оси абсцисс откладываются уникальные буквы в словосочетании, а на оси ординат – их количество в словах). Справа выводится гистограмма строки из базы, а также рассчитанная косинусная дистанция.

По гистограммам данного отчета и по расчету дистанции можно наглядно наблюдать схожесть фраз. Здесь мы видим, что строки «Гвозди, в коробке.» и «Белые болты в коробке» по графику схожи, но по рассчитанной косинусной

дистанции имеют значительное в данном случае расстояние – 0,7 (напомним, что чем меньше это число, тем лучше).

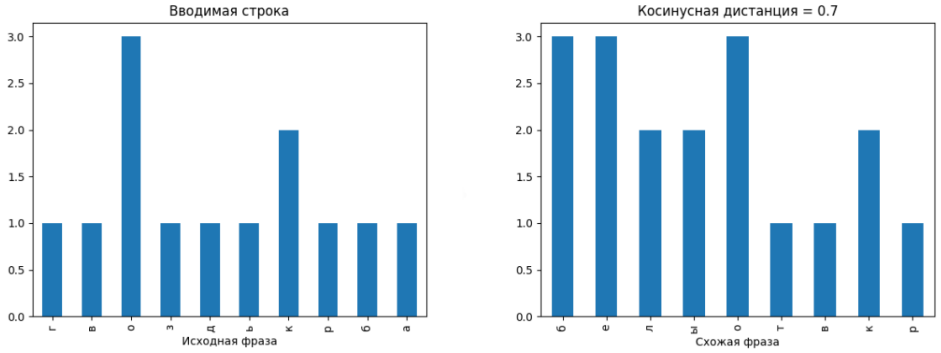

Рисунок 5. Внутренний отчет

Таким образом, с помощью рассмотренного метода можно решить проблему одинаковых слов, занесенных в справочники по-разному; например, одно может быть добавлено полностью, другое с сокращениями, а третье на английском языке и т.д., хотя при этом все они подразумевают один и тот же объект. Подобный способ облегчит работу с программой и поможет исправить ошибки в базе. В перспективе планируется адаптация данной разработки под большую выборку, тестирование на реальных данных, а также внедрение метода в различные организации для производственной эксплуатации.

Список литературы:

1. Архитектура платформы 1С:Предприятие. Справочники [Электронный ресурс]. – Режим доступа https://v8.1c.ru/platforma/spravochniki/ (дата обращения: 31.01.2024)
2. 1C-RarusTechDay 2023 — 6-я открытая техническая конференция для специалистов 1С // 1С-Рарус URL: (https://rarus.ru/events/20230720-1c-rarustechday-2023-587081/, б.д.) (дата обращения: 30.07.2023).
3. Как подключиться к 1С из Python и реализовать обмен данными // Статьи URL: https://dzen.ru/a/XtPGLCpP_TCSfohn (дата обращения: 30.08.2023).
4. Г.И.Фалин, А.И.Фалин. О гистограмме и её свойствах. Математика в профильной школе. ФРАКТАЛ, 2013, №1, стр.44-62
5. Кристиан С. Пероне, «Машинное обучение:: косинусное сходство векторных пространственных моделей (часть III)» в Terra Incognita.