Гапотченко Алёна Юрьевна, студентка,
Кубанский государственный университет,
г. Краснодар

ИСПРАВЛЕНИЕ ОШИБОК В СЛОВАХ РУССКОГО ЯЗЫКА НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ МЕДОИДОВ

Аннотация: В данной статье представлен метод неконтролируемой проверки орфографии, сочетающий в себе инициализацию аномального шаблона и разделение вокруг медоидов (PAM). Описанный подход направлен на сокращение количество раз, когда приходится вычислять расстояния при поиске целевых слов для ошибок в написании.

Ключевые слова: проверка орфографии, целевое слово, инициализация, медоиды, сумма расстояний, кластер.

Методы проверки орфографии были существенно изучены на протяжении многих лет. Миттон [1] указывает, что первая попытка решить проблему восходит к работе Блэра [2] а позже, больше внимания было уделено работе Дамерау [3]. Большинство методов проверки орфографии, описанных в литературе, включая этот, используют словари, как список правильных вариантов написания, которые помогают алгоритмам находить целевые слова. Лишь несколько методов пытаются решить эту проблему без использования словарей [4], например PAM.

Медоиды – это репрезентативные объекты набора данных или кластера внутри набора данных, сумма различий которых со всеми объектами в кластере минимальна. Алгоритм разделения вокруг медоидов (PAM) (Kaufman and Rousseeuw, 1990) делит набор данных Y на K кластеров S = {S1, S2,..., Sk}. Каждый кластер Sk представлен медоидом mk. Последним является объект yi ∈Sk с наименьшим расстоянием до всех остальных объектов, отнесенных к тому же кластеру. PAM создает компактные кластеры путем итеративной минимизации приведенного ниже в формуле 1 критерия.

$$W(S,M)=\sum_{k=1}^{k}\sum_{i\in S_k}\sum_{v\in V}(y_{iv}-m_{kv})^2, \quad (1)$$

где

V представляет особенности набора данных,

M – возвращенный набор медоидов m1, m2,..., mk.

Этот критерий представляет собой сумму расстояний между каждым медоидом mk и каждым объектом yi∈Sk. Минимизация (1) следует алгоритму, приведенному ниже:

1. случайным образом выбрать k медоидов из y, m = {m1, m2,..., mk}, s ← ∅;

2. обновляется s, назначив каждый объект yi∈y кластеру sK, представленному ближайшим к yi медоидом. если это обновление не генерирует никаких изменений в s, остановитесь, выведите s и m;

3. обновляется каждый медоид mk до объекта yi∈sk, который имеет наименьшую сумму расстояний до всех других объектов в том же кластере. возвращение к шагу 2.

PAM – очень популярный алгоритм кластеризации, который использовался в различных сценариях. Однако у него есть известные недостатки, например:

– его окончательная кластеризация сильно зависит от исходных используемых медоидов, а они обычно обнаруживаются случайным образом;

– пользователю необходимо знать, сколько кластеров имеется в наборе данных;

– из-за своей итеративной природы он может попасть в ловушку локальных оптимумов;

– он не учитывает различные особенности, которые могут иметь разную степень значимости.

Здесь мы обращаемся к внутренне связанным слабостям (1). Невозможно определить хорошие начальные медоиды для PAM, не зная, сколько из них следует использовать. Вышеизложенное привело к значительному количеству исследований, посвященных количеству и исходному положению медоидов. Такие усилия породили ряд алгоритмов, решающих одну или обе стороны проблемы, такие как Build (Kaufman and Rousseeuw, 1990), инициализация аномального шаблона (Mirkin, 2005), индекс Хартигана (Hartigan and Wong, 1979) и другие инициализации, основанные по иерархической кластеризации (Миллиган и Исаак, 1980).

Было проведено множество сравнений различных инициализаций в разных сценариях (Чианг и Миркин, 2010; Эмре Селеби и др., 2013; де Аморим, 2012; де Аморим и Комисарчук, 2012), что привело нас к выводу, что трудно назначить единственную инициализацию, которая всегда будет работать. Тем не менее, мы положительно относимся к инициализации аномального шаблона, предложенной Миркиным (2005). Его инициализация затрагивает обе стороны проблемы и исследует наблюдаемые предыдущие успехи в ее использовании (Чанг и Миркин, 2010; де Аморим, 2012; де Аморим и Комисарчук, 2012).

Эта инициализация изначально была разработана для K-Means и получила название интеллектуального K-Means. Ниже будет представлена медоидная версия инициализации аномального паттерна, которую мы использовали в наших экспериментах.

1. устанавливается mc как объект с наименьшей суммой расстояний до всех других объектов в наборе данных y;

2. устанавливается mt для объекта, находящегося дальше всего от mc;

3. применяется pam к y, используя mc и mt в качестве начальных медоидов, mc должен оставаться неизменным во время кластеризации;

4. добавляется mt к m;

5. удаляем mt и его кластер из y. если еще есть объекты для кластеризации, перейдём к шагу 2;

6. применим pam к исходному набору данных y, инициализированному медоидами в m и k = |m|.

На основе вышеизложенного был разработан метод, используемый для поиска целевых слов с ошибками в написании. Этот метод открыт для использования практически любой меры расстояния, допустимой для строк. Наша главная цель с помощью этого метода – сократить количество вычислений расстояний. Для этого мы применяем инициализацию аномального шаблона и PAM, как показано ниже:

1. применяется инициализация аномального шаблона к словарю, найдя количество кластеров k и набор начальных медоидов m init;

2. используя медоиды в m init, применяется pam в словарь, чтобы найти k кластеров. он должен вывести окончательный набор медоидов m = {m1, m2,..., mk};

3. учитывая ошибку в написании w, вычисляется его расстояние до каждого медоида mk∈ m. сохраняется в m∗ медоиды, расстояние до w которых равно найденному минимуму плюс константа c;

4. вычисляет расстояние между w и каждым словом в кластерах, представленных медоидами в m∗, выведя слова, расстояние от которых до w минимально возможно;

5. если появятся еще орфографические ошибки, возвращается к шагу 3.

Очевидно, что большое значение c будет означать большее количество вычислений расстояний. В наших экспериментах с расстоянием Левенштейна (Левенштейн, 1966) мы использовали c = 1.

Основная цель метода – сократить количество вычислений расстояний. Если измерить расстояние между орфографической ошибкой и каждым словом в словаре, эта функция расстояния будет вызвана 57 046 раз, что соответствует размеру словаря. Применяя этот метод к каждой из 34 956 орфографических ошибок в корпусе, который был описали ранее, мера расстояния была рассчитана в среднем 3251,4 раза для каждой орфографической ошибки. Считаем, что это важный результат с вычислительной точки зрения, поскольку мы значительно сокращаем количество вычислений.

Что касается восстановления целевых слов, это во многом зависит от используемой меры расстояния. Мы экспериментировали с популярным расстоянием Левенштейна (Левенштейн, 1966). В 88,42% случаев наш метод возвращал кластер, содержащий целевое слово или слово с меньшим расстоянием до орфографической ошибки, в отличии от целевого. В данном случае мы приписываем некоторые из последних орфографических ошибок, ошибкам реальных слов, что позволяет избежать ряд проблем. Полученные в таблице 1 результаты весьма многообещающие.

Таблица 1

Результаты

Всего ошибок в написании	34 956 слов
Уровень успеха (%)	34 956 слов
Уровень успеха (номинальный)	30 908 слов
Базовый прирост (%)	+ 11,18
Общее количество кластеров	1570 кластеров
Средняя длина кластера	3,78 слов
Расчет среднего расстояния	3251,4

Поскольку метод превосходит базовый уровень на 11,18 процентных пунктов при использовании того же набора данных (это число учитывает, что нам пришлось сократить наши результаты чуть более чем на 3%).

Список литературы:

1. J. Mitton, Fifty years of spellchecking. Writing Systems Research // Bulletin of the IEEE Computer Society Technical Committee. – 2010. – № 2 (1).

2. K. Blair, A program for correcting spelling errors. Information and Control // Conference on empirical methods in natural language processing. – 1960. – № 3.

3. A. Damerau, A technique for computer detection and correction of spelling errors. Communications of the ACM // Paper presented at Naist seminar. – 1964. – № 7 – 171-176 pp.

4. A. Morris, Computer detection of typographical errors // IEEE Transactions on Professional Communication. – 1975. – № 18 – 54-64 pp.