УДК 004.424.62

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ НЕЧЕТКОГО СРАВНЕНИЯ СТРОК ДЛЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОНЛАЙН-ВИКТОРИН

Лямцев Г.К.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Марков А.Н. – магистр техн. наук, ст. преподаватель кафедры информатики

Аннотация. Сравнительный анализ алгоритмов нечеткого сравнения строк. Исследованы алгоритмы Левенштейна, Дамерау-Левенштейна и Джаро-Винклера. Экспериментально определена зависимость F1-меры от порога принятия решения для каждого алгоритма. Установлено, что алгоритм Дамерау-Левенштейна при пороге не больше 2 символа обеспечивает наилучшую F1-меру 92,9%. Для алгоритмов сходства оптимальным является порог 0,8 для Джаро-Винклера (F1-мера равна 88,8%).

Ключевые слова: нечеткое сравнение, метрики строкового расстояния, расстояние Левенштейна, онлайн-викторины, порог принятия решения, F1-мера.

Введение. Современные информационные системы онлайн-обучения активно используют интерактивные викторины для контроля знаний. Традиционные методы проверки текстовых ответов, основанные на точном совпадении строк, не учитывают естественные опечатки пользователей.

Ключевой задачей при применении алгоритмов нечеткого сравнения является выбор оптимального порога принятия решения. Слишком низкий порог для алгоритмов Левенштейна и Дамерау–Левенштейна приводит к отклонению правильных ответов с незначительными опечатками, а слишком высокий – к принятию неправильных ответов. Необходим баланс между точностью и полнотой распознавания [1].

Целью данной работы является проведение сравнительного анализа алгоритмов нечеткого сравнения, экспериментальное определение зависимости F1-меры от порога принятия решения и формулирование рекомендаций по выбору оптимального алгоритма и порога для систем онлайн-викторин.

Основная часть. Для сравнительного анализа было рассмотрено 3 алгоритма: расстояние Левенштейна, расстояние Дамерау-Левенштейна и коэффициент Джаро-Винклера.

Расстояние Левенштейна определяется как минимальное количество операций редактирования, необходимых для преобразования одной строки в другую. Алгоритм использует три типа операций с единичной стоимостью: вставка символа, удаление символа и замена одного символа на другой. Для вычисления расстояния строится матрица размером $(m+1)\times(n+1)$, где m и n – длины сравниваемых строк. Каждая ячейка матрицы содержит минимальное расстояние между префиксами строк. Заполнение происходит динамически: для каждой позиции выбирается минимум из трех возможных операций. Вычислительная сложность алгоритма составляет $O(m\times n)$. Ответ считается правильным, если расстояние не превышает заданного порога θ [2].

Расстояние Дамерау-Левенштейна является расширением алгоритма Левенштейна с учетом транспозиции двух соседних символов [3].

Коэффициент Джаро-Винклера представляет собой метрику сходства, возвращающую значение от 0 (полное несовпадение) до 1 (точное совпадение). Алгоритм работает в два этапа. Сначала вычисляется базовая метрика Джаро, которая учитывает количество совпадающих символов и их транспозиции. Символы считаются совпадающими, если они одинаковы и находятся не дальше определенного расстояния друг от друга. Затем применяется модификация Винклера, добавляющая бонус за совпадение начальных символов строк [4].

Для оценки зависимости эффективности алгоритмов от порога принятия решения был сформирован тестовый набор из 500 пар ответа пользователя и правильного ответа, где каждый ответ пользователя помечался либо правильным, либо неправильным. Были

включены следующие типы ошибок: опечатки (1-2 символа), транспозиции соседних символов, лишние пробелы и знаки препинания.

В качестве метрик качества использовались:

– Точность – доля правильно принятых ответов среди всех принятых.

– Полнота – доля обнаруженных правильных ответов среди всех правильных.

– F1-мера – гармоническое среднее точности и полноты.

Для каждого алгоритма порог варьировался в следующих диапазонах: для Левенштейна и Дамерау-Левенштейна от 0 до 5 символов, для Джаро-Винклера от 0,60 до 0,90.

Результаты экспериментального исследования зависимости проверяемых метрик от порога для алгоритма Левенштейна представлены в таблице 1:

Таблица 1 – Результаты анализа алгоритма Левенштейна

Порог (символов)	**Точность (%)**	**Полнота (%)**	**F1-мера (%)**
0	100,0	45,0	62,1
1	95,2	72,5	82,3
2	87,3	91,5	89,4
3	78,5	96,8	86,8
4	69,2	98,5	81,3
5	61,5	99,2	75,9

Анализ данных показывает, что оптимум достигается при пороге 2 символа с F1-мерой 89,4%.

Результаты экспериментального исследования зависимости проверяемых метрик от порога для алгоритма Дамерау-Левенштейна представлены в таблице 2:

Таблица 2 – Результаты анализа алгоритма Дамерау-Левенштейна

Порог (символов)	**Точность (%)**	**Полнота (%)**	**F1-мера (%)**
0	100,0	48,5	65,3
1	96,8	78,2	86,5
2	92,1	93,8	92,9
3	84,7	97,5	90,7
4	76,3	98,9	86,1
5	68,5	99,4	81,1

Анализ данных показывает, что оптимум достигается при пороге равном 2 символам с F1-мерой 92,9%., также можем заметить, что алгоритм Дамерау-Левенштейна превосходит базовый алгоритм Левенштейна на всех порогах.

Результаты экспериментального исследования зависимости проверяемых метрик от порога для алгоритма Джаро-Винклера представлены в таблице 3:

Таблица 3 – Результаты анализа алгоритма Джаро-Винклера

Порог сходства	**Точность (%)**	**Полнота (%)**	**F1-мера (%)**
0,60	72,5	96,8	82,9
0,70	80,2	93,2	86,2
0,75	85,8	90,5	88,1
0,80	89,5	88,2	88,8
0,85	93,2	83,5	88,1
0,90	96,5	76,2	85,2

Для алгоритмов сходства зависимость обратная: низкий порог (0,5) дает высокую полноту, но низкую точность. С ростом порога точность увеличивается, а полнота падает. Оптимум для Джаро-Винклера находится в диапазоне 0,75-0,85 с максимальной F1-мерой 88,8% при пороге 0,8.

Анализ результатов показывает, что алгоритм Дамерау-Левенштейна превосходит остальные по F1-мере на всех порогах за счет учета транспозиций соседних символов.

Заключение. Проведен сравнительный анализ трех алгоритмов нечеткого сравнения с экспериментальным определением оптимальных порогов принятия решения. Установлено, что алгоритм Дамерау-Левенштейна показывает наилучшую F1-меру 92,9% при пороге не больше двух символов. Алгоритм Левенштейна достигает F1-меру равную 89,4% при том же пороге, уступая на 3,5% из-за отсутствия учета транспозиций. Джаро-Винклер оптимален при пороге не меньше 0,8 с F1-мерой равной 88,8%.

Для системы онлайн-викторин рекомендуется использовать алгоритм Дамерау-Левенштейна с порогом 2 символа, что обеспечивает оптимальный баланс между точностью (92,1%) и полнотой (93,8%). При необходимости приоритизации точности над полнотой следует увеличить порог до 1 символа (точность 96,8%, полнота 78,2%), при приоритете полноты – снизить до 3 символов (точность 84,7%, полнота 97,5%).

Направлениями дальнейших исследований являются: применение методов машинного обучения для автоматической настройки порогов, интеграция семантического анализа для проверки смысловой близости.

Список литературы

1. *A guided tour to approximate string matching / G. Navarro // ACM Computing Surveys. – 2001. – Vol. 33, N 1. – Pp. 31–88.*

2. *Binary codes capable of correcting deletions, insertions, and reversals / V.I. Levenshtein // Soviet Physics Doklady. – 1966. – Vol. 10, N 8. – Pp. 707–710.*

3. *A technique for computer detection and correction of spelling errors / F.J. Damerau // Communications of the ACM. – 1964. – Vol. 7, N 3. – Pp. 171–176.*

4. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage / W.E. Winkler // Proceedings of the Section on Survey Research Methods. – 1990. – Pp. 354–359.*

UDC 004.424.62

COMPARATIVE ANALYSIS OF FUZZY STRING COMPARISON ALGORITHMS FOR AN ONLINE QUIZ INFORMATION SYSTEM

Lyamtsev G.K.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Markov A.N. – Master of Sci. (M. of Sci.), Senior Lecturer at the Department of Informatics

Annotation. A comparative analysis of fuzzy comparison algorithms. The Levenshtein, Damerau-Levenshtein, and Jaro-Winkler algorithms were examined. The dependence of the F1-score on the decision threshold for each algorithm was experimentally determined. It was found that the Damerau-Levenshtein algorithm, with a threshold of no more than two characters, provides the best F1-score of 92.9%. For similarity algorithms, a threshold of 0.8 for the Jaro-Winkler algorithm is optimal (the F1-score is 88.8%).

Keywords: fuzzy comparison, string distance metrics, Levenshtein distance, online quizzes, decision threshold, F1-score.