УДК 004.912

Анализ влияния размера n-грамм на качество выявления текстовых заимствований

А.А. Назаров, А.П. Преображенский

Воронежский институт высоких технологий, Воронеж, Россия

В работе рассматривается влияние длины n-грамм на качество выявления текстовых заимствований. Исследование основано на n-граммном подходе с использованием коэффициента Жаккара для оценки сходства текстов. Эксперименты проводились на корпусе научных и учебных текстов с различными типами заимствований, включая прямые, частичные и перефразированные случаи. Показано, что при малых значениях n повышается полнота обнаружения, но ухудшается точность из-за случайных совпадений. При увеличении n система начинает пропускать часть реальных заимствований. Отдельно отмечено, что наиболее сбалансированные результаты достигаются в диапазоне n = 3–4. В этом случае наблюдается оптимальное соотношение между точностью и полнотой. Полученные результаты могут быть использованы при настройке простых систем поиска заимствований и учебных антиплагиатных модулей.

Ключевые слова: n-граммы, текстовые заимствования, коэффициент Жаккара, токенизация, нормализация текста, метрики качества, F1-мера.

Analysis of the Impact of n-Gram Size on the Quality of Text Plagiarism Detection

A.A. Nazarov, A.P. Preobrazhenskiy

Voronezh Institute of High Technologies, Voronezh, Russia

This paper examines the impact of n-gram length on the quality of text plagiarism detection. The study employs an n-gram-based approach using the Jaccard coefficient to assess text similarity. Experiments were conducted on a corpus of scientific and educational texts containing various types of borrowed content, including verbatim, partial, and paraphrased instances. The results demonstrate that while low n-values increase detection recall, they reduce precision due to accidental matches. Conversely, increasing n causes the system to miss some actual instances of plagiarism. It is noted separately that the most balanced results are achieved in the range n = 3–4. In this case, an optimal balance between accuracy and completeness is observed. These findings can inform the configuration of simple plagiarism detection systems and educational anti-plagiarism modules.

Keywords: n-grams, textual borrowings, Jaccard coefficient, tokenization, text normalization, quality metrics, F1-score.

Одной из приоритетных задач современной научно-образовательной сферы является выявление текстовых заимствований в научных работах. Для решения этой задачи существуют различные методы. При применении простых способов компьютер сравнивает слова или последовательности символов, которые выглядят одинаково. Сложные методы применяют семантические представления, векторные модели и нейросетевые подходы, чтобы анализировать значение информации, а не её внешний вид.

Классическими способами обнаружения плагиата являются методы, в основе которых лежат n-граммы и шинглы. Процесс такого исследования устроен так: программа делит текст на последовательности, которые состоят из n элементов, и затем сравнивает эти наборы [1]. При своей простоте такие методы выдают результаты без резких отклонений, и разработчики часто выбирают их как базовый уровень для сложных систем. Но эти подходы имеют пределы возможностей. К примеру, когда значение n является небольшим, система реагирует на минимальные изменения и фиксирует совпадения в тех местах, где тексты не являются идентичными по содержанию. При больших значениях наоборот – теряется часть совпадений, особенно если текст был немного перефразирован. В итоге качество сильно зависит от выбора параметра n, и это не всегда очевидно заранее.

В связи с этим возникает практический вопрос: какое значение длины n-грамм является наиболее удачным для выявления заимствований в текстах разного типа. В данной работе рассматривается влияние этого параметра на качество обнаружения совпадений. Основная цель – попытаться опытным путём оценить, как меняются метрики качества при разных значениях n и есть ли условно «оптимальный» диапазон. В данной работе рассматривается задача выявления текстовых заимствований на основе n-граммного подхода. В общем виде её можно сформулировать так: даны два текста, необходимо определить степень их сходства и установить, содержит ли один текст фрагменты другого. Задача кажется простой, но на практике всё упирается в детали представления текста и способа сравнения.

Пусть есть два документа D_1 и D_2. Каждый документ разбивается на последовательность n-грамм. Под n-граммой понимается последовательность из n элементов, где элементом может быть слово или символ – в данной работе рассматривается словесный вариант. После разбиения каждый документ представляется как множество или мультимножество таких последовательностей.

Далее вводится мера сходства. Чаще всего используется коэффициент Жаккара, так как он достаточно простой и интерпретируемый. Он определяется как отношение пересечения множеств n-грамм к их объединению:

$$J(D_1, D_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}, \quad (1)$$

где G_1 и G_2 – множества n-грамм для первого и второго документа.

Задачей исследования является подбор оптимального значения параметра n. В ходе эксперимента оно изменяется в диапазоне от двух до шести. При малом значении, когда n = 2, фиксируется большое множество совпадений. Но при этом высока вероятность того, что будут учитываться случайные совпадения. Если n большое, то множество совпадений резко уменьшается, так как даже незначительная переформулировка фразы ведёт к тому, что такие n-граммы считаются системой различными.

Качество проведённого эксперимента оценивалось с помощью стандартных метрик, таких как точность (Precision), полнота (Recall) и F1-мера. Точность вычисляется как отношение правильно найденных заимствований ко всем обнаруженным. Полнота – это доля всех найденных заимствований среди существующих. F1-мера – гармоническое среднее между ними [1].

Методология исследования включает несколько этапов. Сначала выполняется подготовка текстов: приведение к нижнему регистру, удаление пунктуации и токенизация [2]. Иногда при анализе текстов могут дополнительно удаляться стоп-слова. Но в данном исследовании удаление не происходит, так как это действие

может изменить структуру n-грамм. Далее каждый текст преобразовывается в набор n-грамм для разных значений n. После этого вычисляется матрица, которая показывает сходство текстов попарно. И на основе этой матрицы определяется, есть ли заимствования, если уровень сходства выше фиксированного числа. Это число выбирается на основе практического опыта, чтобы результаты были сопоставимы между разными значениями n. Таким образом, методология заключается в том, чтобы сравнить, как параметр n влияет на то, насколько качественно выявляются заимствования, когда способ измерения сходства не меняется. С помощью этого подхода можно заметить закономерности и оценить, при каких значениях n показатели точности и полноты находятся в наиболее выгодном соотношении. Это позволяет выявить устойчивые закономерности и оценить, при каких значениях n достигается оптимальный баланс между точностью и полнотой.

Для проведения экспериментов использовались тексты, которые относятся к разным типам. Проверялось работа n-граммного подхода. В основу были положены статьи из научной сферы и тексты для обучения. К ним также были добавлены работы, написанные студентами, так как именно такие документы проверяются на наличие заимствований наиболее часто. В общем объеме выборки находилось от 60 до 80 документов.

В набор исследуемых документов были включены разные виды заимствований. Это прямые заимствования, когда текст был скопирован почти полностью без поправок. Иногда в нем менялись лишь некоторые слова или знаки препинания, но структура оставалась прежней. Другой вид представлял собой частичные заимствования при котором копировалась только некоторая часть текста, например отдельные абзацы или фразы, которые помещались в новый контекст. С ними работать было труднее, так как границы совпадения текстов становились менее четкими. И третий случай, который является наиболее трудным для анализа, – это заимствования, которые были перефразированы. Текст оставался таким же по смыслу, но его формулировки менялись. Предложения строились иначе, а слова заменялись на синонимы. Для эксперимента множество работ было собрано таким образом, чтобы в них были и случаи с полным совпадением, и варианты, которые требуют сложного анализа. Это позволяет оценить, как изменение длины n-грамм влияет на чувствительность метода в разных ситуациях, а не только в идеальных условиях полного копирования.

Перед тем как применять n-граммный анализ, все тексты были приведены к единому виду. Слова приводились к нижнему регистру. В некоторых вариантах эксперимента знаки пунктуации убирались полностью, в других – оставлялись, чтобы посмотреть, как это влияет на результат. В целом можно сказать, что для словесных n-грамм пунктуация чаще всего играет второстепенную роль, но полностью игнорировать её тоже не всегда правильно. Для того, чтобы не изменялась структура n-грамм, стоп-слова не удалялись. К ним относятся часто встречающиеся слова, такие как «и», «в», «на», «что» и так далее. С одной стороны, их удаление может уменьшить шум и сделать сравнение более «осмысленным». Но, с другой стороны, эти слова часто участвуют в формировании структуры текста, особенно в коротких n-граммах. Поэтому в данной работе стоп-слова не удалялись по умолчанию.

Реализация алгоритма выявления текстовых заимствований в рамках данной работы строится на последовательной обработке текстов и сравнении их n-граммных представлений [3]. Ниже приведено упрощённое описание алгоритма:

Input: texts D1, D2, parameter n, threshold T

```
1. preprocess(D1) -> tokens1
2. preprocess(D2) -> tokens2
3. G1 = build_ngrams(tokens1, n)
4. G2 = build_ngrams(tokens2, n)
5. intersection = G1 ∩ G2
6. union = G1 ∪ G2
7. J = |intersection| / |union|
8. if J >= T:
      return "plagiarism detected"
   else:
      return "no plagiarism"
```

В качестве входных данных алгоритм использует два текста. Сначала эти тексты проходят этап предобработки, который включает в себя токенизацию, нормализацию и приведение к единому виду. После этого каждый документ преобразуется в набор n-грамм, которые имеют заданную длину n. В результате текст становится совокупностью последовательностей, и дальнейшие действия совершаются с этими наборами. Основная часть алгоритма заключается в том, что полученные множества n-грамм сравниваются между собой. Для измерения сходства используется коэффициент Жаккара, который вычисляется по формуле (1). На этом этапе устанавливается количество одинаковых элементов в двух текстах. Если количество общих n-грамм увеличивается, то значение коэффициента становится выше.

Когда сходство вычислено, применяется пороговое правило. Чтобы результаты можно было сравнивать между собой, порог подбирался заранее и был одинаковым во всех экспериментах. Путем проведения опытов величина порога была установлена на уровне 0,3. С помощью этого значения обеспечивается равное соотношение между случаями, когда заимствования являются истинными, и случаями, когда срабатывания являются ложными. Для серии экспериментов порог в данной работе является постоянным, чтобы была возможность корректно сравнивать разные значения n. В ином случае результаты не были бы сопоставимыми. Если значение коэффициента больше, чем заданный порог 0,3, то считается, что пара текстов содержит заимствование. При значении ниже 0,3 считается, что тексты не имеют значительных совпадений.

Была проведена серия экспериментов. Брались различные значения n, исследуемый текст разбивался на n-граммы длиной n и выполнялся полный цикл обработки данных. Каждый раз рассчитывалось сходство документов.

При n = 2 фиксировалось довольно много совпадений. Но эти совпадения были в том числе случайные, так как в текстах присутствовали часто встречающиеся фразы. То есть количество найденных совпадений превышало то количество, которое было на самом деле. Заметное изменение работы системы было зафиксировано при увеличении n до значений 4–5. Количество ложных срабатываний уменьшилось. Сами совпадения оставались достаточно верными. При значении n = 6 и больше качество обнаружения стало падать. При небольшом изменении формулировки или перестановке слов, такие фрагменты уже совпадающими не считались.

В результате выяснилось, что от значения n довольно сильно зависит работа алгоритма. Были вычислены и занесены в таблицу основные метрики качества текста.

Таблица

Зависимость метрик качества от n

n	Precision	Recall	F1-мера
2	0,71	0,88	0,79
3	0,78	0,83	0,80
4	0,84	0,76	0,80
5	0,88	0,68	0,77
6	0,91	0,55	0,68

Из таблицы видно, что при увеличении значения n полнота уменьшается, а точность растёт, так как длинные n-граммы совпадают редко. Даже если текст является плагиатом. При n = 2 система находит много совпадений, включая случайные. Поэтому Recall здесь высокий, но Precision заметно ниже. Можно сказать, что модель в этом случае склонна переоценивать наличие заимствований. При n = 3–4 наблюдается более сбалансированное поведение. В этом диапазоне значения Precision и Recall становятся ближе друг к другу, а F1-мера достигает максимума. Это можно считать наиболее устойчивой зоной работы алгоритма. Дальше, начиная с n = 5, система становится более строгой. Precision продолжает расти, но Recall падает достаточно резко. Особенно это заметно на перефразированных заимствованиях, где структура текста уже сильно отличается, и длинные n-граммы просто не совпадают. При n = 6 ситуация система почти не допускает ложных совпадений, но при этом начинает терять значительную часть реальных заимствований. В итоге F1-мера снижается, что говорит о потере баланса между точностью и полнотой. Можно сделать предварительный вывод, что оптимальный диапазон для данной задачи находится примерно в области n = 3–4.

Полученные результаты в целом подтверждают исходную гипотезу о том, что длина n-грамм существенно влияет на баланс между точностью и полнотой обнаружения заимствований. Малые n-граммы фиксируют локальные совпадения, которые часто встречаются даже в несвязанных текстах [4]. Поэтому система давала избыточное количество срабатываний. Это плохо с точки зрения практической точности. С другой стороны, при увеличении n система начинает учитывать более длинные последовательности слов, которые реже повторяются случайно. Это снижает количество ложных срабатываний, но одновременно приводит к тому, что часть реальных заимствований просто не обнаруживается. Особенно это заметно в случаях перефразирования, где сохраняется смысл, но меняется формулировка.

Отдельно можно отметить, что наиболее устойчивые результаты были получены в диапазоне n = 3–4. Именно здесь наблюдается наиболее сбалансированное поведение метрик. Это, конечно, не универсальное значение, но для данного корпуса текстов оно оказалось наиболее удачным компромиссом. Сам выбор оптимального n зависит от множества факторов. В частности, от языка текстов, их длины, степени формальности и даже предметной области. Например, в более формализованных научных текстах большие n могут работать лучше, чем в студенческих работах, где стиль более свободный.

Также стоит учитывать ограничения проведённого исследования. Рассматриваемый объём документов является небольшим. Если данный алгоритм применять к большому числу документов, то значения метрик может немного отличаться. Кроме того, различные методы сравнения могут давать несколько изменённые результаты. В данном эксперименте применялся коэффициент Жаккара.

В результате проведённых экспериментов выяснилось, что длина n-грамм влияет на качество обнаружения плагиата. Наиболее точные данные были получены при значениях n от трёх до четырёх. При малом n, равном двум, обнаруживалось немалое число ложных совпадений. При слишком большом n, от шести и выше, система хуже находила перефразированные заимствования.

Для более полного исследования текстов на наличие заимствований необходимо сочетать рассмотренный метод с другими способами семантического анализа.

СПИСОК ИСТОЧНИКОВ

1. Янников И.М. Методы и алгоритмы для поиска сходства между текстами / И.М. Янников, М.В. Ершова, А.Н. Исенбаев // Интеллектуальные системы в производстве. – 2024. – Т. 22, № 2. – С. 103–113.

2. Смирнов И.В. Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка / И.В. Смирнов. – М.: ФИЦ ИУ РАН, 2023. – 356 с.

3. Нугуманова А.Б. Автоматизированная обработка текстовых массивов: учебник и практикум для вузов / А.Б. Нугуманова. – 2-е изд. – М.: Издательство Юрайт, 2024. – 82 с.

4. Гасанов Э.Э. Интеллектуальные системы. Теория хранения и поиска информации: учебник для вузов / Э.Э. Гасанов, В.Б. Кудрявцев. – 2-е изд., испр. и доп. – М.: Издательство Юрайт, 2025. – 271 с.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Назаров Антон Андреевич, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Преображенский Андрей Петрович, доктор технических наук, профессор, Воронежский институт высоких технологий, Воронеж, Россия.