ISSN 2226-3780

УДК 519.2: 81

Квашина Ю. А.

МЕТОДЫ ПОИСКА ДУБЛИКАТОВ СКОМПОНОВАННЫХ ТЕКСТОВ НАУЧНОЙ СТИЛИСТИКИ

Данная статья посвящена такой теме, как поиск плагиата. В статье представлена модификация алгоритма шинглов для поиска нечетких дубликатов для скомпонованных документов. Выигрыш в производительности планируется достигнуть за счет уменьшения количества сравнений пар шинглов за счет разбиения скомпонованных тестов на разделы.

Ключевые слова: *хеш-функция, шингл, плагиат, скомпонованный документ, дубликат, блок.*

1. Введение

В современном мире легкий доступ к практически любому виду информации привел к тому, что для того, чтобы сделать что-то новое, не обязательно это придумывать. Разумеется, с появлением интернета стало намного проще находить информацию, время написания любых изданий сократилось в разы, но вместе с этим возникла проблема авторского права. *Авторское право* — институт гражданского права, регулирующий отношения, связанные с созданием и использованием (изданием, исполнением, показом и т. д.) произведений науки, литературы или искусства, то есть объективных результатов творческой деятельности людей в этих областях [1]. Огромные, в большинстве случаев, бесплатные источники информации предоставляют много данных, которые пользователи затем выдают за свои собственные. Особенно это касается студентов и школьников, многие используют чужие дипломы и курсовые для написания своих, преподавателям сложно выявить, оригинальная работа или нет, на что и рассчитывают студенты.

По статистике [2] один из трех студентов использует неправомерно Интернет — ресурсы при выполнении заданий. В ходе опроса 24000 студентов в 70 вузах, Дональд Маккейб выяснил следующее: 58 процентов признались в плагиате; 95 процентов сказали, что они участвовали в той или иной форме обмана, при копировании домашнего задания или написании теста; 36 % студентов и 25 % выпускников признались, что перефразировали или копировали несколько предложений из интернета без сноски и указания автора. Практически столько же 38 % студентов и 24 % выпускников использовали таким же образом информацию из письменных источников. Статистика показывает неутешительные результаты, данные показатели возможно уменьшить только с разработкой программных продуктов или улучшением настоящих, которые способные распознавать плагиатные документы.

2. Анализ средств выявления дубликатов

Существует несколько решений проблемы плагиата, на данный момент, например, такие программные продукты, как «Программа Плагиата.НЕТ» и «Программа Etxt Антиплагиат». Они распространяются бесплатно, поэтому общедоступны и популярны на данный момент. Оба программных продукта используют алгоритм шинглов и его различные вариации. *Шингл* — это небольшой, состоящий из нескольких слов, фрагмент текста, обработанный по специальной методике для анализа [3]. *Алгоритм шинглов* — это алгоритм, разработанный для поиска копий и дубликатов рассматриваемого текста [4].

Существуют также альтернативные методы, реализованные в системах основанных на анализе семантики [5], но из-за большой вычислительной сложности не получившие на данный момент распространения.

Оба приложения, основанные на алгоритме шинглов, являются десктопными, с довольно понятным и дружественным интерфейсом. Из достоинств можно отметить настраиваемые параметры поиска плагиата, возможность поиска не только в интернете, но и в локальных базах данных, красочное оформление результатов. Главным недостатком является скорость, даже при небольших объемах информации проверка происходит очень медленно. Было проведено тестирование данных двух программных продуктов. Для анализа взяли текст, состоящий из 100, 1000 и так далее слов, и засекли время (в минутах) проверки на плагиат в каждой из программ, результаты представлены в таблице (табл. 1).

Также не учитывается структура и тип документа. Это очень важно, так как от этих факторов зависит итоговый результат проверки.

Таблица 1

Сравнительная характеристика скорости проверки документа на плагиат

Количество слов/время (мин)	100	1000	10000	100000
Плагиата.Нет	0,18	3,2	14	46
Etxt Антиплагиат	0,25	3,05	11	35

3. Постановка задачи исследования

Основной характеристикой структуры документа является длина. В зависимости от длины документа различают ультракороткий, короткий текст, обычный и скомпонованный документ. Особое внимание хочется обратить на скомпонованный — документ, состоящий из схожих частей разных документов, он наиболее распространенный. Также существуют, всем известные стили текстов, такие как научный, публицистический, художественный, официально-деловой и разговорный.

Длина и стиль текста не учитывается в программных продуктах, рассмотренных ранее.

Одной из нерешенных задач является анализ документов с учетом вышеперечисленных параметров. Целью статьи является разработка методик определения плагиата с учетом таких параметров, как тип и структура документа.

Рассмотрим такой стиль документов как научный, так как анализ научных документов наиболее востребован. Адекватно оценить и проверить дипломы, статьи, рефераты, курсовые не возможно без анализа на плагиат. Такие научные документы имеют много особенностей, которые нужно учитывать.

В основном научные тексты относятся к типу скомпонованного документа. Так как в каждом таком документе есть определенные разделы с одинаковыми названиями. Первая особенность — изначально определена структура документа. Вторая особенность — в научных текстах часто используется слова и символы из других языков. Расшифровка каких-то терминов, или ссылки на иные ресурсы часто встречаются на латинском или английском языках. При проверке стоит на это обратить внимание, и при обработке удалять подобные вставки. Последнюю особенность, которую стоит выделить, — это шаблонные фразы, в научных документах их довольно много. Такие как: «таким образом», «сделаем выводы», «в заключении», «в целом» «из этого следует» и т. д. Если алгоритм нахождения плагиата будет их учитывать, то результаты будут не совсем корректные. Так как использование всем известным оборотов перечисленных выше, не является плагиатом.

Одной из методик распознавания плагиата является ранее уже упомянутый алгоритм шинглов. Данный алгоритм реализует поиск нечетких дубликатов. *Дубликат* — это копия, второй или следующий экземпляр документа, а *нечеткий дубликат* — документ, в котором основной смысл и стилистика сохранена, а вот полностью он не повторяется [6]. Нечеткие дубликаты позволяют предположить, являются ли два объекта частично одинаковыми или нет [7].

Под объектом будут пониматься текстовые файлы. Задача алгоритма не определить абсолютное значение схожести объектов, а выделить в каждом из объектов схожие части. Он только отвечает на вопрос, являются ли объекты почти дубликатами или нет. Почти дубликат отличается от точных дубликатов, наличием незначительных отличий. Если учесть все особенности научного текста и усовершенствовать алгоритм шинглов, то с использованием него можно определять плагиат научных работ с большей вероятности, чем было ранее.

Целью исследования, результаты которого изложены в статье, является разработка методов поиска нечетких дубликатов скомпонованных документов научной стилистики на основе анализа их разделов.

4. Описание алгоритма поиска дубликатов скомпонованных документов

Разрабатываемый алгоритм условно можно разбить на несколько этапов, представленных на рис. 1. Для выполнения задачи «Вычисление степени плагиатности блока» будет рассмотрено два варианта решения.

Рис. 1. Этапы алгоритма

4.1. Канонизация. Первым шагом в модифицированном алгоритме является канонизация текста, как и во всех алгоритмах шинглов. *Канонизация* — это метод отсечения не несущих смысловой нагрузки слов от всех остальных слов [8], процесс «очищения» текста. Удаляются знаки препинания и табуляции, отступы, лишние пробелы, слова приводятся к одному регистру, с учетом особенности научного текста — к одному языку. Предлоги, союзы и другие служебные части речи удаляются. Также необходимо удалить все цифры, а также специальные слова такие как «год», «месяц». Если не удалять, то плагиатом будут считаться даты, что не совсем корректно.

Следующий шаг — *лемматизация*. Это метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме [9]. Все существительные и прилагательные приводятся к мужскому роду единственного числа в именительном падеже, а глаголы — к инфинитиву. Затем *стемминг* — это поиск основы слова, учитывающий морфологию исходного слова. Стемминг выполняет морфологический разбор слова, находит общую для всех его грамматических форм основу, отсекая суффиксы и окончания [10].

4.2. Сравнение шинглов. Если все эти преобразования с текстом выполнены успешно, то можно приступать к следующему этапу алгоритма — разбиение текста на шинглы. Но, учитывая специфику документа, вводим дополнительный этап. Научные работы имеют схожую структуру, то есть разделение на разделы, чаще всего даже с одинаковыми названиями. Учитывая это, будем не весь текст разбивать на шинглы, и сравнивать с шинглами другого документа, а за основу возьмем сначала раздел, а затем в разделе выделяем непересекающиеся блоки. Длина блока может быть настраиваемым параметром или регулироваться в зависимости от длины абзацев в разделе. Тем самым происходит выделение единицы

документа — блок, с ней работать дальше удобнее, а самое главное быстрее. С этого этапа есть как минимум два пути для дальнейшего сравнивания.

Первое решение — для каждого блока вычислить шинглы и их хэш-функции. Длина шингла вычисляется в зависимости от размера блока (рис. 2). Длина шингла будет изменяемый параметр, выборка производится внахлест, а не встык, таким образом, всего количество шинглов будет равно количеству слов в блоке минус длина шингла плюс 1. После получения всех подпоследовательностей, надо определить их контрольные суммы (хэш, хэш-функция). *Контрольная сумма* — некоторое значение, рассчитанное по набору данных путем применения определенного алгоритма и используемое для проверки целостности данных при их передаче или хранении [11]. Контрольная сумма является уникальным числом, поставленным в соответствие некоторому тексту и/или функция его вычисления. При сравнении двух блоков, записываются в одну последовательность все хэши обоих блоков и сортируются по возрастанию. Затем последовательно сравниваются рядом стоящие хэши, и если количество совпадений превысило пороговое значение 75 %, то анализируемый блок считается плагиатным.

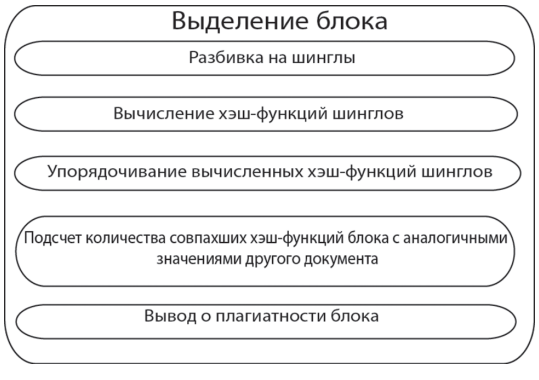

Рис. 2. Выделение блока

Каждый блок сравнивается с каждым блоком, и каждый раздел сравнивается с каждым разделом (рис. 3). Если совпало больше 75 % процентов последовательности, то такой блок считается плагиатным. Данная процедура проделывается со всеми блоками раздела, пока количество плагиатных блоков не превысит пороговое значение — 75 %, затем делается вывод о том, что этот раздел плагиатный, далее переход к следующему разделу и проделывается то же самое. В итоге получаются сведения о плагиатности любого раздела.

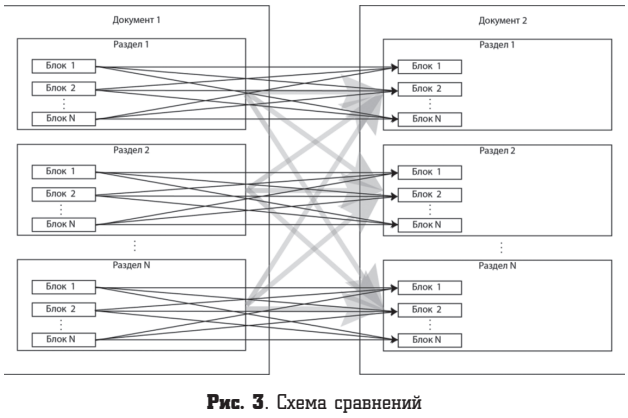

Рис. 3. Схема сравнений

4.3. Сравнение по весу блока. Второй путь состоит в том, чтобы определять вес блока, и брать хэш-функции от него. Вес блока определяется с помощью алгоритма RIDF (Residual IDF) [12]. Строится словарь, ставящий каждому слову в соответствие число разделов, в которых оно встречается хотя бы один раз (df) и определяется суммарная частота каждого слова в тексте (cf). Затем строится частотный словарь раздела и для каждого слова вычисляется его «вес» wt, по формуле:

$$wt = TF * RIDF,$$

где

$$TF = 0.5 + 0.5*tf/tf\_max,$$

$$RIDF = -\log(df/N) + \log(1 - \exp(-cf/N)).$$

Здесь N — число разделов в документе, tf — частота слова в разделе, tf_max — максимальная частота слова в разделе документа. Затем выбираются и сцепляются в алфавитном порядке в строку M (константа для алгоритма, например, 6) слов с наибольшими значениями wt, после вычисляется хэш-функция этой строки, это число и будет хэшем блока. Каждый блок будет ассоциироваться одним числом, которое будет сравниваться. Сравнивая хэш-функции блоков, если они совпали, делается вывод, что блок плагиатный, исходя из этого, делаем вывод и о плагиатности полностью раздела, затем и документа в целом (рис 4).

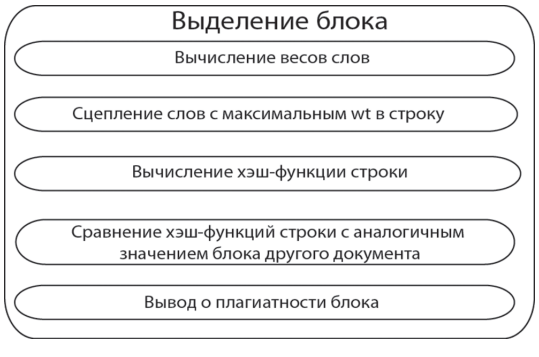

Рис. 4. Выделение блока

Одним из нерешенных вопросов осталось определение алгоритма нахождения хэш-функции. Существует множество видов хэш-функций, наиболее распространенные — md5, crc, sha1, sha2. Для точности результатов иногда используют сразу несколько методов хеширования.

Например, если использовать алгоритм crc, то основная идея его состоит в представлении сообщения в виде большого двоичного числа, делении его на другое фиксированное двоичное число и использовании остатка от этого деления в качестве контрольной суммы. MD5 работает по очень простому алгоритму. Сначала сообщение делится на блоки по 512 бит = 64 байта, затем от каждого блока вычисляется простая в реализации функция и наконец, все результаты функций собираются вместе в 128-битный хеш [13]. A sha1 для входного сообщения произвольной длины (максимум 2^64 — 1 бит) генерирует 160-битное хеш — значение (SHA-2 256/224 или 512/384), называемое также дайджестом сообщения. Для проведения эксперимента можно оставить выбор хеш-функции на усмотрение исследователя.

5. Анализ результатов

Проблема поиска плагиата заключается в количестве сравнений, ведь это напрямую отражается на производительности. Одним из главных недостатков уже реализованных программ поиска плагиата есть именно производительность. Используя и первое и второе решение, можно увеличить производительность в несколько раз по сравнению с уже разработанными методами, за счет того что работа производится не с целым документом, а с блоками.

Сложность простого алгоритма сравнения документов на основе шинглов

$$O(n) = n * m, \quad (1)$$

где n — количество шинглов в первом документе, а m — во втором.

В классическом алгоритме сложность зависела только от n количества шинглов в первом документе и m — во втором. Для расчета аналитической оценки нового алгоритма вводятся дополнительные параметры: b_1 — количество блоков в первом документе, а — b_2 во втором. Аналитическая сложность сортировки хэшей шинглов блока $O(X) = X*\lg X$, где X — количество шинглов блока. Так как потом сравниваются пары хэшей, то к аналитической сложности необходимо прибавить еще X — количество сравнений. Предполагается, что все блоки в документе имеют одинаковую длину X шинглов, равную количеству шинглов в документе, деленному на количество блоков. В итоге получаем формулу:

$$O\left(\sum_{b1*b2}\left(\frac{n}{b_1}+\frac{m}{b_2}+\left(\frac{n}{b_1}+\frac{m}{b_2}\right)*\log\left(\frac{n}{b_1}+\frac{m}{b_2}\right)\right)\right). \quad (2)$$

Сложность же разбиения документа на шинглы и вычисление хэшей шинглов в классическом и модифицированном алгоритмах одинакова.

Ускорение работы алгоритма планируется получить за счет следующих приемов: прекращать процедуру проверки блока на неуникальность при превышении показателя его уникальности 25 % или превышения показателя неуникальности порога в 75 %. В результате количество сравнений уменьшается, а производительность увеличивается.

Достоинством нового алгоритма также является возможность проверить на неоригинальность отдельные разделы документа, в то время как в классическом варианте неоригинальность одного небольшого (по сравнению с размерами всего документа) раздела не сильно влияла оценку всего документа в целом.

Недостатками являются трудоемкость в вычислениях, особенно во втором методе. Также алгоритм не учитывает такие тонкости, например, как цитирование. Выявить цитаты в тексте отдельная сложная задача, поэтому не учитывается в рассмотренных методах. Также не учитывается сноски на использованную литературу, если рядом с текстом стоит сноска, то данные текст не является плагиатом, ведь есть ссылка на автора.

Предлагаемые методы можно использовать для создания программного обеспечения обнаружения дубликатов документов любой архитектуры: десктопное или веб-приложение, сервиса и даже мобильного приложения.

6. Выводы, перспективы развития

При разработке любого из рассмотренных методов удастся ускорить процесс проверки научных работ, облегчить работу преподавателей и в перспективе уменьшить количество плагиата в научных и ученических работах. Также эти способы помогут самим студентам проверять качество их работ. Выявление неоригинальных работ — полностью будет автоматизированным процессом, который не будет занимать у преподавателя много времени. Если студенты будут знать, что их работы проверяются программно, то в написании не будут использовать чужие работы или хотя бы будут указывать автора. Уже давно студенты не считают преступлением заимствование чужих работ, хотя это можно приравнять к воровству, только не материальных вещей. Решение проблемы затруднит и замедлит процесс написания у студентов работ, но зато качество их увеличится.

Итак, разработка модифицированного алгоритма шинглов позволит быстро и качественно выявлять плагиатные работы и преподавателям усовершенствовать проверку научных работ студентов.

Литература

1. Авторское право [Электронный ресурс] — Режим доступа: http://ru.wikipedia.org/wiki/Авторское_право — 21.02.2013 г. — Загл. с экрана.
2. Plagiarism.org [Электронный ресурс] — Режим доступа: http://plagiarism.org/resources/facts-and-stats — 24.06.2011 г. — Перевод контекста.
3. Шингл [Электронный ресурс] — Режим доступа: http://www.webeffector.ru/wiki/Шингл — 24.06.2011 г. — Загл. с экрана.
4. Алгоритм Шинглов [Электронный ресурс] — Режим доступа: http://ru.wikipedia.org/wiki/Алгоритм_шинглов — 13.03.2013 г. — Загл. с экрана.
5. Шевченко, О. Ю. Сравнительный анализ современных систем управления онтологическими базами знаний [Текст] / О. Ю. Шевченко, О. Л. Шевченко // Вісник СевНТУ. Збірник наукових праць. Серія Інформатика, електроніка, зв'язок. — 2012. — № 131. — С. 82—86.
6. Дубликат [Электронный ресурс] — Режим доступа: http://ru.wikipedia.org/wiki/Дубликат — 16.03.2013 г. — Загл. с экрана.
7. Часть 1. Алгоритм шинглов для веб-документов [Электронный ресурс] — Режим доступа: http://www.codeisart.ru/part-1-shingles-algorithm-for-web-documents/ — 16.03.2013 г. — Загл. с экрана.
8. Шингл [Электронный ресурс] — Режим доступа: http://wiki.rookee.ru/SHingl/ — 2013. — Загл. с экрана.
9. SEOPULT [Электронный ресурс] — Режим доступа: http://seopult.ru/library/Лемматизация — 2013. — Загл. с экрана.
10. Стемминг [Электронный ресурс] — Режим доступа: http://wiki.rookee.ru/Stemming/ — 2013. — Загл. с экрана.
11. Контрольная сумма [Электронный ресурс] — Режим доступа: http://ru.wikipedia.org/wiki/Контрольная_сумма — 12.03.2013 г. — Загл. с экрана.
12. Зеленков, Ю. Г. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов [Текст] / Ю. Г. Зеленков, И. В. Сегалович // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL 2007: Сб. работ участников конкурса. — Переславль-Залесский, Россия. — 2007.
13. Как устроены MD5 и SHA1 [Электронный ресурс] — Режим доступа: http://habrahabr.ru/post/112780/ — 29.01.2011 г. — Загл. с экрана.

МЕТОДИ ПОШУКУ ДУБЛІКАТІВ СКОМПАНОВАНИХ ТЕКСТІВ НАУКОВОЇ СТИЛІСТИКИ

Дана стаття присвячена такій темі, як пошук плагіату. У статті представлена модифікація алгоритму шинглів для пошуку нечітких дублікатів для скомпонованих документів. Виграш у продуктивності планується досягти за рахунок зменшення кількості порівнянь пар шинглів за рахунок розбиття скомпонованих тестів на розділи.

Ключові слова: хеш-функція, шингл, плагіат, скомпонований документ, дублікат, блок.

Квашина Юлия Андреевна, *кафедра программной инженерии, Харьковский национальный университет радиоэлектроники, Украина,* ***e-mail: julia.kvashina@gmail.com.***

Квашина Юлія Андріївна, *кафедра програмної інженерії, Харківський національний університет радіоелектроніки, Україна.*

Kvashina Julia, *Kharkiv National University of Radio Electronics, Ukraina,* ***e-mail: julia.kvashina@gmail.com***

УДК 621.391

Корчинский В. В.

МЕТОД ПОВЫШЕНИЯ СТРУКТУРНОЙ СКРЫТНОСТИ ПЕРЕДАЧИ ИНФОРМАЦИИ В СИСТЕМЕ СВЯЗИ МНОГОПОЛЬЗОВАТЕЛЬСКОГО ДОСТУПА

В статье предложен метод формирования группового сигнала для системы связи многопользовательского доступа. Для повышения структурной скрытности информационных двоичных сигналов индивидуальных каналов применена система кодирования на основе многоуровневых таймерных сигнальных конструкций. С целью снижения эффективности дешифрирования перехваченного группового сигнала используется метод разделения индивидуальных сигналов по уровню.

Ключевые слова: *таймерный, сигнал, уровень, конфиденциальный, сигнатура, несанкционированный доступ, скрытность, канал, защита.*

1. Введение

В настоящее время при проектировании конфиденциальных систем связи многопользовательского доступа особый интерес представляют методы передачи, которые обеспечивают не только увеличение пропускной способности канала связи, но и повышают скрытность передаваемой информации [1—6].

Скрытность передачи [4] является одним из важных показателей помехозащищенности и определяет способность системы противостоять действиям, направленным на обнаружение сигнала и измерение его параметров. Не менее важным показателем помехозащищенности является помехоустойчивость, которая характеризует способность системы работать с заданным качеством в условиях воздействия различного рода помех.

Очевидна связь этих двух показателей, так как при решении вопросов, направленных на повышение скрытности синтезируемых сигнальных конструкций, в первую очередь необходимо выполнить условие по обеспечению заданной верности передачи.

В зависимости от требований к показателю скрытности передачи различают следующие виды скрытностей сигнальных конструкций: энергетическая, структурная, информационная и т. д. [1]. Энергетическая скрытность определяет способность системы противостоять мерам, направленным на обнаружение сигнала средствами несанкционированного доступа (НСД). Структурная скрытность должна противостоять мерам НСД, которые направлены на раскрытие формы сигнала и измерение его параметров при условии, что сигнал уже обнаружен и перехвачен. Информационная скрытность [9—12] определяется способностью противостоять мерам, направленным на раскрытие смысла передаваемых сообщений с помощью сигналов информации. Данный вид скрытности реализуется в основном на верхних уровнях эталонной модели OSI [7, 8].

Потенциальная структурная скрытность определяется количеством двоичных измерений (д. из), которое необходимо выполнить для раскрытия структуры сигнала без учета алгоритмов обработки на станции НСД [4]. Общее выражение для потенциальной скрытности имеет вид

$$S = \log_2 A, \qquad (1)$$

где A — ансамбль реализаций, определяемый количеством всех возможных значений каких-либо параметров сигнала. Такими параметрами могут быть несущая частота, структура кода, время прихода сигнала и др. В общем случае скрытность зависит от способа построения конкретного вида сигнала.

В работе [5] рассмотрена возможность увеличения структурной скрытности сигналов в каждом индивидуальном канале системы за счет совместного использования таймерных сигнальных конструкций (ТСК) и псевдослучайных последовательностей. Представляет интерес дальнейшее развитие этого направления для задачи повышения структурной скрытности формируемых сигнальных конструкций группового сигнала.