

Шабуров А. С., Журилова Е. Е.

ОСОБЕННОСТИ РЕАЛИЗАЦИИ АЛГОРИТМОВ МОРФОЛОГИЧЕСКОГО АНАЛИЗА В DLP-СИСТЕМАХ

В статье анализируется проблема выбора оптимального алгоритма морфологического анализа для DLP-систем. Рассматриваются основные алгоритмы, используемые для морфологического анализа: стеммер Портера, Stemka и Mystem, а также алгоритм определения слова по суффиксам и аффиксам. Выявляются их возможности, достоинства и недостатки. Приводятся схемы работы этих алгоритмов и их описание. Рассматривается возможность применения этих алгоритмов в DLP-системах на основе сравнения их характеристик и нахождения оптимальных вариантов. Предлагается структурная модель, определяющая место морфологического анализа в функционировании DLP-системы.

Ключевые слова: утечка информации, DLP-система, морфологический анализ, определение слова, алгоритм, основа слова.

Shaburov A. S., Zhurilova E. E.

FEATURES OF THE MORPHOLOGICAL ANALYSIS ALGORITHMS OF DLP-SYSTEMS

In the article analyzed the problem of choosing resembling algorithm of morphological analysis of DLP-system. It views the main algorithms of morphological analysis: Porter's stemmer, Stemka, Mystem, and also the algorithm of word determination by suffixes and affixes. Reveal its capabilities, strengths and weaknesses. It shows the scheme of this algorithms and its description. Considering the opportunity of using these algorithms in DLP-systems, by comparing its characteristics and finding the optimal variants. It gives a structural model, which determine the place of morphological analysis in functioning of DLP-system.

Keywords: information leak, DLP-system, morphological analyses, word definition, algorithm, stem of a word.

В современных условиях безопасность функционирования информационных систем зависит от многих факторов. Одной из актуальных проблем обеспечения функционирования систем различного назначения является создание защищенной информационной

среды. Защищенность среды зависит в первую очередь от блокирования несанкционированного доступа, а также защиты от утечек информации. На сегодняшний день одним из самых распространенных решений в области борьбы с утечками информации являются

DLP-системы. Рынок средств безопасности предоставляет достаточный выбор решений, в различных ценовых категориях. Основной функционал DLP-систем схож, но каждая компания – разработчик подобных систем использует различные подходы в технологиях обнаружения канала утечки информации и его блокирования. Кроме того, используются разнообразные формы представления полученной информации.

Несмотря на это основой выявления утечек информации в DLP-системах является морфологический анализ текстов. Главным преимуществом этой технологии является универсальность алгоритмов анализа, которые позволяют проводить оценку как сообщений в различных мессенджерах, так и текста электронных документов [1]. Также преимуществами использования этого метода являются возможность работы с содержимым, обучаемость лингвистического алгоритма, масштабируемость и простота настройки. К недостаткам можно отнести зависимость от используемого языка и необходимости применения вероятностного подхода [2].

Морфологический анализ представляет собой процесс определения грамматического значения словоформы и выделения ее основы, или, иными словами, выделение ключевых слов в потоке текста [3].

Любой алгоритм морфологического анализа состоит из двух основных компонентов – декларативного и процедурного. При этом декларативный компонент подразумевает таблицы структурированных данных, требуемых для анализа, а процедурный компонент содержит сами алгоритмы анализа и вспомогательные процедуры [4].

В связи с особенностями русского языка и наличием большого количества слов исключений в нем осуществление морфологического анализа может быть затруднено. Для преодоления этих затруднений существует возможность выбора метода морфологического анализа, среди которых наиболее известными являются три основных.

Первым методом является составление морфологического словаря для конкретного предприятия вручную, с учетом всех ключевых слов, способных указать на утечку информации. Данный способ целесообразно использовать при небольшом объеме возможных ключевых слов, анализируя корень данных слов. Если информационная система предприятия (организации) сложная, содер-

жит большое количество разнообразных ресурсов, то использование данного метода будет затруднительно, особенно на начальных этапах.

Особенностью второго метода является использование алгоритма стемминга, суть которого состоит в выделении основы слова, а не его корня.

Для русского языка наиболее популярными алгоритмами стемминга являются:

- стеммер Портера;
- Stemka;
- Mystem [5].

Стеммер Портера, иначе называемый «snowball», был разработан в 1979 году изначально для английского языка, впоследствии адаптирован под анализ на основе русского языка. При использовании данного алгоритма стемминг происходит на основе множества существующих суффиксов. Сам алгоритм состоит из четырех основных шагов и представлен на рис. 1.

Цифрой 1 обозначен блок операций для отсекаания формообразующих суффиксов. Цифрой 2 обозначен блок операций для отсекаания окончаний «и». Цифрой 3 обозначен блок операций для отсекаания словообразующих суффиксов. Цифрой 4 обозначен блок операций для отсекаания суффиксов превосходной формы, окончаний на «ь» и удвоенных «н».

В результате выполнения алгоритма получается требуемая для опознания часть слова.

Основным достоинством алгоритма Портера является отсутствие словарей основ, что существенно увеличивает быстродействие.

К отрицательным свойствам данного алгоритма можно отнести возможность потери части информации из анализируемого слова. Кроме того, уязвимостью данного алгоритма является возможность ошибки со стороны оператора, задающего правила проверки.

Stemka – русско-украинский идентификатор морфологии, который был создан с помощью специально разработанного морфологического модуля [6]. Алгоритм основан на вероятностной модели. Составляется массив данных, который включает в себя пары «две последние буквы основы» и «суффикс», таким образом, получаются модели различных слов. После этого определяется вероятность появления моделей в тексте, и при вероятности 1/10000 модель отсекается. Результат представляет собой таблицу переходов ко-

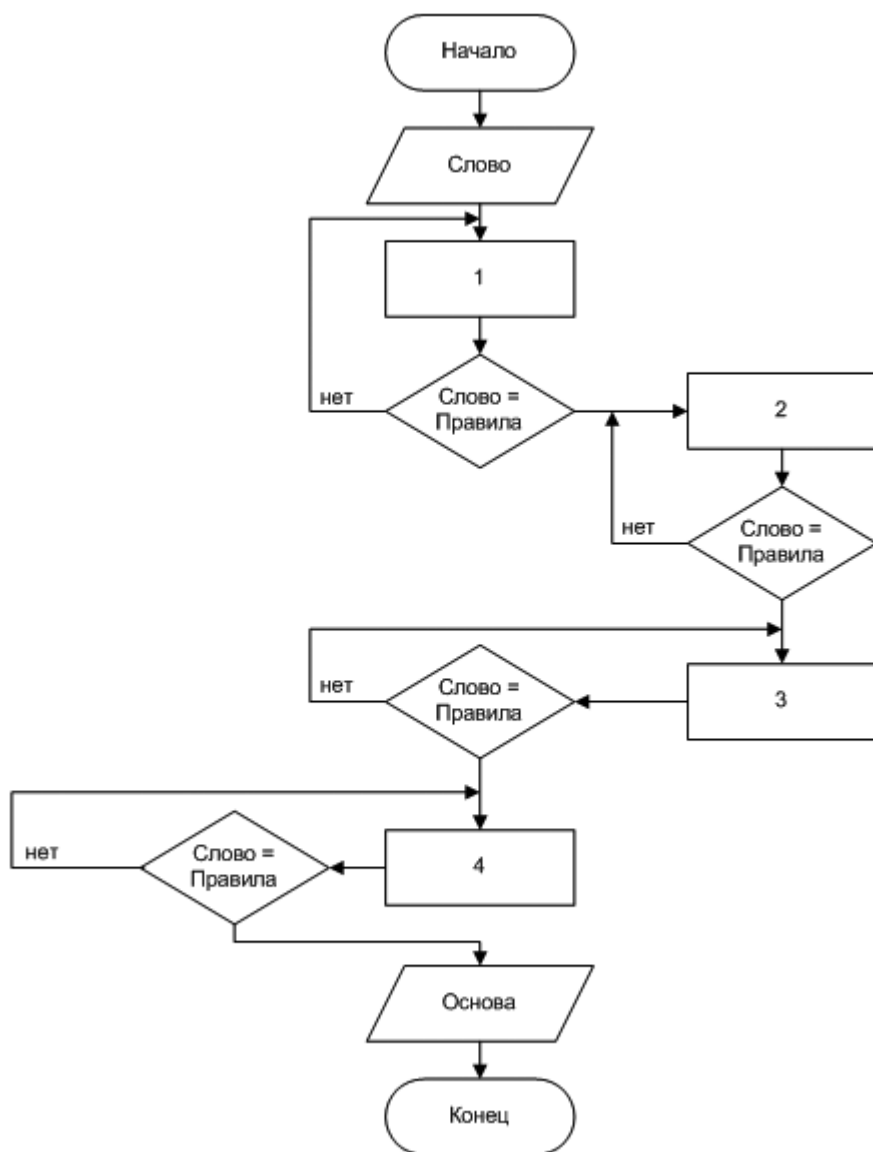


Рис. 1. Алгоритм работы стемминга Портера

нечного автомата, по которым сканируется слово [5].

Для рассмотренных алгоритмов Snowball и Stemka в процессе отладки было сформулировано правило, предусматривающее наличие хотя бы одной гласной буквы в основе.

Mystem был разработан в 1998 Ильей Сегаловичем. Модель строится в виде леса инвертированных префиксных деревьев суффиксов и инвертированного префиксного дерева для основ, для этого используется словарь с перечислением всех грамматических форм (парадигмы) слова [5].

На рис. 2 представлен алгоритм Mystem, функционирование которого заключается в следующем. Очередное анализируемое сло-

во подвергается разделению на стемму и суффикс. Такое разделение производится программой на основе уже имеющегося дерева суффиксов. Далее происходит сопоставление получившейся основы с уже имеющимися в словаре (блок 2) для нахождения соответствий. Если соответствие найдено, алгоритм заканчивает работу, и результатом является гипотеза для словарного слова.

Если же соответствие найти не удалось, то алгоритм продолжает работу по поиску нужной гипотезы. Для этого происходит генерация гипотетической модели слова, базирующейся на основе слова, суффиксе и ближайшей основе из имеющегося словаря (блок 3). После этого полученная модель снова сверя-

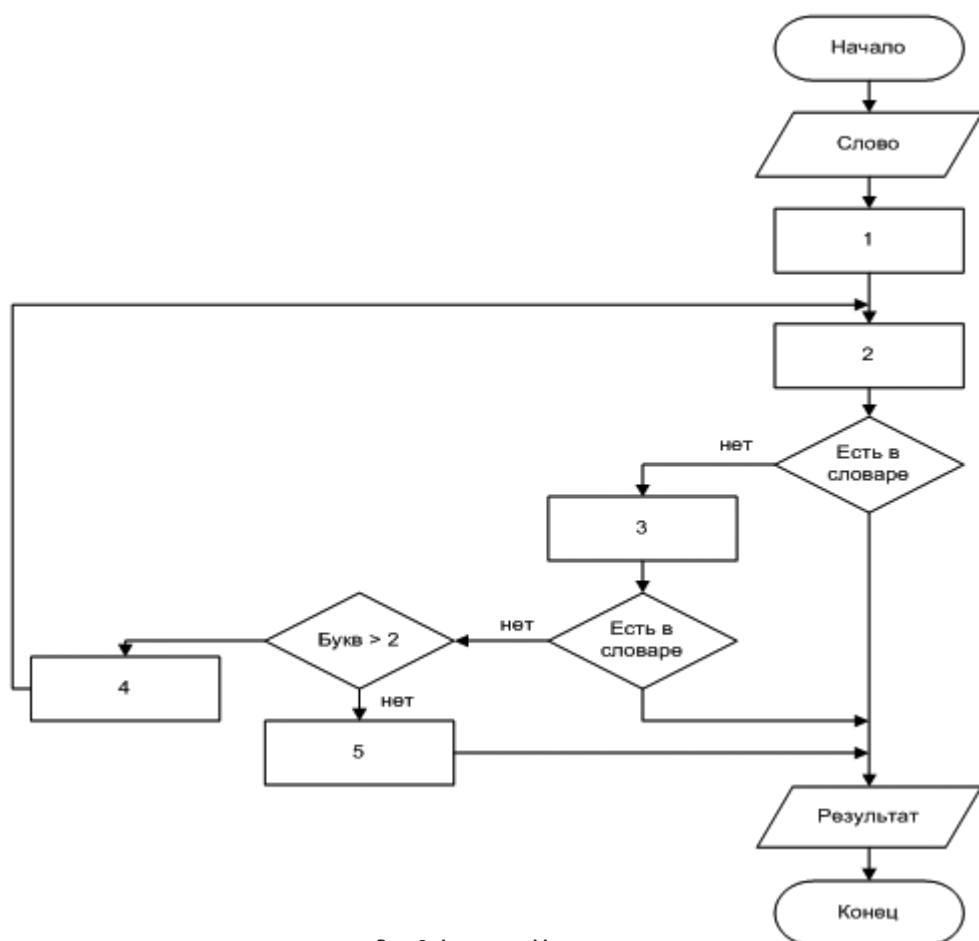


Рис. 2. Алгоритм Mystem

ется со словарем. При положительном результате алгоритм заканчивает работу.

Если результат снова отрицателен и совпадение не найдено, то алгоритм продолжает производить вышеуказанные действия, постепенно уменьшая основу на одну букву (блок 4) до тех пор, пока основа не будет найдена, либо пока количество букв не сократится до двух. В этом случае происходят ранжирование все полученных в ходе работы алгоритма гипотез (блок 5) по продуктивности и отсекаание менее продуктивных. В результате на выходе алгоритма получается набор гипотез для несуществующего в имеющемся словаре слова.

Преимуществами данного алгоритма являются простота реализации и словарей, а также возможность определения форм слова, отсутствующих в словаре. К недостаткам относятся ориентация только на русский язык и анализ по окончанию.

Третьим методом является определение слова по его суффиксу и аффиксу и приведение слова к его начальной форме. Данный

способ является наиболее рациональным благодаря особенностям русского языка, однако для повышения качества работы алгоритма необходимо добавить как можно больше слов-исключений [7].

Таким образом, использование одного из вышеперечисленных алгоритмов морфологического анализа позволяет распознать конфиденциальную информацию в потоке перехватываемой информации.

Существуют несколько критериев, влияющих на выбор алгоритма морфологического анализа для DLP-систем. К ним относятся точность определения слова при морфологическом анализе, возможность обучаемости, временные затраты на настройку системы. Согласно исследованиям точность определения должна быть не ниже 95–97%. Возможность обучаемости позволяет симитировать на этапе настройки системы опасные ситуации, тем самым позволив системе самостоятельно адаптироваться под необходимые параметры, что существенно сокращает время ввода системы в эксплуатацию. Для сотрудни-

Сравнение алгоритмов морфологического анализа

Название алгоритма	Точность определения слова при морфологическом анализе	Возможность обучаемости	Временные затраты на настройку
Составление морфологического словаря	80–85%	Отсутствует	1–2 дня
Стеммер Портера	85–90%	Отсутствует	3–5 часов
Stemka	87–95%	Присутствует	1–1,5 дня
Mystem	92–97%	Присутствует	2–3 дня
Определение по суффиксу и аффиксу	90–96%	Отсутствует	1–2 дня

ка, не имеющего навыков аналитика, создание словаря для морфологического анализа может занять достаточно большое количество времени, что существенно скажется на скорости ввода системы в эксплуатацию. Проанализируем вышеуказанные алгоритмы по этим критериям (табл. 1).

Исходя из данных, представленных в таблице, наиболее подходящим алгоритмом морфологического анализа для использования в DLP-системах является алгоритм Mystem. Несмотря на длительность настройки, алгоритм имеет самую высокую из перечисленных алгоритмов вероятность определения слова и обладает важным свойством обучаемости.

На рис. 3 представлено место морфологического анализа в структуре работы DLP-системы. Информация, отправляемая поль-

зователем, перехватывается и передается на сервер, где подвергается морфологическому анализу. В процессе анализа определяются источники информации, конечные получатели, а также иные характеристики, необходимые для принятия решения о принадлежности информации к конфиденциальной. При наличии признаков возможной утечки информации ограниченного доступа администратору информационной безопасности передается отчет о нарушении, в который также включаются данные, выявленные при анализе. На основании политики информационной безопасности принимается решение по выявленному инциденту.

Таким образом, морфологический анализ является основой алгоритма выявления утечек информации в DLP-системах. Качество



Рис. 3. Место морфологического анализа в структуре DLP-системы

реализации анализа определяет эффективность функционирования всей системы. В процессе исследования были рассмотрены различные алгоритмы морфологического анализа, используемые для выявления каналов утечки информации в DLP-системах. Анализ характеристик и функциональных воз-

можностей позволил выявить оптимальный алгоритм, наиболее подходящий для использования, основным преимуществом которого является возможность получения не только слова, имеющегося в словаре, но и нескольких гипотез для слова, которое в словаре отсутствует.

Примечания

1. Давлетханов М. Современные технологии обнаружения утечек [Электронный ресурс]. – URL: <https://www.anti-malware.ru/node/8578#part2> (дата обращения: 6.04.2016).
2. Жарников М. Обзор технологий и вендоров «классического» DLP // Презентация компании ИТКС Информационная безопасность. – Екатеринбург, 2014.
3. Шабуров А. С., Журилова Е. Е., Лужнов В. С. Технические аспекты внедрения DLP-системы на основе Falcongaze Secure Tower // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – Пермь, 2015. – № 16. – С. 57–67.
4. Пруцков А. В., Розанов А. К. Методы морфологической обработки текстов // Прикаспийский журнал: управление и высокие технологии. Обработка сигналов и данных, распознавание образов, выявление закономерностей и прогнозирование. – Астрахань, 2014. № 3 (27). URL: <http://prutzkow.com/pdf/114.pdf> (дата обращения: 6.04.2016).
5. Астапова О. П. Исследование и разработка методов нормализации слов русского языка : курсовая работа. – М., 2012. URL: <http://seminar.at.ispras.ru/wp-content/uploads/2012/10/Astapova-thesis.pdf> (дата обращения: 6.04.2016).
6. Сегалович И. Быстрый морфологический алгоритм подбора неизвестного для поисковой системы слова с помощью словаря [Электронный ресурс]. – М., 2014. URL: <http://wseob.ru/seo/morphological-algorithm> (дата обращения: 6.04.2016).
7. Жаринов Р. Ф. Метод защиты от перлюстрации в DLP-системах // Доклады ТУСУРа. – Томск, 2012. – № 1 (25). URL: <http://www.tusur.ru/filearchive/reports-magazine/2012-25-2/126.pdf> (дата обращения: 6.04.2016).
8. Левцов В. Контроль подмены символов в системах борьбы с утечками конфиденциальных данных [Электронный ресурс]. URL: http://www.leta.ru/press-center/publications/article_487.html (дата обращения: 10.04.2016).

Шабуров Андрей Сергеевич, кандидат технических наук, доцент кафедры автоматизации и телемеханики Пермского национального исследовательского политехнического университета. 614990, Пермь, Комсомольский пр., 29. E-mail: shans@at.pstu.ru

Журилова Елена Евгеньевна, студент кафедры автоматизации и телемеханики Пермского национального исследовательского политехнического университета. 614990, Пермь, Комсомольский пр., 29. E-mail: ele11485995@yandex.ru

Shaburov Andrey Sergeevich, PhD of Technical Sciences at the Department of Automation and Telemechanics, Perm National Research Polytechnic University. 614990, 29, Komsomolsky prospect, Per. E-mail: shans@at.pstu.ru

Zhurilova Elena Evgen'evna, student at the Department of Automation and Telemechanics, Perm National Research Polytechnic University. 614990, 29, Komsomolsky prospect, Perm. E-mail: ele11485995@yandex.ru