УДК 004.9

ПОДХОДЫ К АЛГОРИТМИЗАЦИИ ПОСТРОЕНИЯ ТАБЛИЧНОГО РЕФЕРАТА НА ОСНОВЕ РАБОТЫ КИНЕТИЧЕСКОЙ МАШИНЫ КИРДИНА

Корпусова Н.Е., Личаргин Д.В.
научный руководитель канд. тех. наук Личаргин Д.В.
Сибирский федеральный университет

На сегодняшний день проблема реферирования текстов актуальна в связи с лавинообразным увеличением объема информации из различных источников. В связи с быстрыми темпами развития науки и техники в каждой научной сфере появляется все больше и больше публикаций, диссертаций и учебных пособий. Следовательно, чтобы успеть рассмотреть хотя бы одну отрасль, человеку необходимо ежедневно читать как минимум тысячу страниц текста, причем не на одном, а на десятках языков. Для упрощения обработки больших объемов литературы и были созданы различные способы их компактного представления без потери смысловой целостности, в том числе и с помощью компьютеров путём создания множества программных продуктов, помогающих представить текст в виде таблиц, диаграмм, наборов ключевых слов.

Основные задачи данной работы заключаются:

1) в решении проблемы составления табличного реферата на основе принципов работы кинетической машины Кирдина;

2) в рассмотрении методов подстановки данных в шаблоны из пространства слов и их представления в форме полноценного реферата, который можно адекватно привести к виду, близкому к первоначальному тексту по смыслу.

Метод табличного реферирования заключается в представлении исходной информации в виде структурированной таблицы. Он ориентирован на выделение основной информации из текста, будь то ключевые фразы или целые предложения. Основная цель – упрощение восприятия текста. Результатом работы является таблица с кратким изложением исходной научной работы или публикации. С помощью алгоритма построения табличного реферата мы можем сфокусировать внимание на основных пунктах и обязательных фактах источника, пренебрегая незначительными данными и тем самым обеспечивая отсутствие избыточности информации в своём реферате.

Метод может быть использован в целях реферирования научных и технических статей, научно-популярных текстов и инструкций, так как метод опускает избыточную информацию, ориентируясь на ключевые темы источника.

В методе табличного реферата используются логически объединенные строки и столбцы. Каждый объект может быть представлен как строка или столбец таблицы.

Табличный реферат может быть использован на широком наборе текстов и для различных целей. Вид конечной таблицы зависит от преследуемой человеком-составителем цели. Рассмотрим пример использования таблицы для классификации и описания объектов.

Предположим, что строки являются объектами, а столбцы – свойствами или атрибутами этих объектов. Тогда на пересечении строки и столбца будет отметка, при условии, что объект обладает данным свойством.

Метод табличного реферата ускоряет процесс запоминания основных положений текста, а также процесс восприятия особо важной информации. Также данный метод помогает в структурировании полученных знаний.

Кинетическая машина Кирдина используется наряду с нормальным алгоритмом Маркова, машиной Колмогорова и Тьюринга или схемой Поста в целях обеспечения процесса пошагового вычисления. Объект, подвергающийся обработке, это – некоторый ансамбль слов T. Каждое слово в этом ансамбле может быть представлено в нескольких экземплярах, количество которых мы обозначим как s. Начало обработки

характеризуется воздействием на эти слова некоторых правил-шаблонов, которые притягивают к себе некоторое количество слов на основе коэффициента «энергетической выгодности».

При применении команд *P*, которые являются лексической или грамматической трансформацией слова из ансамбля *T*, функция *f(T)* – притяжения к другой структуре изменит свои значения и примется решение о занесении слова или словосочетания в структуру, в нашем случае, структуру табличного реферата. Чем больше трансформаций *P*, тем меньше энергетическая выгодность. Следовательно, семантические единицы не будут изыматься из текста, и переноситься на носитель (табличный реферат).

Таблица 1.

Пример табличного реферата

Malware	**Virus / worm / Trojan horse is**	**Malware is shaped into**	**Aims of malware are**	**Defense methods against malware are**	**Self-preservation techniques are**	**Malware transport mechanisms are**
Viruses	Self-replicating program that spreads by attaching itself to other programs	Executable files, boot sectors, documents	Deleting files, corrupting data, displaying messages on the victim's screen.	Virus signatures, heuristics and integrity verifications, principle of least privilege, user education	Polymorphic techniques	Removable storage, e-mail attachments, web downloads and shared directories
Worms	A self-replicating program that spreads via networks	Separate program modules	Planting a distributed denial of service flood agent, opening up a backdoor	System patching, arbitrary outbound connections blocking		Buffer overflow exploits, file-sharing services, e-mail
Trojan horses	Non-replicating program that includes hidden malicious functionality	Useful program	Opening up backdoors and sniffers	Showing which programs are listening on TCP and UDP network ports, MD5 hashes	Steganography methods, polymorphism	E-mail and web-site downloads

Рассмотрим принципы работы кинетической машины Кирдина для нашего исследования в более простом виде. Описывая неформально её работу, мы можем говорить о банке с некоторым количеством слов (в нашем случае это будут слова исходного текста о вредоносном программном обеспечении), в которую опущены некоторые шаблоны-правила – *P1*, *P2* и *P3*. Они, сталкиваясь с цепочкой содержащихся семантических единиц, способствуют либо их распаду, либо синтезу или даже замене слов в этой цепочке. Для составления табличного реферата, опираясь на специфику работы этих шаблонов, нам необходимо выделить ключевые фразы из «словесного» набора, то есть в нашем случае необходим синтез пар слов. Каждый шаблон – название столбца, к которому будут притягиваться слова из нижеприведенного текста: «Malware», «Virus/worm/trojan horse + is …», «Malware + is shaped into/can be/can be classified as

…»
и т.д.

Далее рассмотрим в качестве примера фрагмент текста о вредоносном программном обеспечении. «Computer security expert Edward Skoudis and technical writer William Stallings covered the theme of computer malicious software. Since computer game «Darwin», which was the first self-replicating program developed in 1962, computer industry has faced with explosive rise malicious programs that pursue illegal purposes. … That was done in order to teach the reader to basic defensive skills against this type of crimes. Viruses are historically first kind of malware. Appeared in the sixties, it became widespread in the end of XX century that resulted in virus epidemic in computer network everywhere. Although, in some experts sigh, viruses are outdated, the most scientists note that it can still present a danger for security systems. … The simplest way is searching code fragments of known viruses in tested files (which is called virus signature verification), but it cannot always be effective. … The authors advised to patch user's system in a timely manner to protect you against the worms».

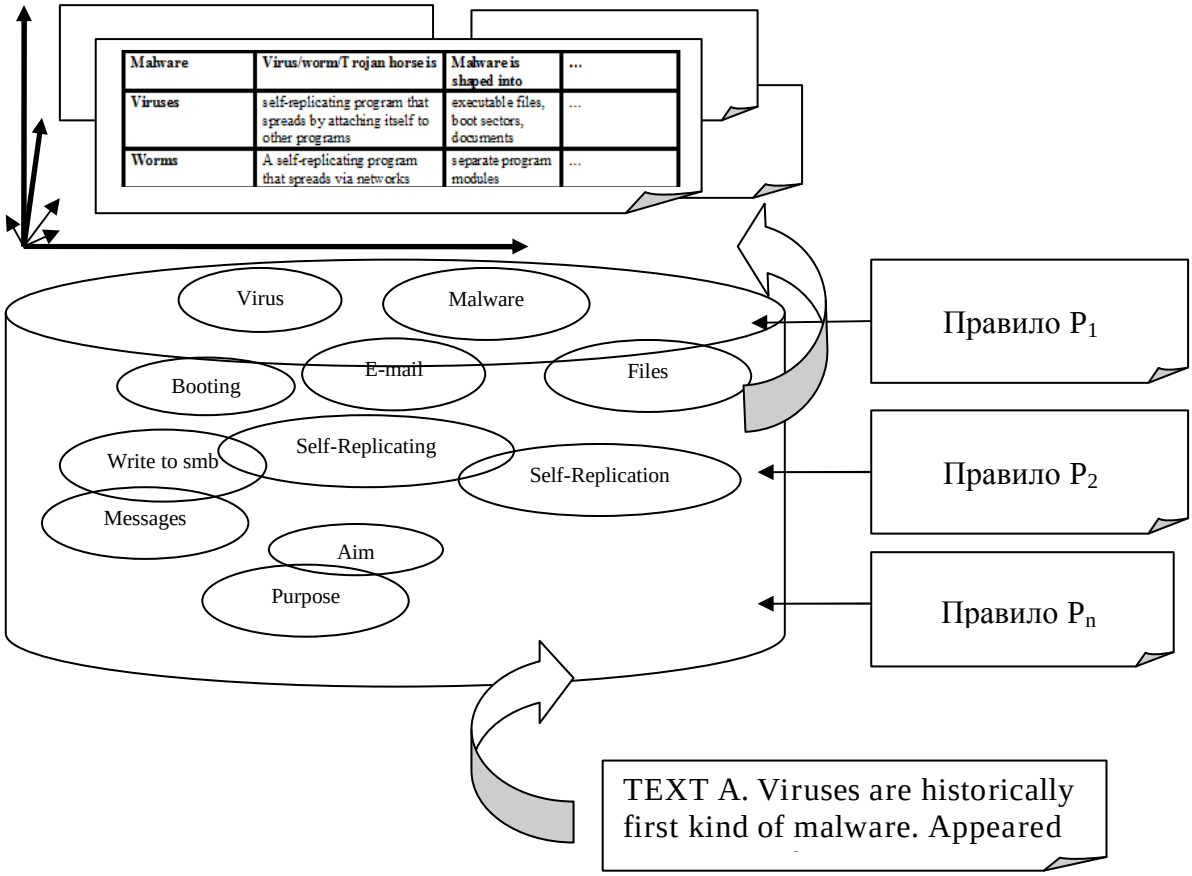

Рисунок 1. Модель лексико-грамматического пространства на основе рассматриваемого текста «вредоносное программное обеспечение»

В ходе преобразований по принципу работы машины Кирдина получим следующий табличный реферат (см. Таблицу 1). Общий вид модели преобразования текста в табличный реферат можно представить в форме схемы (см. Рисунок 1). Рассмотрим некоторые правила катализаторы машины Кирдина, обеспечивающих преобразования: «текст – семантическая сеть – смысловой шаблон».

Классы множества правил применяемых к словам языка. Для решения проблемы генерации табличного реферата предлагается использовать определенные классы правил-катализаторов в рамках этой модели, для получения на выходе табличного реферата приемлемого качества (что рассматривается в настоящий момент чисто

теоретически). Здесь под правилами будем подразумевать преобразование слова S в ряд слов S_i ассоциативно близких к первому по тому или иному критерию. Правила преобразования терминов и общих по значению слов имеют следующий вид с учетом примеров лексических и иных лингвистических трансформаций:

1. Определения: стремление получить, эмоциональная тенденция, направленность в будущее;
2. По частям речи: желание, желать, желательный, желательно;
3. По позиции в предложении: что-то как желание, быть желанием, что-то желания, с учетом желания;
4. Ряды дифинонимов (разные элементы одной речевой ситуации или явления): желающий, объект желания, страстный, вожделенный;
5. Ряд близких синонимов: хотение;
6. Широкий синонимический ряд (с общей 1-3 семами): стремление, нужда, необходимость;
7. Группы слов контекстуальных синонимов (от 1 и более общей семы): важность / значимость / насущность, цель / задача / план, идея / проект / концепция, динамика / направленность / сосредоточение;
8. Метафорические употребления: жажда, алчность, жадность, вожделение;
9. По категориям, например, множественного числа: с желаниями;
10. «Атрибуции»: с множеством желаний, с кучей желаний, с чувством желания, с чувством теплоты / прикосновения / огня желания, с эмоциональным / душевным желанием;
11. Перевод на другой язык: a desire, a wish, to want.

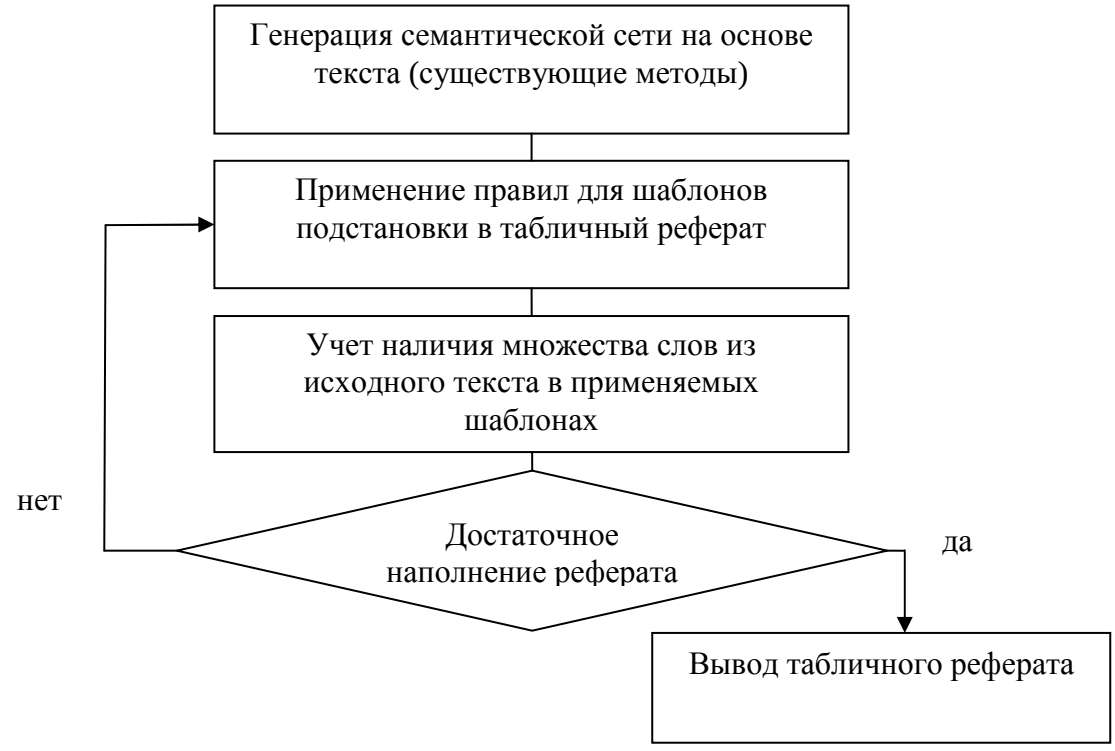

Рисунок 2. Блок-схема алгоритма генерации табличного реферата

Таким образом, мы можем сделать вывод о возможной принципиальной разрешимости проблемы составления табличного реферата на основе принципа работы кинетической машины Кирдина в некотором приближении, что требует привлечения определенных человеческих и информационных ресурсов, способных составить необходимые базы данных, знаний и правил.

Выводы. В работе выполнен анализ проблемы формализации процесса построения табличного реферата, предложено общее направление в решении этой проблемы с привлечением такой абстрактной модели как машина Кирдина.