- Масштабируемость – скрапер должен поддерживать возможность увеличения производительности за счет добавления дополнительных вычислительных узлов, на которых он исполняется;
- Производительность и эффективность – скрапер должен обеспечивать эффективное использование системных ресурсов, включая процессор, память и полосу пропускания сети;
- Качество – скрапер должен уметь отделять спам-страницы от полезных и извлекать последние;
- Актуальность – скрапер должен поддерживать обновление собранных данных;
- Расширяемость – скрапер должен быть модульным, т.е. позволять добавлять новую функциональность, для анализа новых форматов данных, протоколов и т.д.

Помимо описанных общих требований для скраперов, можно обозначить основные требования, для конкретной задачи исследования:

Скрапер должен быть кросплатформенным, чтобы его можно было одинаково настраивать и конфигурировать на вычислительных узлах с разными операционными системами;

Скрапер должен обеспечивать производительность обработки порядка 100 стр/сек, чтобы время сбора описанного выше объема данных составляло часы, а не дни. В том случае если окажется, что данных для сбора и анализа больше предполагаемого, скрапер должен предоставлять возможность легко увеличить его производительность путем выделения ему для работы большего числа потоков или добавления дополнительных вычислительных узлов;

Скрапер должен быть интегрирован с базой данных для хранения собранной информации и полнотекстовым индексом, позволяющим быстро извлекать данные для последующего анализа, отвечающие указанным условиям;

Требуется скрапер для сбора данных в ширину и вертикального поиска, так как в указанной задаче необходимо извлечь информацию о конкретной предметной области, а не узкое множество фактов;

В настоящее время существует множество готовых решений веб-скраперов, но готового решения для данной задачи исследования нет, поэтому, для реализации поставленной задачи был разработан собственный веб-скрапер.

Созданный скрапер является эффективным инструметом для поиска в Вебе, ядро написано на C++ с которым взаимодействует Ruby-оболочка., поддерживает граф связей узлов, различные парсеры, фильтры и нормализаторы URL. Он позволяет использовать различные хранилища данных, такие как Cassandra, Hbase и др. Скрапер также является масштабируемым (до 100 узлов в кластере и легко настраивается и расширяется, в полной мере является “вежливым”.

Списокиспользованныхисточников:

1. PAPAVASSILIOU V., PROKOPIDIS P., THURMAIR G. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web // Proceedings of the 6th Workshop on Building and Using Comparable Corpora. — 2013.
2. AHUJA M.S., BAL J.S., VARNICA Web Crawler: Extracting the Web Data // International Journal of Computer Trends and Technology. — 2014.
3. YADAV M., GOYAL N. Comparison of Open Source Crawlers-A Review// International Journal of Scientific & Engineering Research. — 2015.

КРИПТОГРАФИЧЕСКИЕ АЛГОРИТМЫ ШИФРОВ ЗАМЕНЫ

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

Сидоренко К.А., Приловский Е.В., Усенко Д.В.
Стройникова Е. Д. – ассистент кафедры информатики

Шифры замены являются наиболее часто используемыми шифрами на сегодняшний день. Они характеризуются тем, что отдельные части сообщения (слова, буквы) заменяются на другие буквы, числа, символы и т. д. Но, при этом замена осуществляется так, чтобы через шифрованное сообщение можно было восстановить передаваемое сообщение. Несмотря на вытеснение шифров подстановки блочными шифрами одноразовые блокноты ещё остаются применимыми на государственном уровне. Они используются для обеспечения сверхсекретных каналов связи. Так, по некоторым данным, телефонная линия между главами США и СССР шифровалась при помощи одноразового блокнота и вполне возможно, что подобные линии существуют до сих пор. Одноразовые блокноты применяются шпионами в различных государствах для сокрытия важной информации. Такие сообщения невозможно расшифровать, если отсутствует ключ, записанный в блокноте, независимо от вычислительной мощности ЭВМ.

Шифр Цезаря

Шифр Цезаря является шифром подстановки, который работает следующим образом: все символы циклически заменяются символами, которые расположены на определенном числе позиций в любом направлении от них в алфавите. Рассмотрим следующий пример: в шифре со сдвигом вправо на 3 происходит замена A на D, B на E,…, Z на C и т. д. (рис.1). Способ шифрования с помощью шифра Цезаря есть составляющая часть более сложных шифров, например такого, как шифр Виженера. Но, т. к. шифр Цезаря является моноалфавитным, его легко разгадать и он не практичен в использовании.

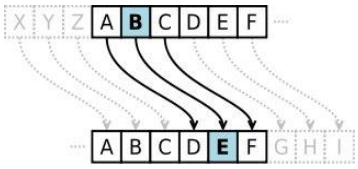

Рис. 1 – Шифр Цезаря с величиной циклического сдвига 3

Шифр Атбаш

Атба́ш – простой шифр подстановки для алфавитного письма. Правило шифрования состоит в замене i-й буквы алфавита буквой с номером n – i + 1, где n – число букв в алфавите. Ниже, на рис. 2, дан пример для английского алфавита:

Исходная буква	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Зашифрованная буква	Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A

Рис. 2 – Ключевая таблица шифра Атбаш для английского алфавита

Шифр Виженера

Для реализации шифра воспользуемся следующей таблицей для английского алфавита (рис. 3):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Рис. 3 – Таблица Виженера для английского алфавита

Как видим, таблица Виженера составляется из строк по 26 символов, причём каждая строкациклически сдвинута относительно строки, находящейся над ней, на одну букву влево. Таким образом, в таблице получается 26 различных шифров Цезаря. На каждом этапе шифрования используются разные алфавиты, их выбирают в зависимости от символа ключевого слова. Например, пусть дан исходный текст «ATTACKATDAWN». Человек, посылающий сообщение, записывает ключевое слово «LEMON» циклически до того момента, пока не достигнет конца исходного текста. В нашем случае получится последовательность «LEMONLEMONLE». Первый символ исходного текста «A» зашифрован с помощью символа «L», который является первым символом ключа. Первый символ «L» шифрованного текста находится на пересечении строки, соответствующей «L», и столбца, соответствующего «A», в квадрате Виженера. Аналогично для второго символа исходного текста используется второй символ ключа; т. е. второй символ шифрованного текста «X» получается на пересечении строки, соответствующей «E», и столбца, соответствующего «T». Аналогично шифруется остальная часть исходного текста. Таким образом, имеем:

ATTACKATDAWN – исходный текст;

LEMONLEMONLE – ключевая последовательность;

LXFOPVEFRNHR – зашифрованный текст.

На основе алгоритма шифра Виженера разработана учебная программа на языке программирования C#. Сначала требуется ввести текст, который подлежит шифрованию. Затем высчитываем символ шифра по формуле c[i] = (alphabet[i] + keyword[i]) % N, где i = {0…n}, n – количество символов, c[i] – символ в строке, alphabet[i] – символ в алфавите, keyword[i] – ключ. Получаем зашифрованный текст и выводим его на экран. Расшифровка производится по аналогичному принципу.

Шифровальные устройства

1. Энигма – переносная шифровальная машина, которая применялась для шифрования и дешифрования. Шифрование производилось за счет комбинаций нажатия клавиш.

2. Шифровальное устройство М-94 – принцип работы основывался на механизме крутящихся дисков,

на которых были написаны буквы и цифры.

3. Шифровальное колесо Болтона – работало по принципу простой замены одной буквы на другую.

4. Шифровальная машина Конвертер М-209 – зашифрованное сообщение распечатывалось на бумаге в виде пятизначных групп.

5. Шифровальная машина Лоренц – принцип работы был основан на поточном шифре Вернама.

Подготовлен доклад с целью ознакомления студентов с основами криптографии. Данная информация может быть полезна студентам, обучающимся по специальности “Защита информации”, а также тем, кто умеет или учится профессионально программировать, интересуется сжатием данных или занимается исследованием современных средств шифрования. В наши дни криптография используется практически во всех сферах, работающих с приёмом и передачей информации, обеспечивает работу сверхсекретных каналов связи, а также эта наука успешно применяется в банковской деятельности.

Список использованных источников:

1. Василенко, О. Н. Теоретико-числовые алгоритмы в криптографии / О. Н. Василенко. – М. : МЦНМО, 2003. – 326 с.

2. Введение в криптографию / В. В. Ященко [и др.] ; под общ. ред. В. В. Ященко. – 3-е изд., перераб. – М. : МЦНМО, 2003. – 400 с.

3. Математические и компьютерные основы криптологии : учеб. пособие / Ю. С. Харин [и др.]. – Минск : Новое знание, 2003. – 382 с.

4. Стройникова, Е. Д. Основы прикладной алгебры : учеб.-метод. пособие / Е. Д. Стройникова. – Минск : БГУИР, 2010. – 120 с.

5. Шифровальные устройства [Электронный ресурс]. – Режим доступа: http://kryptography.narod.ru/mashiny.html.

ДИАГНОСТИКА НА РАК ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ

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

Пунько В.В., Приходько В.С.

Волорова Н. А. – к.т.н., доцент

Популяция людей на планете с каждым днем увеличивается, как и увеличивается число болезней, не подлежащих лечению и очень трудных в диагностике. Одним из таких заболеваний является рак. Около 13 % всех смертей в мире происходит из-за онкологических заболеваний, которые сегодня считаются самой распространенной патологией после инсульта и ишемии.

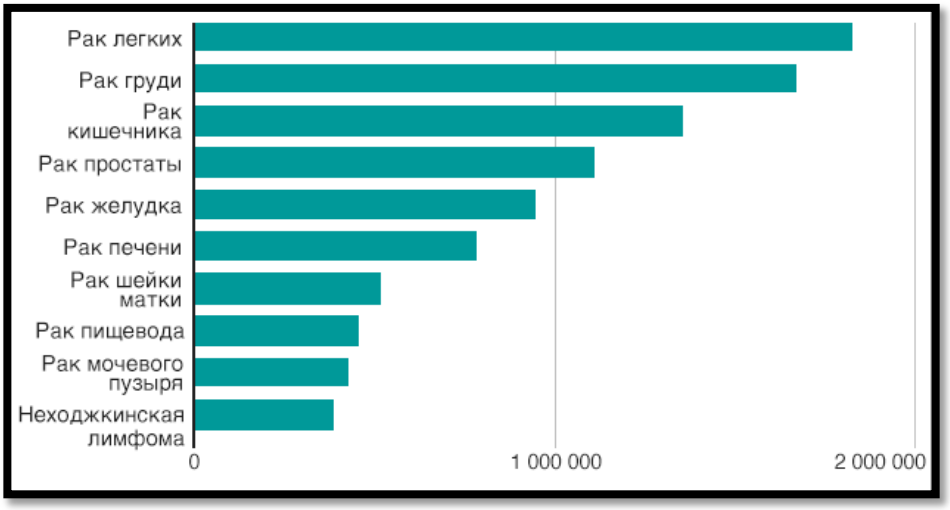

Рис. 1 – Статистика раковых заболеваний на 2012 год

Дело в том, что онкологических заболеваний (того самого рака) — огромное количество разновидностей. Например, раков молочной железы существует более 20 видов, и, кроме этого, у каждого вида рака молочной железы есть характеристики, влияющие на стратегию лечения. Заболеваний, которые называют лимфомы — тысячи видов и подвидов. И, опять же, существует принципиальная разница в их лечении. От правильности постановки диагноза в итоге зависит успешность или не успешность лечения. Пациента можно лечить сколь угодно хорошо, но если его лечат от другого вида рака — лечение не имеет существенного эффекта.

Про онкологическую диагностику много говорят, и, несмотря на это, очень мало знают в среде непрофессионалов. Во-первых, скрининг (ранняя диагностика) и диагностика — это совершенно разные вещи. Во-вторых, от этапа, когда пациент впервые попадает к онкологу, и онколог подозревает у пациента