УДК 004.89

05.00.00 Технические науки

МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОЦЕССА ОБНАРУЖЕНИЯ СВЕДЕНИЙ КОНФИДЕНЦИАЛЬНОГО ХАРАКТЕРА В ЭЛЕКТРОННЫХ ДОКУМЕНТАХ

Птицын Андрей Александрович
SPIN-код = 7006-1394

Краснодарское высшее военное училище, Краснодар, Россия

В статье разрабатывается математическое обеспечение процесса обнаружения сведений конфиденциального характера на основе технологии баз знаний. Представлен алгоритм обнаружения сведений конфиденциального характера в электронных документах, передаваемых за пределы защищаемой информационной системы, за счет применения лингвистических технологий глубокого анализа текста. Произведена оценка вычислительной сложности разработанного алгоритма. Разработанные структуры данных и алгоритм реализованы на языке программирования C++. Представлены результаты серии экспериментов, подтвердившие работоспособность разработанного алгоритма. Проведенные экспериментальные исследования разработанного алгоритма показали его практическую применяемость. Выполнена оценка качества обнаружения сведений конфиденциального характера. Полученные оценки показателей качества обнаружения показали, что разработанные структуры данных и алгоритм обеспечиваю более эффективное и качественное решение задач обнаружения конфиденциальной информации в электронных документах, за счет применения технологии баз знаний, где учитывается предметная область анализируемой информации. Отличительной особенностью разработанного алгоритма обнаружения сведений конфиденциального характера является использование быстрого доступа к хэшированным онтографам понятий и параллельное выполнение правил базы знаний, позволяющего повысить показатели полноты и точности обнаружения. Областью применения разработанного математического обеспечения являются средства защиты информации, предназначенные для выявления передачи информации содержащей сведения конфиденциального характера в электронных документах за пределы защищаемой информационной системы с нарушением политики безопасности

Ключевые слова: МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ, СВЕДЕНИЯ

UDC 004.89

Technical sciences

MATHEMATICAL SOFTWARE FOR DETECTING CONFIDENTIAL DATA IN ELECTRONIC DOCUMENTS

Ptitsin Andrey Aleksandrovich
RSCI SPIN-code = 7006-1394

Krasnodar higher military school, Krasnodar, Russia

In the article, we develop the software for process of confidential data detection based on the knowledge bases technology. The algorithm of detection of confidential data in the electronic documents transferred outbound of protected information system, due to application of linguistic technologies of the thorough text analysis is presented. The estimation of computing complexity of the developed algorithm is made up. The developed data structures and algorithm are realized in the programming language C++. Results of the experiments, confirmed workability of the developed algorithm are presented. The performed experimental researches of the developed algorithm have shown its practical applicability. The estimation of quality of confidential data detection is made up. The obtained estimations of detection quality have shown, that the developed structures of data and algorithm provide more effective and qualitative solution of problems of the confidential information detection in electronic documents, at due to application of knowledge bases technology where the subject domain of the analyzed information is considered. Distinctive feature of the developed algorithm of confidential data detection is the use of rapid access to hashed concept ontographs simultaneous implementation of knowledge base rules, which allows raising indicators of completeness and accuracy of detection. A scope of application of the developed software is the protection frames of the information intended for revealing of an information transfer containing data of confidential character in electronic documents outbound of protected information system with violation of security policy

Keywords: SOFTWARE, CONFIDENTIAL DATA, DLP, DATA LEAK PREVENTION, ELECTRONIC

КОНФИДЕНЦИАЛЬНОГО ХАРАКТЕРА, DLP, DOCUMENT
ЗАЩИТА ИНФОРМАЦИИ ОТ УТЕЧЕК,
ЭЛЕКТРОННЫЙ ДОКУМЕНТ

Введение

Одним из важнейших направлений обеспечения безопасности информации любой организации является решение широкого круга вопросов, связанного с предотвращением нарушения установленного порядка обработки и передачи защищаемой информации по информационно-телекоммуникационной сети (ИТКС). Данная проблема в ИТКС носит комплексный характер и требует разработки и совершенствование способов, методов и средств защиты информации [1].

Известно [2], что одним из направлений развития средств защиты информации (СЗИ), предназначенных для предотвращения утечек защищаемой информации, в настоящее время являются DLP – системы, использующие технологии анализа данных, пересекающих периметр защищаемой информационной системы организации, на основе признаков сообщения и анализа контента.

Однако как показывает анализ результатов выполненных исследований в данной области, сложность практической реализации технологий автоматического анализа текстов, отсутствие моделей описания предметной области в данных СЗИ на уровне семантики, существенно влияют на показатели качества обнаружения сведений конфиденциального характера в электронных документах [3].

Таким образом, в рассматриваемой предметной области существует объективное противоречие, связанное с одной стороны требованиями нормативных правовых актов в области информационной безопасности по предотвращению утечки сведений конфиденциального характера, а с другой стороны необходимостью совершенствования математического обеспечения процесса обнаружения сведений конфиденциального характера, в вышеописанных СЗИ.

Целью настоящего исследования является разработка математического обеспечения процесса обнаружения сведений конфиденциального характера, с использованием технологии баз знаний.

Функциональная структура базы знаний интеллектуальной системы

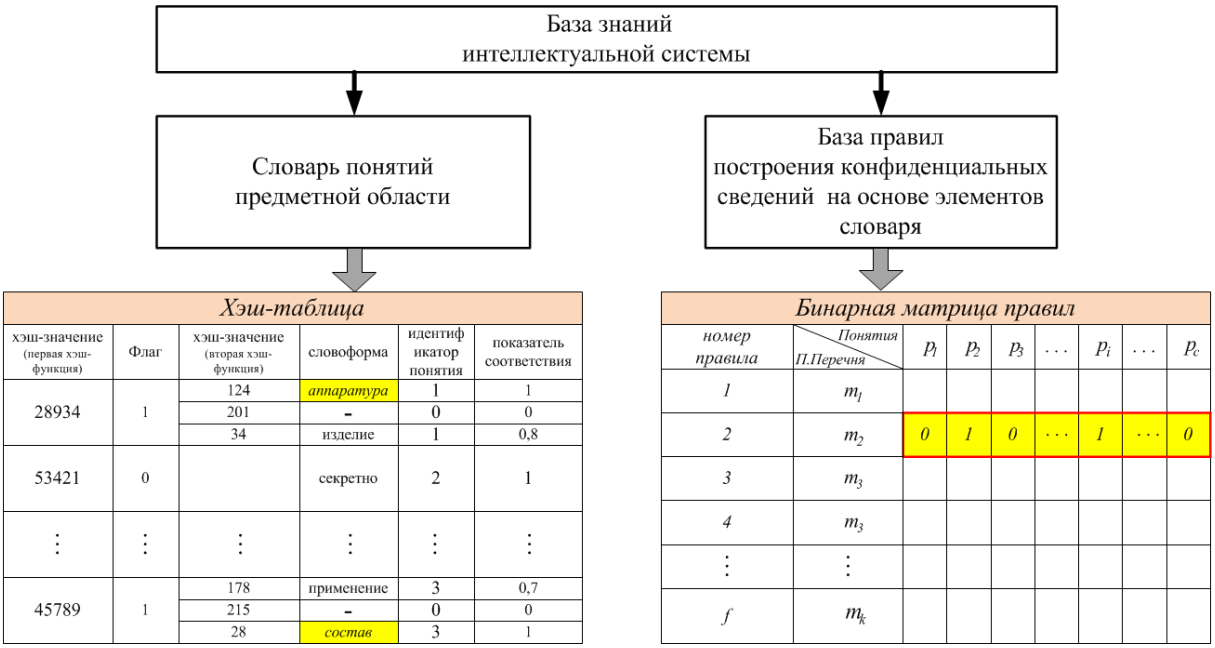
Для практической реализации разработанного математического обеспечения обнаружения сведений конфиденциального характера, предложена разработка интеллектуальной системы (ИС).

Под ИС в настоящем исследовании понимается технология обнаружения сведений конфиденциального характера в реальном контексте на основе собственной базы знаний (БЗ), содержащей актуальные знания о предметной области (ПрО). БЗ ИС представляет собой формальную систему понятий и связей между ними, описанную в Перечне сведений, конфиденциального характера (далее – Перечень) [4].

Таким образом, прикладной предметной областью (ПрО) разработанной ИС является Перечень. БЗ ИС состоит из двух основных компонентов [5] (рис. 1):

словарь ПрО (ассоциативный массив) – структура данных, реализованная хэш-таблицей;

база правил, реализованная в виде бинарной матрицы M размером $f \times c$.



Хэш-таблица

хэш-значение (первая хэш-функция)	Флаг	хэш-значение (вторая хэш-функция)	словоформа	идентификатор понятия	показатель соответствия
28934	1	124	*аппаратура*	1	1
		201	-	0	0
		34	изделие	1	0,8
53421	0		секретно	2	1
⋮	⋮	⋮	⋮	⋮	⋮
45789	1	178	применение	3	0,7
		215	-	0	0
		28	*состав*	3	1

Бинарная матрица правил

номер правила	*Понятия* / *П.Перечня*	p_1	p_2	p_3	…	p_i	…	p_c
1	m_1							
2	m_2	*0*	*1*	*0*	…	*1*	…	*0*
3	m_3							
4	m_3							
⋮	⋮							
f	m_k							

Рис. 1 Структура БЗ ИС

Данные в хэш-таблице имеют следующие типы:

1) первое ключевое поле – значение первой хэш-функции от словоформы; тип данных беззнаковое целое число;

2) флаг коллизии – булевый тип;

3) второе ключевое поле – значение второй хэш-функции; тип данных беззнаковое целое число;

4) словоформа понятия; тип данных – строковый;

5) идентификатор понятия; тип данных – беззнаковое целое число;

6) показатель соответствия лексемы словаря понятию; тип данных – беззнаковое вещественное число $[0…1]$.

В базе правил каждая строка матрицы M представляет собой бинарную строку наличия понятий словаря в j-м пункте Перечня, где p_i – понятие из словаря, $1 \le j \le c$; c – количество понятий; m_j – пункты Перечня, $1 \le j \le k$; k – общее количество пунктов Перечня; f – количество правил. Каждый пункт Перечня описывается одним или более числом правил.

Алгоритм обнаружения сведений конфиденциального характера

В результате экспериментальных исследований документов содержащих сведения конфиденциального характера было выявлено, что значительная доля сведений выражается понятиями, локализованными в пределах одного предложения, а их количество колеблется от 3 до 4 понятий [6]. Поэтому приняты допущения:

областью поиска сведений конфиденциального характера анализируемого документа является одно предложение;

не используется грамматический анализ текста, так как снижается скорость поиска сведений конфиденциального характера, при этом автоматический грамматический анализ русских текстов неоднозначен.

Общий алгоритм обнаружения сведений конфиденциального характера в электронных документах представлен на рисунке 2:

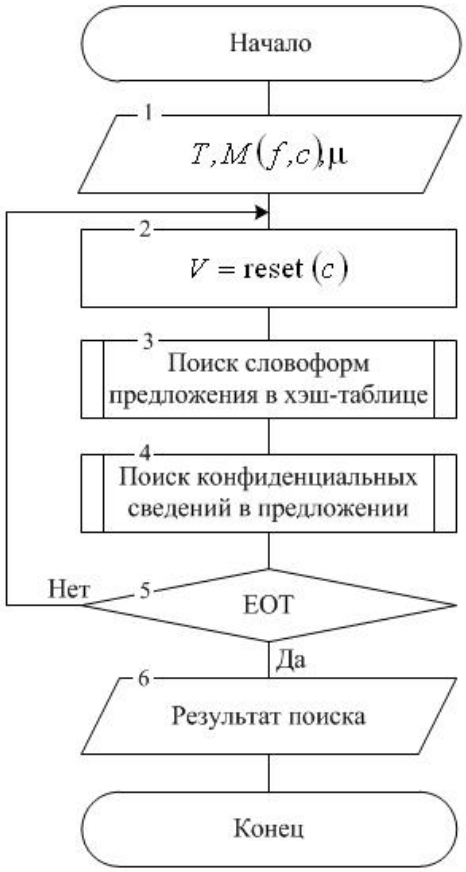


Рис. 2 Блок-схема алгоритма обнаружения сведений конфиденциального характера

Алгоритм обнаружения конфиденциальных сведений состоит из следующих шагов:

1. сброса битов бинарной строки предложения;

2. поиск словоформ предложения в хэш-таблице;

3. поиск конфиденциальных сведений в предложении;

4. повторять пока не достигнут конец текста документа.

Рассмотрим каждый из шагов работы алгоритма.

На вход ИС поступает электронное сообщение (документ), текст T которого подлежит анализу. В свою очередь текст T разбивается на предложения, которые представляются в виде бинарных строк $V = v_1, v_2, \ldots, v_i, \ldots, v_c$, при $v_i = 0$ искомое понятие p_i отсутствует в предложении, при $v_i = 1$ – присутствует.

Таким образом, входными данными алгоритма являются:

T – текст, проверяемого электронного сообщения (документа);

$M(f, c)$ – бинарная матрица правил, где f – количество правил, c – количество понятий;

μ – количество потоков параллельных вычислений в зависимости от архитектуры ЭВМ.

На первом шаге осуществляется выполнение сброса битов бинарной строки предложения V, то есть инициализация ее нулевыми значениями. Далее алгоритм состоит из двух предопределенных процессов, работа которых, будет описана ниже.

Поиск словоформ предложения в хэш-таблице

Программа последовательно считывает словоформы d_i из проверяемого предложения текста, до признака конца предложения (сепараторы «.», «!», «?»). Далее производится вычисление первой хэш-функции $H_1 = h_1(d_i)$ и проверяется актуальность второй хэш-функции $F(H_1)$, то есть признака возникновения коллизии. Если $F(H_1) = 0$, то

считывается кортеж $(s_i, N, R_i) = \text{read}(H_1)$, где s_i – словоформа из хэш-таблицы, N – идентификатор словоформы, R_i – показатель соответствия; если $F(H_1) = 1$, то вычисляется значение второй хэш-функции $H_2 = h_2(d_i)$. При $N = 0$ просматривается следующая словоформа из предложения, если же $N \neq 0$, то считается, что словоформа предложения находится в хэш-таблице.

Блок-схема алгоритма поиска словоформ предложения в хэш-таблице представлена на рисунке 3:

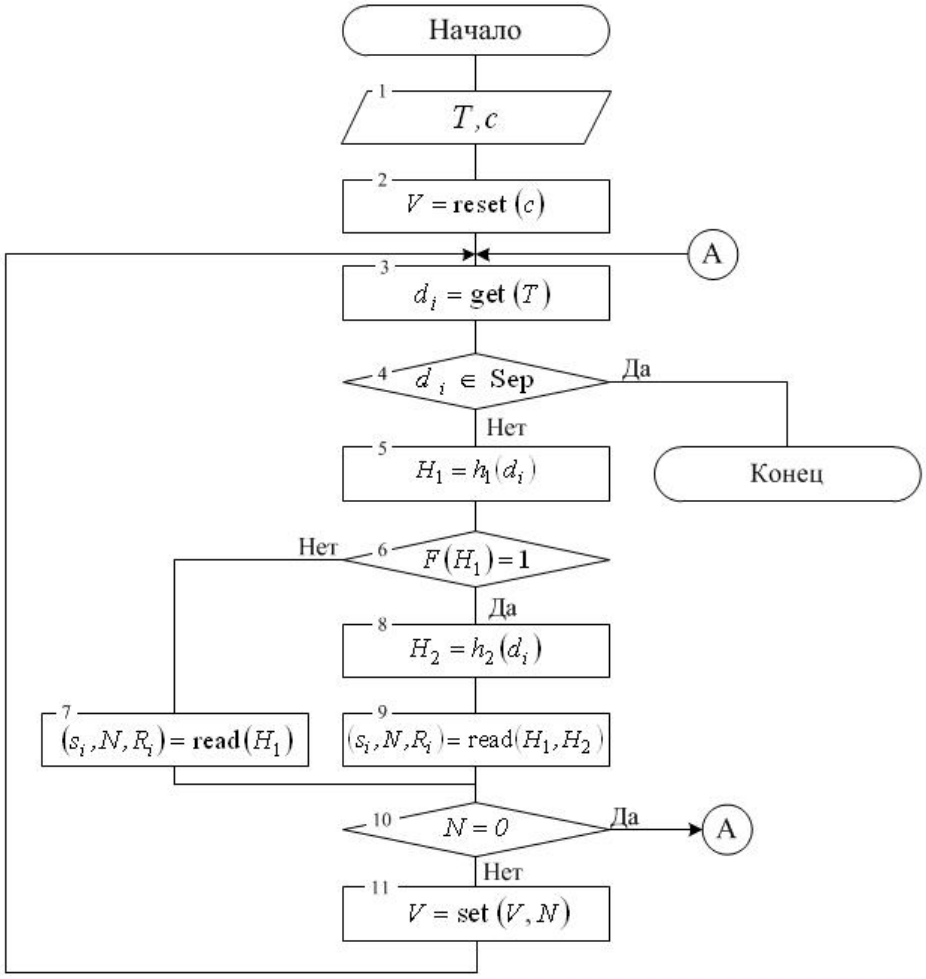


Рис. 3 Блок-схема алгоритма поиска словоформ предложения в хэш-таблице

Поиск конфиденциальных сведений в предложении

На первом шаге осуществляется вычисление количества строк бинарной матрицы $h_{\max} = \left\lceil \frac{c}{\mu} \right\rceil$, которые должен обработать каждый из μ потоков. Далее производится распараллеливание потоков в зависимости от архитектуры ЭВМ и вычисление выражения $q_\mu = V_i \,\&\, B_{h\mu+\mu} \oplus B_{h\mu+\mu}$, для каждой строки бинарной матрицы. При $q_\mu = 0$ – признак j-го пункта Перечня в j-м предложении полагается найденным.

На третьем шаге с целью минимизации ложных срабатываний вычисляется показатель уверенности правила θ для c понятий:

$$\theta = \sum_{i=1}^{C_c^1} R_i - \sum_{i<j}^{C_c^2} R_i R_j + \sum_{i<j<k}^{C_c^3} R_i R_j R_k - \ldots + (-1)^{c-1} \sum_{i=1}^{C_c^c} \prod^{c} R_i ,$$

где

R_i – показатель соответствия словоформы понятию, который задается экспертно при формализации Перечня в диапазоне значений $0 < R_i \le 1$.

Если $\theta \ge \theta_0$, то правило выполнено успешно, и мы получаем доказательство наличия сведений конфиденциального характера в тексте, где θ_0 – порог уверенности заключения правила, заданный экспертно в диапазоне значений $0 < \theta_0 \le 1$. Блок-схема алгоритма поиска конфиденциальных сведений в предложении представлена на рисунке 4:

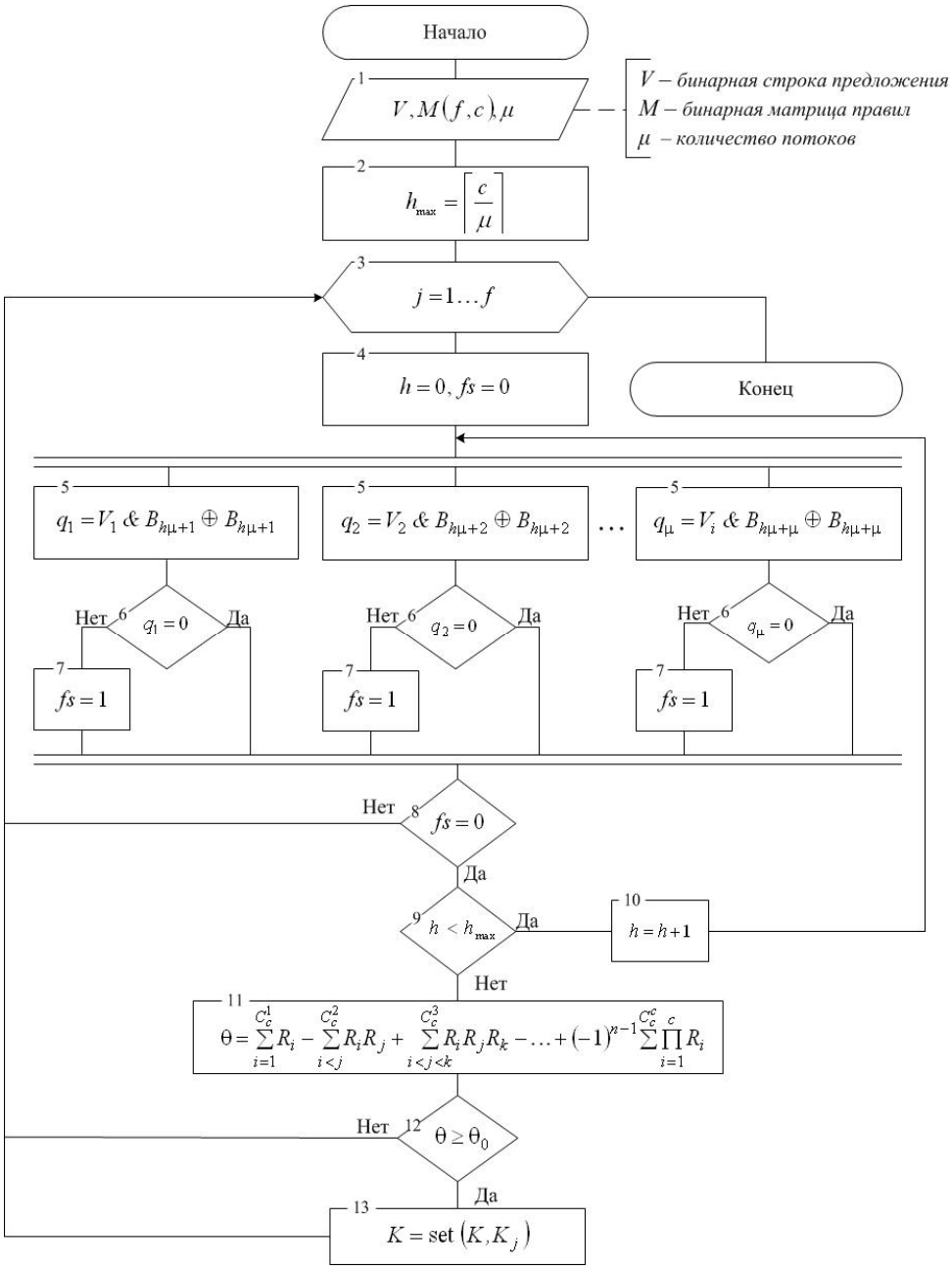


Рис 4. Блок-схема алгоритма поиска конфиденциальных сведений в предложении

Оценка вычислительной сложности алгоритма обнаружения сведений, конфиденциального характера

Для оценки вычислительной сложности алгоритма введем следующие параметры:

η – количество слов в предложении;

f – количество правил в базе знаний;

u – количество предложений в тексте.

Первым шагом работы алгоритма является поиск словоформ одного предложения проверяемого текста в хэш-таблице. Вычислительная сложность первого шага O_1 оценивается следующим образом:

$$O_1(\eta) = \text{Const}_{\text{хэш}} \times \eta \quad (1)$$

На втором шаге осуществляется поиск предложений в БЗ ИС посредством битовых операций над строками матрицы. Вычислительная сложность второго шага O_2 оценивается следующим образом:

$$O_2(\eta, f) = \text{Const}_{\text{хэш}} \times \eta \times f \quad (2)$$

Таким образом, итоговая оценка вычислительной сложности алгоритма обнаружения сведений, конфиденциального характера во всем тексте электронного документа оценивается следующим образом:

$$O(\eta, f, u) = \text{Const}_{\text{хэш}} \times \eta \times f \times u \quad (3)$$

Оценка показателей качества обнаружения сведений конфиденциального характера

Методы и алгоритмы, применяемые в DLP-системах, идентичны специальному математическому обеспечению, используемому информационно-поисковыми системами. Учитывая связь подобных систем с информационным поиском, для оценки показателей качества обнаружения сведений конфиденциального характера, целесообразно использовать показатели полноты $R_{\text{полн}}$ и точности $P_{\text{точн}}$ [7].

Значение показателя полноты определяется формулой:

$$R_{\text{полн}} = \frac{tp}{tp + fn}, \quad (4)$$

где

tp – количество обнаруженных системой документов, содержащих конфиденциальные сведения;

fn – количество не обнаруженных системой документов, содержащих сведения конфиденциального характера.

При $R_{\text{полн}} = 1$ – все документы обнаруженные системой содержат сведения конфиденциального характера, то есть «ошибки второго рода отсутствуют». Значения показателя полноты находятся в диапазоне значений $R_{\text{полн}} \in [0\ldots1]$.

Значение показателя точности определяется формулой:

$$P_{\text{точн}} = \frac{tp}{tp + fp}, \quad (5)$$

где

tp – количество обнаруженных системой документов, содержащих конфиденциальные сведения;

fp – количество не конфиденциальных документов обнаруженных системой и идентифицированных как сведения конфиденциального характера.

При $P_{\text{точн}} = 1$ – все документы обнаруженные системой содержат сведения конфиденциального характера, то есть «ошибки первого рода» отсутствуют. Значения показателя точности находятся в диапазоне значений $P_{\text{точн}} \in [0\ldots1]$.

В качестве результирующего показателя оценки алгоритма, позволяющего найти баланс между показателями полноты и точности, была выбрана F-мера (мера Ван Ризбергена) [7]:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (6)$$

Здесь $\beta^2 = \frac{1-\alpha}{\alpha}$, $\alpha \in [0,1]$, где $\beta^2 \in [0,\infty]$. Для оценки качества выбраны одинаковые веса показателей полноты и точности посредством

установки параметров $\alpha = 1/2$, или $\beta = 1$. При этом выражение (6) упрощается:

$$F_1 = \frac{2PR}{P+R}, \quad (7)$$

Для оценки качества обнаружения сведений конфиденциального характера в электронных документах разработанного математического обеспечения были реализованы две серии тестов. С целью сравнительной оценки полученных результатов был сгенерирован словарь ключевых слов и словосочетаний, использующийся в работе вышеописанных СЗИ.

В качестве базы для проведения эксперимента по оценке качества обнаружения сведений конфиденциального характера с применением разработанного математического обеспечения и сгенерированного словаря использовалась подобранная коллекция документов:

документы, содержащие сведения конфиденциального характера (300 документов);

документы, не содержащие сведений конфиденциального характера (200 документов).

Первый тест состоял в оценке полноты обнаружения сведений конфиденциального характера. Данный показатель характеризует способность СЗИ обнаруживать все релевантные документы. При его выполнении были выбраны 250 документов содержащих сведения конфиденциального характера. Результаты тестирования представлены на графике (рис. 5):

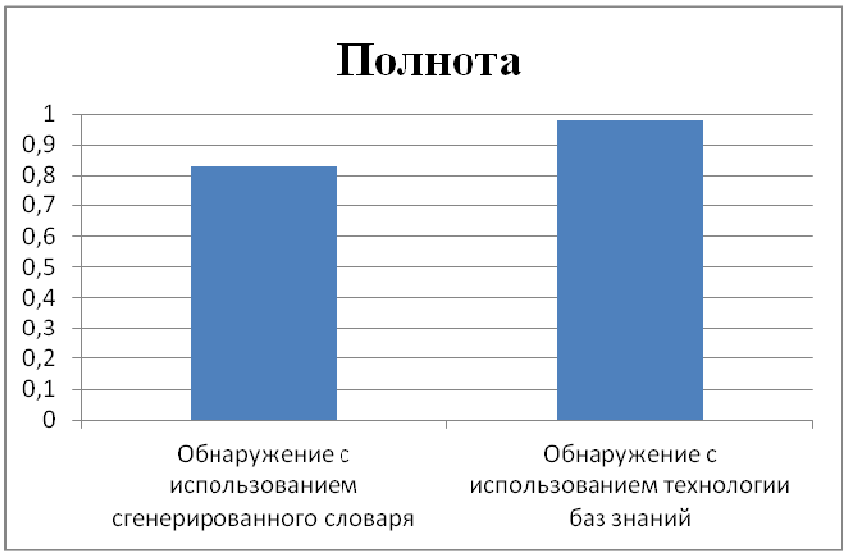


Рис. 5 Результаты оценки показателей полноты

Второй тест эксперимента состоял в оценке показателя точности обнаружения. Данный показатель характеризует способность СЗИ отсеивать нерелевантные документы, то есть показывает количество ложных срабатываний. Для проведения теста были отобраны 300 документов, содержащих сведения конфиденциального характера и 200 документов, не содержащих конфиденциальную информацию. Результаты тестирования представлены на графике (рис. 6)

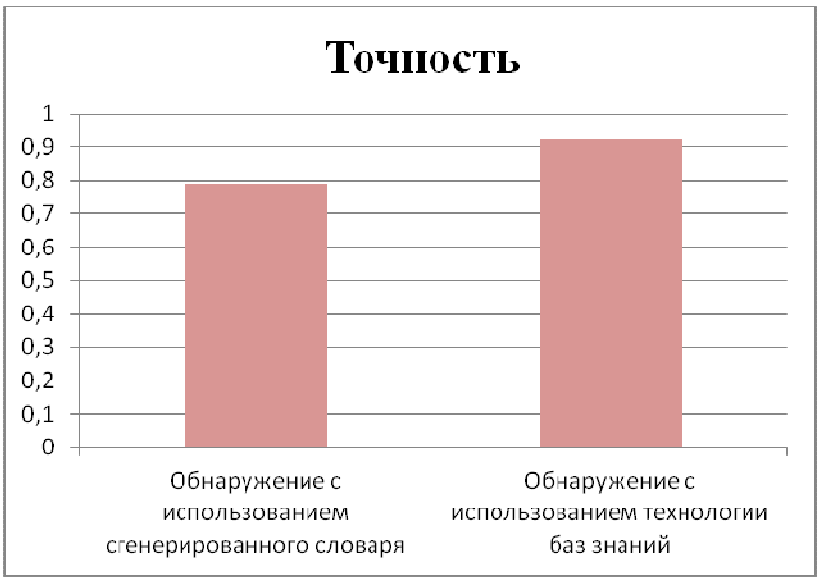


Рис. 6 Результаты оценки показателей точности

Таким образом, эксперимент показал, что при использовании разработанного математического обеспечения с использованием

технологии баз знаний полнота обнаружения возросла на 15%, а точность на 17%.

Выводы

В настоящем исследовании был представлен алгоритма обнаружения сведений конфиденциального характера использующий быстрый доступ к хэшированным онтографам понятий и параллельное выполнение правил обнаружения сведений конфиденциального характера, позволяющий повысить качество обнаружения по показателям полноты и точности. Произведена оценка вычислительной сложности разработанного алгоритма. Проведен эксперимент по тестированию интеллектуальной системы.

Список литературы:

1. Доктрина информационной безопасности Российской Федерации от 09.09.2000 № Пр-1895. — М. : 2000.

2. Сравнение систем защиты от утечек (DLP) 2014 – часть 2 [Электронный ресурс] / Режим доступа: http://www.anti-malware.ru/comparisons/data_leak_protection_2014_part2#part34.

3. Птицын, А.А. Анализ средств, предотвращения утечки защищаемой информации / А.А. Птицын, И.В. Савельев // Информационная безопасность – актуальные проблемы современности. Совершенствование образовательных технологий подготовки специалистов в области информационной безопасности: Сб.трудов VIII – IX Всерос. НТК, г. Геленджик-Краснодар: ФВАС, 2014. — С.216-221.

4. Указ Президента РФ от 06.03.1997 № 188 (с изм. и доп., вступившими в силу с 23.09.2005) «Об утверждении перечня сведений конфиденциального характера» // НПП ГАРАНТ —2014.

5. Птицын, А.А. Разработка интеллектуальной системы предотвращения утечки защищаемой информации с использованием технологий баз знаний / А.А. Птицын // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2015. – №04(108). – IDA [article ID]:1081504042. – Режим доступа: http://ej.kubagro.ru/2015/04/pdf/42.pdf, 0,688 у.п.л.

6. Птицын, А.А. Обоснование структуры правил базы знаний интеллектуальной системы / А.А. Птицын, И.В. Савельев // Информационная безопасность – актуальные проблемы современности. Совершенствование образовательных технологий подготовки специалистов в области информационной безопасности: Сб.трудов X – XI Всерос. НТК, г. Геленджик-Краснодар: КВВУ, 2015. — С.145-149.

7. Кристофер, Д. Маннинг Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. – пер. с англ. – М.: ООО «И.Д. Вильямс», 2011. – 528 с. Ил.

References

1. Doktrina informacionnoj bezopasnosti Rossijskoj Federacii ot 09.09.2000 № Pr-1895. — M. : 2000.

2. Sravnenie sistem zashhity ot utechek (DLP) 2014 – chast' 2 [Jelektronnyj resurs] / Rezhim dostupa: http://www.anti-malware.ru/comparisons/data_leak_protection_2014_part2#part34.

3. Pticyn, A.A. Analiz sredstv, predotvrashhenija utechki zashhishhaemoj informacii / A.A. Pticyn, I.V. Savel'ev // Informacionnaja bezopasnost' – aktual'nye problemy sovremennosti. Sovershenstvovanie obrazovatel'nyh tehnologij podgotovki specialistov v oblasti informacionnoj bezopasnosti: Sb.trudov VIII – IX Vseros. NTK, g. Gelendzhik-Krasnodar: FVAS, 2014. — S.216-221.

4. Ukaz Prezidenta RF ot 06.03.1997 № 188 (s izm. i dop., vstupivshimi v silu s 23.09.2005) «Ob utverzhdenii perechnja svedenij konfidencial'nogo haraktera» // NPP GARANT —2014.

5. Pticyn, A.A. Razrabotka intellektual'noj sistemy predotvrashhenija utechki zashhishhaemoj informacii s ispol'zovaniem tehnologij baz znanij / A.A. Pticyn // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2015. – №04(108). – IDA [article ID]:1081504042. – Rezhim dostupa: http://ej.kubagro.ru/2015/04/pdf/42.pdf, 0,688 u.p.l.

6. Pticyn, A.A. Obosnovanie struktury pravil bazy znanij intellektual'noj sistemy / A.A. Pticyn, I.V. Savel'ev // Informacionnaja bezopasnost' – aktual'nye problemy sovremennosti. Sovershenstvovanie obrazovatel'nyh tehnologij podgotovki specialistov v oblasti informacionnoj bezopasnosti: Sb.trudov X – XI Vseros. NTK, g. Gelendzhik-Krasnodar: KVVU, 2015. — S.145-149.

7. Kristofer, D. Manning Vvedenie v informacionnyj poisk / Kristofer D. Manning, Prabhakar Raghavan, Hajnrih Shjutce. – per. s angl. – M.: OOO «I.D. Vil'jams», 2011. – 528 s. Il.