УДК 004.89

05.00.00 Технические науки

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПРЕДОТВРАЩЕНИЯ УТЕЧКИ ЗАЩИЩАЕМОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ БАЗ ЗНАНИЙ

Птицын Андрей Александрович
SPIN-код = 7006-1394

Филиал Военной академии связи (г. Краснодар), Краснодар, Россия

В статье рассматривается задача повышения эффективности контроля и предотвращения утечки информации конфиденциального характера циркулирующей как внутри организации, так и при взаимодействии с информационными сетями общего пользования посредством разработки интеллектуальной системы прагматической идентификации конфиденциальной информации. Основной целью разработки данной системы является своевременное выявление, предотвращение и локализация (минимизация) угроз, связанных с нарушением установленного порядка обработки, хранения и передачи сведений конфиденциального характера. Представлен подход к организации базы знаний интеллектуальной системы прагматической идентификации информации конфиденциального характера, использующей две модели представления знаний. Определена предметная область, на решение задач, которой ориентирована интеллектуальная система. Описана онтологическая модель представления знаний, использованная для формального представления понятий предметной области в базе знаний. Выбран показатель точности (соответствия) передачи смысла лексемы смыслу понятия предметной области. Описан показатель уверенности каждого правила из базы правил, который характеризует меру правдоподобия того или иного заключения при выполнении правила. Приводится общая структура базы правил, условия выполнения каждого правила логического вывода. Так же предлагается для ускорения работы интеллектуальной системы применить параллельный поиск правил в базе знаний. В заключении описан алгоритм параллельного поиска

Ключевые слова: ЗАЩИТА ИНФОРМАЦИИ, ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА, ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ, КОНФИДЕНЦИАЛЬНАЯ ИНФОРМАЦИЯ, БАЗА ЗНАНИЙ, ЛОГИЧЕСКИЙ ВЫВОД

UDC 004.89

Technical sciences

DEVELOPMENT OF INTELLECTUAL SYSTEM PREVENTING LEAKAGE OF PROTECTED INFORMATION USING KNOWLEDGE BASES

Ptitsin Andrey Aleksandrovich
SPIN-code = 7006-1394

Branch of the Military Academy of connection (Krasnodar), Krasnodar, Russia

The article deals with the problem of increasing the efficiency of the control and prevention of leakage of confidential information circulating within the organization, and in cooperation with the information networks of general use through the development of intelligent system of pragmatic identification of confidential information. The main purpose of this system is to develop early detection, prevention and localization (minimization) of the risks associated with violation of the order of processing, storage and transmission of confidential information. The article presents an approach to the organization of the knowledge base of intelligent system of pragmatic identification of information of a confidential nature, using two models of knowledge representation. We have defined a subject field to meet the challenges of which the intelligent system was meant for. The article describes the ontological model of knowledge representation used for formal representation of domain concepts in the knowledge base. It presents an indicator of exactness of meaning reference to a concept of domain meaning. The article describes a confidence index for each rule from the rule base that characterizes the measure of the likelihood of a conclusion when a rule implemented. It presents a general structure of the rules base, conditions of execution of each rule of conclusion. To accelerate the work of intelligent system a parallel search for the rules in the knowledge base is to be applied. In conclusion, the algorithm of parallel search is described

Keywords: PROTECTION INFORMATION, INTELLECTUAL SYSTEM, ONTOLOGICAL MODEL, CONFIDENTIAL INFORMATION, KNOWLEDGE BASE, LOGIC CONCLUSION

1. Введение

Интеграция информационных ресурсов, содержащих сведения конфиденциального характера, в единое информационное пространство неизбежно ведет к увеличению угроз безопасности информации и к усложнению систем защиты информации. Согласно [1,2], угрозами безопасности информационных и телекоммуникационных средств и систем является нарушение установленного порядка обработки, хранения и передачи информации конфиденциального характера. В результате утечки данной информации наносится значительный ущерб не только собственнику информации, но и государству в целом.

В связи с этим всё более актуальной становится задача разработки интеллектуальных систем (ИС) защиты информации, предназначенных для контроля и предотвращения утечки защищаемой информации.

Для повышения эффективности контроля и предотвращения утечки информации конфиденциального характера за пределы организации предлагается разработка ИС прагматической идентификации конфиденциальной информации. Цель разработки такой системы:

применение ИС для содержательного анализа циркулирующей информации как внутри организации, так и при взаимодействии с информационными сетями общего пользования на предмет конфиденциальности;

повышение эффективности мониторинга информации, передаваемой за пределы защищаемой информационной системы, за счет применения лингвистических технологий глубокого анализа текста.

Под прагматической идентификацией информации будем понимать идентификацию конфиденциальной информации в реальном контексте на основе собственной базы знаний (БЗ), содержащей актуальные знания о предметной области (ПрО).

Важнейшим элементом ИС является формализация сведений конфиденциального характера организации и представление их в БЗ. Таким образом, предметной областью БЗ ИС является Перечень сведений конфиденциального характера, утвержденный Указом Президента Российской Федерации от 06.03.1997 г. № 188[3].

В ИС были использованы две модели [4,5]:

онтологическая модель представления знаний, используется для формального представления понятий предметной области (ПрО);

продукционная модель правил логического вывода.

2. Формальное представление понятий предметной области

Онтологическая модель ПрО представляется в виде сетевой структуры, в которой семантика каждого понятия определяется через его отношения с другими понятиями (структурированный словарь ПрО). Формально записанные знания в онтологической модели составляют семантическую основу БЗ для компьютерного анализа информации.

Под формальной моделью онтологии O понимается упорядоченная тройка вида [4]:

$$O =< X, R, F >, \quad (1)$$

где

X — множество концептов (понятий) ПрО, которую представляет онтология O;

R — множество отношений между концептами (понятиями) ПрО, представленное множеством антецедентов продукционных правил БЗ;

F — конечное множество функций интерпретации (аксиоматизации), представленное продукциями БЗ.

Перечень сведений конфиденциального характера, содержит семантические признаки конфиденциальной информации. Семантические признаки представляют собой описания групп понятий ПрО. Полное

количество разделов указанного Перечня составляют полную совокупность ПрО. Каждая статья раздела описывает группу понятий.

Каждое понятие p_i ПрО представляется в словаре БЗ в виде совокупности лексем L_i, синонимичных понятию. Лексема l_j, наиболее полно передающая смысл понятия, является именем этого понятия. Каждая лексема l_j представляется совокупностью словоформ, составляющих словоизменительную парадигму лексем и имеющих разные грамматические значения.

Формализация словарных понятий Перечня сведений конфиденциального характера осуществляется специалистом по информационной безопасности с привлечением экспертов из отделов и служб организации.

3. Продукционная модель правил логического вывода

Образец правила имеет вид:

$$p_{i_1} \,\&\, p_{i_2} \,\&_{\ldots}\, \&\, p_{i_j} \,\&_{\ldots}\, \&\, p_{i_s} \rightarrow I_r, \qquad (2)$$

где

индекс i_j — номер понятия, $1 \le j \le s$;

s — количество понятий, описывающих I_r;

I_r — информация конфиденциального характера, $1 \le r \le n$;

r — номер пункта Перечня;

n — количество пунктов Перечня.

Антецедент правила $p_{i_1} \,\&\, p_{i_2} \,\&_{\ldots}\, \&\, p_{i_j} \,\&_{\ldots}\, \&\, p_{i_s}$ определяет отношение между понятиями $p_{i_1}, p_{i_2}, \ldots, p_{i_j} \cdot \ldots \cdot p_{i_s}$. Имя этого отношения выражается консеквентом I_r.

Каждая лексема l_j в дереве понятий передаёт смысл этого понятия с различным показателем точности (соответствия) B_i передачи смысла

лексемы смыслу понятия, где $0 < B_i \le 1$, при $B_i = 1$— лексема абсолютно точно передаёт смысл понятия.

В нашем случае показателем точности (соответствия) B_i является семантическое расстояние между лексическими единицами-количественная оценка близости понятий по смыслу.

Анализ документов конфиденциального характера показал, что класс лексических единиц можно интерпретировать, как n- мерное пространство, в котором каждое из значений лексем L_i можно задать в виде точки или вектора. Для пары лексем расстояние определяется через число совпадающих или различающихся семантических признаков в их значениях.

Показатель точности B_i задаётся экспертно.

Показатель уверенности C правила (3) характеризует меру правдоподобия того или иного заключения при выполнении этого правила. Значение показателя уверенности C описывается в виде нормированной суммы:

$$C = B_1 + B_2 - B_1 B_2. \qquad (3)$$

При выполнении правила, содержащего более трёх понятий, показатель уверенности C вычисляется рекурсивно, $0 \le C \le 1$.

Конъюнкцию понятий правила (3) можно представить в виде бинарной строки m_i (рис.1):

P_i / m_i	P_1	P_2	P_3	...	P_s
	0	1	0	...	0

Рис. 1— Бинарная строка конъюнкции понятий

При $p_i = 0$ понятие не входит в описание пункта Перечня m_i, при $p_i = 1$ входит.

Базу правил ИС можно представить в форме бинарной матрицы M размером $m \times n$ (рис.2). Каждый пункт перечня описывается одним или более числом правил.

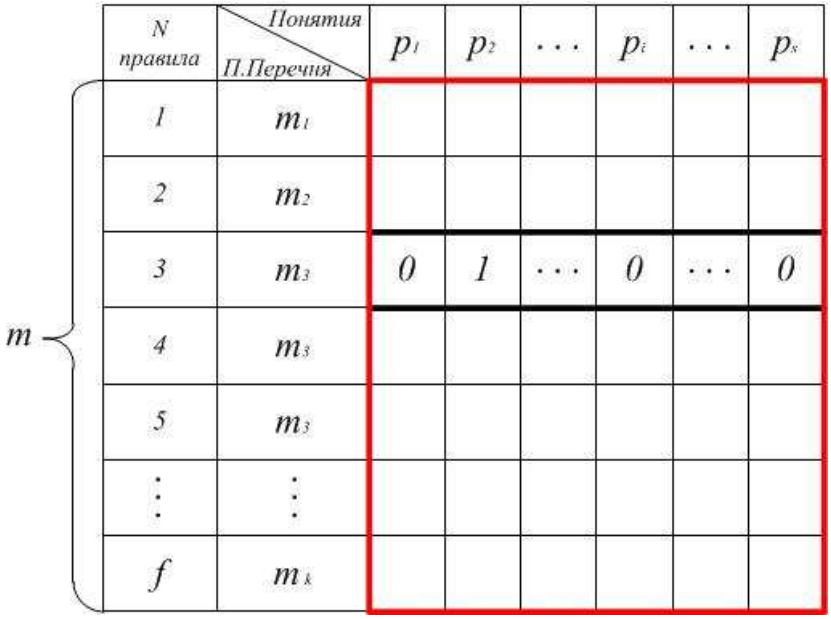

Рис. 2 —Бинарная матрица правил

где

p_i — понятие из базы понятий, $1 \le i \le s$;

s — количество понятий;

m_j — пункты Перечня, $1 \le j \le k$;

j — номер пункта перечня;

k — общее количество Пунктов перечня;

f — номер правила для j-го пункта Перечня.

Результат поиска понятий в предложении анализируемого текста можно представить в виде бинарной строки v_i (рис. 3):

p_i / v_i	p_1	p_2	p_3	...	p_s
	0	1	0	...	0

Рис. 3 - Бинарная строка результата поиска

При $p_i = 0$ понятие не найдено в предложении, а при $p_i = 1$ найдено.

Выполнение одного правила осуществляется путём вычисления выражения:

$$q = v_i \ \& \ m_i \oplus m_i \qquad (4)$$

При поиске во всей БЗ выражение имеет вид:

$$q = v_i \ \& \ M \oplus M, \qquad (5)$$

где

M — бинарная матрица правил БЗ.

Так как текст документа представляется больше чем одним предложением, то выполнение каждого правила в БЗ осуществляется путём вычисления значения выражения столбца Q:

$$Q = V \ \& \ M \oplus M, \qquad (6)$$

где

V — результат поиска понятий в каждом предложении анализируемого текста, представленный матрицей, составленной из бинарных строк v_i.

При $Q = 0$ — признак j-го пункта Перечня в i-ом предложении найден, при $Q \neq 0$ — не найден.

Значение показателя уверенности C, полученное согласно формуле (3), сравнивается с порогом уверенности заключения правил θ, так же заданным экспертно. Показатель уверенности C принимает значение в диапазоне $0 \leq C \leq 1$.

Заключение правила Y может принимать одно из двух значений 0 или 1, которое формируется следующим образом:

$$\begin{cases} C \le \theta, Y = 1, \\ C < \theta, Y = 0, \end{cases} \quad (7)$$

где

θ — порог уверенности заключения правила.

При $Y = 1$ правило выполнено успешно и мы получаем формальное доказательство наличия информации конфиденциального характера соответствующей пункту Перечня m_i, при $Y = 0$ правило не выполнено.

Порог уверенности θ заключения правила задаётся экспертно в диапазоне $0 < \theta < 1$. Каждому правилу порог уверенности задаётся разный в зависимости от часто используемых понятий в документах конфиденциального характера.

В представленном способе организации БЗ большое значение имеет количество операций, которое нужно выполнить для проверки одного правила в БП, так как ИС будет функционировать в режиме приближённом к масштабу реального времени.

Природа задачи позволяет использовать параллельный поиск правил в БП. Алгоритм поиск в БП можно представить в виде следующих шагов:

Исходные данные:

v_i — бинарная строка предложения;

M — бинарная матрица правил;

d — количество потоков.

Шаг 1:

выполняется операция $h_{\max} = \left\lceil \frac{k}{d} \right\rceil$ вычисления количества строк бинарной матрицы, которые должен обработать каждый поток, где k - количество колонок бинарной матрицы;

Шаг 2:

распараллеливание потоков в зависимости от архитектуры ЭВМ и выполнение одного правила путём вычисления выражения $q_d = v_i \,\&\, m_{hd+d} \oplus m_{hd+d}$;

Шаг 3:

при $q_d = 0$—признак j-го пункта Перечня в i-ом предложении найден, при $q_d \neq 0$— не найден, и цикл продолжает поиск правил в БП;

Блок-схема алгоритм параллельного поиска правил в БП представлен на (рис 4):

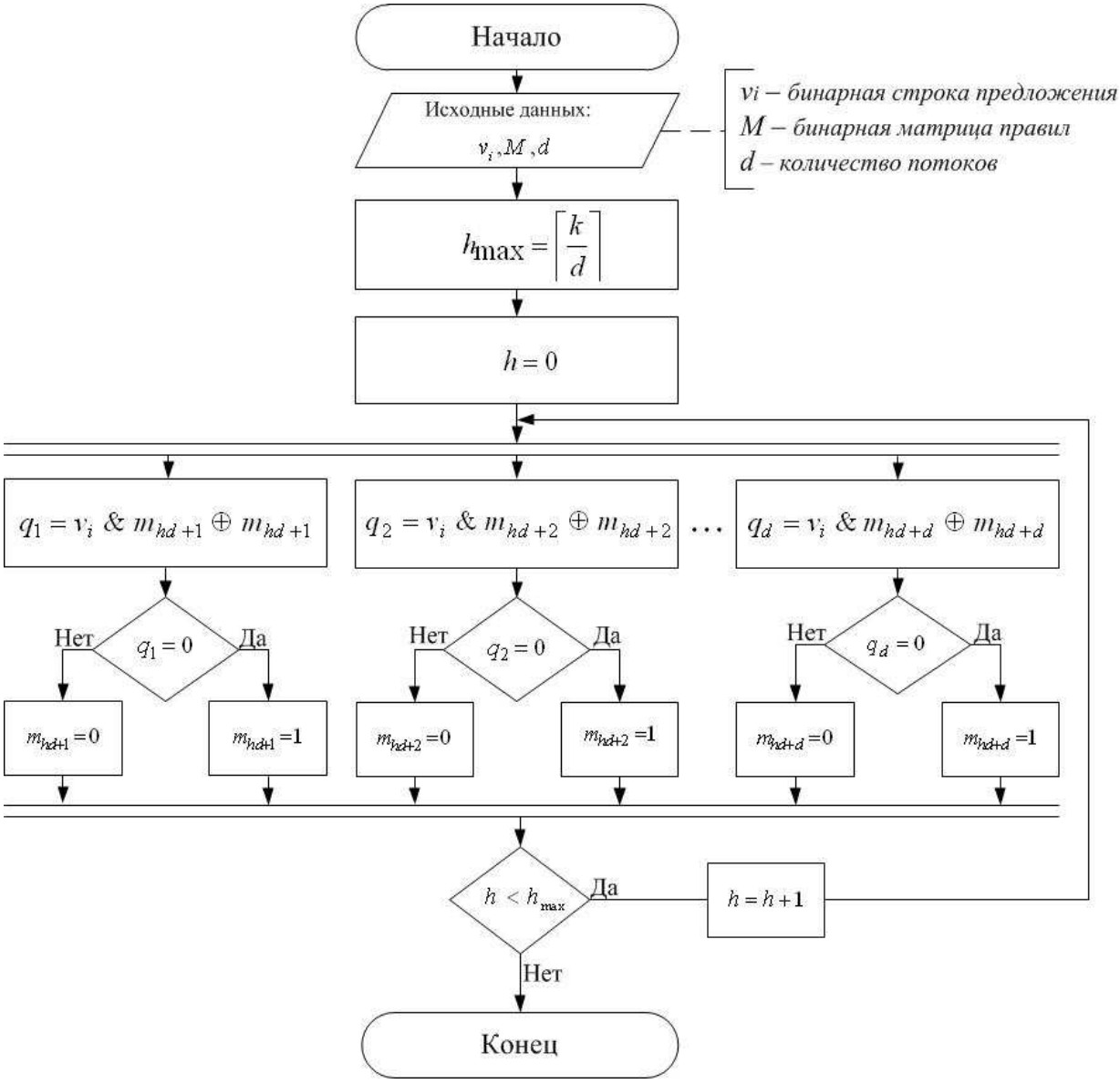

Рис. 4 –Алгоритм распараллеливания поиска правил в БП

Выводы

Таким образом, предлагаемая к разработке ИС прагматической идентификации информации конфиденциального характера позволит повысить эффективность контроля и предотвращения утечки защищаемой информации посредством экспертной формализации словарных понятий разделов Перечня сведений конфиденциального характера организации и представления их в БЗ ИС.

Список литературы:

1. Доктрина информационной безопасности Российской Федерации от 09.09.2000 № Пр-1895. — М. : 2000.
2. Указ Президента РФ № 351 от 17.03.2008 «О мерах по обеспечению информационной безопасности Российской Федерации при использовании информационно-телекоммуникационных сетей международного информационного обмена». — М. : 2008.
3. 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.
4. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учебное пособие.— Москва: Изд-во МГТУ им. Н.Э. Баумана, 2006.—304 с.
5. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем.— Санкт-Петербург: Изд-во Питер, 2001.— 384 с.

References

1. Doktrina informacionnoj bezopasnosti Rossijskoj Federacii ot 09.09.2000 № Pr-1895. — M. : 2000.
2. Ukaz Prezidenta RF № 351 ot 17.03.2008 «O merah po obespecheniju informacionnoj bezopasnosti Rossijskoj Federacii pri ispol'zovanii informacionno-telekommunikacionnyh setej mezhdunarodnogo informacionnogo obmena». — M. : 2008.
3. 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.
4. Bashmakov A.I., Bashmakov I.A. Intellektual'nye informacionnye tehnologii: Uchebnoe posobie.— Moskva: Izd-vo MGTU im. N.Je. Baumana, 2006.—304 s.
5. Gavrilova T.A., Horoshevskij V.F. Bazy znanij intellektual'nyh sistem.—Sankt-Peterburg: Izd-vo Piter, 2001.— 384 s.