DOI 10.36622/1682-7813.2025.28.2.007
УДК 004.457

МЕТОД МОНИТОРИНГА НАЛИЧИЯ СЕМАНТИЧЕСКИХ ПРИЗНАКОВ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ В УЧЕБНЫХ ПРОЕКТАХ

З.Т. Аралбаев, Т.З. Аралбаев, Г.Г. Аралбаева

Цель исследования состоит в снижение риска утечки конфиденциальной информации, содержащейся в проектной документации в вузе на основе мониторинга ее семантических признаков. В работе построена концептуальная модель процесса мониторинга, разработана математическая модель обнаружения семантических признаков и распознавания вида конфиденциальной информации в тексте учебного проекта с использованием двух модификаций метода, разработано программное инструментальное средство мониторинга, проведены вычислительные эксперименты, анализ их результатов, а также определены рекомендации по применению результатов исследований. Предложенный метод позволяет контролировать процесс подготовки проектной документации как руководителям проектов, так и разработчикам на стадиях проектных работ, связанных с конфиденциальной информацией. Использование двух модификаций метода обеспечивает возможность сопоставления их результатов и повышение достоверности принимаемых решений. Программно-аппаратная реализация метода позволяет получить высокую производительность на основе параллельного сопоставления семантических признаков исследуемого проекта разным классам конфиденциальной информации и синхронного вычисления оценок близости текстов проектов. Предложенный метод обеспечивает возможность профилактики предупреждения утечки информации и может быть использован, как учебный материал, при подготовке специалистов по направлению «Информационная безопасность».

Ключевые слова: семантические признаки, конфиденциальная информация, мониторинг, учебные проекты.

Актуальность проведенных исследований и полученных результатов работы определяется необходимостью снижения риска утечки конфиденциальной информации (КИ), содержащейся в отчетной, учебной, научно-исследовательской, проектной документации учащихся в учебной организации. Поскольку в каждом из перечисленных документов содержатся сведения о конкретно поставленных задачах, методах исследований, полученных результатах и рекомендаций по их применению, эти документы объединены в работе под общим названием «учебные проекты». Качество подготовки специалистов во многом определяется умением решать производственные задачи инновационного типа, что неизбежно обуславливает использование и наличие в учебных проектах сведений конфиденциального характера, представленных в регламентирующих документах [1 - 7], полученные учащимися в организациях с различной формой собственности при прохождении различных видов практики, участии в научно-исследовательских, хоздоговорных и грантовских работах образовательной организации.

Актуальность рассматриваемой тематики подчеркивается в Государственном стандарте [2], в котором отмечается, что меры защиты информации должны приниматься при управлении проектом независимо от его типа для идентификации и обработки рисков в рамках проекта.

К отличительным особенностям учебных проектов, как объектов контроля, следует отнести:

- необходимость одновременного обеспечения требуемого уровня оригинальности результатов и исключения сведений инновационного типа для предотвращения утечки информации, составляющей предмет коммерческой тайны;
- сжатые графиком учебного процесса сроки выполнения проектов, что не всегда позволяет опубликовать результаты проектов в открытой печати или провести

исследования по данным из открытых источников информации;

- специфика тематики учебных проектов по направлению подготовки «Информационная безопасность», непосредственно связанная с обработкой сведений конфиденциального типа, а также необходимостью представления проектов в процессе их обсуждения и внедрения.

Следует отметить, что в учебных организациях, как и на любом объекте информатизации, принимаются меры, отраженные в документах по проведению политики информационной безопасности. Однако широкое внедрение в настоящее время технологий распределенного проектирования, новых форм дистанционного обмена и хранения информации, предъявляют повышенные требования к методам и средствам мониторинга информационных потоков с проектной документацией, отличающихся повышенной оперативностью и достоверностью в получении результатов.

Суть рассматриваемой задачи мониторинга заключается в обеспечении надлежащего контроля содержания учебных проектов на всех стадиях их жизненного цикла на основе программных средств выявления семантических признаков конфиденциальной информации и оценки интенсивности их проявления.

Судя по перечню и числу публикаций, этой теме уделяется в настоящее время большое внимание отечественными и зарубежными исследователями.

В частности, авторы работ: [8-10] - отмечают повышении числа прецедентов утечки конфиденциальной информации в вузах в связи с принятием новых технологий дистанционного обучения. В ряде публикаций, связанных по тематике, с исследованиями настоящей работы, представлены результаты, касающиеся определения понятия (сущности) «конфиденциальная информация» и семантических форм ее представления, в частности, в работе [11]. Методам автоматизированного построения тезаурусов любой предметной области и кластеризации текстов на основе задания семантических признаковых пространств (мешков слов) посвящены работы [12] и [13]. Принципы выделения семантических отношений и поиска семантических признаков описаны в статье [14], а результаты определения признаков, описывающих патентные исследования, представлены в работе [15]. В работе [16], приводятся аналитические выражения для численной оценки близости текстов на основе их семантических характеристик. В рассмотренном перечне научной литературы большое внимание уделено решению задач теоретического исследования семантики языка для описания текстов, в частности, с использованием биграмм [17], а также вопросам экспериментального анализа содержания текстов на основе принципов машинного обучения [18]. Особый интерес представляют работы [19-20], выполненные на уровне запатентованных результатов отечественных разработок систем определения текстов, содержащих конфиденциальные сведения, а также статья [21], представляющая перспективы работ по определению контекста содержания документов на основе больших языковых моделей и методов искусственного интеллекта.

Анализ перечисленных работ показал, что несмотря на высокий уровень теоретической проработки вопросов и полученных практических результатов в области информационного поиска и внедрения методов машинного обучения в учебный процесс, в области защиты проектной документации имеются уязвимости организационно-технического типа, нейтрализация которых позволит снизить риск утечки конфиденциальной информации на основе дополнительных мер своевременного ее обнаружения и предупреждения несанкционированного использования.

В качестве дополнения ниши недостающих сведений по рассматриваемой тематике авторами проведены экспериментальные исследования по разработке метода обнаружения наличия конфиденциальной информации в учебных проектах по семантическим признакам ее описания и определению рекомендаций по его использованию.

Цель работы: снижение риска утечки конфиденциальной информации, содержащейся в проектной документации в вузе на основе мониторинга ее семантических признаков.

Для достижения цели в работе решены следующие задачи:

- построена концептуальная модель процесса мониторинга;
- построена математическая модель обнаружения семантических признаков и распознавания вида КИ в тексте проекта;
- разработано программное инструментальное средство мониторинга;
- проведены вычислительные эксперименты и анализ их результатов;
- определены рекомендации по применению результатов исследований.

В основу концепции метода мониторинга положен принцип распознавания класса конфиденциальной информации в исследуемом проекте из множества $\boldsymbol{B}$ на основе меры близости $\boldsymbol{W}$ его текста с классами текстов-эталонов из множества $\boldsymbol{A}$, определенным документом [1]. В зависимости от принципа формирования признакового пространства проектов (ППП) элементов множества $\boldsymbol{A}$ в работе предложены две модификации метода: **M**1 и **M**2.

В соответствии с модификацией **M1** в качестве элементов множества-эталонов ППП для $\boldsymbol{A}$ использовались базы семантических признаков проектов (БПП) с конкретной КИ, причем каждая из БПП была сформирована на множестве текстов проектов с КИ, относящихся к одному классу КИ, например, с коммерческой информацией инновационного характера. В качестве контролируемых семантических признаков использованы уникальные парные сочетания слов (биграммы), характерных для каждого класса КИ. Выбор уникальных биграмм в качестве признаков распознавания обусловлен меньшей вероятностью их появления в тексте, и, как следствие, большей информативностью в пользу распознаваемого класса, а также стремлением сократить объем обрабатываемых данных при реализации метода. При выборе уникальных биграмм учитывались такие свойства уникальных биграмм, как особенность изменения их состава при изменении объема БПП, а также концепция их обновления в соответствии с принципом о неполноте К. Геделя.

Перечень признаков распознавания определялся на основе исследования семантического признакового пространства (ПП) классов образов с КИ и классов без КИ. Эти классы определены на основе полученной статистики проектов, множества семантических признаков КИ и накопленного опыта экспертов по разделению проектов на классы.

Для устранения биграмм, не характерных для БПП с КИ, производилось предварительное отсеивание (фильтрация) этих биграмм из БПП на основе процедуры их выявления и устранения, описанной ниже в выражении (6).

В соответствии с модификацией **M2** использовались локальные семантические признаковые пространства (ЛПП), аналогичные «мешкам слов» [15] для этих же классов КИ, сформированные экспертом на основе априорного опыта работы с проектами. Принципиальным отличием модификаций метода является процедура формирования ППП. В первом случае объем БПП определялся имеющейся базой проектов конкретного класса КИ и техническими возможностями используемой компьютерной техники. Во второй модификации содержимое ЛПП определялось предпочтениями эксперта с учетом специфики задачи мониторинга, при этом объем ЛПП был на два порядка меньше объема БПП.

В основу принципа распознавания класса текста с КИ положена модель, описанная в авторском пособии [22]. Для формализации задачи использованы следующие условные обозначения:

- $\boldsymbol{A = \{A_1, A_2, \dots, A_n, \dots, A_N\}}$ - исходное множество заданных признаковых пространств текстов проектов, используемых в качестве пространств – эталонов, с которыми в процессе распознавания будут сравниваться исследуемые тексты проектов. Каждое пространство из $\boldsymbol{A}$ описывается своим списком признаков $\boldsymbol{S_n} \in \boldsymbol{S}$; $\boldsymbol{n}$=1, $\boldsymbol{N}$.

- $\boldsymbol{B=\{B_1, B_2, \dots, B_m, \dots, B_M\}}$ – множество текстов проектов для мониторинга; каждый проект задается своим списком признаков $\boldsymbol{C_m}$

из множества списков ***C***; ***m***=1, ***M.*** Число элементов в каждом списке – произвольно.

Для определения характеристик семантического пространства исследуемого текста проекта производится поэлементное сравнение семантических характеристик признаков c_k из списка ***C*** с элементами всех списков из множества ***S***. При этом для каждого исследуемого текста формируется множество ***W*** сумм совпадений уникальных элементов списков ***C*** и ***S***, в котором:

$$W_n = \sum_{l=1}^{L} z_{nl}, n = 1, N; \qquad (1)$$

где индекс ***l*** соответствует номеру элемента из списка ***S, L*** – число уникальных биграмм в этом списке.

Математическая модель для определения меры близости списка проекта из ***B*** со списками ***A*** имеет следующий вид:

$$z_{nl}\{c_k, S_n\} = \begin{cases} 1, & \text{если} \quad c_k \in S_n; \\ 0, & \text{если} \quad c_k \notin S_n; \end{cases} \qquad (2)$$

$$M\{S_n, C_m\} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1l} & \dots & z_{1L} \\ z_{21} & z_{22} & \dots & z_{2l} & \dots & z_{2L} \\ z_{n1} & \dots & \dots & z_{nl} & \dots & z_{nL} \\ z_{N1} & z_{N2} & \dots & z_{Nl} & \dots & z_{NL} \end{pmatrix}. \qquad (3)$$

Матрица ***M*** из выражения (3) является матрицей результатов пересечения уникальных элементов списков ***S*** и списка контролируемого текста проекта из ***C***, рассчитанных по выражению (2). Общее число единиц матрицы для конкретного эталонного признакового пространства определяет меру близости его к исследуемому тексту проекта по выражению (1). Проранжированный в порядке убывания ряд значений W_n (n=1, N) определяет порядок убывания степени близости исследуемого текста проекта по заданным признакам ПП. Процедура ранжирования может быть реализована с использованием устройства с повышенной производительностью, представленного в авторской работе [23]. Принятие решения о наличии и характере признаков КИ принимает эксперт, специалист по руководству проектами, по установленным правилам принятия решения, в частности, с использованием пороговых значений W_n, по характеристикам обнаруженных признаков КИ, либо по совокупности полученных результатов.

Качество метода мониторинга КИ определяется оценками вероятности обнаружения семантических признаков КИ и определения класса конфиденциальной информации. Оно обеспечивается выбранной концепцией метода мониторинга, используемыми принципами по формированию семантических признаковых пространств и инструментальными средствами реализации метода.

В процессе экспериментальных исследований рассматривались принципы формирования ППП на основе общей теории кластеризации и распознавания образов. В частности, были исследованы множества текстов проектов для различных типов проектной документации, выделены множества информативных признаков и дана оценка качества распознавания текстов. Предметом исследования являлись проекты, содержащие технические решения инновационного типа (проекты класса ***A2***), сведения по защите персональных данных (проекты класса ***A1***), а также проекты по обработке статистических данных (проекты класса ***A0***), наличие конфиденциальных данных в которых является не обязательным и зависит от условий конкретной задачи. Для построения ПП, расчета и анализа полученных результатов было разработано специальное программное обеспечение в среде программирования языка Питон. Выбор данного языка обусловлен наличием

модулей программ, позволяющих выполнение различных операций с файлами проектов, со списками биграмм. В частности, средства языка позволяют определять списки уникальных биграмм, поиск пересечений по элементам списков, определять частоту встречаемости семантических признаков общего вида в тексте с одновременным их ранжированием, выполнять оформление отчетов, а также обеспечивают удобный интерфейс с различными офисными приложениями компьютера.

В процессе работы были исследованы более 100 различных проектов учащихся бакалавриата и магистратуры по техническим и экономическим направлениям, а также отдельные технические документы, используемые при оформлении отчетов по грантовским разработкам. При этом рассматривались обе модификации метода. Максимальный объем выборок БПП ограничивался техническими характеристиками компьютера (до 15 Мбайт текстовых файлов в формате docx). Процедура определения наличия семантических признаков конфиденциальной информации проводилась с использованием выражения (2). При этом в зависимости от модификации метода текст исследуемого проекта сравнивался последовательно с БПП и ЛПП классов ***A0***, ***A1*** и ***A2***. Решение о наличии признаков и оценок размеров пространства КИ принималось экспертом по максимальной мере близости, рассчитанной с использованием выражения (1).

Параметры оценки величины пересечения ППП для первой и второй модификаций метода, ***IS***1 и ***IS***2, рассчитывались по выражениям (4) и (5).

$$IS1 = \frac{V1}{V3} \cdot 100\%; \qquad (4)$$

$$IS2 = \frac{V2}{V4} \cdot 100\%, \qquad (5)$$

где ***V1*** и ***V2*** числа уникальных биграмм в областях пересечения для первой и второй модификаций метода, соответственно,

V3 и ***V4***, соответственно, числа уникальных биграмм в БПП и ЛПП. Величина ***V***3 из (4) соответствует параметру ***E*** из выражения (6), в котором: ***H*** – мощность множества семантических признаков в БПП с КИ, ***D*** – соответственно, для БПП без КИ.

$$E = H \setminus H \cap D. \qquad (6)$$

Графическая модель для выражения (6) представлена на рис. 1, а.

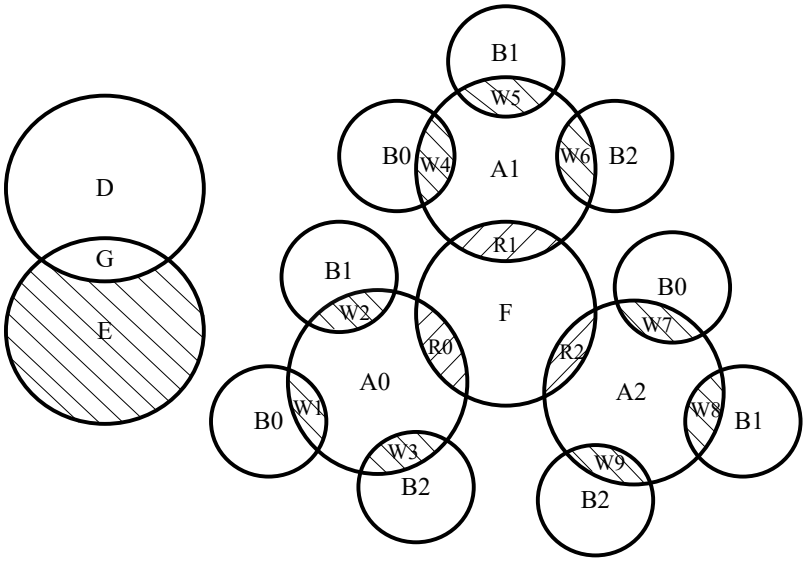

Рис. 1. Графические модели выявления областей пересечения исследуемых семантических признаковых пространств проектов

Знаки наклонного отрезка и объединения в (6) использованы как, соответственно, знаки вычитания и определения пересечения пространств (в частности, пространства ***G*** на рисунке). Величина ***V***4 в выражении (5) соответствует числу биграмм в ЛПП («мешке биграмм»).

На рис. 1, б представлена графовая модель для модификации метода **M**1, описывающая процессы формирования областей пересечения множества ***F***, не содержащего КИ, с множествами: ***A0***, ***A1*** и ***A2*** – при их фильтрации, а также процессы получения областей пересечения множеств ***A0***, ***A1*** и ***A2*** при идентификации множеств исследуемых экспериментальных текстов проектов: ***B0***, ***B1*** и ***B2***. Число текстов проектов класса ***B*** может быть произвольно. В качестве пространств признаков в областях ***D*** и ***F*** в экспериментах были использованы ППП проектов гуманитарного и экономического типов.

Применение данной модели позволяет отсеивать из множества ***H*** неинформативные биграммы, по данным результатов экспериментов, до 10%, что сокращает время дальнейшей обработки биграмм множества ***V3***.

ЛПП для классов: ***A0***, ***A1*** и ***A 2*** – формировались экспертным методом с учетом следующих особенностей и рекомендаций:

- анализ предметной области тематики проектов производился с учетом учебно-методических и законодательных материалов, требуемых компетенций учебных рабочих программ, научной и учебной тематики конкретного учебного подразделения, деятельности организаций-контрагентов, а также результатов разработок, содержащих конфиденциальную информацию по каждому классу проектов;

- перечень биграмм признакового пространства проектов формировался в объеме 100-300 сочетаний пар слов (***V4*** в выражении (5)). Этого объема было достаточно для определения перечня семантических признаков КИ и отнесения исследуемого текста к одному из классов распознаваемых образов. При этом в синтезируемое ППП включались пары слов с использованием, как именительного, так и родительного падежей. При наличии в исследуемых текстах ***n***-грамм с различным значением ***n*** (в частности, при использовании метода «нанизывания падежей») применялся способ разбиения их на последовательности биграмм. Например, для включения в ЛПП текстовой фразы: «угроза удаления информационных ресурсов», перечень биграмм имел следующий вид: «угроза удаления», «удаления информационных», «информационных ресурсов»;

- обеспечивалась способность адаптации ППП под конкретную задачу мониторинга. Это касается возможностей гибко и оперативно изменять состав ППП в зависимости от сведений, более актуальных на данный момент времени. В частности, обеспечивалась возможность ввода изменений в ППП по кодам классификации документации, географическим и почтовым адресам, спискам телефонов, наименованиям организаций, названиям законодательных документов, персональным данным, по характеристикам предметной области исследований, ключевым словам, программным кодам и другим сведениям;

- при выборе множества стоп-слов производился учет значимости их в контексте проектов. В частности, это касалось таких семантических элементов: «**И**», «**ИЛИ**», «**НЕ**» – используемых в описании теорем и аксиом при разработке компьютерной техники;

- для более полного представления результатов мониторинга разрабатывались по несколько модификаций ППП, учитывающих специфику конкретной задачи, например, разработку специальных видов аппаратно-программных средств, видов машиностроительных конструкций, баз данных, компьютерных сетей. При этом, мониторинг проводился последовательно от ППП общего типа, с семантическими признаками типа гиперонимов, к специализированным ЛПП;

- при создании нескольких ЛПП для различных классов проектов проводилась проверка вновь создаваемых ЛПП на близость с существующими ЛПП. Отличие признаковых пространств для надежного распознавания проектов должно быть не менее 50 %. Проверка ЛПП на близость

(сходство) проводилась с использованием разработанного программного обеспечения.

В табл. 1 представлены результаты экспериментов по исследованию представленного метода мониторинга семантических признаковых пространств трех классов текстов проектов: ***A0*** – ***A2***.

Строка табл. 1 «**M1**-***IS 1*** – max» содержит числовые характеристики проверки множеств проектов классов ***A0***, ***A1*** и ***A2*** модификации **M1**, соответствующие максимальным значениям параметра ***IS***, рассчитанным по выражению (4), при проверке текстов ***B0-B2*** с использованием базовых признаковых пространств текстов проектов. Строка таблицы «***M1-IS 1*** – min» содержит результаты, соответствующие минимальным значениям параметра ***IS***.

Аналогичные строки таблицы для модификации **M2** позволяют оценить верхнюю и нижнюю границы диапазона количества обнаруженных семантических признаков.

Существенное отличие по значениям оценок ***IS*** для класса с персональными данными в колонках ***A1*** и ***B1*** объясняется превалирующим объемов уникальных персональных данных над данными других классов КИ.

Таблица 1

Результаты экспериментов

модификация метода (M)			ПРИЗНАКОВЫЕ ПРОСТРАНСТВА ПРОЕКТОВ								
			A0			A1			A2		
M1	IS 1	max	10	1.5	2.8	1.7	20.0	6	1.1	2.1	10
		min	7	1.0	1.0	1.6	18.0	5	0.9	1.4	5
M2	IS 2	max	11	2	2	2.2	35	17	1.0	12	16
		min	3.6	0.5	1.3	0.9	13	11	0.1	1.5	11
оценки пересечения признаковых пространств (IS, %)			B0	B1	B2	B0	B1	B2	B0	B1	B2
			КЛАССЫ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ								

Как показывают результаты экспериментов, предлагаемый метод позволяет определить принадлежность исследуемого текста проекта к одному из классов эталонов согласно разработанной концепции и математической модели.

Например, при идентификации текста с классом КИ ***B0*** для обеих модификаций метода в соответствии с экспериментальными данными получены максимальные значения ***IS*** в колонке ***B0***, соответствующей ***A0***.

При идентификации текста с классом КИ ***B1*** для обеих модификаций метода в соответствии с экспериментальными данными получены максимальные значения ***IS*** в колонке ***B1***, соответствующей ***A1***.

Выбор модификации метода определяется целями, задачами и требованиями к мониторингу семантических признаков учебных проектов. Применение обеих модификаций целесообразно при возникновении неопределенностей в определении класса конфиденциальной информации. Такая ситуация возможна, если проекты содержат семантические признаки из близких по решаемым задачам классов. В данном случае окончательное решение принимается по величинам верхних и нижних уровней ***IS***, либо на основе уточнения содержимого базового и локального признакового пространств.

Перспективное использование метода позволит накопить статистические данные результатов для получения вероятностных оценок распределения погрешностей и уточнения требований к выбору исходных признаковых пространств для решения задачи распознавания классов конфиденциальной информации.

На основании проведенных исследований сделаны выводы и определены следующие рекомендации:

- предложенный метод позволяет контролировать процесс подготовки учебных проектов, как руководителям работ, так и разработчикам документации на основе мониторинга семантических признаков на стадиях проектирования, связанных с конфиденциальной информацией;

- использование двух модификаций метода позволяет сопоставлять результаты и повышать достоверность принимаемых решений;

- аппаратная реализация метода в перспективе позволит получить высокую производительность на основе параллельного сопоставления семантических признаков исследуемого проекта разным классам конфиденциальной информации и синхронного вычисления оценок близости текстов проектов;

- метод также обеспечивает возможность профилактики предупреждения утечки информации из учебных проектов, не связанных с обработкой конфиденциальной информацией с целью исключения случайных прецедентов нарушения политики информационной безопасности в организации на этапе прохождения экспертизы публикационного материала;

- алгоритмическое, программное и информационное обеспечение метода могут быть использованы, как вспомогательный инструмент при аудите выпускных квалификационных работ и, как учебный материал, при подготовке специалистов по направлению «Информационная безопасность» по дисциплинам, связанным с распознаванием образов на основе семантического анализа текстов.

Список литературы

1. Указ Президента Российской Федерации от 06.03.1997 г. № 188. Об утверждении перечня сведений конфиденциального характера. (В редакции указов Президента Российской Федерации от 23.09.2005 № 1111; от 13.07.2015 № 357). pravo.gov.ru.

2. ГОСТ Р ИСО/МЭК 27002-2021. Национальный стандарт Российской Федерации. Информационные технологии. Методы и средства обеспечения безопасности. Свод норм и правил применения мер обеспечения информационной безопасности. (утв. и введен в действие приказом Федерального агентства по техническому регулированию и метрологии от 20 мая 2021 г. N 416-ст).

3. Федеральный закон "О коммерческой тайне" от 29.07.2004 N 98-ФЗ. РФ. Принят Государственной Думой 9.07.2004 г. // sudact.ru/law/federalnyi-zakon-ot-29072004-n-98-fz-o/ (дата обращения 11.05.2025).

4. О персональных данных: Федеральный закон от 27 июля 2006 г. № 152-ФЗ // СПС КонсультантПлюс. URL: https://www.consultant.ru/document/cons_doc_LAW_61798 (дата обращения 24.02.2025).

5. Базовая модель угроз безопасности персональных данных при их обработке в информационных системах персональных данных. ФСТЭК РФ. Утверждена 15.02.2008.

6. ГОСТР 51583- 2014. Защита информации. Порядок создания автоматизированных систем в защищенном исполнении. Общие положения. Введ. 2014–09–01. М.: Стандартинформ, 2014. 15 с.

7. Банк угроз ФСТЭК – Федеральная служба по техническому и экспортному контролю. Государственный научно-исследовательский испытательный институт проблем технической защиты информации // Москва, 2025. URL: https://bdu.fstec.ru/threat (дата обращения 24.02.2025).

8. Нечай А.А. Подходы к выявлению конфиденциальной информации /А.А. Нечай, С.А. Краснов, И.В. Першина// Экономика и социум. 2015 г. №1(14). С.26-31.

9. Alexei Arina, Alexei Anatolie. Cyber Security Threat Analysis In Higher Education Institutions As a Result of Distance Learning. International Journal of scientific & technology research. Volume 10, ISSUE 03, march 2021.

10. Wahab Akanni Adeniyi, Johnson O. Fatokun. The risk of data leakages as data grows in information technology-driven tertiary education institutions. // European Journal of Mathematics and Computer Science. Vol. 5. No. 2, 2018, ISS 2059-995/ -p. 33-40.

11. Бурдов С.Н. Сущность и отличительные признаки конфиденциальной информации / С.Н. Бурдов // Правовое государство: теория и практика. №4(34). 2013. С. 150-157.

12. Пархоменко П.А. Обзор и экспериментальное сравнение методов кластеризации текстов / П.А. Пархоменко, А.А. Григорьев, Н.А. Астраханцев // Труды ИСП РАН, том 29. Вып. 2. 2017. С. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6.

13. Маркелов К.С. Методы автоматизированного синтеза тезаурусов предметных областей на основе анализа текстов / К.С. Маркелов, А.Б. Нейман // Альманах современной науки и образования Тамбов: Грамота, 2012. № 7 (62). С. 81-86.

14. Лагутина Н.С. Методические аспекты выделения семантических отношений для автоматической генерации специализированных тезаурусов и их оценки / Н.С. Лагутина К.В., Э.И. Мамедов, И.В. Парамонов // Моделирование и анализ информационных систем. Т. 23, № 6 (2016). С. 826–840.

15. Информационно-поисковая система SciFinder: учебно-методическое пособие / И.В. Зибарева // Новосибирский государственный университет, 2015. 120 с.

16. Бермудес С.Х.Г. О методе определения текстовой близости, основанном на семантических классах / С.Х.Г. Бермудес, С.У. Керимова //Инженерный вестник Дона. №4 (2016). С.17-29. ivdon.ru/ru/mfgazine/arhive/n4y2016/3832/.

17. Петюшко А.А. Биграммные языки // Автореферат дис. ... кандидата физ.-матем. наук : 01.01.09/А.А. Петюшко / Моск. гос. ун-т им. М.В. Ломоносова. Москва, 2015. 17 с.

18. Марцинкевич В.И. Анализ возможностей парсинга электронных текстовых документов для автоматизации нормоконтроля / В.И. Марцинкевич [и др.] // Научно-технический журнал. 2021, №3. С. 45-58.

19. Пат. 2665915 **C**1. Российская Федерация, МПК G06F 17/27. Система и способ определения текста, содержащего конфиденциальные данные. / Дорогой Д.С. (РФ); Заявка №2017121122 от 16.06.2017. Опубл. 04.09.2018. Б.И. №25.

20. Заявка на изобретение № 2013136905/08 Российская федерация, МПК G06F 17/27. Способ автоматизированного сравнения текстов на естественном языке / Харламов А.А. (РФ); // Опубл. 20.02.2015. Бюл. №5.

21. Луцкович А.И. Система анализа индикаторов компрометации на основе методов искусственного интеллекта / А.И. Луцкович, А.Д. Ахметова, А.М. Вульфин, А.Д. Кириллова // Информационные технологии интеллектуальной поддержки принятия решений (памяти проф. Н.И. Юсуповой) ITIDS'2024. Труды X Международной научной конференции. УУНиТ. Уфа, 2024. С. 148-154.

22. Ассоциативно-мажоритарный подход к решению задач распознавания образов в системах защиты информации: учебно-методическое пособие для обучающихся по образовательным программам высшего образования по направлениям подготовки 09.03.01 Информатика и вычислительная техника и 10.03.01 Информационная безопасность / Т.З. Аралбаев [и др.]; М-во науки и высш. образования Рос. Федерации,

Федер. гос. бюджет. образоват. учреждение высш. образования "Оренбург. гос. ун-т". Оренбург: ОГУ, 2023. ISBN 978-5-7410-3150. 150 с.

23. Пат. 2792182 МПК **С**1. Российская федерация. Устройство для ранжирования чисел. / Т.З. Аралбаев, Г.Г. Аралбаева [и др.]; Заявитель и патентообладатель: Оренбургский государственный университет. Заявка № 2022131995 от 07.12.2022. Опубл. 20.03.2023, Бюл. № 8, 2023. 1 с.

Оренбургский государственный университет
Orenburg State University

Поступила в редакцию 3.03.25

Информация об авторах

Аралбаев Закарья Ташбулатович - руководитель направления, ООО «Газпромнефть-Снабжение». Санкт-Петербург, e-mail: aralbaev.ZT@gazpromneft.ru

Аралбаев Ташбулат Захарович – д-р техн. наук, профессор, Оренбургский государственный университет, e-mail: atz1953@gmail.com

Аралбаева Галия Галаутдиновна – д-р экон. наук, профессор, Оренбургский государственный университет, e-mail: galia55@mail.ru

A METHOD FOR MONITORING THE PRESENCE OF SEMANTIC FEATURES OF CONFIDENTIAL INFORMATION IN EDUCATIONAL PROJECTS

Z.T. Aralbaev, T.Z. Aralbaev, G.G. Aralbaeva

The purpose of the study is to reduce the risk of leakage of confidential information contained in the university's project documentation by monitoring its semantic features. The paper builds a conceptual model of the monitoring process, develops a mathematical model for detecting semantic features and recognizing the type of confidential information in the text of an educational project using two modifications of the method, develops a software monitoring tool, conducts computational experiments, analyzes their results, and defines recommendations for the application of research results. The proposed method allows you to control the process of preparing educational project documentation for both project managers and developers at the stages of project work related to confidential information. Using two modifications of the method allows you to compare their results and increase the reliability of decisions. The hardware implementation of the method allows for high performance based on parallel comparison of semantic features of the studied project to different classes of confidential information and synchronous calculation of estimates of the proximity of project texts. The proposed method provides the possibility of information leakage prevention and can be used as an educational material in the training of specialists in the field of Information Security.

Keywords: semantic features, confidential information, monitoring, educational projects.

Submitted 3.03.25

Information about the authors

Zakarya Tashbulatovich Aralbayev – Head of the department, Gazpromneft-Supply LLC. Saint Petersburg, e-mail: aralbaev.ZT@gazpromneft.ru

Aralbayev Tashbulat Zakharovich – Dr. Sc. (Technical) Sciences, Professor, Orenburg State University, e-mail: atz1953@gmail.com

Aralbayeva Galiya Galautdinovna – Dr. Sc. (Economics), Professor, Orenburg State University, e-mail: galia55@mail.ru