УДК 004.4

Г. Э. Яхъяева, **О. В. Ясинская**, **А. А. Карманова**

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

gul_nara@mail.ru, yasinskaya.olga@gmail.com, anast.karmy.aa@gmail.com

ВЕРОЯТНОСТНАЯ ВОПРОСНО-ОТВЕТНАЯ СИСТЕМА В ОБЛАСТИ КОМПЬЮТЕРНОЙ БЕЗОПАСНОСТИ [*]

Описывается вопросно-ответная система *QA-RiskPanel*, использующая в качестве источника знаний базу данных по компьютерной безопасности. Система *QA-RiskPanel* основана на прецедентном подходе к моделированию предметных областей и позволяет пользователю задавать вероятностные вопросы с целью определения и прогнозирования различных рисков, связанных с компьютерными атаками. Излагаются теоретико-модельные основы, приводится классификация вопросных шаблонов, описывается программная реализация разработанного подхода.

Ключевые слова: информационная безопасность, компьютерная атака, прецедент компьютерной атаки, обобщенная нечеткая модель, обобщенный прецедент, вопрос, суждение.

Введение

С момента создания первого поколения компьютеров зарождаются различные направления исследований в области формального представления знания и методов автоматической обработки этого знания. Эти исследования год от года увеличиваются и к концу 1980-х гг. уже составляют отдельную область информатики, которую сегодня принято называть инженерия знаний. Решающие факторы увеличения интереса к данным исследованиям – беспрецедентный рост доступной цифровой информации и растущее число пользователей, у которых есть доступ ко всей этой информации. В связи с этим возникает необходимость в разработке информационных систем, позволяющих упростить локализацию, извлечение и манипулирование этими огромными объемами данных. Одним из возможных решений в данном направлении являются вопросно-ответные системы.

Исследования в области вопросно-ответных систем ведутся в двух направлениях: информационно-поисковые системы (*information retrieval systems*) и интеллектуальные системы (*intelligence support systems*) [1].

Вопросно-ответные системы, развивающиеся в рамках информационно-поискового подхода, нацелены на поиск ответов на задаваемые вопросы в виде текстовых отрывков, доступных в сети Интернет. Традиционно в данном направлении вопросно-ответные системы делятся на два типа: системы с открытым доменом (*open domain*) и системы с ограниченным доменом (*restricted domain*) [2; 3].

[*] Исследование выполнено при финансовой поддержке Минобрнауки России, задание № 2014/139 на выполнение государственных работ в сфере научной деятельности в рамках базовой части, а также при финансовой поддержке РФФИ в рамках научного проекта № 14-07-00903_а.

Яхъяева Г. Э., *Ясинская О. В.*, *Карманова А. А.* Вероятностная вопросно-ответная система в области компьютерной безопасности // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 3. С. 132–145.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2014. Том 12, выпуск 3

Системы с открытым доменом относятся к *общим* вопросно-ответным системам, которые ориентированы на обработку вопросов в любой предметной области. В свою очередь системы с ограниченным доменом относятся к *специализированным* вопросно-ответным системам и рассчитаны на обработку вопросов узкой тематики, т. е. вопросов по определенной предметной области.

Однако все они фокусируются на разработке методов обработки естественного языка (*natural language processing*) и анализа текстов (*text mining*). Принципиальное отличие заключается лишь в степени проработки онтологии.

Вопросно-ответные системы с открытым доменом используют общие онтологии естественного языка (например, онтологию WordNet). Вследствие этго они вынуждены решать такие проблемы естественного языка, как синонимия, омонимия, полиморфизм и т. п. Как правило, современные вопросно-ответные системы с открытым доменом способны формализовать для последующей обработки лишь некоторые предопределенные классы вопросов – фактографические вопросы. Фактографические вопросы можно разделить на следующие: вопросы о персонах, вопросы о времени, вопросы о географических топонимах, вопросы о списках чего-либо, вопросы об определениях и т. д. [4].

Вопросно-ответные системы с ограниченным доменом используют узкоспециализированные онтологии. Это позволяет, с одной стороны, решать проблему с многозначностью слов естественного языка, а с другой стороны, дает возможность получать ответы на более специфические вопросы в данной предметной области.

Исследования в области искусственного интеллекта привели к созданию вопросно-ответных систем, основанных на знаниях (*knowledge based QA-systems*), которые в качестве источника знаний используют различные базы знаний. Очевидно, что такие вопросно-ответные системы являются системами с ограниченным доменом. Заметим, что эти системы могут давать ответы только в связи с информацией, ранее закодированной в базе знаний, т. е. являются менее гибкими при формировании вопросов.

Однако главное преимущество этого подхода заключается в том, что наличие концептуальной модели предметной области, представленной в структуре базы знаний, позволяет использовать такие передовые технологии обработки структурированной информации, как логический вывод, рассуждения по аналогии и т. п. Это в свою очередь приводит к смещению целей, для которых создаются данные системы. Такие системы в основном нацелены не на поиск и локализацию запрашиваемой информации, а на выявление скрытых закономерностей, анализа критических ситуаций, описание рисков в данной предметной области.

Так, система L&C [5], разработанная в медицинской области, решает задачи интеграции авторитетного медицинского знания с информацией о конкретных пациентах. Вопросно-ответная система Демир-Фушмана [6] основана на использовании статистических методов в клинической медицине. Система WEBCOOP [7] использует процедуры логического вывода для генерации ответов в области туризма.

В данной работе мы описываем вопросно-ответную систему «QA-RiskPanel», разработанную в рамках программного комплекса «RiskPanel» [8] для предметной области информационной безопасности. Данная система основана на прецедентном подходе к моделированию предметных областей и позволяет пользователю задавать вероятностные вопросы с целью определения и прогнозирования различных рисков, связанных с компьютерными атаками. Для описания базы знаний мы рассматриваем конечное множество прецедентов компьютерных атак и исходя из этих прецедентов оцениваем вероятности различных утверждений, имеющих отношение к безопасности корпоративной информационной системы. Также мы будем учитывать, что наши знания о каждом конкретном прецеденте могут быть неполными. Поэтому каждый отдельный прецедент компьютерной атаки будет описываться при помощи модели (названной нами *обобщенным прецедентом*), на которой значениями истинности являются 1 (истинно), 0 (ложно) и [0,1] (неопределенно). Тогда вся база знаний будет моделироваться в виде *обобщенной нечеткой модели*, являющейся произведением обобщенных прецедентов.

Статью начинает краткое теоретико-модельное описание базы знаний, формализованной на основе прецедентного подхода к описанию предметной области. Более подробное описание этого подхода можно найти в работах [9–11].

Очевидно, выбор формализации базы знаний накладывает определенные ограничения на типы вопросов, которые может обрабатывать разрабатываемая вопросно-ответная система. Поэтому мы опишем классификацию вопросных шаблонов данной системы. Классификация проведена в духе эротетической логики, основополагающие идеи которой можно найти в работе [12].

Завершит статью описание программной реализации разработанной вопросно-ответной системы.

Теоретико-модельное описание базы знаний

Рассмотрим предметную область Δ. Разработка вопросно-ответной системы с ограниченным доменом начинается с описания онтологии предметной области Δ. С теоретико-модельной точки зрения описание онтологии заключается в задании сигнатуры предметной области Δ (т. е. описании множества понятий данной предметной области) и задании аналитической теории предметной области Δ (т. е. описании явных и неявных определений данной предметной области) [13].

Итак, пусть у нас имеется сигнатура σ. Мы подразумеваем, что сигнатура σ не содержит функциональных символов. Это обобщение допустимо, так как любой *n*-местный функциональный символ заменяется *n* + 1-местным предикатным символом (см., например, [14]).

Рассмотрим множество $C \subseteq \sigma$ констант сигнатуры σ. Заменим в сигнатуре σ все константы на одноместные предикаты. Это обобщение позволит нам в дальнейшем рассматривать суждение, субъектом которого является константа, как частный случай суждения с предикатным субъектом. Рассмотрим сигнатуру $\sigma' = [\sigma]_{P_c}^{c \in C}$, где P_c – одноместный предикатный символ, не принадлежащий сигнатуре σ.

Обозначим через $\mathbb{K}(\sigma')$ класс моделей сигнатуры σ', отвечающих следующим условиям:

1) для любой модели $\mathfrak{A} \in \mathbb{K}(\sigma')$ имеем $|\mathfrak{A}| = C$;

2) для любого предикатного символа $P_c \in \sigma' \setminus \sigma$ и для любой модели $\mathfrak{A} \in \mathbb{K}(\sigma')$ имеем $\mathfrak{A} \vDash P_c(c)$ и $\mathfrak{A} \nvDash P(d)$ для любого $d \neq c$.

Каждый прецедент предметной области Δ описывается некоторой моделью $\mathfrak{A} \in \mathbb{K}(\sigma')$. Однако заметим, что не каждая модель класса $\mathbb{K}(\sigma')$ является прецедентом предметной области Δ. Знание о том, какие именно модели формализуют прецеденты предметной области Δ и является основополагающим при формировании базы знаний данной предметной области.

Все понятия предметной области Δ можно разделить на две группы: *аналитические* и *эмпирические* [15]. С помощью аналитических понятий описываются все определения данной предметной области, поэтому на всех прецедентах предметной области они должны трактоваться однозначно. Интерпретация эмпирических понятий может меняться на разных прецедентах предметной области.

Так, в предметной области компьютерной безопасности понятие «*вредоносная программа*» является аналитическим понятием и не меняется от прецедента к прецеденту. Понятие «*быть использованным в данной атаке*» – эмпирическое понятие, характеризующее то, чем один прецедент отличается от другого.

Таким образом, сигнатуру σ' предметной области Δ разделим на два непересекающихся множества: σ_a – аналитическая сигнатура и σ_e – эмпирическая сигнатура. Заметим, что $\{P_c | c \in C\} \subseteq \sigma_a$, т. е. все константы являются аналитическими понятиями.

Тогда аналитическая теория $\mathrm{Th}(\Delta)$ предметной области Δ будет описываться дедуктивно-замкнутым подмножеством множества всех предложений $S(\sigma_a)$ аналитической сигнатуры σ_a.

Введем обозначение

$$\mathbb{K}(\mathrm{Th}(\Delta)) \rightleftharpoons \{\mathfrak{A} \in \mathbb{K}(\sigma') \,|\, \mathfrak{A} \models \mathrm{Th}(\Delta)\}.$$

Таким образом, множество $E(\Delta)$ всех прецедентов предметной области Δ является собственным подмножеством класса $\mathbb{K}(\mathrm{Th}(\Delta))$, т. е. $E(\Delta) \subset \mathbb{K}(\mathrm{Th}(\Delta))$.

Определение 1. *Пусть* $E \subseteq E(\Delta)$ *– конечное множество прецедентов предметной области* Δ. ***Нечеткой моделью*** *множества прецедентов* E *назовем алгебраическую систему* $\mathfrak{A}_{\mu_E} \rightleftharpoons \langle C, \sigma', \mu_E \rangle$, *где для любого предложения* $\varphi \in S(\sigma)$ *имеем*

$$\mu_E(\varphi) = \frac{\left\| \{\mathfrak{A} \in E \,|\, \mathfrak{A} \models \varphi\} \right\|}{\|E\|}.$$

Заметим, что нечеткую модель $\mathfrak{A}_{\mu_{E(\Delta)}}$ можно было бы рассматривать как теоретико-модельную формализацию базы знаний предметной области Δ. Однако в подавляющем большинстве источников, описывающих конкретные прецеденты предметной области, информация о прецеденте является четкой, но не полной. Иначе говоря, для каждого конкретного прецедента мы не обладаем полной информацией обо всех описанных в сигнатуре понятиях. Для разрешения этой проблемы было предложено использовать методологию семантики открытого мира, широко применяемую в системах логики описаний (Description Logic) [16]. Основная идея данного подхода заключается в рассмотрении всех возможных интерпретаций неизвестной информации.

Итак, пусть $S_a(\sigma)$ – множество всех атомарных предложений сигнатуры σ. Знание об истинности / ложности всех предложений множества $S_a(\sigma)$ на данном прецеденте однозначно определяет математическую модель прецедента $\mathfrak{A} \in E(\Delta)$. Допустим теперь, что мы обладаем только частичной информацией об истинности / ложности предложений множества $S_a(\sigma)$ на рассматриваемом прецеденте. Тогда существует такое подмножество предложений $U \subseteq S_a(\sigma)$, значение истинности которых на рассматриваемом прецеденте нам точно известно. Следовательно, для математического описания такого прецедента нам необходимо рассмотреть все модели, *согласованные* с этим знанием.

Определение 2. *Рассмотрим множество* $U \subseteq S_a(\sigma)$ *и означивание* $v : U \to \{0,1\}$. *Будем говорить, что прецедент* $\mathfrak{A} \in E(\Delta)$ ***согласуется*** *с означиванием* v, *если для любого предложения* $\varphi \in U$ *имеем*

$$\mathfrak{A} \models \varphi \Leftrightarrow v(\varphi) = 1.$$

Определение 3. *Рассмотрим множество* $U \subseteq S_a(\sigma)$ *и означивание* $v : U \to \{0,1\}$. Класс прецедентов

$$K(v) \rightleftharpoons \{\mathfrak{A} \in E(\Delta) \,|\, \mathfrak{A} - \text{согласуется с } v\}.$$

будем называть ***обобщенным прецедентом***, *порожденным означиванием* v.

Определение 4. *Конечное множество* $K(v_1), ..., K(v_n) \subseteq E(\Delta)$ *обобщенных нечетких прецедентов, описывающих предметную область* Δ, *назовем* ***базой знаний*** *предметной области* Δ.

Определение 5. *Пусть* $KB(\Delta) = \{K(v_1), ..., K(v_n)\}$ *– база знаний предметной области* Δ. ***Обобщенной нечеткой моделью*** *предметной области* Δ *назовем алгебраическую систему* $\mathfrak{A}_\Delta = \langle C, \sigma', \xi \rangle$, *где для любого предложения* $\varphi \in S(\sigma')$ *имеем*

$$\xi(\varphi) = \{\mu_E(\varphi) \,|\, E \in K(v_1) \times ... \times K(v_n)\}.$$

В работе [17] показано, что значениями истинности на модели $\mathfrak{A}_\Delta$ являются интервалы, определенные на множестве

$$Q^n = \left\{0, \frac{1}{n}, ..., \frac{n-1}{n}, 1\right\},$$

где n – число прецедентов, занесенных в базу знаний.

Заметим, что в строго математическом смысле мы не будем получать интервальную модель. Однако при $n \to \infty$ значения истинности предложений на модели $\mathfrak{A}_\Delta$ будут стремиться к интервалам на множестве $[0,...,1] \cap \mathbb{Q}$. Таким образом, на практике, имея дело с достаточно большим множеством прецедентов, мы можем воспринимать значения истинности на модели $\mathfrak{A}_\Delta$ как интервалы рациональных чисел.

Также в работе [17] описан алгоритм подсчета значения истинности на модели $\mathfrak{A}_\Delta$ для любого бесквантворного предложения сигнатуры σ.

Формализация и классификация вопросных шаблонов

Вопрос выступает одним из важнейших элементов процесса получения новых знаний. Правильная постановка вопроса есть результат сложной мыслительной деятельности. Вопрос логически следует из всего предшествующего анализа предмета [18]. Однако сам вопрос не является суждением. Его можно рассматривать как требование отыскать ответ, являющийся истинным суждением.

Таким образом, при разработке классификации вопросов мы будем учитывать структуру суждений, являющихся ответами на поставленные вопросы. В классической, аристотелевской силлогистике логическая структура суждения состоит из четырех элементов: субъекта, предиката, связки и кванторного слова [19; 20]. Субъект суждения выражает знание о предмете суждения, т. е. то, о чем говорится в данном суждении. Предикат суждения выражает знание о свойствах субъекта. Связка устанавливает связь между субъектом и предикатом. Она может быть утвердительной или отрицательной. Кванторное слово отражает объем суждения. Если соединить вместе все четыре элемента, то получится следующая формула суждения:

Все (некоторые) *S* есть (не есть) *H*,

где *S* – субъект суждения и *H* – предикат суждения.

Традиционно все суждения делятся на истинные и ложные на заданной модели предметной области. С теоретико-модельной точки зрения это подразумевает, что суждения рассматриваются на классических моделях предметных областей. В нашем подходе вместо классических моделей рассматриваются обобщенные нечеткие модели. Значения истинности на таких моделях являются интервалами рациональных чисел из отрезка [0, 1]. Данные интервалы отражают объективную вероятность событий в рассматриваемой предметной области [21]. В связи с этим мы будем рассматривать суждение в более широком смысле, т. е. понимать под пятым компонентом суждения его вероятность (будем обозначать буквой *P*). Таким образом, под формулой суждения будем понимать следующее выражение:

Все (некоторые) *S* есть (не есть) *H* с вероятность *P*.

Рассмотрим теперь теоретико-модельную формализацию суждения. Пусть рассматриваемая предметная область Δ описывается предикатной сигнатурой σ'. База знаний $KB(\Delta)$ данной предметной области формализована в виде обобщенной нечеткой модели $\mathfrak{A}_\Delta$.

Дадим формальное определение суждения.

Определение 6. ***Атомарными суждениями*** сигнатуры σ' будем называть выражения вида

$$\boldsymbol{S_1} : \exists x\big(S(x) \,\&\, H(x)\big)[\alpha, \beta];$$
$$\boldsymbol{S_2} : \forall x\big(S(x) \to H(x)\big)[\alpha, \beta];$$
$$\boldsymbol{S_3} : \exists x\big(S(x) \,\&\, \neg H(x)\big)[\alpha, \beta];$$

$$\boldsymbol{S_4} : \forall x\big(S(x) \to \neg H(x)\big)[\alpha,\beta],$$

где $S, H \in \sigma'$ и $\alpha,\beta \in [0,1] \cap \mathbb{Q}(\alpha \le \beta)$.

Часть выражения в $\boldsymbol{S_1}$–$\boldsymbol{S_4}$, *написанную до квадратных скобок, будем называть* ***телом*** *атомарного суждения и обозначать греческими буквами* φ *или* ψ. *Выражение в квадратных скобках будем называть* ***вероятностной характеристикой*** *суждения.*

Замечание 1. *Нетрудно заметить, что суждения типа* $\boldsymbol{S_3}$ *являются логическими отрицаниями соответствующих суждений типа* $\boldsymbol{S_2}$, *а суждения типа* $\boldsymbol{S_4}$ *являются логическими отрицаниями суждений типа* $\boldsymbol{S_1}$.

Замечание 2. *Если сигнатурный символ, используемый для формализации субъекта суждения, является константным предикатом, т. е.* $S = P_c \in \sigma' \setminus \sigma$, *то суждения типа* $\boldsymbol{S_1}$ *и* $\boldsymbol{S_2}$ *становятся эквивалентными выражению*

$$\boldsymbol{S_5} : H(c)[\alpha,\beta],$$

а суждения типа $\boldsymbol{S_3}$ *и* $\boldsymbol{S_4}$ *становятся эквивалентными выражению*

$$\boldsymbol{S_6} : \neg H(c)[\alpha,\beta].$$

Если рассматривать частный случай, когда в роли субъекта и предиката атомарного суждения выступает один и тот же сигнатурный символ, т. е. $S = H$, то получим следующие частные случаи суждений:

$$\boldsymbol{S_1} : \exists x S(x)[\alpha,\beta];$$
$$\boldsymbol{S_2} : \top[\alpha,\beta];$$
$$\boldsymbol{S_3} : \bot[\alpha,\beta];$$
$$\boldsymbol{S_4} : \forall x \neg S(x)[\alpha,\beta],$$

где $\top$ и $\bot$ – тождественно истинное и тождественно ложное предложения. С точки зрения вопросно-ответной системы рассмотрение таких суждений нецелесообразно. Поэтому мы будем полагать, что при формировании субъекта и предиката атомарного суждения всегда будут использоваться различные предикатные символы.

В предыдущем параграфе мы отмечали, что все сигнатурные символы можно разделить на аналитические и эмпирические предикаты предметной области Δ. Заметим, что если и субъект и предикат суждения будут описываться при помощи аналитических сигнатурных символов, то мы будем получать *аналитические суждения* о предметной области Δ. Значениями истинности таких суждений на модели $\mathfrak{A}_\Delta$ являются либо 1 (истинно), либо 0 (ложно). Например, в предметной области компьютерной безопасности суждение «*Каждый вирус является вредоносной программой*» является истинным аналитическим суждением и принадлежит аналитической теории данной предметной области. С другой стороны, суждение «*Каждая вредоносная программа является вирусом*» – ложное аналитическое суждение, так как в данной предметной области мы полагаем, что *вредоносные программы* делятся на *вирусы*, *трояны* и *черви*.

Так как при разработке вопросно-ответной системы с ограниченным доменом Δ мы полагаем, что аналитическая теория $\mathrm{Th}(\Delta)$ полностью определена, то при формировании шаблонов вопросов нет необходимости рассматривать аналитические суждения данной предметной области.

Таким образом, при формировании вопросов мы будем рассматривать только эмпирические атомарные суждения, т. е. суждения, в которых либо субъект, либо предикат формализуется с помощью эмпирического сигнатурного символа.

Пусть φ_1 и φ_2 – тела некоторых суждений. Тогда для любых $\alpha,\beta \in [0,1] \cap \mathbb{Q}$ таких, что $\alpha \le \beta$, выражения

$$(\varphi_1 \,\&\, \varphi_2)[\alpha,\beta];\ (\varphi_1 \vee \varphi_2)[\alpha,\beta];\ (\varphi_1 \to \varphi_2)[\alpha,\beta];\ (\neg\varphi_1)[\alpha,\beta]$$

также являются суждениями.

Замечание 3. *Сложные суждения можно также получать, если в качестве субъекта или предиката суждения рассматривать не атомарные формулы, а произвольные формулы*

данной сигнатуры с одной свободной переменной. Однако рассмотрение таких суждений выходит за рамки данной работы.

Определение 7. *Суждение* $\varphi[\alpha,\beta]$ *является* ***верным*** *на обобщенной нечеткой модели* $\mathfrak{A}_\Delta$, *если* $\xi_\Delta(\varphi)\subseteq[\alpha,\beta]$. *В противном случае суждение* $\varphi[\alpha,\beta]$ *будет* ***неверным*** *на модели* $\mathfrak{A}_\xi$.

Теперь, когда описано формальное определение суждения, можно переходить к описанию типов вопросов к данным суждениям.

Традиционно в вопросно-ответных системах рассматриваются два типа вопросов: «ли-вопросы» и «какой-вопросы». Так как наши суждения носят вероятностный характер, то мы будем рассматривать еще и третий тип вопросов – «вероятностные вопросы».

«Ли-вопросы» направлены на выяснение того, верно ли данное суждение или нет. Например «*Верно ли, что вероятность использования в компьютерной атаке вредоносной программы больше 0,7?*» Ответом на такой вопрос является либо «да», либо «нет». Вопросы такого типа образуются по следующей схеме:

Верно ли, что $\langle$суждение$\rangle$?

«Вероятностные вопросы» направлены на выявление вероятностной характеристики тела суждения. Например: *«Какова вероятность того, что компьютерная атака будет направлена на взлом операционной системы?»* Ответом на такой вопрос будет интервал рациональных чисел из отрезка [0, 1]. Вопросы такого типа образуются по следующей схеме:

Какова вероятность того, что $\langle$тело суждения$\rangle$?

Заметим, что в более общем случае, если рассматривать интерпретацию вопросов на моделях с истинностными функциями различной природы (например, булевозначные модели или прецедентные модели [10]), вместо вероятностных вопросов мы получим *оценочные вопросы*. Общая схема таких вопросов будет следующая:

Каково значение истинности того, что $\langle$тело суждения$\rangle$?

В случае интерпретации вопроса на классической модели оценочный вопрос отождествляется с «ли-вопросом».

И, наконец, «какой-вопросы» направлены на выявление объема субъекта вопроса. Например: *«Какие последствия компьютерной атаки наступят с вероятностью более 0,7?»* Ответом на такой вопрос является подмножество множества A объектов модели $\mathfrak{A}_\Delta$.

С логической точки зрения «какой-вопросы» могут порождать как атомарные суждения, так и более сложные суждения. Однако с лингвистической точки зрения формулировка сложных «какой-вопросов» весьма затруднительна и требует дополнительных исследований. В работах по эротетической логике [12; 18] для этих целей вводится понятие ремы (или предпосылки) вопроса и рассматриваются многородовые-однородовые ремы и многообъектные-однообъектные ремы. Теоретико-модельная формализация таких вопросов выходит за рамки данной работы.

Таким образом, для формализации «какой-вопросов» мы будем использовать следующие схемы:

Какие $x\in A$ такие, что $\big(S(x)\,\&\,H(x)\big)[\alpha,\beta]$?

Какие $x\in A$ такие, что $\big(S(x)\,\&\,\neg H(x)\big)[\alpha,\beta]$?

Описание базы знаний вопросно-ответной системы *QA-RiskPanel*

На основе методологии прецедентного подхода в НГУ была разработана программная система *RiskPanel* [8]. В этой системе сигнатура предметной области компьютерной безопасности описывается одним эмпирическим понятием H: «*имело место в данной атаке*» и множеством аналитических понятий, которое насчитывает более 50 понятий и может пополняться по мере появления новых видов компьютерных атак. Такая особенность сигнатуры накладывает одно семантическое ограничение на структуру атомарных суждений. В данной предметной области мы будем рассматривать только эмпирические суждения,

субъектами которых являются различные аналитические понятия, а в роли предиката всегда выступает эмпирическое понятие H.

Множество аналитических понятий в базе данных системы *RiskPanel* разделено на шесть категорий: «*Симптомы*», «*Угрозы*», «*Уязвимости*», «*Последствия*», «*Потери*» и «*Контрмеры*». Каждая из шести категорий представлена в базе данных в виде древовидной структуры. Так, например, на рис. 1 изображен фрагмент дерева категории «*Угрозы*». Вершинами дерева являются аналитические понятия предметной области. Тупиковые вершины интерпретируются как константы данной сигнатуры (например, «*Фишинг*», «*Вирус*», «*Троянский конь*», «*Бэкдор*», «*Программа-шпион*»). Детальное описание структуры базы данных можно найти в работе [22].

Каждый прецедент компьютерной атаки в базе данных системы *RiskPanel* характеризуется наличием определенных понятий из каждой категории. На основе этой базы данных формируется база знаний разрабатываемой вопросно-ответной системы *QA-RiskPanel*. При обращении к конкретному прецеденту происходит считывание из базы данных множества понятий, которыми он обладает.

Таким образом, мы получаем знание об истинности на данном прецеденте некоторых атомарных суждений типа $\boldsymbol{S_1}$. Например, если в данном прецеденте компьютерной атаки была использована вредоносная программа (см. рис. 1), то на этом прецеденте будет истинно атомарное суждение *«Существует вредоносная программа, которая была использована в данной атаке»*. По полученному множеству истинных атомарных суждений мы вычисляем значения истинности для всех остальных атомарных суждений. Делается это путем расстановки символов «+», «–» и «?» во всем дереве категории, где «+» означает, что прецедент обладает понятием, «–» – не обладает, «?» – не известно, обладает или нет. Для каждого типа суждений ($\boldsymbol{S_1}$–$\boldsymbol{S_4}$) разработаны следующие алгоритмы «означивания» деревьев понятий.

I. *Алгоритм подсчета значений истинности для атомарных суждений типа* $\boldsymbol{S_1}$. Все понятия вверх по дереву от понятия, обладание которым задано в базе данных, помечаются символом «+» (т. е. как понятия, которыми прецедент обладает). Все понятия вниз по дереву от понятия, обладание которым задано в базе данных, помечаются символом «?» (т. е. как понятия, обладание которыми для данного прецедента не определено). Остальные понятия

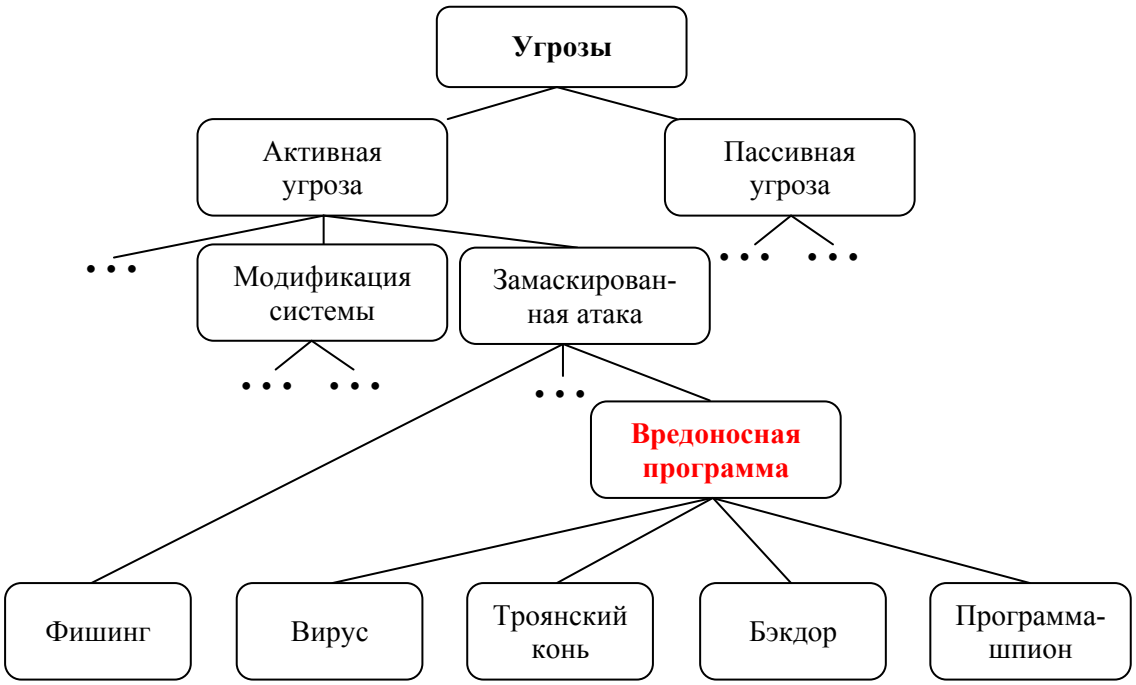

Рис. 1. Фрагмент дерева понятий категории «*Угрозы*»

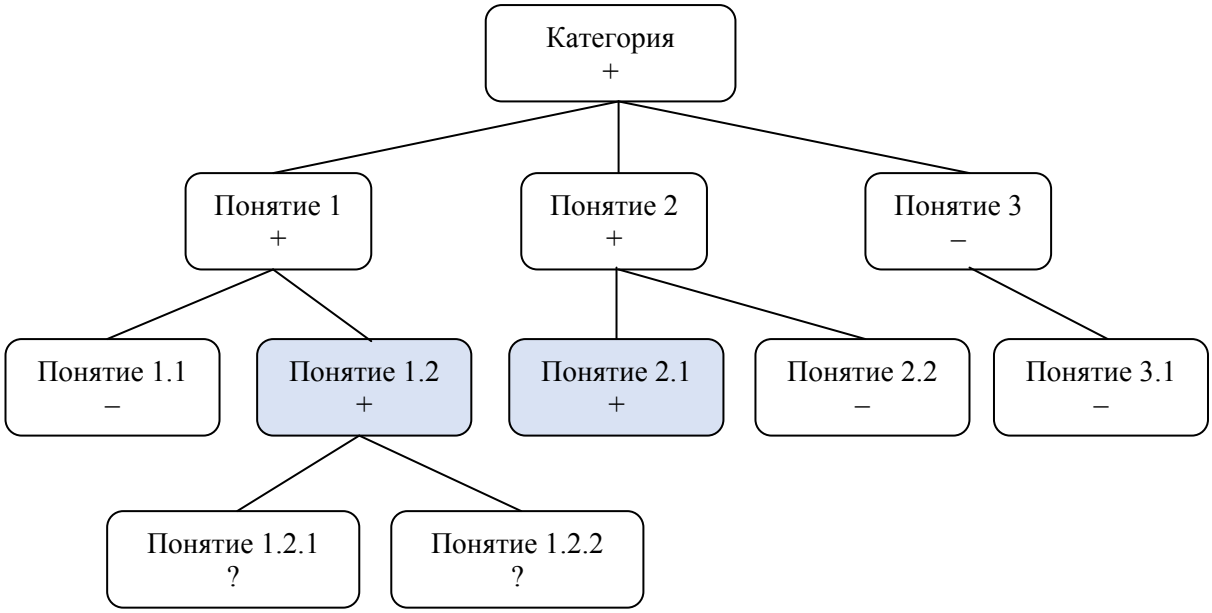

Рис. 2. Принцип работы алгоритма по определению истинностного значения суждения вида $\exists x\left(S\left(x\right)\&H\left(x\right)\right)$ на прецеденте компьютерной атаки

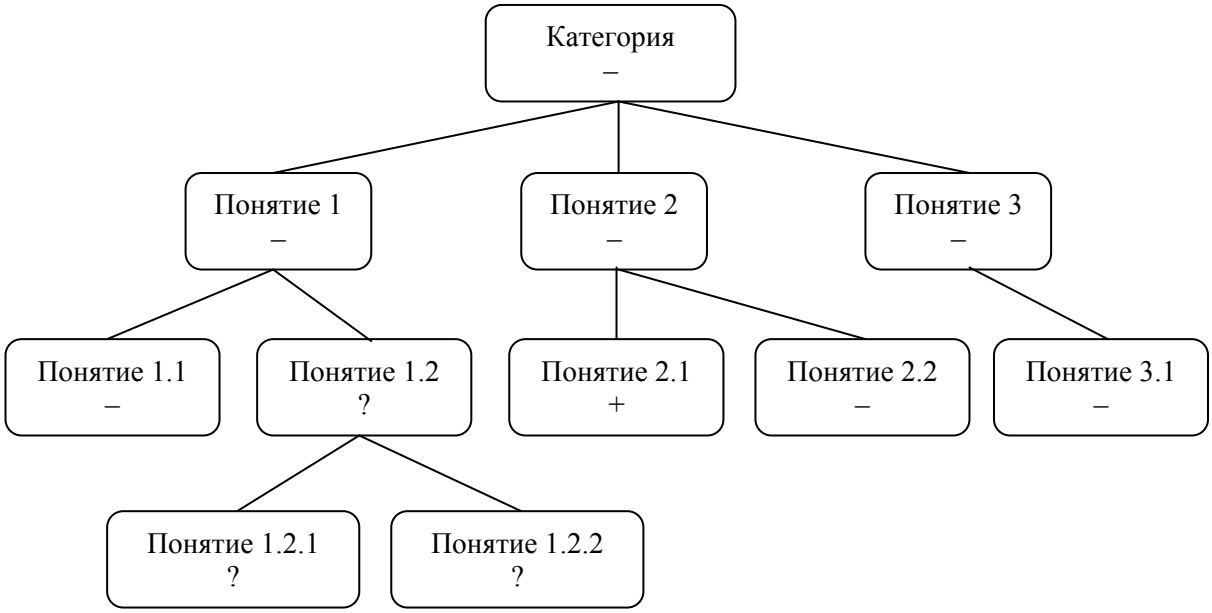

Рис. 3. Принцип работы алгоритма по определению истинностного значения суждения вида $\forall x\left(S\left(x\right)\to H\left(x\right)\right)$ на прецеденте компьютерной атаки

помечаются символом «–» (т. е. как понятия, которыми прецедент не обладает). На рис. 2 представлен пример работы алгоритма. Серым цветом выделены понятия, факт обладания которыми внесен в базу данных.

II. *Алгоритм подсчета значений истинности для атомарных суждений типа* $\boldsymbol{S_2}$. Значения понятий, находящихся в тупиковых вершинах, приравниваются значениям, полученным при означивании дерева понятий для суждений типа $\boldsymbol{S_1}$ (см. Замечание 2). Далее алгоритм идет вверх по дереву от тупиковых вершин, означивая оставшиеся понятия по следующему принципу. Если среди дочерних понятий есть хотя бы одно понятие, помеченное «–», то родительское понятие помечается «–». Если все дочерние понятия помечены «+», то родительское понятие также помечается «+». Иначе родительское понятие помечается символом «?». На рис. 3 представлен пример работы алгоритма.

III. *Алгоритм подсчета значений истинности для атомарных суждений типа* $\boldsymbol{S_3}$. В дереве, построенном для подсчета значений истинности суждений типа $\boldsymbol{S_2}$, символы «+» меняются на символы «–» и символы «–» меняются на символы «+» (см. Замечание 1).

IV. *Алгоритм подсчета значений истинности для атомарных суждений типа* $\boldsymbol{S_4}$. В дереве, построенном для подсчета значений истинности суждений типа $\boldsymbol{S_1}$, символы «+» меняются на символы «–» и символы «–» меняются на символы «+» (см. Замечание 1).

По полученным значениям истинности атомарных суждений на всех прецедентах предметной области считаются интервальные значения истинности как данных суждений, так и сложных суждений на обобщенной нечеткой модели $\mathfrak{A}_\Delta$. Описание алгоритмов подсчета данных значений истинности можно найти в работе [17].

Описание пользовательского интерфейса программной системы QA-RiskPanel

С точки зрения пользователя вопросы в вопросно-ответных системах должны быть приближены к естественному языку, насколько это возможно. С точки зрения системы поиск ответов в базе знаний происходит только при однозначном переводе введенного пользователем вопроса в некоторое фиксированное логическое представление. Таким образом, формирование вопросов в программной системе преследует две противоположные цели, и задача разработки программного интерфейса превращается в задачу поиска компромисса между формальностью запроса и естественностью языка.

Разрабатываемая вопросно-ответная система способна давать ответы на вопросы разных типов. В настоящее время в системе реализованы алгоритмы обработки «ли-вопросов», «вероятностных вопросов» и «какой-вопросов». В последующих версиях планируется расширять этот список другими вопросными типами.

Конструктор «ли-вопросов» состоит из трех функциональных блоков, заполняемых пользователем: конструктор суждений, конструктор отношений и блок выбора вероятностного интервала.

Конструктор суждений предназначен для ввода атомарных суждений. Пользователь последовательно выбирает квантор, субъект и связку суждения. Так как в данной предметной области описано только одно эмпирическое понятие, нет необходимости для ввода предиката суждения. На рис. 4 показано поле построения атомарного суждения.

Поскольку «ли-вопрос» может содержать несколько атомарных суждений, в конструкторе суждений реализована функциональность добавления нового суждения. При нажатии на кнопку «Добавить» появляется новое поле формирования суждения (рис. 5).

Конструктор отношений позволяет строить логические выражения из построенных атомарных суждений. Введенные пользователем данные проверяются на корректность, после чего преобразуются в представление, удобное для извлечения ответа из базы знаний.

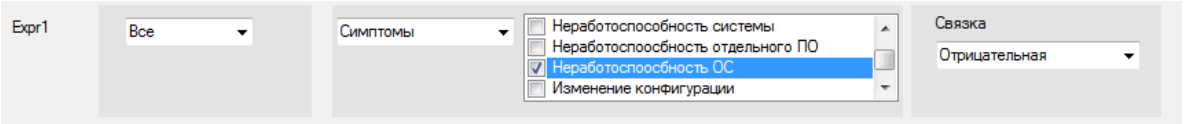

Рис. 4. Поле построения атомарного суждения

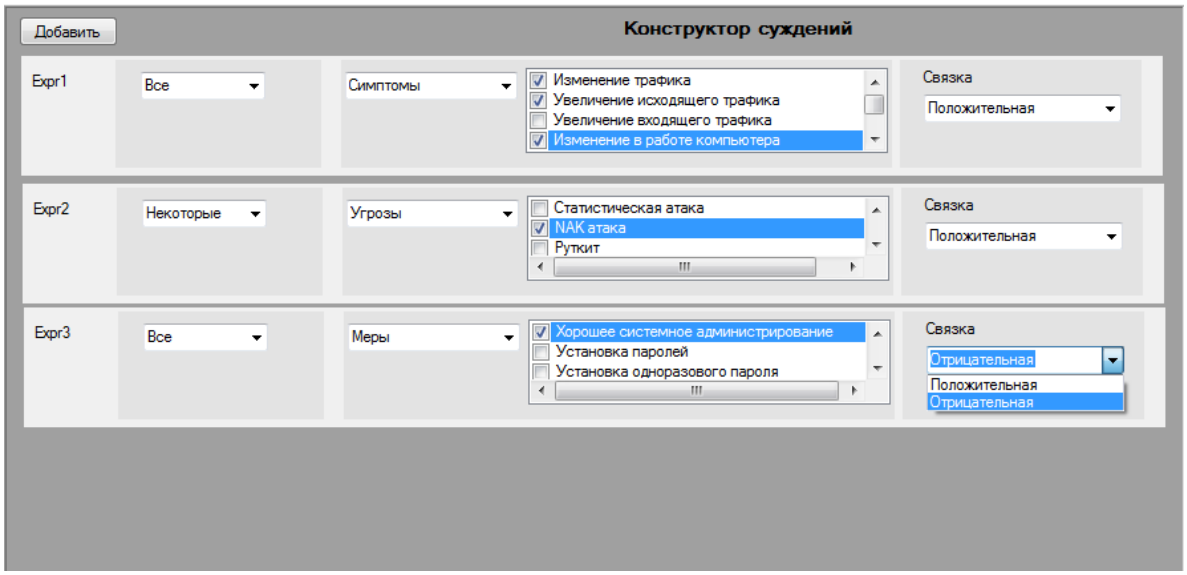

Рис. 5. Конструктор суждений

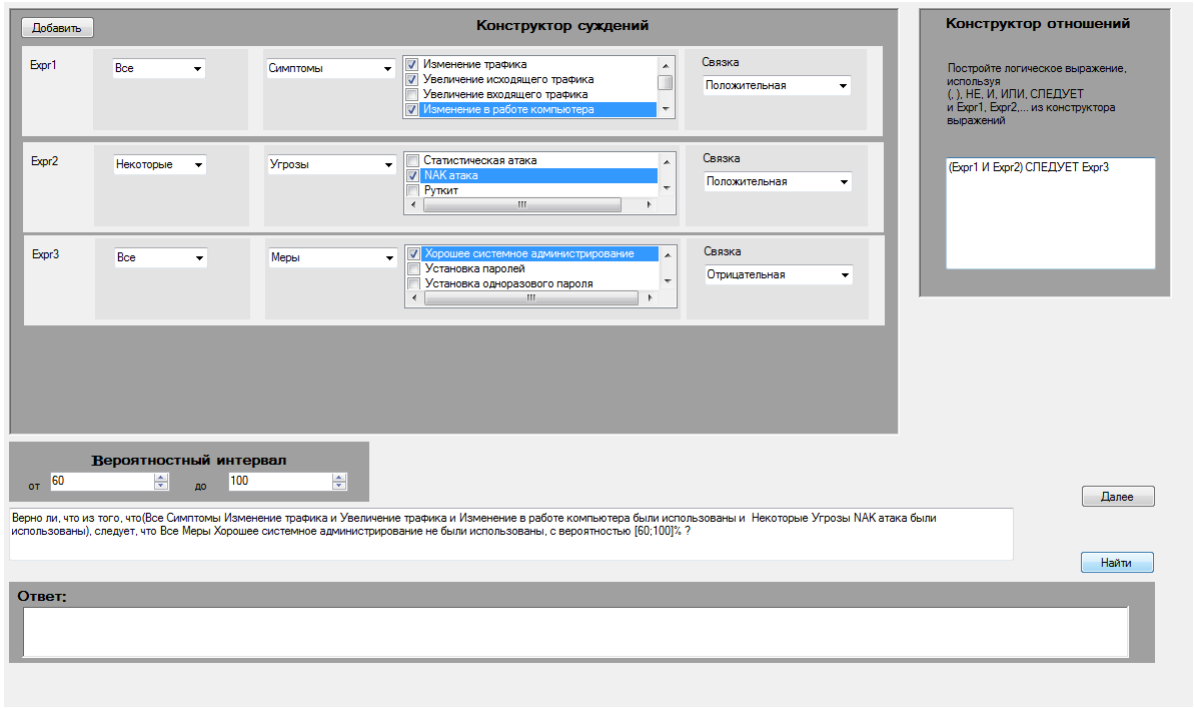

Рис. 6. Интерфейс вкладки формирования «ли-вопроса»

Блок выбора вероятностного интервала предназначен для задания вероятностного параметра «ли-вопроса».

После того, как пользователь составит выражение из суждений при помощи конструкторов, он нажимает кнопку «Далее», и в текстовой области появляется вопрос, сгенерированный из данных, введенных им. Пользователь нажимает кнопку «Найти», и начинают работать алгоритмы поиска ответов. Общий вид вкладки «ли-вопроса» представлен на рис. 6.

Конструктор «вероятностных вопросов» отличается от конструктора «ли-вопросов» отсутствием блока выбора вероятностного интервала, так как отыскание вероятностного интервала и является сутью вопросов такого типа.

Конструктор «какой-вопросов» состоит из двух функциональных блоков: конструктора суждений и блока выбора вероятностного интервала. Более того, в конструкторе суждений отсутствует поле выбора квантора суждения и не реализована функциональность добавления нового суждения.

Заключение

В работе описан математический аппарат и программная реализация вопросно-ответной системы *QA-RiskPanel*, являющейся одним из модулей программного комплекса *RiskPanel.*

Базируясь на прецедентном подходе к формализации неполного знания о предметной области, данная вопросно-ответная система позволяет пользователю получать информацию о вероятности наступления того или иного события в области информационной безопасности. Данный функционал может быть использован специалистом для прогнозирования информационных рисков и описания мер по их ликвидации или минимизации.

В настоящее время в системе реализован модуль обработки безусловных вероятностных вопросов. Данный модуль ориентирован на обработку информации в ситуации, когда о начавшемся компьютерном нападении ничего неизвестно. В дальнейшем планируется дополнить данную вопросно-ответную систему модулем обработки условных вероятностных вопросов, т. е. модулем, работающим в условиях, когда уже имеется некоторая вероятностная информация о начавшейся компьютерной атаке.

Список литературы

1. *Mollá D.*, *Vicedo J. L.* Question Answering in Restricted Domains: An Overview // Computational Linguistics. 2007. Vol. 33. No. 1. P. 41–61.

2. *Allam M. N.*, *Haggag M. H.* The Question Answering Systems: A Survey // International Journal of Research and Reviews in Information Sciences. 2012. Vol. 2. No. 3. P. 211–221.

3. *Sanjay K. Dwivedi*, *Vaishali Singh.* Research and Reviews in Question Answering System // Procedia Technology. 2013. No. 10. P. 417–42.

4. *Xin Li*, *Dan Roth.* Learning question classifiers // Proc. of the 19th International Conference on Computational Linguistics (COLING 2002). P. 556–562.

5. *Werner C.*, *Smith B.*, *Van Mol M.* Using ontology in query answering systems: Scenarios, requirements and challenges // Proc. of the 2nd CoLogNET-Elsnet Symposium. Amsterdam, 2003. P. 5–15.

6. *Demner-Fushman D.*, *Lin J.* Answering Clinical Questions with Knowledge-Based and Statistical Techniques // Computational Linguistics. 2007. Vol. 33. No. 1. P. 63–103.

7. *Benamara F.* Cooperative question answering in restricted domains: The WEBCOOP Experiments // Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004). Barcelona, 2004. P. 31–38.

8. *Пальчунов Д. Е.*, *Яхъяева Г. Э.*, *Хамутская А. А.* Программная система управления информационными рисками RiskPanel // Программная инженерия. 2011. № 7. С. 29–36.

9. *Palchunov D. E.*, *Yakhyaeva G. E.* Interval fuzzy algebraic systems // Proc. of the Asian Logic Conference 2005. World Scientific Publishers. 2006. P. 23–37.

10. *Пальчунов Д. Е.*, *Яхъяева Г. Э.* Нечеткие алгебраические системы. // Вестн. Новосиб. гос. ун-та. Серия: Математика, механика, информатика. 2010. Т. 10, вып. 3. С. 75–92.

11. *Яхъяева Г. Э.*, *Ясинская О. В.* Применение методологии прецедентных моделей в системе риск – менеджмента, направленного на раннюю диагностику компьютерного нападения // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2012. Т. 10, вып. 2. С. 106–115.

12. *Белнап Н.*, *Стил Т.* Логика вопросов и ответов. М.: Прогресс, 1981. 288 с.

13. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Ч. 1: Теоретико-модельная формализация онтологии и рефлексии // Философия науки. 2006. № 4 (31). С. 86–114.

14. *Мальцев А. И.* Алгебраические системы. М.: Наука, 1970. 392 с.

15. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Ч. 2: Онтологии и формализации понятий // Философия науки. 2008. № 2. С. 62–99.

16. The Description Logic Handbook / Ed. by F. Baader. N. Y.: Cambridge Univ. Press, 2003. 555 p.

17. *Яхъяева Г. Э.*, *Ясинская О. В.* Методы согласования знаний по компьютерной безопасности, извлечѐнных из различных документов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 3. С. 63–73.

18. Логика: Учебник / Под ред. А. И. Мигунова, И. Б. Микиртумова, Б. И. Федорова. М.: Проспект, 2011. 680 с.

19. Логика: Учебник / Под ред. Д. П. Горского, П. В. Таванца. М., 1956. 280 с.

20. *Светлов В. А.* Логика: Учеб. пособие. СПб.: Питер, 2011. 320 с.

21. *Gillies D.* Philosophical Theories of Probability. Routledge, 2012. 240 p.

22. *Мирзагитов А. А.*, *Пальчунов Д. Е.* Методы разработки онтологии по информационной безопасности, основанные на прецедентном подходе // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 3. С. 37–46.

Материал поступил в редколлегию 29.10.2014

G. E. Yakhyaeva, **O. V. Yasinskaya**, **A. A. Karmanova**

Novosibirsk State University
2 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

gul_nara@mail.ru, *yasinskaya.olga@gmail.com*, *anast.karmy.aa@gmail.com*

PROBABILISTIC QUESTION-ANSWERING SYSTEM IN THE FIELD OF COMPUTER SECURITY

This paper is devoted to the description of the *QA-RiskPanel* question-answering system that uses computer security database as a source of knowledge. *QA-RiskPanel* is a case-based system and allows the user to set the probabilistic questions to determine and predict the various risks associated with the computer attacks. The paper presents model-theoretic foundations and classification of question templates and describes a software implementation of the developed approach.

Keywords: information security, computer attack, case of computer attack, generalized fuzzy model, generalized case, question, statement.

References

1. Mollá D., Vicedo J. L. Question Answering in Restricted Domains: An Overview. *Computational Linguistics*, 2007, vol. 33, no. 1, p. 41–61.

2. Allam M. N., Haggag M. H. The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences*, 2012, vol. 2, no. 3, p. 211–221.

3. Sanjay K. Dwivedi, Vaishali Singh. Research and Reviews in Question Answering System. *Procedia Technology*, 2013, no. 10, p. 417–42.

4. Xin Li, Dan Roth. Learning question classifiers. *Proc. of the 19th International Conference on Computational Linguistics* (*COLING* 2002), 2002, p. 556–562.

5. Werner C., Smith B., Van Mol M. Using ontology in query answering systems: Scenarios, requirements and challenges. *Proc. of the 2nd CoLogNET-Elsnet Symposium*. Amsterdam, 2003, p. 5–15.

6. Demner-Fushman D., Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 2007, vol. 33, no. 1, p. 63–103.

7. Benamara F. Cooperative question answering in restricted domains: The WEBCOOP Experiments. *Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL-2004*). Barcelona, 2004, p. 31–38.

8. Palchunov D. E., Yakhjyaeva G. E., Hamutskya A. A. Software system for information risk management «RiskPanel». *Programmnaya ingeneriya*, 2011, no. 7, p. 29–36. (in Russ.)

9. Palchunov D. E., Yakhyaeva G. E. Interval fuzzy algebraic systems. *Proc. of the Asian Logic Conference 2005*. World Scientific Publishers, 2006, p. 23–37. (in Russ.)

10. Palchunov D. E., Yakhyaeva G. E. Fuzzy algebraic systems. *Vestnik of Novosibirsk State University. Series*: *Mathematics*, *Mechanics and Informatics*, 2010, vol. 10, no. 3, p. 75–92. (in Russ.)

11. Yakhyaeva G. E., Yasinskaya O. V. The application of precedent models methodology in the risk-management system aimed at early detection of computer attacks. *Vestnik of Novosibirsk State University. Series*: *Information Technologies*, 2012, vol. 10, no. 2, p. 106–115. (in Russ.)

12. Belnap N. D., Steel T. B. *The Logic of Questions and Answers*. London, 1976, 176 p.

13. Palchunov D. E. Simulation of thinking and formalization of reflection. Part 1. Model-theoretic formalization of ontologies and reflection. *Filosofiya nauki*, 2006, no. 4 (31), p. 86–114. (in Russ.)

14. Malcev A. I. *Algebraic Systems*. Moscow, Nauka, 1970, 392 p. (in Russ.)

15. Palchunov D. E. Simulation of thinking and formalization of reflection. Part 2. Ontologies and formalization of concepts. *Filosofiya nauki*, 2008, no. 2 (37), p. 62–99. (in Russ.)

16. Baader F. (ed.) *The Description Logic Handbook*. New York, Cambridge Univ. Press, 2003. 555 p.

17. Yakhyaeva G. E., Yasinskaya O. V. Matching methods in computer security knowledge learned from multiple documents. *Vestnik of Novosibirsk State University. Series*: *Information Technologies*, 2013, vol. 11, no. 3, p. 63–73. (in Russ.)

18. Migynov A. I., Mikirtumov I. B., Fedorov B. I. (eds.) *Handbook of Logic*. Moscow, Prospect, 2011, 680 p. (in Russ.)

19. Gorski D. P., Tavanc P. V. (eds.) *Logic*. Moscow, 1956, 280 p. (in Russ.)

20. Svetlov V. A. *Logic*: Book for students. St.-Petersburg, 2011, 320 p. (in Russ.)

21. Gillies D. *Philosophical Theories of Probability*. Routledge, 2012, 240 p.

22. Mirzagitov A. A., Palchunov D. E. Methods of the ontology of information security development based on the precedent approach. *Vestnik of Novosibirsk State University. Series*: *Information Technologies*, 2013, vol. 11, no. 3, p. 37–46. (in Russ.)