Анализ гетерогенных слабо структурированных данных в задачах обеспечения информационой безопасности

М. А. Полтавцева[1], П. Д. Зегжда[2]
Санкт-Петербургский политехнический университет Петра Великого
[1]poltavtseva@ibks.spbstu.ru, [2]zeg@ ibks.spbstu.ru

Е. А. Зайцева
ООО «Необит»
elizavetazaytceva@mail.ru

***Аннотация*. При решении задач информационной безопасности возникает необходимость анализа и оценки подобия наборов гетерогенных слабо структурированных объектов. Авторы предлагают использовать для этого подход на базе прецедентного анализа. Данные представляются в виде множества объектов (прецедентов), каждый из которых описывается переменным набором свойств. В работе предложены меры оценки близости отдельных объектов и прецедентов, представлены результаты экспериментального тестирования метода. Авторами приведена архитектура системы анализа и предложен метод обучения такой системы.**

Ключевые слова: анализ гетерогенных слабо структурирвоанных объектов; прецедентный анализ; информационная безопасность

I. Введение

Информационная безопасность является мульти дисциплинарной областью, сочетающей в себе организационные меры и работу с техническими средствами защиты информации. Для оценки защищенности объектов информатизации специалисты по защите информации используют различные методы и средства, включая не только анализ технических показателей и результаты мониторинга, но и, зачастую, совместную оценку социальных и организационных данных. Примерами таких задач могут служить: оценка наличия утечек и данных в открытых источниках, тестирование на проникновение, анализ сетевого обмена.

В силу разнообразия предметных областей исследователю приходится иметь дело не только со слабо формализованными или неполными данными, но и с данными различной семантики, или гетерогенными данными (техническими, социальными и т.д.), производить их совместный анализ. Таким образом, задача анализа гетерогенных объектов является высоко актуальной для современного положения в области обеспечения информационной безопасности.

При финансовой поддержке Министерства образования и науки Российской Федерации в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы» (Соглашение № 14.578.21.0231, уникальный идентификатор соглашения RFMEFI57817X0231)

Данная статья посвящена вопросу построения аналитической системы для работы с гетерогенными слабо структурированными данными, включая вопросы их совместного анализа, оценки схожести и обучения.

II. Анализ данных в задачах защиты информации

Решение различных аналитических задач в области информационной безопасности является высоко актуальной задачей. За последние двадцать лет ей посвящено достаточно большое число работ, включая последние работы в области больших данных[1] и искусственного интеллекта [2].

С точки зрения представления базы знаний и анализа информации можно выделить целый ряд подходов, методов и техник [3].

Во-первых, широко применяются системы, основанные на правилах. К этому классу относится большинство экспертных систем [4, 5] в том числе, в области информационной безопасности [6]. В таких системах экспертно задается база знаний в виде набора правил, онтологии, фреймов или другого, сводимого к системе правил представления. Аналитический вывод базируется на применении исходных данных к предикатам правил базы знаний и формирования цепочек выводов.

Во-вторых, нужно отметить системы машинного обучения. Это системы, основанные на различных методах Data Mining в последнее время широко применяются при решении частных задач защиты информации [7, 8]. Методы Data Mining заключаются в применении больших наборов обучающих данных, для выявления закономерностей предметной области.

В-третьих, прецедентные системы [9]. В области информационной безопасности можно отметить работу [10] по анализу инцидентов безопасности, однако описанные в ней методы не были распространены на анализ сложных систем и тестирование защищенности.

Анализируя подходы к построению аналитических систем в области информационной безопасности, можно отметить широкое применение систем, основанных на экспертных знаниях. В последние время также популярны

подходы с использованием машинного обучения. Прецедентный анализ, в свою очередь, востребован не так широко, что обуславливается сложностью формализации мульти дисциплинарных прецедентов.

При выборе подхода построения аналитической системы необходимо отметить следующее:

Многообразие современных устройств, программного обеспечения, а главное – вариантов атак, развитие APT – атак, обуславливают невозможность формализации и поддержания в актуальном состоянии экспертом сколько ни будь значительной предметной области.

Во многих задачах, оперирующих гетерогенными слабо структурированными данными, отсутствует достаточное число обучающих наборов (например, в области тестирования на проникновение [11]).

На основании этих свойств предлагается использовать прецедентный подход к построению аналитической системы с использованием элементов математической статистики и моделирования.

III. Подход к построению системы поддержки принятия решений при оценке защищенности

Входными данными системы анализа гетерогенных слабо структурированных данных являются результаты сбора информации об объектах предметной области. В зависимости от конкретной задачи, это могут быть различные комбинации результатов работы технических устройств, систем мониторинга и сканирования, анализа социальных сетей, интернет ресурсов, сообщений, сетевого трафика и других источников. Основными свойствами входных данных являются:

1. Семантическая и синтаксическая неоднородность (гетерогенность).
2. Слабая степень структуризации (в случае учета естественно-языковых текстов, графических изображений, видео и других подобных видов информации).
3. Неполнота.

Семантическая неоднородность порождается неоднородностью источников данных, слабая структуризация – как источником, так и самой природой поступающих сведений. Неполнота обуславливается различием в наборах данных (в том числе формализованных) собранных из одного источника при различных обстоятельствах. Например, количество сведений о распределенной системе, полученной в результате сканирования, может существенно отличаться в зависимости от ее ткущих настроек и внешних по отношению к ней обстоятельств.

Данные такого рода, исходя из их свойств в аналитической системе, предлагается формализовать в виде набора объектов, характеризующихся простыми свойствами [11]. Согласно [12], в этом случае множество объектов задается как $O = \{o_1, o_2, \ldots, o_m\}$, где $o_j = \{c \mid c_i = Unknown \cup c_i = Value \cup c_i = RefO\}$, где O – множество объектов; C – свойства объектов; Unknown – известен факт наличия свойства, но его значение не определено; Value – известное значение свойства; RefO – свойство – ссылка на объект. В данном случае можно говорить о представлении объекта как исключительно набора свойств («bag of attributes»).

Выходными данными аналитической системы должны являться рекомендательные предложения по рассматриваемой ситуации, в зависимости от специфики задачи: наличие опасной ситуации, возможные уязвимости системы и другие. В общем случае задача заключается в поиске в базе знаний и построении выводов на основе хранимой информации и поступающих сведений.

С этой точки зрения прецедент $p = \{o \mid o_j \in O\}$, где O – множество объектов. То есть, под прецедентом понимается набор объектов описывающий конкретный пример предметной области, ситуацию, атаку и т.д. Этот подход, по аналогии с анализом естественно-языковых текстов [13] можно назвать «bag of objects».

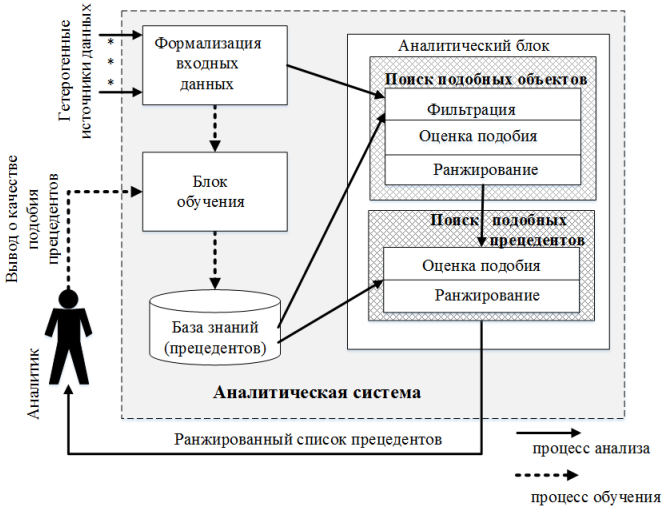

Рис. 1. Структура системы поддержки принятия решений при оценке защищенности

Структура системы, построенной в соответствии с выбранным подходом, включает основные блоки в виде базы знаний, блока формализации входных данных, блока анализа и блока обучения (рис. 1).

IV. Анализ данных и оценка схожести прецедентов

База знаний аналитической системы формируется исходя из известного набора прецедентов. Для задач, в которых получение такого набора затруднено, предлагается применить методы моделирования прецедентов на основе экспертных оценок и статистических данных с последующей коррекцией в процессе применения через блок обучения. Корректность применения моделирования в данном случае обоснована в [14].

Формализация входных данных в соответствии с выбранной моделью их объектного представления осуществляется соответствующим ETL-модулем при поступлении в систему.

Блок анализа осуществляет два этапа аналитической обработки. На первом этапе проводится предварительная фильтрация свойств объектов (например, на основе оценки значимости свойства [11]) и проводится оценка схожести объектов при помощи интегральной меры сходства.

Для расчета меры подобия объектов, для них определяются бинарные вектора свойств, где 1 – наличие у объекта соответствующего свойства, 0 – его отсутствие. Координаты вектора определяются как упорядоченные свойства, соответствующие пересечению множество свойств объектов:

$$\left.\begin{array}{l} o_1 \to C_1 = \{c_1^1, \ldots c_1^{k1}\} \\ o_2 \to C_2 = \{c_2^1, \ldots c_2^{k2}\} \end{array}\right\} \to \overrightarrow{C_{1,2}} = C_1 \cup C_2$$

o1, o2 – объекты, описанные наборами свойств;

c_j^i – i-е свойство j-го объекта;

Схожесть двух объектов определяется как степень близости описывающих их бинарных векторов. Таким образом формируется оценка соответствия двух объектов по наличию свойств $s^1 \in \{0,1\}$, где 1 – объекты полностью совпадают, а 0 – полностью не совпадают.

Если оба объекта имеют свойство c^i, и в обоих случаях его значение определено, можно сравнить объекты по значению этого свойства. На данном этапе учитывалось только совпадение свойств. В результате для любого свойства может быть получена оценка $s_i^2 \in (0,1)$, где 1 – свойства полностью совпадают, а 0 – полностью не совпадают. Если у объектов есть n общих определенных свойств, то оценка соответствия двух объектов по значению свойств определяется как:

$$s^2 = \frac{\sum_{i=1}^{n} s_i^2}{n}$$

Мера схожести объектов определяется как $s_i = \alpha * s_i^1 + \beta * s_i^2$, где сумма весовых коэффициентов $\alpha + \beta = 1$. Для свойств, которые нельзя сравнить по значению, $\beta = 0$. Общая мера схожести двух объектов $s = \frac{\sum_{i=1}^{m} s_i}{m}$, где $m = \left|\overrightarrow{C_{1,2}}\right| = |C_1 \cup C_2|$.

Подобие прецедентов определяется на основании схожести их объектов. Так как каждый прецедент – это неупорядоченный набор объектов, на данном этапе схожесть прецедентов оценивается в зависимости от задачи:

1. Схожесть прецедентов для оценки защищенности путем тестирования на проникновение.

Так как наиболее важным в этом случае является возможность проникновения в систему, схожесть прецедентов определяется по наиболее схожим объектам, которые могут быть входной точкой атаки, то есть vмера схожести прецедентов Pi и Pj: $sim_{i,j} = \max(s_{k,l})$, где $s_{k,l}$ – мера схожести объектов $o_k \in P_i, o_l \in P_j$.

2. Схожесть прецедентов для оценки подобия между объектами информатизации.

В этом случае важно общее сходство систем, вычисляемое как средняя мера схожести объектов, с учетом числа объектов анализируемого прецедента:

$$sim_{i,j} = \frac{\sum s_{k,l}}{K}$$

где $s_{k,l}$ – мера схожести объектов $o_k \in P_i, o_l \in P_j$, K – число объектов во входном прецеденте. Число объектов в прецеденте из базы знаний не учитывается, так как важно не только общее полное сходство между системами, но и совпадение входного случая с какой-либо частью другого прецедента.

Результаты апробации блока анализа приведены в таблице 1. Типы объектов прецедентов не учитываются при анализе и даны для показания качества оценки схожести объектов. Стоит отметить, что на верху ранжированных списков схожести при экспериментальной проверке всегда находились объекты одного типа с исходным. Также в таблице приведены обе меры близости, рассчитанные для данного примера.

Важным вопросом является обучение системы в процессе использования необходимо для поддержания ее актуальности и соответствия современной картине предметной области в сфере информационной безопасности. Однако, для выбранного подхода обучение – сложная задача, в силу разнородности объектов; неполноты и, зачастую, субъективности задания прецедента. Для решения этой проблемы авторы предлагают использовать оценку важности конкретного значения (или важности значения) свойства каждого объекта. Для этого каждое свойство представляется как кортеж *<Measure, Value>*, где *Measure∈[0,1]*. При успешном сопоставлении текущего прецедента Pg, описывающего анализируемую систему, с имеющимся в базе знаний некоторым прецедентом Pa, то для всех объектов прецедента Pa, сходных с хотя бы одним объектом Pg, для всех сходных по наличию атрибутов будет выполняться увеличение важности их значения:

*Measure = Measure+(1- Measure)*0.5*

Для различающихся значений свойств предлагается снижение меры:

*Measure = Measure*0.5.*

Значение считается значимым при *Measure* > σ где σ – пороговое значение. Ведется тестирование этого подхода на наборе реальных прецедентов при оценке защищенности.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ПОИСКА ПОДОБНЫХ ПРЕЦЕДЕНТОВ

Номер преце-дента	**Тип объекта в тестовом наборе**	**Тип объекта в прецеденте**	**S**	**Sim (тип 1)**	**Sim (тип 2)**
P5	IP	IP	1,00	1,00	0,33
P0	PC	PC	0,85	0,93	0,86
	PC	PC	0,85		
	Application	Application	0,93		
	Application	Application	0,83		
	Application	Application	0,85		
	Application	Application	0,79		
	Network	Network	0,85		
	E-mail	E-mail	0,89		
P2	PC	PC	0,75	0,93	0,57
	PC	PC	0,75		
	Application	Application	0,93		
	Application	Application	0,83		
	Application	Application	0,85		
	Application	Application	0,79		
	Network	Network	0,85		
P1	PC	PC	0,75	0,85	0,78
	PC	PC	0,75		
	Network	Network	0,85		

V. ЗАКЛЮЧЕНИЕ

Авторами был разработан подход к анализу гетерогенных объектов при решении задач оценки защищенности. В качестве базы для построения аналитической системы был выбран прецедентный анализ. Авторами разработаны меры схожести объектов и прецедентов. Структура системы поддержки принятия решений на базе предложенного подхода включает компоненты предварительной обработки данных, анализа и обучения. Предложенные меры схожести хорошо показали себя при проведении экспериментальной проверки подхода.

Дальнейшими направлениями работы авторы видят доработку блока обучения системы поддержки принятия решений, совершенствование методов оценки подобия объектов и прецедентов и введение в оценку схожести прецедентов в явном виде учет их структуры.

СПИСОК ЛИТЕРАТУРЫ

[1] Tianmu L., Leiming Y. SIEM Based on Big Data Analysis // Cloud Computing and Security. 2017. Vol. 10602. P. 167-175. DOI https://doi.org/10.1007/978-3-319-68505-2_15

[2] Lee S., Shin Y. The Direction of Information Security Control Analysis Using Artificial Intelligence // Advances in Computer Science and Ubiquitous Computing. 2017. vol 474. P. 872-877 DOI https://doi.org/10.1007/978-981-10-7605-3_138

[3] Artificial Intelligence Techniques: A Comprehensive Catalogue / editor Bundy A. Berlin: Springer, 1997. 129 p. DOI https://doi.org/10.1007/978-3-642-60359-4

[4] Jackson P. Introduction to Expert Systems (3rd ed.). USA. Boston, MA,. Addison-Wesley Longman Publishing Co., 1998. 542 p.

[5] Осипов Г.С. Динамические интеллектуальные системы // Искусственный интелелкт и принятие решений. № 1, 2008. С. 47-54.

[6] Ажмухамедов И.М., Марьенков А.Н. Обеспечение информационной безопасности компьютерных сетей на основе анализа сетевого трафика // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2011. №1. С. 137-141.

[7] Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. NY.: Taylor and Francis Group, LLC. 2011. 248 p. DOI: 10.13140/RG.2.2.35197.26085

[8] Цветкова О.Л., Айдинян А.Р. Интеллектуальная система оценки информационной безопасности предприятия от внутренних угроз // Вестник компьютерных и информацонных технологий. №: 8. 2014. С. 48-53.

[9] Карпов Л.Е., Юдин В.Н. Адаптивное управление по прецедентам, основанное на классификации состояний управляемых объектов // Труды ИСП РАН. 2007. №2. С. 37-57.

[10] Жуков В.Г., Шаляпин А.А. Прецедентный анализ инцидентов информационной безопасности // Сибирский журнал науки и технологий. 2013. №2 (48). С. 19-24.

[11] Poltavtseva M. A., Pechenkin A. I. Intelligent Data Analysis in Decision Support Systems for Penetration Tests. // ISSN 0146-4116, Automatic Control and Computer Sciences, 2017, Vol. 51, No. 8, P. 985–991. DOI https://doi.org/10.3103/S014641161708017X

[12] Heterogeneous semi-structured objects analysis / Poltavtseva M.A., Zegzhda P.D. // Intelligent Systems Conference 2018 6-7 September 2018 / (unpublished)

[13] Wallach H.M. Topic modeling: beyond bag-of-words. // Proc. of the 23rd int. Conf. on Machine learning. Pittsburgh. ACM. New York 2006. P. 977–984. DOI: https://doi.org/10.1145/1143844.1143967

[14] Зегжда П.Д., Полтавцева М.А., Печенкин А.И., Лаврова Д.С., Зайцева Е.А. Проецедентный анализ гетерогенных слабо структурированных объектов в задачах информационной безопасности // Проблемы информационной безопасности. Компьютерные системы. №1. 2018. С. 17-32.