4. Айвазян С. А. Прикладная статистика. Основы эконометрики. Т.2. — М.: ЮНИТИ–ДАНА, 2001. — 432 с.
5. https://www.youtube.com/watch?v=vmEHCJofslg
6. https://proglib.io/p/web-scraping/
7. https://python-scripts.com/beautifulsoup-html-parsing

UDC 004.386

RESEARCH OF AUTOMATED DEVELOPMENT APPROACHES OF DOMAIN ONTOLOGIES

Anuar Turganbekov
turganbekovanuar@gmail.com

Master student, L. N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

Supervisor – A. M. Kankenova

Introduction

Starting from the beginning of the Information age, humanity advanced its technologies in multiple domains and number of unstructured and unconnected data has being growing exponentially. Different domain experts and information sources could not follow standards and come to a consensus [1]. In order to solve this significant problem, the concept of the Semantic Web, a hypertext web extension, was introduced. It implies that any data can be shared and integrated within the World Wide Web, having a form understandable by both humans and machines [2, 3].

Ontologies are considered as a further implementation of the Semantic Web concept and a connecting link for information systems with different structures and applications. They define a semantic structure of terms, which describe a data, and relations between those terms. Ontologies also certify that data content is consistent, shared and understandable for both human experts and machines [2, 4].

Manual and automated development approaches exist for constructing ontologies from sets of data. The first approach is considered time and resource consuming, and so most of ontology developers switched to the latter. Automated development of ontologies, also known as ontology learning, is based on methods from information technologies fields such as natural language processing (NLP), machine learning, data mining, information retrieval and knowledge representation. In order for an unstructured data to transform into an ontology, an ontology developer should apply ontology learning methodology in order to process that data and further evaluate developed ontologies [1].

Ontology learning methodology

1. Linguistics methods

Linguistics methods adapt methodology of linguistics onto formal language representation of ontologies and they are used in conjunction with statistical methods and inductive logic programming. Linguistics include pre-processing, terms and concepts extraction, and relations extraction methods categories [1].

1.1. Pre-processing

1.1.1. Part of speech tagging

Part of speech tagging allows matching words with their representative parts of speech and labelling them with tags [1]. As stated by [1], there were three implementations of this method reviewed: Brill Tagger, TreeTagger and Stanford CoreNLP API.

1.1.2. Sentence parsing

Sentence parsing is a part of syntactic analysis that discovers dependencies between every words and arranges them in a parsing tree structure [1]. The authors in [1] reviewed applications of this method in Principar, Minipar, Link Grammar Parser, Stanford Parser, GATE and Apache OpenNLP implementations.

1.1.3. Lemmatization

Using this method words may be reduced to their normal forms [1]. As observed in [1], this method was applied in Cornel API for a set of textual data and used with WordNet library.

1.2. Terms and concepts extraction

1.2.1. Syntactic analysis

In this method, data goes through several steps. Firstly, part of speech tagging is applied to sentences, and then syntactic structures are extracted and analyzed in order to gain terms [1].

1.2.2. Subcategorization

Subcategorization is a concept, which claims that a specific number of certain forms are selected and evaluated. Only those words are considered to create a concept [1].

1.2.3. Usage of seed words

Seed words are considered as base words, specific to a domain, which gives an opportunity for other implementations to extract similar domain-related terms. Only relevant and semantically close words are chosen [1].

1.3. Relations extraction

1.3.1. Dependency analysis

This method allows finding relations between terms in a corresponding parsing tree using their dependency data [1].

1.3.2. Lexicosyntactic patterns

According to this approach, regular expressions are used in order to extract similar phrases according to some patterns [1].

2. Statistical methods

Statistical methods do not consider semantics layer of ontologies, but apply probabilities on early stages of ontology development [1].

2.1. Terms and concepts extraction

2.1.1. C/NC value

C/NC value method evaluate multi-word terms and assign them scores. Scores depend on C value and NC Value. The first value corresponds to valid group of terms in the corpus. NC value is a modification of C value and used to find lengthier strings and group of terms that appear more frequently. The longest groups forms a set of concepts [1].

2.1.2. Latent semantic analysis (LSA)

The latent semantic analysis algorithm applies singular value decomposition on terms matrix in order to reduce its size while keeping the similarity structure. Terms which are part of one phrase are considered to have a similar meaning [1].

2.2. Relations extraction

2.2.1. Formal concept analysis (FCA)

According to this approach, object or concepts are related with their attributes or properties. Using this method, an attribute matrix of an object is used in order to find all clusters of attributes and objects. The result of this approach is a hierarchical structure of both objects and their properties [1].

2.2.2. Hierarchical clustering

The hierarchical clustering method allows to group data elements into clusters using an appropriate similarity measure such as cosine or Jaccard similarities [1]. Two strategies in the paper [1] are revised: agglomerative clustering and divisive clustering.

3. Inductive logic programming (ILP).

Inductive logic programming is considered as a machine learning subfield that allows acquiring hypotheses from existing knowledge and examples cluster using logic programming. ILP is applied at the last stages of ontology development and considers its terms as an initial data [1].

Literature

1. M. N. Asim, M. Wasim, M. U. G. Khan, et al. A survey of ontology learning techniques and applications. // Database. 2018. Vol. 2018. P. 1-15.
2. D. Vrandecic. Ontology Evaluation. // 2010. Karlsruhe Institute of Technology. P. 11-15, 190-195.
3. T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. // Scientific American. 2001. Vol. 5.

4. T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. // International Journal of Human-Computer Studies. 1995. Vol. 43 (5/6). P. 907-928.

УДК 519.876.2

ГРНТИ 81.93.29

МОДЕЛЬ РАСПОЗНАВАНИЯ АНОМАЛЬНОГО ПОВЕДЕНИЯ ПРИ VPN ПОДКЛЮЧЕНИЯХ К ИНФРАСТРУКТУРЕ

Тынымбаев Болат Айткожинович
tynymbaevba@gmail.com

Докторант 2 курса, ЕНУ им. Л.Н. Гумилева, Нур-Султан, Казахстан

Научный руководитель – А.А. Адамов

Введение. В условиях введения чрезвычайного положения, в марте-апреле 2020 года многими организациями на территории Республики Казахстан было принято решение перевести своих работников на удалённый режим работы. При этом наиболее удобным и обеспечивающим безопасное соединение является VPN подключение, при условии выполнения отдельных правил и требований информационной безопасности. Однако, модель угроз организации увеличивается при данном виде подключении работников к критичным информационным активам. В данной работе будет описан ландшафт угроз, а также представлена модель для защиты инфраструктуры организации с применением машинного обучения при анализе VPN подключений работников, а также классификации пользователей-работников на отдельные виды группы

Модели угроз удалённого подключения. Угрозы, которые должна рассматривать организация, при предоставлении удалённого доступа своим работникам можно разделить согласно двум видам активов:

- рабочие станции работников, подключающихся удалённо;

- удалённое сетевое соединение.

Подобные модели угроз описаны в следующих статьях [1-2]. Угрозы, объединенные в группы по схожести типов атак, представлены в таблице 1.

Таблица 1. Модель угроз при удалённом подключении

Угроза	Описание
Угроза взлома, присутствия	Поскольку на личные рабочие станции работников