UDC 004.82:004.62 **doi: 10.32620/aktt.2026.2.11**

Serhii DANOV, Igor SHOSTAK

National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine

ANALYSIS OF THE EFFICIENCY OF ONTOLOGY-ORIENTED APPROACHES TO BUSINESS INFORMATION EXTRACTION FROM UNSTRUCTURED WEB SOURCES RELATING TO AEROSPACE PRODUCTION ORGANIZATION

*The **subject** of this research is ontology-oriented approaches to extracting business information from unstructured web sources. The **aim** of the article is to analyze the effectiveness of modern ontology-oriented approaches for extracting business information from unstructured web sources and to substantiate their feasibility for decision support systems. Such approaches are particularly important for the information and analytical support of aerospace enterprises, where activities require processing significant volumes of heterogeneous external information on cooperative relations, supplies, technical product support, regulatory requirements, and the market environment. **Tasks** include: analyzing the main challenges of collecting, mining, and processing business information from unstructured web sources; determining the feasibility of using ontologies to extract and integrate business information; and performing a comparative analysis of modern ontology-oriented approaches and identifying promising areas for future application. The study employed **methods** for analysis and generalization scientific sources, systems analysis, and comparative analysis, along with approaches to semantic and ontological modeling. The findings establish that the main factors complicating web-based business information extractions are the heterogeneity of data formats and structures, the ambiguity of natural language, the dynamism of the information environment, and the incompleteness or inconsistency of information. It is demonstrated that the use of ontologies enables the semantization, structuring, logical coordination, and integration of business information within a corporate knowledge base, while providing a foundation for improving the quality of analytical data processing. Modern ontology-oriented approaches are categorized into template-based methods, deep linguistic analysis, and machine learning. A comparative analysis reveals that hybrid approaches, which combine the advantages of various methodologies to ensure greater completeness, flexibility, and semantic consistency, are the most promising for decision support systems. **Conclusions**. The scientific novelty of the obtained results lies in the generalization and comparison of modern ontology-oriented approaches to extracting business information from unstructured web sources, considering their suitability for semantic data coordination, the formation of corporate knowledge bases, and their application in decision support systems.*

Keywords: *business information; unstructured web sources; ontology; information mining; semantic integration; corporate knowledge base; decision support systems*.

1. Introduction

Digitalization of economic processes and rapid growth of data volumes in the web space lead to an increase in the role of business information as a strategic management resource. In the modern information environment, information that is significant for enterprises is formed not only in internal information systems, but also in the external digital circuit, which includes corporate web resources, news platforms, open reports, industry portals, electronic documents and other sources of business communication. Under such conditions, the problem of obtaining, structuring and interpreting business information becomes interdisciplinary and combines the tasks of information search, natural language processing, semantic data integration and knowledge management [1, 2]. This problem is especially relevant for aerospace companies, since decision-making in such organizations depends not only on internal production and technical data, but also on external sources, including supplier notifications, technical bulletins, regulatory documents, analytical market reviews, materials on cooperative relations and maintenance of complex equipment. Under such conditions, timely detection, semantic coordination and integration of external information become an important prerequisite for supporting managerial, logistical and technological decisions.

A significant feature of business information coming from the web space is its predominantly unstructured or semi-structured nature. In most cases, relevant information is not presented in a form directly suitable for automated analytical processing, but is distributed between text fragments, tables, web elements, PDF documents, dynamic pages and other types of content that require prior extraction, normalization and semantic alignment. Under such conditions, traditional approaches based on

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

lexical search or template extraction of fragments do not provide a sufficient level of accuracy, completeness and semantic consistency of results, especially when it comes to identifying entities, events and relationships in heterogeneous sources [3, 4].

In this context, ontologically oriented approaches to information extraction are of particular relevance, which allow us to move from superficial syntactic detection of text elements to the construction of formalized models of the subject area. In this approach, ontology serves as a conceptual framework that defines classes of entities, types of their relationships, attribute characteristics, interpretation constraints, and rules of logical agreement, creating the prerequisites for transforming disparate text information into structured knowledge. This, in turn, opens up opportunities for the formation of corporate knowledge bases, increasing the interpretability of analytics results, and using extracted information in decision support systems. From the standpoint of software engineering, this approach also creates a basis for building software systems in which the processes of collecting, integrating, updating, and using business information can be implemented as a coordinated interaction of functional components and multi-agent means of managing information and intellectual resources [2, 5].

1.1. Motivation

The motivation for the study is that in modern conditions, the effectiveness of decision-making support largely depends on the timely receipt, interpretation and integration of business information from a large number of external web sources. However, such information, as a rule, is distributed between disparate platforms, differs in presentation format, structure, completeness and reliability, which significantly complicates its use as part of corporate information and analytical systems. Under these conditions, the problem of creating software solutions that are capable not only of collecting data, but also of ensuring their further semantic coordination and integration into a single knowledge environment becomes particularly relevant [1, 6].

From the standpoint of software engineering, the main difficulty lies in the fact that existing solutions often implement only individual functions of working with business information: source search, fact extraction, knowledge graph construction or analytical interpretation of results. Such fragmentation complicates the construction of holistic systems in which the processes of knowledge collection, filtering, semantic analysis, integration, accumulation and updating should be organized as a coordinated interaction of components of a single architecture. That is why there is a need to use ontological models as a formal basis for representing the subject area and corporate knowledge base, which makes it possible to reduce semantic ambiguity and ensure interoperability of subsystems [3, 4].

An additional motivating factor is the variability of the web environment, which requires the software system to be able to adapt to new sources, new types of entities, new data formats and new requirements for analytics. In such conditions, a promising combination of ontologically-oriented approaches with multi-agent tools that allow distributing the functions of collecting, analyzing, coordinating, updating and controlling knowledge between specialized components is promising. These circumstances indicate the constructiveness of the outlined approach to organizing effective management of information and intellectual resources of an enterprise in conditions of heterogeneity of sources and variability of the environment and determine the feasibility of further research in this direction [5, 7].

In the context of the aerospace industry, these problems are exacerbated by the high complexity of cooperation processes, dependence on component suppliers, the need to take into account technical and operational documentation, as well as increased requirements for the relevance and consistency of knowledge used in decision-making. That is why the development of tools for semantic processing and integration of external information is important not only from the standpoint of general business analytics, but also for the tasks of information support of aerospace enterprises.

1.2. Publication Analysis

Analysis of current research shows that in recent years the problem of extracting business information from unstructured web sources has been considered mainly within the framework of several interrelated areas: generative information extraction, ontology engineering, semantic data integration, knowledge graph construction, knowledge-based analytics and retrieval-augmented generation. At the same time, most of the existing works focus either on separate algorithms for extracting entities, relationships and events, or on building semantic models for already prepared data. This does not allow to fully form a holistic software engineering approach to collecting, coordinating, integrating and further using business information in corporate decision support systems [3], [8].

In the article [3], large language models are considered as the basis of generative information extraction and approaches to solving the tasks of named entity recognition, relation extraction, event extraction and other forms of structured knowledge extraction from text are systematized. The authors show that the generative paradigm significantly expands the possibilities of processing unstructured sources, but the main emphasis of the work is on the classification of models and subtasks of

information extraction. However, the work still leaves unresolved the issue of integrating the extracted information into corporate knowledge bases and its coordination with the formalized model of the subject area.

The article [5] considers the use of knowledge graphs to support descriptive business analytics, in particular to capture provenance, preprocessing stages, and the context of analytical procedures. The authors argue that knowledge graphs can serve as an environment for accumulating and explaining knowledge use in analytical systems. However, the article still leaves unresolved the issue of automated collection and primary extraction of business information from heterogeneous web sources, which is critically important for building corporate decision support systems.

The work [8] contains a technology for semi-automatic construction of ontologies from unstructured texts based on a combination of NLP procedures, knowledge graph and ontology version control tools. The authors demonstrate that even for non-expert users it is possible to significantly reduce the complexity of constructing a semantic model of a subject area by automating the stages of entity recognition, relationships and ontology enrichment. However, in this work, the issue of adapting such an approach to the continuous collection of business information from the web space and its further integration into the corporate knowledge base in conditions of source variability still remains unresolved.

In the article [9], trusted knowledge extraction for operations and maintenance intelligence tasks is considered and 16 NLP tools and LLM solutions are evaluated on real data from the technical maintenance domain. The authors show that building a knowledge extraction pipeline for critical applications requires taking into account trust in the tools, confidentiality constraints and real readiness of technologies for implementation. However, the work still leaves unresolved the issue of formalized ontological representation of extracted knowledge and its integration into a single corporate model suitable for long-term use in business analytics tasks.

In paper [10], the authors proposed a knowledge graph-driven framework for multi-party synergistic interaction of operation and maintenance participants in complex products. The authors developed a formal ontological method for representing heterogeneous data of several stakeholders and supplemented it with a navigation model to reduce information ambiguities. However, it still leaves unresolved the issue of using such an approach specifically for collecting and extracting business information from the open web space, rather than from existing domain data.

The authors of the article [11] propose a knowledge intelligence management method for manufacturing enterprises, which includes automatic knowledge graph construction, its refinement and a natural language interaction mechanism based on LLM. The authors show that the combination of LLM, knowledge graph and agent functions can improve the quality of knowledge management within the enterprise and simplify access to knowledge in new business scenarios. However, the issue of transferring this approach to the task of semantic reconciliation of business information from heterogeneous external web sources and its integration into a corporate knowledge base in a more general software engineering context remains unresolved in the work.

The article [12] considers a framework that combines knowledge graph construction and retrieval-augmented generation for complex information support in a crisis environment. The authors demonstrate that the integration of knowledge graph with the retrieval-augmented generation approach allows to increase the completeness and contextual relevance of answers, reducing the dependence on isolated text search. However, the issue of building an ontological model of business information and organizing a software system capable of supporting its continuous replenishment, coordination and use in decision-making processes remains unresolved.

Thus, the conducted analysis of publications gives grounds to assert that modern research confirms the feasibility of using ontological models, knowledge graph, LLM and related tools to solve individual problems of information extraction and interpretation. At the same time, in most works, the issue of building a holistic software system that would combine the collection of business information from unstructured web sources, its semantic coordination, integration into a corporate knowledge base and further use with the involvement of multi-agent tools for managing information and intellectual resources remains insufficiently addressed. This is what determines the feasibility of further research within the chosen topic.

1.3. State of the Art

Despite the development of modern data processing tools, the problem of automated extraction of business information from unstructured web sources remains insufficiently solved at the level of building holistic software systems. The main difficulty lies in the fact that business-significant information comes from sources that differ in structure, presentation formats, access methods, level of completeness and update speed. As a result, software tools built on the basis of only local search or isolated fact extraction algorithms do not provide proper consistency of results and do not form a single environment for the further use of knowledge in decision support [1].

Solving this problem is complicated by the fact that unstructured web sources contain semantically ambiguous, fragmented, and context-dependent information that

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

cannot be directly integrated into corporate information systems. Even in cases where individual methods allow to detect entities, events, or relationships, the results of such extraction remain weakly connected to each other if the system lacks a formalized model of the subject domain. This reduces the interoperability of software components, complicates the reuse of knowledge, and does not allow to ensure the proper level of semantic data consistency [4].

From the standpoint of software engineering, the disadvantage of most existing approaches is their fragmented implementation, in which the modules for collecting, analyzing, integrating, storing and updating information function as loosely coupled subsystems. Under such conditions, it is difficult to scale the system, adapt it to new sources, maintain the relevance of the corporate knowledge base and control the consistency of the accumulated information. This necessitates the use of ontological models as the basis for an integrated representation of business information and the use of multi-agent tools to coordinate the processes of collecting, semantic matching, updating and using knowledge [5].

Thus, the scientific and applied tasks is to develop ontological models, methods of collecting, extracting, semantic matching and integrating business information in the web space using multi-agent means of managing information and intellectual resources, which ensure the formation and updating of the corporate knowledge base in the conditions of heterogeneity of sources and variability of the environment to increase the efficiency of decision-making support processes, in particular in the information and analytical circuits of enterprises in the aerospace industry, where the tasks of monitoring the external environment, cooperative relations, supply, maintenance of technical documentation and assessment of the state of production and operational processes are important. It is the solution of this problem that determines the logic of further research and the need to develop appropriate models, methods and architectural solutions.

1.4. Objectives and Tasks

The **purpose** of the article is to analyze the effectiveness of ontologically-oriented approaches to extracting business information from unstructured web sources and to justify the feasibility of their use as a basis for building software systems focused on semantic data reconciliation, the formation of a corporate knowledge base, and decision support.

To achieve the goal, within the framework of this publication it is necessary to solve the following **tasks**:

1. Analyze the main problems of collecting, extracting, and processing business information from unstructured web sources.
2. To identify the possibilities of using ontologies in the tasks of extracting and integrating business information from unstructured web sources.
3. To assess the effectiveness of modern ontology-oriented approaches to extracting business information from unstructured web sources and identify promising areas of their use in decision support systems.

2. Ontological Models, Methods and Tools for Extracting Business Information from Unstructured Web Sources

2.1. Features of Collecting, Extracting and Processing Business Information from Unstructured Web Sources

Automated collection and extraction of business information from the web space is a complex scientific and applied task, which is due to the fundamental properties of the web content itself. Unlike data stored in internal corporate systems according to relatively stable schemes, web sources form a dynamic, distributed and heterogeneous information environment [1]. For business analytics, this means that significant information about companies, markets, products, financial events, personnel changes or partnerships can be presented in different formats, in different contexts and with different degrees of completeness [4]. As a result, the task of information extraction itself goes beyond the usual search for text fragments and requires a comprehensive approach to their detection, interpretation, coordination and integration. This is especially true in the aerospace industry, where external information can relate not only to the market and partnerships, but also to product support, changes in regulatory requirements, the status of component supplies, and cooperation between enterprises.

One of the key problems is the heterogeneity of web data. It manifests itself both at the level of formats and at the level of structure and semantics of sources:

1. Heterogeneity of formats: information can be presented in the form of:

- HTML pages: news articles, press releases, corporate blogs that require cleaning of markup, navigation elements, and advertising;
- PDF documents: annual reports, analytical studies, official documents, which often contain complex tables and graphs; for the aerospace industry, such sources also include technical bulletins, test reports, certification documents, specifications, and component catalogs - extracting information from which requires specialized tools;
- structured data: JSON responses from social media APIs (Twitter, LinkedIn), XML news feeds (RSS), microdata (Schema.org) for describing products or organizations; in applied tasks of the aerospace industry,

these can also be supplier registers, catalogs of components and assemblies, as well as metadata of technical products;

- plain text: comments on forums, reviews on websites.

2. Structural heterogeneity: even within the same format, the structure can vary dramatically, for example, two news sites may have completely different HTML structures for presenting the same news content. This makes it difficult to create universal "wrappers" for data extraction.

3. Semantic heterogeneity: the same concept can be expressed in different ways, for example, a company's profit can be called "revenue", "sales", "turnover" or "income". The system must understand that all these terms refer to the same concept.

4. Language heterogeneity: business information is global, for example, important news about a European company may first appear in German and then be translated into English. The system must be able to handle multilingual content.

The next significant problem is the ambiguity of natural language, which is the main carrier of business information in the web space. Ambiguity manifests itself on several levels:

− at the lexical level, a single word can have multiple meanings depending on the context, for example, the term “Apple” can refer to a technology corporation, a fruit, or a personal name. In the business domain, a typical example is the word “Ford,” which can refer to an automobile company, the actor Harrison Ford, or a geographical concept;

− at the level of synonymy, different words and phrases can represent the same entity, as is the case with company names, positions or brands, for example, “International Business Machines”, “IBM” and “Blue Giant” are designations of the same corporation. Similarly, the expressions “CEO”, “Chief Executive Officer” and “chief executive officer” designate the same managerial position;

− at the level of anaphoric connections, instead of direct mention, pronouns (“he”, “she”, “company”) or descriptive constructions (“tech giant”, “the company from Cupertino”) are used to refer to previously mentioned objects. All this complicates the correct correlation of textual mentions with specific entities of the subject area;

− at the level of syntactic ambiguity, several possible interpretations of relationships between objects can be created, which directly affects the quality of the extracted facts. For example, in the sentence “Company X sold a division of Company Y with its assets,” it is difficult to unambiguously determine which party to the transaction the assets belong to.

The problem of ambiguity is also closely related to the dynamism of the web environment. Unlike static text corpora, the web space is constantly changing: new companies, products, technologies and events appear, the characteristics of already known entities change, rebranding, personnel changes and transformation of market relations occur. As a result, dictionaries, reference books and previously built knowledge resources quickly lose their relevance. In addition, the use of terms itself changes, that is, conceptual drift occurs when the previously established meaning of a certain token is transformed under the influence of a new business context. This means that the business information extraction system should be focused not only on one-time extraction of information, but also on constant updating of knowledge, adaptation to new entities and rethinking of existing descriptions.

Another important factor is incompleteness and inconsistency of data. In a real web environment, information about the same entity or event often turns out to be distributed among several sources, each of which contains only a part of the necessary information. For example, one message may record the fact of the appointment of a new manager, but not name the person, while another source provides a name, but does not specify the position or context of the appointment. To form a holistic picture of the event, it is necessary to integrate these fragments into a single knowledge structure. At the same time, different sources may provide incompatible or mutually exclusive data, in particular regarding financial indicators, time characteristics of events or the status of corporate relations. Under such conditions, the system must not only extract facts, but also perform procedures for their coordination, assessment of reliability and resolution of conflicts between sources.

Thus, the collection and extraction of business information from unstructured web sources is complicated by the combined effect of several interrelated factors: the heterogeneity of formats and structures, the semantic ambiguity of natural language, the dynamism of the information environment, as well as the incompleteness and inconsistency of data (Fig. 1). All this indicates that an effective solution to such a problem is impossible within the framework of only traditional search or template approaches. It is necessary to transition to models and methods that provide semantic interpretation, integration of information from different sources and the formation of a consistent representation of knowledge suitable for further use in decision support systems.

2.2. Using Ontologies in Business Information Extraction and Integration Processes

The ontology of the subject area acts as a formal model of knowledge, which creates the basis for the transition from syntactic processing of text to its semantic in-

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

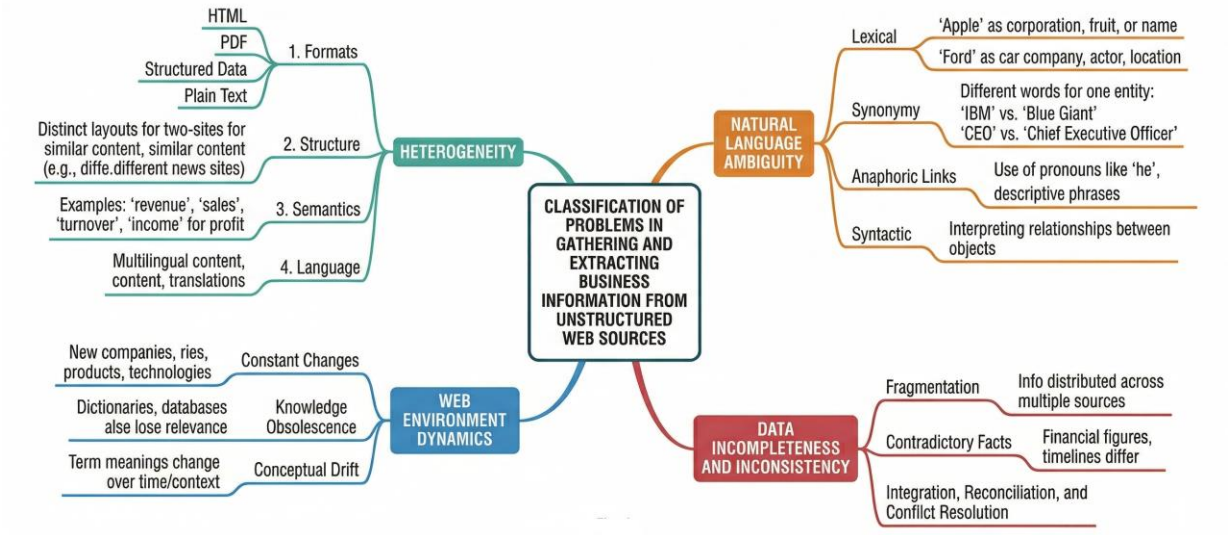

Fig. 1. Classification of problems in gathering and extracting business information from unstructured web sources

terpretation. In the tasks of extracting business information, this is of fundamental importance, since it allows working not only with individual words or fragments of text, but with formalized entities, relations, attributes and events relevant to the subject area [13]. If unstructured text is a stream of symbols and language constructs, then the ontology defines a conceptual scheme through which the system can interpret this stream as an ordered knowledge environment. That is why it is advisable to consider ontologies as one of the key means of semantizing business information in the web space [14].

The formal nature of an ontology is determined by its structural components. These include classes that represent types of objects in the domain; properties that describe relationships between objects or their attributes; individuals as specific instances of classes; and axioms that specify logical rules and constraints. Unlike traditional directories or classifiers, an ontology not only fixes a list of concepts, but also determines the way they are linked within a single knowledge space [15]. This creates the conditions for the formation of a coherent model of the domain, in which business entities can be linked not only by the fact of mentioning them, but also on the basis of logically justified semantic relations.

One of the most important tasks solved with the help of ontologies is the elimination of ambiguity during the interpretation of textual references. In business information, this problem arises constantly, since one and the same lexical unit can have several meanings, and one and the same entity can be denoted by different names, abbreviations or descriptive constructions [16, 17]. In this context, an important role is played by the grounding procedure, or entity linking, which consists in comparing the mention of an entity in the text with its unique identifier in the knowledge base. Such a process involves first generating a set of possible candidates, and then choosing the correct option based on the context. Thanks to this, the system is able to establish that different forms of mention can refer to the same company, person or product, and therefore correctly integrate information from different sources.

After disambiguation, ontology performs the next important function - knowledge structuring. It is not just about the accumulation of facts, but about the organization of these facts in the form of a logically connected system [18]. Taxonomic relations allow you to build hierarchies of concepts, which creates the opportunity to carry out search and analysis at different levels of generalization. At the same time, relations of the "part–whole" type allow you to model the internal structure of complex objects, in particular companies and their divisions, and the characteristics of properties in ontological description languages expand the expressive capabilities of the model. Thanks to this, ontology becomes not just a means of ordering terms, but a full-fledged tool for formalizing complex relationships between entities, which is especially important in the business domain.

Another fundamental advantage of ontologies is the support of logical inference. Based on already recorded facts and defined rules, the system can automatically obtain new, previously implicit knowledge [19]. This allows not only to classify objects by indirect features, but also to detect logical contradictions and establish indirect connections between entities. In the business analytics context, this possibility is especially valuable, since real data often comes in fragmented form, and important conclusions are formed not from a single message, but from a set of several facts. Thus, logical inference turns the ontological model from a passive knowledge repository into an active tool for analytical interpretation.

The final and most applicable aspect of using ontologies is data integration through a common semantic model [20]. In a real web environment, facts about the same entity or event can be distributed across multiple sources, expressed in different words, and presented in different formats. An ontology allows these facts to be associated with the same concepts and relationships, resulting in a single, consistent knowledge base. This means that disparate messages, reports, tables, and textual references can be integrated into a common semantic representation that is machine-readable, structured, and amenable to further analysis.

Thus, ontologies perform a complex function in the processes of working with business information: they provide semantics of text data, disambiguation, knowledge structuring, logical inference and integration of information from various sources. This makes them not just an auxiliary tool for describing a subject area, but a fundamental basis for building corporate knowledge bases and software systems focused on supporting decision-making. In the context of this article, this gives grounds to consider ontologies as a key element of ontology-oriented approaches to extracting business information from unstructured web sources.

2.3. Comparative Analysis of Ontology-Oriented Approaches to Extracting Business Information from Unstructured Web Sources

Modern ontology-oriented approaches to extracting business information from unstructured web sources should be considered as a set of methods that combine natural language processing tools with formal domain knowledge models [21, 22]. In general, such approaches can be divided into three main groups: template-based methods, methods based on deep linguistic analysis, and methods based on machine learning. Each of these groups has its own advantages, limitations, and areas of effective application, and their comparison allows us to assess the suitability of the corresponding solutions for use in decision support systems. Hybrid approaches should be distinguished separately, which combine the strengths of several groups of methods [7]. A generalized classification of such approaches is presented in Fig. 2.

The main groups of ontologically oriented approaches include:

- template-based methods;
- methods based on deep linguistic analysis;
- methods based on machine learning.

Template-based methods are among the first in the development of information mining. They involve the use of predefined linguistic or syntactic templates that are associated with ontology concepts and allow the detection of certain types of facts in the text. Such approaches can be implemented through manual rule creation by domain experts or through semi-automatic pattern training using bootstrapping procedures. In the first case, the system receives highly accurate but rigidly fixed rules; in the second, the share of manual work is reduced, but the risk of accumulating errors increases during the iterative expansion of the set of templates.

The advantages of template methods are as follows:

- high accuracy within well-formalized scenarios;
- complete transparency and clarity of the system's operation;
- the possibility of expert control of extraction rules.

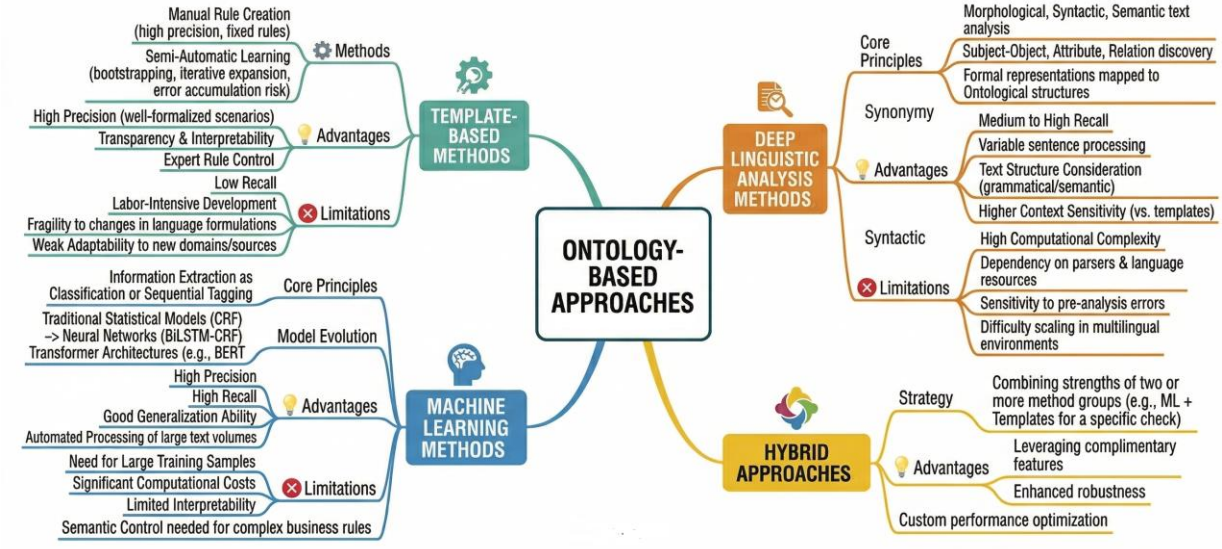

Fig. 2. Classification of ontology-based approaches to business information mining

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

The main disadvantages of template methods: low completeness of results; high complexity of development; fragility regarding changes in language formulations; weak adaptability to new domains and sources.

In contrast to template-based approaches, methods based on deep linguistic analysis rely on detailed morphological, syntactic and semantic parsing of the text. They allow to obtain formal representations of sentences, which are then compared with ontological structures. In practice, this means that the system not only records the sequence of words, but also tries to determine which objects are subjects of an action, which are its objects, what attributes accompany the event and how these elements can be related to the relations defined in the ontology. This approach provides greater flexibility and allows working with more complex language constructs than is possible within the framework of rigid templates.

The methods of deep linguistic analysis are characterized by:

- medium or high completeness of extraction;
- better ability to work with variable sentences;
- taking into account the grammatical and semantic structure of the text;
- higher context sensitivity compared to template methods.

However, such methods also have significant limitations: high computational complexity; dependence on the quality of parsers and language resources; sensitivity to errors in previous stages of analysis; the difficulty of scaling in a multilingual environment.

Machine learning methods treat the problem of information extraction as a classification or sequential labeling problem. It is this group of approaches that today demonstrates the best quality indicators in the problems of named entity extraction, relationship classification and event detection. Starting from classical statistical models, such as CRF, the development of this direction has led to the use of neural networks of the BiLSTM-CRF type and transformer architectures, in particular BERT-like models. The advantage of these methods is the ability to take into account a wide context, generalize to new texts and effectively work with large volumes of data in which events and relationships are described by various linguistic means.

Strengths of machine learning methods:

- high accuracy;
- high completeness;
- good ability to generalize;
- suitability for automating the processing of large amounts of text.

The main limitations of these methods are: the need for large training samples; significant computational costs; limited interpretability of results; difficulty of use without additional semantic control.

A comparison of the above groups of methods shows (Table 1) that none of them can be considered a universal solution. Template approaches are the most transparent, but the least flexible. Deep linguistic analysis methods better reflect the structure of the text, but require a complex linguistic infrastructure. Machine learning methods provide the best results in terms of accuracy and completeness, but at the same time reduce the level of explainability and require significant resources for training and support. That is why it is advisable to evaluate the effectiveness of modern approaches not by a single indicator, but by a set of characteristics, among which the key ones are accuracy, completeness, complexity of implementation, requirements for input data, adaptability and interpretability.

In view of the above, the most promising approaches for decision support systems are hybrid approaches that combine the strengths of different classes of methods. In such systems, machine learning models can be used for the initial extraction of entities and relations, deep linguistic analysis tools for specifying structural dependencies, and ontology and rules based on it for typing, filtering, validation, and semantic matching of results. It is this organization that allows combining high completeness of extraction with interpretability, logical consistency, and suitability for integration into a corporate knowledge base. Therefore, promising areas for further use of modern ontologically oriented approaches in decision support systems are primarily associated with the development of hybrid software solutions, in which the ontology acts as a semantic framework for collecting, integrating, and using business information.

3. Results and Discussion

As a result of the conducted research, it was found that the extraction of business information from unstructured web sources is a complex multi-component task, the quality of the solution of which is significantly affected by the heterogeneity of formats, the semantic ambiguity of natural language, the dynamism of the web environment, as well as the incompleteness and inconsistency of data. Unlike traditional information retrieval, in this case it is not enough to simply identify relevant text fragments, since the further use of such information in decision-making support systems requires its normalization, semantic coordination, integration and structured representation within a single knowledge environment. This gives grounds to consider the task of extracting business information not as a separate NLP procedure, but as an element of a broader software engineering problem associated with the construction of integrated information and analytical systems.

Table 1

Comparative characteristics of modern ontology-oriented approaches to business information mining

No.	Criterion	Template-based methods	Methods based on deep linguistic analysis	Machine learning-based methods	Hybrid approaches
1	Basic principle	Using rules and templates related to ontology	Morphological, syntactic and semantic analysis of the text with subsequent comparison with the ontology	Training models for recognizing entities, relationships, and events	Combining ML models, linguistic analysis, rules, and ontology
2	Precision	High in narrow scenarios	Medium or high	High	High
3	Completeness	Low	Medium	High	High
4	Interpretability	High	Medium	Low	Medium or high
5	Adaptability	Low	Medium	High	High
6	Resource requirements	Expert rules and manual support	Linguistic resources, parsers, dictionaries	Large training samples and computational resources	Combined resources
7	Implementation complexity	Medium	High	High	High
8	Suitability for integration into the knowledge base	Limited	Moderate	Moderate	High
9	Suitability for semantic matching	Limited	Moderate	Moderate	High
10	Feasibility for decision support systems	Partially	Moderate	High	Highest

The analysis showed that the use of ontologies is appropriate for solving the above problems, since it is the ontological model that allows us to move from fragmentary processing of textual references to a formalized representation of business entities, events, attributes and relationships between them. The paper summarizes that ontologies provide several fundamentally important functions: eliminating ambiguity by matching textual references with specific entities, structuring knowledge using taxonomies and formal relations, supporting logical inference, and integrating facts from disparate sources into a common semantic model. It is thanks to these properties that ontologies can be considered as a basic tool for forming a corporate knowledge base suitable for further use in analytical and management tasks.

The study also systematizes modern ontology-oriented approaches to business information mining and compares them. It is found that template methods provide high accuracy and transparency in well-formalized domains, but are inferior in completeness and adaptability. Deep linguistic analysis methods better take into account sentence structure and context, but are difficult to implement and depend on the quality of language resources. Machine learning methods, especially transformative models, demonstrate the best results in terms of accuracy and completeness, but are characterized by high data requirements and limited interpretability. Therefore, none of the considered approaches is universal, and their practical value is determined not only by the quality of extraction, but also by their suitability for integration into decision support software systems.

Comparative analysis allows us to assert that the most promising are hybrid approaches that combine the strengths of different groups of methods. In such solutions, machine learning models can provide the initial detection of entities, relations and events, linguistic analysis methods can specify structural dependencies in the text, and ontology can perform the functions of normalization, typing, logical validation and integration of results into the corporate knowledge base. It is this organization that allows us to coordinate the requirements for accuracy, completeness, interpretability and suitability of results for further use in decision support systems. Thus, the work substantiates that the further development of ontologically oriented approaches should be associated not with the isolated improvement of individual methods, but with the construction of hybrid software solutions capable of integrating heterogeneous web sources into a

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

single, homogenized knowledge environment. In the applied aspect, the obtained results can be used in the creation of information and analytical subsystems for enterprises in the aerospace industry. In particular, we are talking about the tasks of monitoring the external environment in order to determine the market situation, analyze cooperative relations, assess the status of component supplies, maintain technical documentation, identify critical changes in the regulatory framework, and support management decisions based on integrated knowledge. This allows us to consider ontologically oriented and hybrid approaches not only as theoretically justified, but also as practically relevant for the information support of complex high-tech industries.

The results obtained are also important in the broader context of future research, as they create a basis for the formation of an ontological model of integrated representation of business information, the development of methods for its collection, coordination and integration, as well as for further substantiation of means of managing information and intellectual resources. In this sense, the article serves as a conceptual basis for the next stages of research, in which the emphasis will already shift from comparative analysis to the construction of our own models, methods and architectural solutions.

4. Conclusions

The article analyzes the main problems of collecting, extracting and processing business information from unstructured web sources. It is established that the complexity of this task is due to the heterogeneity of data formats and structures, the ambiguity of natural language, the dynamism of the web environment, as well as the incompleteness and contradictions of information coming from different sources. It is shown that under such conditions, traditional approaches to search and local fact extraction do not provide the proper level of semantic coordination and data integration necessary for further use in decision support systems.

The possibilities of using ontologies in the tasks of extracting and integrating business information from unstructured web sources are determined. It is substantiated that ontologies should be considered as a means of semantizing text information, eliminating ambiguity, structuring knowledge, logical inference and integrating facts within a single corporate knowledge base. It is the ontological model that provides a formal representation of the subject area, which makes it possible to transition from fragmentary text references to a coherent knowledge environment suitable for analytical use.

As a result of a comparative analysis of modern ontologically-oriented approaches to business information extraction, it was found that template methods, methods of deep linguistic analysis and machine learning methods have different advantages and limitations, and therefore cannot be considered as interchangeable universal solutions. It is shown that the most promising for decision support systems are hybrid approaches that combine high completeness of machine learning, structurality of linguistic analysis and semantic reliability of ontological matching. This gives grounds to consider hybrid ontologically-oriented solutions as the most appropriate direction for the further development of systems for extracting business information from unstructured web sources.

The practical significance of the results obtained lies in the possibility of their use in the development of software tools for semantic coordination and integration of external information in information and analytical systems of aerospace enterprises. In particular, the proposed generalizations regarding the use of ontologies, knowledge graphs and hybrid approaches can be useful for building information and analytical subsystems for monitoring suppliers, analyzing cooperation relationships, supporting technical documentation, processing messages about the technical condition of products and supporting decision-making in the management processes of complex aerospace systems.

Contributions of authors: conceptualization, methodology – **Igor Shostak;** formulation of research goals and objectives – **Igor Shostak;** conducting research on the current state – **Serhii Danov;** interpretation of results – **Serhii Danov, Igor Shostak.**

Conflict of Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, author ship or otherwise, that could affect the research and its results presented in this paper.

Author **Igor Shostak** is a member of the Editorial Board of this journal. He were not involved in the peer review, handling, or decision-making process for this manuscript.

Financing

This study was conducted without financial support.

Data Availability

The work has associated data in the data repository.

Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence methods while creating the presented work.

All the authors have read and agreed to the published version of this manuscript.

References

1. Hoseini, S., Theissen-Lipp, J., & Quix, C. A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. *Journal of Web Semantics*, 2024, vol. 81, article no. 100819. DOI: 10.1016/j.websem.2024.100819.

2. Zeroual, S., Nessah, D., & Bakhouche, A. A systematic literature review on ontology-driven business intelligence components. *The Electronic Journal of Knowledge Management*, 2026, vol. 24, no. 1, pp. 101–118. DOI: 10.34190/ejkm.24.1.4277.

3. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 2024, vol. 18, article no. 186357. DOI: 10.1007/s11704-024-40555-y.

4. Scannapieco, S., & Tomazzoli, C. Cnosso, a novel method for business document automation based on open information extraction. *Expert Systems with* Applications, 2024, vol. 245, article no. 123038. DOI: 10.1016/j.eswa.2023.123038.

5. Staudinger, S., Schütz, C. G., Schrefl, M., & Neuböck, T. Knowledge graph support for descriptive business analytics. *Decision*, 2025, vol. 52, no. 3, pp. 285–306. DOI: 10.1007/s40622-025-00432-4.

6. Arslan, M., Munawar, S., & Cruz, C. Business insights using RAG–LLMs: a review and case study. *Journal of Decision Systems*, 2024, pp. 1–30. DOI: 10.1080/12460125.2024.2410040.

7. Shim, M., Kim, H., Choi, Y., Kim, J., & Lee, J. OmEGa(Ω): Ontology-based information extraction framework for constructing task-centric knowledge graph from manufacturing documents with large language model. *Advanced Engineering Informatics*, 2025, vol. 64, article no. 103001. DOI: 10.1016/j.aei.2024.103001.

8. Amdouni, E., Belfadel, A., Gagnant, M., Renault, I., Kierszbaum, S., Carrion, J., Dussartre, M., & Tmar, S. Semi-Automatic Building of Ontologies from Unstructured French Texts: Industrial Case Study. *Data Science and Engineering*, 2025, vol. 10, no. 3, pp. 339–361. DOI: 10.1007/s41019-025-00284-z.

9. Mealey, K. P., Karr Jr, J. A., Moreira, P. S., Brenner, P. R., & Vardeman II, C. F. Trusted knowledge extraction for operations and maintenance intelligence. *Natural Language Processing Journal*, 2025, vol. 13, article no. 100187. DOI: 10.1016/j.nlp.2025.100187.

10. Zhao, X., Wang, R., Ren, S., Zhang, G., & Zhang, Y. A knowledge graph-driven framework of multi-stakeholder synergistic operation and maintenance for complex products: design, implementation and industrial validation. *Advanced Engineering Informatics*, 2025, vol. 68, article no. 103746. DOI: 10.1016/j.aei.2025.103746.

11. Liu, Z., Hu, B., Feng, Y., Lu, C., & Tan, J. Making manufacturing knowledge graph more intelligent: A knowledge intelligence management method for manufacturing enterprises. *Advanced Engineering Informatics*, 2026, vol. 71, article no. 104264. DOI: 10.1016/j.aei.2025.104264.

12. Yao, L., Ren, F., Du, K., & Du, Q. From knowledge graph construction to retrieval-augmented generation: a framework for comprehensive earthquake emergency support. *Geo-spatial Information Science*, 2025 vol. 29, no. 1, pp. 509–529. DOI: 10.1080/10095020.2025.2514813.

13. Golubeva, A. A rapid review on ontology- and data-driven business process modelling. *New Trends in Computer Sciences*, 2025, vol. 3, no. 2, pp. 83–99. DOI: 10.3846/ntcs.2025.24801.

14. Stănescu, G., & Oprea, S.-V. Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling. *Electronics*, 2025, vol. 14, no. 7, article no. 1313. DOI: 10.3390/electronics14071313.

15. Muppasani, B. C., Gervet, C., & De Giacomo, G. Building a planning ontology to represent and exploit planning domain knowledge. *Discover Artificial Intelligence*, 2025, vol. 5, article no. 93. DOI: 10.1007/s44248-025-00093-9.

16. Scharpf, P. Entity linking with Wikidata: a systematic literature review. *ACM Computing Surveys*, 2024, vol. 56, no. 12, article no. 302, pp. 1–38. DOI: 10.1145/3617696.

17. Gohourou, D., & Kuwabara, K. Knowledge Graph Extraction of Business Interactions from News Text for Business Networking Analysis. *Machine Learning and Knowledge Extraction*, 2024, vol. 6, no. 1, pp. 126–142. DOI: 10.3390/make6010007.

18. Ambalavanan, R., Snead, R. S., Marczika, J., Towett, G., Malioukis, A., & Mbogori-Kairichi, M. Ontologies as the semantic bridge between artificial intelligence and healthcare. *Frontiers in Digital Health*, 2025, vol. 7, article no. 1668385. DOI: 10.3389/fdgth.2025.1668385.

19. Liu, H., Yang, S., Shi, G., & Miao, Z. Knowledge graph reasoning: Mainstream methods, applications and prospects. *Engineering Applications of Artificial Intelligence*, 2025, vol. 159, part B, article no. 111625. DOI: 10.1016/j.engappai.2025.111625.

20. Parsanasab, E., Ahmadipour, A., & Mehraeen, E. Utilization of Ontology to Develop Artificial Intelligence Systems in the Healthcare Industry. *Healthc Inform Res*, 2025, vol. 31, no. 4, pp. 320-330. DOI: 10.4258/hir.2025.31.4.320.

21. Hadji, A., & Kholladi, M. K. Disaster Information Extraction: Evaluation of NLP Techniques Using

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

JAPE Rules, Ontologies and Machine Learning Approaches. *In: Proceedings of the 3rd International Conference on Computer Science's Complex Systems and Their Applications*, 2025, pp. 294–313. DOI: 10.1007/978-3-031-90758-6_22.

22. Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. A Survey of Information Extraction Based on Deep Learning. *Applied Sciences*, 2022, vol. 12, no. 19, article no. 9691. DOI: 10.3390/app12199691.

Отримано 11.01.2026, отримано у доопрацьованому вигляді 05.03.2026
Дата ухвалення 15.04.2026, дата публікації 22.04.2026

АНАЛІЗ ЕФЕКТИВНОСТІ ОНТОЛОГІЧНО-ОРІЄНТОВАНИХ ПІДХОДІВ ДО ВИДОБУВАННЯ БІЗНЕС-ІНФОРМАЦІЇ З НЕСТРУКТУРОВАНИХ ВЕБДЖЕРЕЛ

С. О. Данов, І. В. Шостак

Предметом дослідження є онтологічно-орієнтовані підходи до видобування бізнес-інформації з неструктурованих вебджерел. **Метою** статті є аналіз ефективності сучасних онтологічно-орієнтованих підходів до видобування бізнес-інформації з неструктурованих вебджерел та обґрунтування їх доцільності для використання в системах підтримки прийняття рішень. **Завдання**: проаналізувати основні проблеми збору, видобування та оброблення бізнес-інформації з неструктурованих вебджерел; визначити можливості використання онтологій у задачах видобування та інтеграції бізнес-інформації; виконати порівняльний аналіз сучасних онтологічно-орієнтованих підходів та визначити перспективні напрями їх подальшого використання. У ході дослідженні застосовано **методи** аналізу та узагальнення наукових джерел, системного аналізу, порівняльного аналізу, а також підходи семантичного та онтологічного моделювання. У **результаті** встановлено, що основними чинниками, які ускладнюють видобування бізнес-інформації з вебпростору є гетерогенність форматів і структур даних, неоднозначність природної мови, динамічність інформаційного середовища, неповнота та суперечливість відомостей. Обґрунтовано, що використання онтологій забезпечує семантизацію, структуризацію, логічне узгодження й інтеграцію бізнес-інформації в межах корпоративної бази знань, а також створює підґрунтя для підвищення якості аналітичного опрацювання даних. Систематизовано сучасні онтологічно-орієнтовані підходи. За результатами порівняльного аналізу встановлено, що найбільш перспективними для систем підтримки прийняття рішень є гібридні підходи, які поєднують переваги різних груп методів, забезпечуючи гнучкість результатів. **Висновки**. Наукова новизна отриманих результатів полягає в узагальненні та порівнянні сучасних онтологічно-орієнтованих підходів до видобування бізнес-інформації з неструктурованих вебджерел з урахуванням їх придатності до семантичного узгодження даних, формування корпоративної бази знань та використання в системах підтримки прийняття рішень.

Ключові слова: бізнес-інформація; неструктуровані вебджерела; онтологія; видобування інформації; семантична інтеграція; корпоративна база знань; системи підтримки прийняття рішень.

Данов Сергій Олександрович – асп. каф. інженерії програмного забезпечення, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Шостак Ігор Володимирович – д-р техн. наук, проф., проф. каф. інженерії програмного забезпечення, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Serhii Danov – PhD Student at the Department of Software Engineering, National Aerospace University “Kharkiv Aviation Institute”, Kharkiv, Ukraine,
e-mail: s.o.danov@student.khai.edu, ORCID: 0009-0002-6502-7565.

Igor Shostak – Doctor of Technical Sciences, Professor, Professor at the Department of Software Engineering, National Aerospace University “Kharkiv Aviation Institute”, Kharkiv, Ukraine,
e-mail: i.shostak@khai.edu, ORCID: 0000-0002-3051-0488.