Journal of Brilliant Engineering 1 (2023) 4798
Brilliant Engineering
journal home: www.acapublishing.com/journals/4/ben

Research Article

Examining Knowledge Extraction Processes from Heterogeneous Data Sources

Serdar Kürşat Sarıkoz*,[1], Muhammet Ali Akcayol[2]
[1] Department of Computer Science, Informatics Institute, Gazi University, Ankara, Turkey
[2] Department of Computer Engineering, Gazi University, Ankara, Turkey

Keywords

Hetoregenous information Networks,
Knowledge bases,
Information extraction,
Veracity of information.

Abstract

In the last 20 years, e-mail, instant messaging, documents, blogs, news, text communication in the transfer of information over the web, as a result of the presentation and transmission of information as a result of the Web the dramatic increase in the amount of data in digital environments has increased the importance of studies in the field of knowledge extraction from unstructured data. Since the 2000s, one of the primary goals of researchers in the field of artificial intelligence has been to extract knowledge from heterogeneous data sources on the World Wide Web, including real-life entities and semantic relationships between entities, and to display them in machine-readable format. Advances in natural language processing and information extraction have increased the importance of large-scale knowledge bases in complex applications, resulting in scalable information extraction from semi-structured and unstructured heterogeneous data sources on the Web, and the detection of entities and relationships; It enabled the automatic creation of prominent knowledge bases in this field such as DbPedia, YAGO, NELL, Freebase, Probase, Google Knowledge Vault, IBM Watsons, which contain millions of semantic relationships between hundreds of thousands of entities, and displaying the created information in machine-readable format. Within the scope of this article; Web-scale(end-to-end) knowledge extraction from heterogeneous data sources, methods, challenges and opportunities are provided.

1. Introduction

Since the first day of its use, the Internet continues to develop impressively and take its place in many parts of our lives by bringing along technological opportunities. The internet and internet technologies have had an extraordinary impact on people around the world, contributing to innovation, growth and the development of new business models with the developments in communication technologies, playing a central role in promoting human entrepreneurship and creativity, and making unique contributions to the development of societies. It is seen that it has become an indispensable part of twenty-first century (21st century) civilization in many fields, such as information sharing, communication, transportation, health, education, finance, security.

Ever-increasing noisy data content and resources since the World Wide Web was introduced in 1991 (www) are today accepted as the world's largest heterogeneous information source in need of discovery with its unique content.

The world's most extensive heterogeneous data and information network, the noisy and low-quality, unstructured data and media content on the Web, and the fact that valuable scientific, cultural and general life information prevent people from discovering technological information-sharing platforms such as Freebase and Wikipedia brought with it research opportunities.

A lot of research has been done about natural language processing (NLP), knowledge representation (KR), knowledge extraction (KE), knowledge representation (KR), information extraction (IE), scalable information extraction from the web (Open Information Extraction From Web - OIE), semantic networks (SN), automatic creation of knowledge bases (KB) enriched with semantic relations, knowledge graphs enriched with semantic relations (Knowledge Graphs - KGs), and question answering (QA) for the extraction of information from unstructured data on the web, creation and display of knowledge bases, in the last twenty years, almost all of them in English and have been published in the literature [1-6].

Knowledge bases with enriched information content have received increasing attention and importance in both industry and academic studies; search engines, question-answering systems, personalized recommendation systems, machine learning, natural language processing, etc. have become critical for a wide variety of knowledge-based cognitive applications [5,6].

In line with Tim Bernars Lee's vision of the semantic web [7], research has been conducted on knowledge bases in a machine-readable format containing millions of entities and hundreds of millions of relationships between entities. Academic research such as KnowItAll [8], TextRunner [9], DbPedia [10], Yago [11], Freebase [12], Nell [13], Wikidata [14], IBM Watson through data in unstructured heterogeneous sources on the web Industrial knowledge bases such as [15] and Google Knowledge Vault [16] have been created.

Within the scope of this article, the processes for generating information from heterogeneous data sources are discussed. In the 2nd chapter, heterogeneous information networks and their properties; in the 3rd chapter knowledge bases and concepts; in the 4th chapter the concept and components of information extraction (IE – Information Extraction), which is critical in the creation of the knowledge base, in the 5th chapter in the 5th chapter veracity of information, in the 6th and the last chapter discussion and conclusions are included.

*Corresponding Author: serdar.sarikoz@btk.gov.tr
Received 05 Feb 2023; Revised 07 Feb 2023; Accepted 07 Feb 2023
https://doi.org/10.36937/ben.2023.4798

2. Heterogenous Information Networks

Although almost all of the real systems used today consist of components of different object types that interact with each other, it is seen that the interactions between different objects and different connection types are modeled as homogeneous information networks. Heterogeneous Information Networks (HINs) emerge as complex network structures used in the representation and presentation of multiple relationships between different types of entities or object types [17-19].

Unlike the representation of homogeneous information networks with traditional data structures and applications, HINs are represented by a uniform entity-relationship model; It differs from different data sources in terms of displaying multiple types of relationships between different types of assets, about to real-life facts and information [17-20].

HINs are used to represent and analysis of information in many different domains, such as knowledge discovery, e-commerce, social media, health, general information networks, fraud, anomaly detection, decision support systems and forecasting. While HINs are used in a social networking application to represent likes, posts and sharing between assets and assets, in the study to be conducted on pharmaceutical or biological networks in the health ecosystem, it has the ability to represent together genes, proteins, chemical and molecular structures, and relationships between diseases (Sun et al. (2013) [17]).

Intended for HINs; It is seen that researches on similarity, clustering, classification, ordering, link prediction, recommendation and combination and presentation of information from different sources continue (Shi et al. (2017) [20]).

3. Knowledge Bases and Concepts

Equipping machines with comprehensive knowledge of the entities in the world and the relationships between them has been a long-standing goal of artificial intelligence. As a result of this, in the last ten years, important studies have been carried out for the automatic creation of large-scale knowledge bases from web content and text sources, and today it has taken its place in different applications for information analytics (Nakashole et al. (2011) [24]).

It is seen that most of the data on the web is in an unstructured format. Where text data and speech content on websites, social media applications, news portals, etc., heterogeneous data sources can contain important information, extracting real-world relationships between assets and assets from these data in a meaningful and highly accurate machine-readable format and their integration with existing information systems can be applied to big data analytics. It is considered that it will bring opportunities for different applications (Weikum et al. (2010) [6]).

3.1. Definition

A general expression knowledge bases (KB - knowledge bases) emerges as the technology used to represent, store and display for the presentation, storage and display of the relationship between assets and assets over structured, semi-structured and unstructured data [10-14].

Information extraction methods (IE-Information Extraction), including natural language processing algorithms, statistical models and machine learning algorithms (ML), are used in creating of knowledge bases; with the aforementioned methods, web et al. It is aimed to determine the relationships between assets and assets through data on unstructured, semi-structured and structured resources [9,15,16,23,24,33].

KB is generally defined as follows [15,18,25,33]:

KB= {E, R, X}, labeled and directed KB,

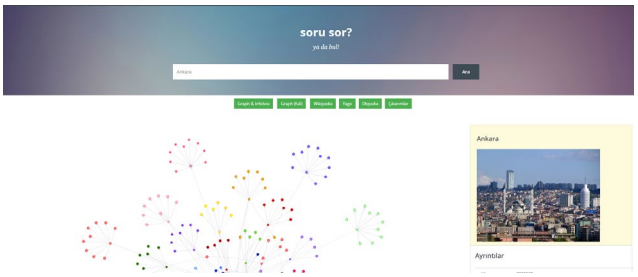
- $E=\{e_1,e_2,\ldots e_N\}$ set of entities (nodes),
- $R=\{r_1,r_2,\ldots r_N\}$ set of relations, (edges)
- $X_{ijk}=(e_i, r_k, e_j)$ triple expression
- $Y_{ijk} \in \{0,1\}$ If $Y_{ijk} \in \{0,1\}$, it takes the value "1" if X_{ijk} is present, and "0" otherwise.
- $Y \in \{0,1\}^{N_e N_e N_r}$ is represented,
- $Y_{ijk} = \begin{cases} 1, if\ the\ triple\ (e1, rk, ej)\ exists \\ 0, otherwise \end{cases}$

Can be formulated.

3.2. Application Areas Using Knowledge Bases

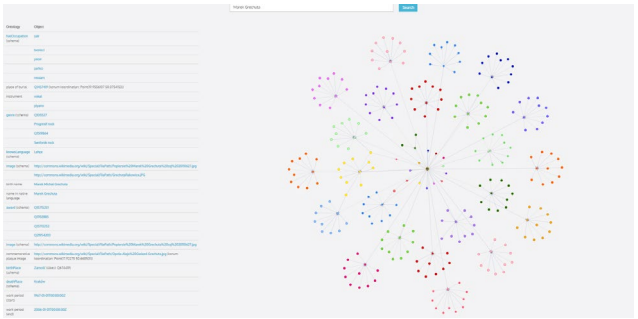
Application areas of knowledge bases are given below [25-33]:

Semantic Search and Question Answering Systems: It is aimed to create, display and present information in machine-readable format within the scope of interpreting and exhibiting users' information needs in terms of assets and relationships [27,30].



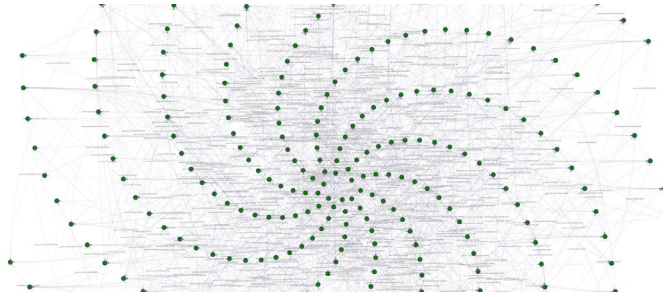
Picture 1. Semantic Search

- Preparation of summary data in the machine-readable format: Recent developments and increasing data size have led to the need for users to access fast and summary data from large data stacks and data summarization systems such as users' need for extensive summary information about entities and relationships emerge in this context [8-16,25,27-30].



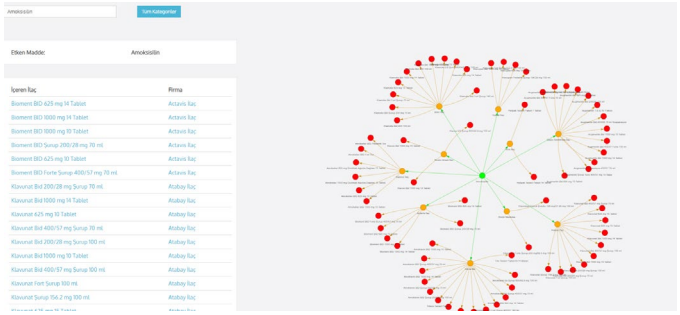
Picture 2. Entity & Entity Relationships in RDF Format

- Big Data analytics: Daily news, social media, academic publications etc. It emerges within the scope of making data analytics on web contents and making sense of the content and the relations between entities [27,75].



Picture 3. Entity & Entity Relationships in RDF Format

Drug design: After the COVID-19 pandemic in 2019, it is seen that research has been carried out to develop decision support systems for drug design through the demonstration of interactions such as disease-drug-gene-molecule-protein etc. [76-80].



Picture 4. Drug, Disease, Active Ingredient, Manufacturer, interaction

3.3. Classification of Knowledge Bases

In the literature, it is seen that knowledge bases differ according to the purpose of use, the method of creation, and the way the content is updated [1,4,9-16,25,33].

3.3.1. Knowledge bases by purpose of use

a) General purpose knowledge bases:

Knowledge bases such as Freebase, Google Knowledge Graph, YAGO, DBpedia, and Wikidata aim to display information including all entities and relationships in different domains. Knowledge bases in this structure, question answering, search engine infrastructure, semantic applications etc., appear to be used for this purpose [9-14].

b) Special purpose knowledge bases:

Special purpose knowledge bases such as Google Scholar, Amazon Product Services, and Microsoft Probase and Entity Cube are intended to display information containing entities and relationships for a particular domain [1,4,25,27].

3.3.2 Knowledge Bases by Creation Method

In the studies in the literature, there are generally four methods in the creation of the technical infrastructure of knowledge bases [8-16]:

- Knowledge bases built on semi structured resources: YAGO, YAGO2, YAGO4, DBpedia and Freebase, Wikidata [10-14],

- Knowledge bases created “without the use of schema” by scanning open sources all over the Web: Reverb, Ollie, Prismatic

- Knowledge bases created using a fixed ontology or “schema” by scanning open sources all over the Web; Projects where information extraction methods such as Nell, Prospera, DeepDive, and Elementary are applied,

- Knowledge bases created using IS-A structure: Probase.

3.3.3. Knowledge Bases According to System Maintenance, Update and Operation

- Closed World Assumption (CWA)
- Open World Assumption (OWA)

In the CWA approach, entities or relationships that are not generally accepted are not added to the knowledge base as new results. This method is also known as canonicalization in the literature [16].

In the OWA approach, newly acquired entities and relationships are added directly during the data collection.

3.4. Data Collection Processes in the Creation of Knowledge Bases

The data collection process is named in different studies in the literature as data acquisition, and data collection, data harvesting emerges as a data collection process from various heterogeneous data sources. The data sources required to create the knowledge base vary according to the type of knowledge base to be made [6,9,12,16,25,27,31].

Collecting data from different sources will create opportunities for Institutions or Organizations to gain a more comprehensive understanding of their operational processes, customers and sectors, and with this information, trends, improvement processes, patterns, anomalies and segmentation.

The data required to build general purpose knowledge bases are mostly made through structured resources.

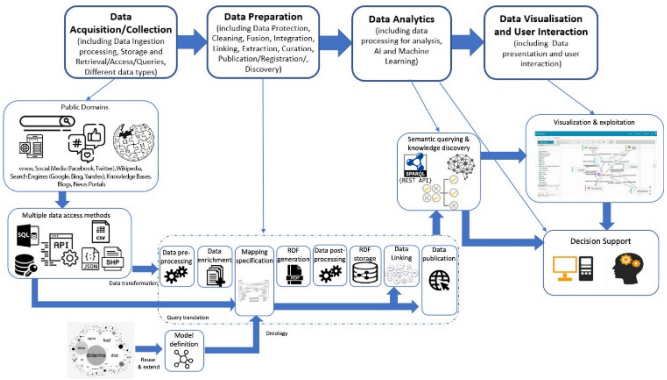


Figure 1. Components of Web Scale KE-KB-KG System [74].

4. INFORMATION EXTRACTION

Automatic extraction of information from unstructured sources and integration with existing databases has attracted the attention of researchers in the field of artificial intelligence since the 1980s. Due to the complexity of natural language and the uncertainties it brings, information extraction (IE) studies are considered the most critical component in the construction of knowledge bases and knowledge graphs.

While the studies in the field of IE were limited when the reflections in the field of IE were first started, it is seen that today information extraction studies are aimed at Web-scale, high-accuracy and OPEN IE [1,3,8,28,38,41,43,44].

4.1. Definition

From unstructured heterogeneous data sources, semi-structured and structured data sources with information extraction methods, assets, and relationships between assets, eliminating uncertainty and extracting structured information are aimed [36-40].

The templates, objects and relations to be inferred must include processes that will be embodied by the IE system in the text processing process.

The main problem to be solved is extracting information such as entities, properties, and relationships from open data sources.

Critical technologies in the information extraction process include identification of entities, the determination of types and classes of entities, elimination of ambiguities, and inference of relationships and ontologies.

4.2. Previous Studies

It has been observed that there have been four major changes in IE studies in the past 30 years [35-38,47]:

- Rule-based
- Studies to understand the message content (MUC – Message Understanding Conferences (1987-1998)
- ACE (Automatic Content Extraction) and CoNLL (Conference on Natural Language Learning)
- Knowledge Base Population (KBP, 2009-)

MUC (Message Understanding Conference): Since its inception in 1988, it has made significant contributions to the development of IE studies and NLP technologies [35,36,47].

Precision, recall and F-Measure metrics, which are widely used to measure the performance of information extraction studies, have emerged with MUC conferences for the identification of named entities and association extraction [36,47].

ACE (ACE-Automatic Content Extraction): Automatic content extraction studies it is aimed to extract general relations and events [36,39].

KBP (Knowledge Base Population): It is aimed to store named entities, relationships and ontologies [38-43].

4.3. Key Components of Information Extraction Studies

Natural Language Processing studies combine advanced theories, methods and technologies from different fields of artificial intelligence. Natural Language Processing aims to design and build algorithms that will analyze, understand, and generate the languages that humans naturally use [81].

NLP has an important place in knowledge extraction studies. NLP studies are focused on developing effective algorithms for processing texts and making them accessible for computer applications.

Lexical Analysis: The analysis process that covers all of the studies to determine the paragraphs, sentences and words by separating the punctuation marks of the language [82,83].

Syntax analysis: It covers analysis studies to check syntactic accuracy with the help of grammatical rules and dictionaries of the language. At this stage, the correct sequence of the word sequence and the representation of the relationships between different words are also provided [82,83].

Semantic Analysis: It covers the process of determining the meaning of sentences by syntactic analysis. Statements that do not fit syntactic analysis are ignored [82,83,84].

4.4. Basic Stages of Information Extraction Studies

The main stages of Information Extraction studies are presented below [35]:

-Named Entity Recognition.

-Named Entity Classification (NEC) and Disambiguation (NED).

-Relation Extraction (RE).

4.5. Open IE

OIE systems are used to detect assets from heterogeneous data sources on the web, eliminate the uncertainties of assets, detect relationships in triple format; It aims to identify and extract the relationships between entities such as the subject, object and predicate of the sentence and to present this information in a structured form as a knowledge base or knowledge graph [40-45].

One of the main advantages of OIE is that it enables quick and easy extraction of information from large volumes of unstructured text data which will significantly contribute to the data analysis of unstructured data and structured data of institutions or organizations in the field of business intelligence.

An OIE system extracts different triples (arg1, rel, arg2) from each sentence in a text, usually in RDF format, that represent key propositions or claims [44-48].

It is seen that the first work in the OIE area is TextRunner (Yates et al. (2007) [9]).

However, when Open IE studies are examined in general, it is seen that these studies are also divided into four classes according to the technique they use [46-48]:

- Supervised OIEs,
- Rule Based OIEs,
- Clause Based OIEs,
- Hybrid OIEs

5. VERACITY OF INFORMATION

With the development of internet technologies, the volume of data produced, transmitted and shared over the web has reached incredible proportions. With the dynamism of big data circulation, users' opinions, comments, corrections, etc. The contributions on the data cause the data to appear in front of the users with its changed form in different sources since the first source of the data. Studies conducted under the discipline of data science for the "discovery of reality" have gained importance in recent years to prevent the negativities caused by false information, erroneous content and misleading data [59,69,85].

Data quality has become more critical in the big data lifecycle. Big data is typically; volume, speed and diversity, and recently the concept of “veracity” – accuracy has emerged as the fourth “V” and has started to take its place as one of the main challenges in big data studies [85].

The importance of the accuracy and dynamics of information on the Internet seems to have led to research that arouses great interest not only from academia and the web industry, but also from government agencies and news agencies for its direct application [85].

5.1. Problem Definition

The "accuracy of big data" is a topical topic of particular interest to the data science community in general. It is seen that these studies take place as "information reliability", "trust management", "information validation", "data fusion" or "information aggregation" in some studies [55,59,60,85].

It is seen that researches on the veracity of information aim to investigate the accuracy of data obtained from different data sources on the Web, to eliminate and verify the problems caused by noisy information.

Unlike voting/average scoring approaches that treat all sources of information equally, it aims to calculate the reliability of the sources from which reliable information can be discovered and the accuracy of the data.

In studies on truth discovery methods, source reliability, fact reliability is calculated over each other. If the fact values obtained from a source contain high confidence values, the source or sources containing the fact positively affect its reliability.

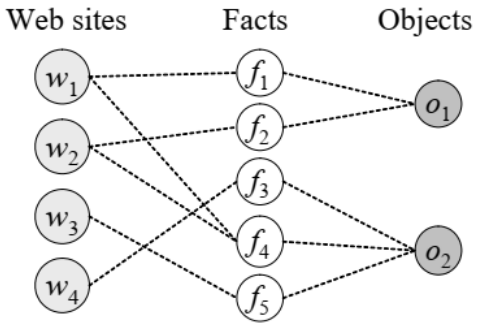


Figure 2. Source– Fact – Object Relationship [55,69].

The trust value of a source or a website depends on the accuracy, not the number of fact values it has produced. However, the reliability of a fact depends on the reliability of the sources it is supported, not on the fact that it is supported by a large number of sources.

5.2. Prominent Methods in Knowledge Verification Studies

Given the data or claims produced by multiple sources, the ultimate goal of online truth discovery is to classify each alleged information as true or false, calculating the credibility and accuracy of the relevant source.

Truth discovery studies involve complex tasks such as selection of heterogeneous data sources, information extraction from structured and unstructured content, detection and analysis of assets, data integration, and factual verification based on evidence.

There is an interdependence relationship between the data sources and the data itself produced. It is possible that the data provided by the trusted provider will be considered accurate, and the sources on which the reliable data depend is considered reliable.

One of the biggest challenges in discovering the truth in accurate data is that sources can duplicate each other, so errors can quickly spread and lead to inaccurate results. The identification of copy-producing sources and duplicate data in truth-discovery studies is essential for truth-discovery studies; It will help users more accurately assess the quality of resources and help users identify the most appropriate resources for their needs.

It is seen that the studies in the literature are generally gathered in four categories [59,65,85]:

- Iterative Methods [55,69]
- Agreement-Based Methods: Basically, it is based on counting the number of agreeing/contrasting sources for each data item [85].
- Analytical Methods: It treats the problem as a graph optimization problem to solve the truth discovery [59,85].
- Probabilistic Methods: Calculation of source and value accuracy is based on Bayesian probability models [59,85].

5.3. Other Components of Knowledge Verification Studies

In the process of information validation, the data coming from the source should be evaluated with different aspects before preprocessing [59]:

-Input Data: the data from the source should be evaluated with different aspects before preprocessing.

-Source Reliability: The selection of sources, their independence, whether they are affected by other sources will affect information verification studies.

-Evaluation of the Reliability of Assets: It is seen that the reliability of the objects is calculated through the calculation of the trust value of the objects and the reliability of the source, and it is increased or penalized together.

-Output Values: It includes issues that need to be evaluated, such as encountering more than one real value and reflecting the result.

5.4. Featured Algorithms in the Literature

TruthFinder: The calculation is made on the iterative application of source reliability and fact confidence value by applying Bayesian analysis. The algorithm in question is the first study in the literature, and similarity - implication introduced by this study, which inspired other studies [55,59,63].

AccuSim: The AccuSim algorithm applying Bayesian Analysis is proposed. The “implication” function has been adapted to the studies for similarity detection [56,59].

AccuCopy: In this method, whether there is copying over the similarity ratio between the sources, and if it is evaluated that there is copying, the source weight value is reduced [57].

2-Estimates: Reliability calculation with the "complementary vote" method has been proposed with the approach that an object has only TRUE fact value [58,59].

6. Conclusion and Discussion

Within the scope of this article, all the stages that need to be addressed in order to extract knowledge from heterogeneous data sources are included from our own perspective.

It is seen that the knowledge bases in the literature are designed for general or commercial purposes. However, there seems to be a need for the integration of general-purpose knowledge bases, commercial knowledge bases and non-confidential public data together, where all assets are represented individually.

In studies on knowledge bases; It is seen that the researches for updating the knowledge bases through reliable sources, inconsistent information, incomplete information, and completing the missing relations continue.

An information study to be conducted to knowledge extraction from heterogeneous data sources; it should provide compatibility with accuracy, trustworthiness, consistency, relevancy, timeliness, and interoperability other knowledge bases.

It is seen that the knowledge bases created for general purposes in the literature update the search engine infrastructure through a certain number of known sources.

In the current studies, it is seen that the results such as the up-to-dateness and accuracy of the knowledge base and the time interval in which the presented information is valid are insufficient. This issue appears to be causing irrelevant results to be displayed in search results.

This situation shows that research should be done in displaying the results for entities or relationships that are not included in the irrelevant results or that cannot be semantic matched.

In the CWA approach, entities or relationships that are not generally accepted are not added to the knowledge base as new results. This method is also known as canonicalization in the literature.

In the OWA approach, newly acquired entities and relationships are added directly during the data collection process.

Declaration of Conflict of Interests

The authors declare that there is no conflict of interest. They have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1.] Barbosa, D., Wang, H., & Yu, C. (2013, April). Shallow information extraction for the knowledge web. In 2013 IEEE 29th International Conference on Data Engineering (ICDE) (pp. 1264-1267).

[2.] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (2000). Learning to construct knowledge bases from the World Wide Web. Artificial intelligence, 118(1-2), 69-113.

[3.] Doan, A., Gravano, L., Ramakrishnan, R., & Vaithyanathan, S. (2008). Introduction to the special issue on managing information extraction. ACM Sigmod Record, 37(4), 5.

[4.] Dong, X. L., & Srivastava, D. (2015, May). Knowledge curation and knowledge fusion: challenges, models and applications. In Proceedings of the 2015 acm sigmod international conference on management of data (pp. 2063-2066).

[5.] Wang, K. (2015, May). The knowledge Web meets big scholars. In Proceedings of the 24th International Conference on World Wide Web (pp. 577-578).

[6.] Weikum, G., & Theobald, M. (2010, June). From information to knowledge: harvesting entities and relationships from web sources. In Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 65-76)

[7.] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.

[8.] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., ... & Yates, A. (2004, May). Web-scale information extraction in KnowItall: (preliminary results). In Proceedings of the 13th international conference on World Wide Web (pp. 100-110).

[9.] Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., & Soderland, S. (2007, April). TextRunner: open information extraction on the web. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 25-26).

[10.] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. Journal of web semantics, 7(3), 154-165.

[11.] Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from Wikipedia and wordnet. Journal of Web Semantics, 6(3), 203-217.

[12.] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250).

[13.] Zimmermann, A., Gravier, C., Subercaze, J., & Cruzille, Q. (2013, May). Nell2RDF: Read the Web, and Turn it into RDF. In KNOW@ LOD (pp. 2-8).

[14.] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

[15.] Ferrucci, David A. Introduction to "This Is Watson". IBM Journal of Research and Development, 2012, 56.3.4: 1: 1-1: 15.

[16.] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).

[17.] Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. Acm Sigkdd Explorations Newsletter, 14(2), 20-28.

[18.] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2016). A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 29(1), 17-37.

[19.] Xie, Y., Yu, B., Lv, S., Zhang, C., Wang, G., & Gong, M. (2021). A survey on heterogeneous network representation learning. Pattern recognition, 116, 107936.

[20.] Shi, C., & Philip, S. Y. (2017). Heterogeneous information network analysis and applications. Cham: Springer International Publishing. Pp:5-24

[21.] Hu, B., Fang, Y., & Shi, C. (2019, July). Adversarial learning on heterogeneous information networks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 120-129).

[22.] Bordes, A., Weston, J., Collobert, R., & Bengio, Y. (2011, August). Learning structured embeddings of knowledge bases. In Twenty-fifth AAAI conference on artificial intelligence.

[23.] Niu, F., Zhang, C., Ré, C., & Shavlik, J. (2012). Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. International Journal on Semantic Web and Information Systems (IJSWIS), 8(3), 42-73.

[24.] Nakashole, N., Theobald, M., & Weikum, G. (2011, February). Scalable knowledge harvesting with high precision and high recall. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 227-236).

[25.] Weikum, G., Hoffart, J., & Suchanek, F. M. (2016). Ten Years of Knowledge Harvesting: Lessons and Challenges. IEEE Data Eng. Bull., 39(3), 41-50.

[26.] Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36, 1-22.

[27.] Suchanek, F., & Weikum, G. (2013, June). Knowledge harvesting in the big-data era. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 933-938).

[28.] Wu, F., & Weld, D. S. (2010, July). Open information extraction using wikipedia. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 118-127).

[29.] Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., De Melo, G., & Weikum, G. (2011, March). YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In Proceedings of the 20th international conference companion on World wide web (pp. 229-232).

[30.] Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., & Weikum, G. (2012, July). Natural language questions for the web of data. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 379-390).

[31.] Deshpande, O., Lamba, D. S., Tourn, M., Das, S., Subramaniam, S., Rajaraman, A., ... & Doan, A. (2013, June). Building, maintaining, and using knowledge bases: a report from the trenches. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 1209-1220).

[32.] Sheth, A., Padhee, S., & Gyrard, A. (2019). Knowledge graphs and knowledge networks: the story in brief. IEEE Internet Computing, 23(4), 67-75.

[33.] Weikum, G., Dong, X. L., Razniewski, S., & Suchanek, F. (2021). Machine knowledge: Creation and curation of comprehensive knowledge bases. Foundations and Trends® in Databases, 10(2-4), 108-490.

[34.] Pellissier Tanon, T., Weikum, G., & Suchanek, F. (2020, May). Yago 4: A reason-able knowledge base. In European Semantic Web Conference (pp. 583-596). Springer, Cham.

[35.] Internet: Heng J., Information Extraction: Techniques, Advances and Challenges, https://blender.cs.illinois.edu/paper/IE_2012.pdf, Last Access 17/01/2023

[36.] Internet: Grishman R., "Information Extraction: Capabilities and Challenges", http://www.cs.nyu.edu/grishman/tarragona.pdf, Last Access 17/01/2023

[37.] Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

[38.] Grishman, R. (1997). Information extraction: Techniques and challenges. In Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14–18, 1997 (pp. 10-27). Springer Berlin Heidelberg.

[39.] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004, May). The automatic content extraction (ace) program-tasks, data, and evaluation. In Lrec (Vol. 2, No. 1, pp. 837-840).

[40.] Sarawagi, S. (2008). Information extraction. Foundations and Trends® in Databases, 1(3), 261-377

[41.] Grishman, R. (2015). Information extraction. IEEE Intelligent Systems, 30(5), 8-15.

[42.] Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2012, April). Dependency-based open information extraction. In Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP (pp. 10-18).

[43.] Gamallo, P. (2014). An overview of open information extraction (invited talk). In 3rd Symposium on Languages, Applications and Technologies. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[44.] Etzioni, O., Fader, A., Christensen, J., & Soderland, S. (2011, June). Open information extraction: The second generation. In Twenty-Second International Joint Conference on Artificial Intelligence.

[45.] Wu, F., & Weld, D. S. (2010, July). Open information extraction using wikipedia. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 118-127).

[46.] Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A survey on open information extraction. arXiv preprint arXiv:1806.05599.

[47.] Grishman, R. (2019). Twenty-five years of information extraction. Natural Language Engineering, 25(6), 677-692.

[48.] Muhammad, I., Kearney, A., Gamble, C., Coenen, F., & Williamson, P. (2020). Open information extraction for knowledge graph construction. In Database and Expert Systems Applications: DEXA 2020 International Workshops BIOKDD, IWCFS and MLKgraphs, Bratislava, Slovakia, September 14–17, 2020, Proceedings 31 (pp. 103-113). Springer International Publishing.

[49.] Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: theory, practice & open challenges. Proceedings of the VLDB Endowment, 5(12), 2018-2019.

[50.] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery in Data, 1(1), 2007

[51.] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: a generic approach to entity resolution. VLDB Journal, 18(1), 2009.

[52.] P. Christen. Data Matching. Springer, 2012.

[53.] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In SIGMOD, 2005.

[54.] H. Köpcke, A. Thor, and E. Rahm. Evaluation of entity resolution approaches on real-world match problems. PVLDB, 3(1):484–493, 2010

[55.] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), pages 1048-1052, 2007

[56.] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. PVLDB, 2(1):550-561, 2009.

[57.] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In Proc. of the ACM International Conference on Web Search and Data Mining (WSDM'10), pages 131-140, 2010.

[58.] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In Proc. of the International Conference on Computational Linguistics (COLING'10), pages 877-885, 2010.

[59.] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., ... & Han, J. (2016). A survey on truth discovery. ACM Sigkdd Explorations Newsletter, 17(2), 1-16.

[60.] Lehmann, J., Gerber, D., Morsey, M., & Ngomo, A. C. N. (2012, November). Defacto-deep fact validation. In International semantic web conference (pp. 312-327). Springer, Berlin, Heidelberg.

[61.] Esteves, D., Rula, A., Reddy, A. J., & Lehmann, J. (2018). Toward veracity assessment in RDF knowledge bases: An exploratory analysis. Journal of Data and Information Quality (JDIQ), 9(3), 16.

[62.] Liu, W., Liu, J., Duan, H., Zhang, J., Hu, W., & Wei, B. (2017, April). TruthDiscover: resolving object conflicts on massive linked data. In Proceedings of the 26th International Conference on World

Wide Web Companion (pp. 243-246). International World Wide Web Conferences Steering Committee.

[63.] Ba, M. L., Berti-Equille, L., Shah, K., & Hammady, H. M. (2016, April). VERA: A platform for veracity estimation over web data. In Proceedings of the 25th international conference companion on world wide web (pp. 159-162).

[64.] ESTEVES, D., RULA, A., REDDY, A. J., & LEHMANN, J. (2018). Towards Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. Journal of Data and Information Quality (JDIQ), 51.

[65.] Zhao, Z., Cheng, J., & Ng, W. (2014, November). Truth discovery in data streams: A single-pass probabilistic approach. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1589-1598).

[66.] Lamine Ba, M., Berti-Equille, L., & Hammady, H. M. (2016, March). Discovering the Truth on the Web Data: One Facet of Data Forensics. In Qatar Foundation Annual Research Conference Proceedings (Vol. 2016, No. 1, p. ICTPP3179). Qatar: HBKU Press.

[67.] Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014, June). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 1187-1198).

[68.] Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., ... & Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. Proceedings of the VLDB Endowment, 8(4), 425-436.

[69.] Gupta, M., & Han, J. (2011). Heterogeneous network-based trust analysis: a survey. ACM SIGKDD Explorations Newsletter, 13(1), 54-71.

[70.] Azzalini, F., Piantella, D., & Tanca, L. (2019, June). Data Fusion with Source Authority and Multiple Truth. In SEBD.

[71.] Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., ... & Zhang, W. (2015). Knowledge-based trust: Estimating the trustworthiness of web sources. arXiv preprint arXiv:1502.03519.

[72.] Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. Proceedings of the VLDB Endowment, 2(1), 550-561.

[73.] Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2015). Truth finding on the deep web: Is the problem solved. arXiv preprint arXiv:1503.00303.

[74.] Internet: Palma R., A Knowledge Graph for Agri-Foos Sector, https://blog.metaphacts.com/a-knowledge-graph-for-the-agri-food-sector, Last Access: 06/02/2023

[75.] Janev, V., Graux, D., Jabeen, H., & Sallinger, E. (2020). *Knowledge graphs and big data processing* (p. 209). Springer Nature pp:12-35.

[76.] Bonner, S., Barrett, I. P., Ye, C., Swiers, R., Engkvist, O., Bender, A., ... & Hamilton, W. L. (2022). A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, *23*(6).

[77.] Gogineni, A. K. (2022). Analysis of Drug repurposing Knowledge graphs for Covid-19. arXiv preprint arXiv:2212.03911.

[78.] Zeng, X., Tu, X., Liu, Y., Fu, X., & Su, Y. (2022). Toward better drug discovery with knowledge graph. Current opinion in structural biology, 72, 114-126.

[79.] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 46(D1), D1074-D1082.

[80.] Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., & Sadoghi, M. (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Journal of Web Semantics, 44, 104-117.

[81.] Cambria, E., Schuller, B., Xia, Y., & White, B. (2016). New avenues in knowledge bases for natural language processing. Knowledge-Based Systems, 108(C), 1-4.

[82.] Ranjan, N., Mundada, K., Phaltane, K., & Ahmad, S. (2016). A Survey on Techniques in NLP. International Journal of Computer Applications, 134(8), 6-9.

[83.] Adalı, E. (2012). Doğal Dil İşleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).

[84.] Etzioni, O., Banko, M., & Cafarella, M. J. (2006, July). Machine Reading. In AAAI (Vol. 6, pp. 1517-1519).

[85.] Berti-Equille, L., & Borge-Holthoefer, J. (2015). Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics. Synthesis Lectures on Data Management, 7(3), 1-8.

How to Cite This Article

Sarıkoz S.K., Akcayol M.A., Examining Knowledge Extraction Processes from Heterogeneous Data Sources, Brilliant Engineering, 1(2023), 4798.
https://doi.org/10.36937/ben.2023.4798.