MANTECH Publications
Journal of Data Mining and Knowledge Engineering
Volume 1, Issue 1, January-June 2016

Automated Knowledge Discovery Using Large Language Models

Sudhansu Thakur
Associate Professor
Department of Computer Applications
Sharda Institute of Technology, Jaipur, India
Email: *Sudhanshuthakur15@gmail.com*

Abstract

Automated knowledge discovery has become a critical task in modern digital environments where massive textual and multimodal data are generated continuously. Traditional data mining and knowledge discovery approaches rely heavily on structured databases and manual feature engineering, which limits scalability and semantic understanding. Large language models (LLMs) provide a new paradigm for extracting knowledge directly from unstructured sources such as scientific literature, technical reports, social media, and enterprise documents. These models demonstrate strong capabilities in representation learning, semantic reasoning, summarization, and relationship extraction, enabling automated discovery of concepts, patterns, and insights.

This paper reviews the role of large language models in automated knowledge discovery, covering architectures, methodologies, application domains, and evaluation strategies. We discuss how LLM-based pipelines perform knowledge extraction, entity discovery, relation mapping, and hypothesis generation. Advantages such as contextual awareness, cross-domain generalization, and interactive learning are examined along with limitations including hallucination, bias, and explainability challenges. A conceptual framework for LLM-driven knowledge discovery is presented along with comparative analysis of classical and LLM-based methods.

The review highlights emerging research directions including autonomous scientific discovery, domain-adapted LLMs, hybrid symbolic-neural reasoning, and trustworthy knowledge systems. Automated knowledge discovery using

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

LLMs is expected to transform scientific research, engineering analytics, healthcare intelligence, and industrial decision support.

Keywords: *Large language models, knowledge discovery, automated reasoning, text mining, semantic analysis, knowledge graphs*

INTRODUCTION

The exponential growth of digital information has created a strong demand for automated methods that can discover useful knowledge from large and heterogeneous data sources. Scientific publications, patents, engineering documents, and online communications collectively form vast repositories of human knowledge. However, much of this information remains under-utilized because it is stored in unstructured textual form rather than structured databases.

Knowledge discovery traditionally involves processes such as data collection, preprocessing, pattern detection, interpretation, and validation. Classical approaches like data mining and natural language processing require domain-specific rules, feature extraction pipelines, and supervised training data. These requirements restrict scalability and adaptability across domains.

Large language models (LLMs) have recently emerged as powerful tools for understanding and generating natural language. Trained on massive corpora using transformer architectures, LLMs learn contextual semantic representations that encode linguistic and conceptual relationships. This capability allows them to infer knowledge patterns without explicit programming.

Automated knowledge discovery using LLMs refers to the process of extracting structured insights, concepts, relationships, and hypotheses from unstructured or semi-structured data using large language models. Unlike conventional text mining, LLM-based discovery can integrate reasoning, summarization, and cross-document synthesis in a unified framework.

The objective of this paper is to review current approaches and methodologies that enable automated knowledge discovery using LLMs. The paper also analyses applications, advantages, and challenges while proposing a general framework for LLM-driven knowledge

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

discovery systems.

FUNDAMENTALS OF KNOWLEDGE DISCOVERY

1. Traditional Knowledge Discovery Process

Knowledge discovery in databases (KDD) consists of several sequential stages:

- Data selection
- Data preprocessing
- Transformation
- Data mining
- Interpretation

Traditional techniques rely on structured datasets where features and labels are clearly defined. Algorithms such as clustering, classification, and association rule mining identify statistical patterns.

However, knowledge present in textual data requires linguistic understanding, semantic interpretation, and contextual reasoning. Conventional NLP approaches such as bag-of-words and rule-based parsing capture limited semantics and often fail to generalize across domains.

2. Knowledge Discovery from Text

Text-based knowledge discovery aims to extract:

- Entities (concepts, objects, terms)
- Relationships between entities
- Events or processes
- Trends or patterns
- Hypotheses

This process is often called text mining or knowledge extraction. It typically involves named entity recognition, relation extraction, topic modeling, and summarization.

Limitations of classical text mining include:

- Dependence on annotated datasets
- Domain-specific rules
- Limited contextual understanding
- Difficulty integrating multi-document knowledge

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

These limitations motivate the adoption of LLM-based approaches.

LARGE LANGUAGE MODELS FOR KNOWLEDGE DISCOVERY

Large language models (LLMs) have emerged as a transformative technology for automated knowledge discovery due to their capability to learn semantic representations from vast textual corpora. Unlike traditional machine learning systems that rely on handcrafted features and domain-specific rules, LLMs learn generalizable language representations that capture relationships between concepts, entities, and contexts. This enables them to identify hidden patterns, extract relationships, and synthesize knowledge across documents and domains.

LLMs function not only as language generators but also as implicit knowledge bases. Their training on large-scale corpora allows them to encode factual information, conceptual associations, and reasoning patterns in their internal parameters. When applied to knowledge discovery tasks, they can extract entities, infer relations, summarize evidence, and even propose new hypotheses by combining learned associations.

The ability of LLMs to perform knowledge discovery arises mainly from their transformer architecture and distributed knowledge representation mechanisms, which are described in the following subsections.

1. Architecture of Large Language Models

Large language models are deep neural networks built on the transformer architecture, which relies on self-attention mechanisms to model relationships between tokens in a sequence. This architecture allows the model to consider all words in a sentence simultaneously rather than sequentially, enabling deep contextual understanding.

In traditional recurrent or convolutional language models, long-range dependencies are difficult to capture due to vanishing gradients or limited receptive fields. Transformers overcome this limitation by computing attention weights between every pair of tokens, allowing the model to learn global context across long text spans such as paragraphs, documents, or even collections of documents.

A typical LLM consists of multiple stacked transformer layers containing attention heads and

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

feed-forward networks. Each layer refines token representations by integrating contextual information from the entire sequence. Through large-scale pretraining on diverse corpora, these layers gradually learn linguistic patterns, semantic relationships, and world knowledge.

From a knowledge discovery perspective, several architectural characteristics of LLMs are particularly important.

Contextual embeddings.

LLMs generate contextualized word representations, meaning that the embedding of a word changes depending on its surrounding context. For example, the word “stress” in materials engineering and psychology will be represented differently based on context. This allows LLMs to correctly interpret domain-specific meanings and extract accurate knowledge from technical texts. Contextual embeddings also help identify entities and concepts that vary across domains but share semantic roles.

Semantic reasoning capability.

Self-attention layers enable the model to learn associations between concepts even when they appear far apart in text. During training, the model observes co-occurrence patterns and logical structures such as cause-effect or condition-outcome relationships. As a result, LLMs can infer relationships between entities without explicit rules. For knowledge discovery tasks, this means the model can connect related concepts across sentences or documents and identify implicit relations.

Long-range dependency modeling.

Knowledge discovery often requires integrating information distributed across large text spans, such as linking a chemical property mentioned in one section of a paper to an application described later. Transformer attention allows LLMs to capture such dependencies across long sequences. This capability supports multi-sentence relation extraction, document-level summarization, and cross-document synthesis, which are essential for automated discovery from scientific or technical literature.

Cross-domain generalization.

LLMs are typically trained on heterogeneous datasets containing multiple domains including

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

science, engineering, medicine, and general knowledge. The learned representations therefore capture universal semantic structures that transfer across domains. For knowledge discovery, this allows models trained on general corpora to analyze specialized texts with minimal adaptation. Cross-domain generalization is especially valuable when labeled data is scarce, as often occurs in scientific discovery tasks.

Few-shot and zero-shot learning capability.

Unlike classical models that require large labeled datasets, LLMs can perform new tasks using only instructions or a few examples. This ability arises from their large-scale pretraining and general language understanding. In knowledge discovery, few-shot learning enables rapid deployment of extraction tasks such as identifying material properties or biological interactions without extensive annotation. This significantly reduces development effort and allows dynamic discovery pipelines.

LLMs encode knowledge implicitly in parameters learned from massive text corpora. Accessing this knowledge can be achieved through several mechanisms.

Prompting involves providing instructions or queries that guide the model to retrieve or infer knowledge from its internal representations. Prompt-based extraction can identify entities, relations, or summaries directly from text.

Fine-tuning adapts the pretrained model to specific domains or discovery tasks using smaller labeled datasets. Domain-adapted models improve accuracy in specialized areas such as biomedical or engineering knowledge extraction.

Retrieval-augmented generation integrates external knowledge sources such as databases or document collections with the LLM. Retrieved information is provided as context to the model, enabling more accurate and up-to-date knowledge discovery.

Through these architectural and training properties, LLMs provide a flexible and scalable foundation for automated knowledge discovery across diverse textual domains.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

2. Knowledge Representation in LLMs

Knowledge in large language models is represented in distributed neural form across the model's parameters rather than as explicit symbolic structures such as rules or graphs. Each concept or fact is encoded as patterns of activation distributed across many neurons and layers. This distributed representation enables the model to capture complex semantic associations and generalize beyond explicitly observed examples.

Instead of storing isolated facts, LLMs learn probabilistic relationships between tokens, contexts, and concepts. During training, the model predicts words based on surrounding text, gradually learning which concepts tend to co-occur and how they relate linguistically and semantically. Over billions of training examples, this process results in a rich internal representation of world knowledge and domain knowledge.

Several categories of knowledge are encoded within LLM representations.

Factual knowledge.

LLMs store factual associations such as definitions, properties, and known relationships. For example, a model trained on scientific text may encode that "graphene has high electrical conductivity" or "fatigue causes material failure." These facts are not stored as explicit statements but as strong probabilistic associations between tokens and contexts. When prompted, the model retrieves such knowledge through generation.

Conceptual relationships.

Beyond isolated facts, LLMs capture relationships between concepts such as hierarchy, causality, similarity, and dependency. For instance, the model may learn that corrosion leads to structural degradation or that temperature influences reaction rate. These conceptual relationships enable knowledge discovery because the model can infer connections not explicitly stated in one place but implied across texts.

Linguistic patterns.

LLMs learn syntactic and semantic structures of language, including how relations are expressed in sentences. Patterns such as "X causes Y," "X consists of Y," or "X interacts with Y" are implicitly learned. This linguistic knowledge allows the model to recognize relation

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

expressions even when phrased differently. Consequently, relation extraction and knowledge mapping become more robust across writing styles and domains.

Procedural knowledge.

Procedural knowledge refers to understanding of processes, sequences, and methods described in text. Scientific and engineering documents often describe procedures such as manufacturing steps or experimental methods. LLMs can encode these sequences and reproduce or summarize them. In knowledge discovery, procedural understanding helps identify workflows, operational dependencies, and process optimization insights.

Domain insights and contextual associations.

When trained on domain-specific corpora, LLMs internalize specialized terminology and conceptual frameworks. For example, biomedical LLMs capture gene–disease associations, while engineering LLMs capture material–property relationships. These domain insights allow discovery of new connections by combining known associations across contexts.

A key advantage of distributed knowledge representation is the ability to infer relationships even when they are not explicitly stated. Because related concepts share similar embedding patterns, the model can generalize associations. For example, if training data indicates that high temperature reduces strength in several alloys, the model may infer similar behavior for a new alloy mentioned in text. This capability supports hypothesis generation and predictive knowledge discovery.

However, distributed representation also introduces challenges. Knowledge is not stored in interpretable symbolic form, making it difficult to trace the exact source or reasoning behind a discovered relationship. Additionally, probabilistic associations may lead to incorrect or hallucinated connections if training data is sparse or biased. Therefore, combining LLM representations with external knowledge graphs or validation mechanisms is often necessary for reliable discovery.

Despite these challenges, the distributed and probabilistic representation of knowledge in LLMs provides flexibility, generalization, and inference capability that traditional symbolic systems cannot easily achieve. This makes LLMs highly suitable for automated knowledge

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

discovery from complex and heterogeneous textual data.

AUTOMATED KNOWLEDGE DISCOVERY PIPELINE USING LLMS

Automated knowledge discovery with LLMs generally follows a multi-stage pipeline integrating neural and symbolic processing.

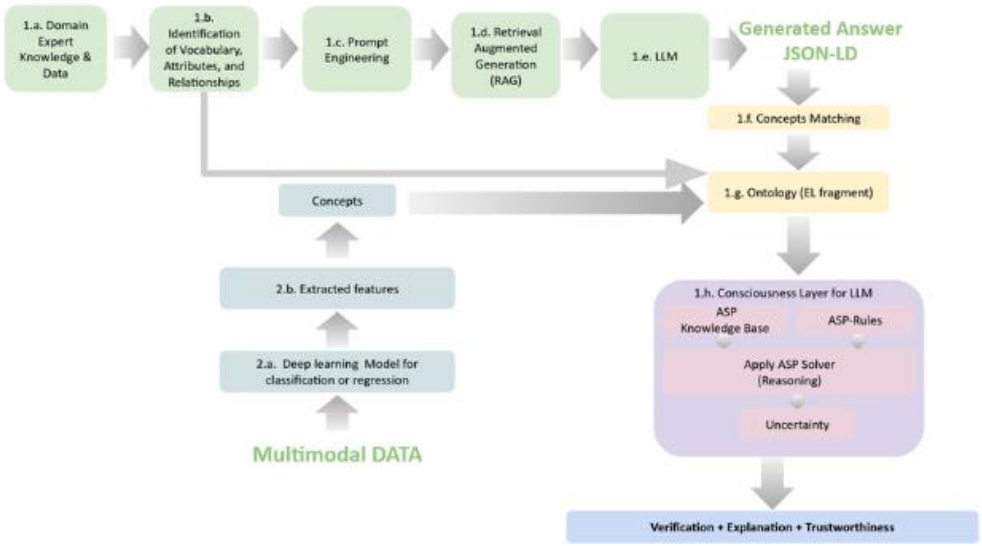


Figure 1: Conceptual Pipeline for LLM-based Knowledge Discovery

1. Data Acquisition and Preprocessing

Sources include:

- Scientific literature
- Technical reports
- Patents
- Web data
- Enterprise documents

Preprocessing tasks:

- Text normalization
- Document segmentation
- Noise removal
- Metadata extraction

LLMs can assist in preprocessing by detecting language, summarizing sections, and filtering irrelevant content.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

2. Entity and Concept Extraction

LLMs can identify entities and concepts through prompt-based extraction or fine-tuned models. Unlike traditional named entity recognition, LLM extraction is context-aware and adaptable.

Example outputs:

- Materials and properties in engineering papers
- Genes and diseases in biomedical text
- Components and faults in industrial reports

LLMs can also infer implicit entities not directly mentioned.

3. Relationship and Pattern Discovery

Relation extraction identifies associations between entities such as cause-effect, part-of, or interaction relationships.

LLMs perform this through:

- Semantic reasoning
- Cross-sentence inference
- Multi-document synthesis

For example, an LLM can infer that:

"Material A improves strength of alloy B under temperature C."

This relationship may be derived from multiple sentences across documents.

4. Knowledge Graph Construction

Extracted entities and relations can be organized into knowledge graphs. LLMs assist in:

- Ontology generation
- Relation normalization
- Entity linking
- Graph completion

Knowledge graphs enable structured querying and visualization of discovered knowledge.

5. Hypothesis and Insight Generation

One of the most promising capabilities of LLM-based discovery is generating new hypotheses. By combining information from diverse sources, LLMs can propose potential relationships not yet experimentally validated.

Examples:

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- Drug repurposing suggestions
- Material property predictions
- Engineering failure causes

This moves knowledge discovery from descriptive to predictive and creative stages.

COMPARISON OF CLASSICAL AND LLM-BASED KNOWLEDGE DISCOVERY

Table 1: Classical vs LLM-based Knowledge Discovery

Aspect	Classical Methods	LLM-based Methods
Data type	Structured or semi-structured	Unstructured text
Feature engineering	Manual	Automatic
Context understanding	Limited	Deep semantic
Cross-document inference	Weak	Strong
Domain adaptation	Difficult	Few-shot
Knowledge representation	Explicit rules	Neural embeddings
Hypothesis generation	Rare	Possible
Scalability	Moderate	High

The comparison shows that LLM-based approaches overcome many limitations of traditional systems.

APPLICATION DOMAINS

1. Scientific Literature Mining

LLMs can analyze large volumes of research papers to identify:

- Emerging trends
- Research gaps
- Concept relationships
- Method comparisons

Automated literature discovery reduces manual review workload and accelerates research synthesis.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

2. Healthcare and Biomedical Discovery

Biomedical knowledge discovery involves extracting relationships between genes, diseases, drugs, and pathways.

LLMs enable:

- Drug-target discovery
- Disease association mapping
- Clinical guideline extraction
- Medical knowledge graphs

These applications support precision medicine and clinical decision systems.

3. Engineering and Industrial Intelligence

Engineering domains generate large textual datasets such as maintenance logs, design documents, and failure reports.

LLM-based discovery can identify:

- Fault patterns
- Root causes
- Design improvements
- Safety risks

This supports predictive maintenance and reliability engineering.

4. Business and Market Intelligence

Organizations use LLMs to analyze:

- Customer feedback
- Market reports
- Competitor information
- Financial documents

Knowledge discovery helps detect trends, opportunities, and risks for strategic decisions.

5. Cybersecurity Knowledge Discovery

LLMs can process security reports and threat intelligence feeds to identify:

- Attack patterns
- Vulnerability relationships
- Malware behaviors

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- Threat actor profiles

Automated discovery supports proactive cybersecurity defense.

ADVANTAGES OF LLM-BASED KNOWLEDGE DISCOVERY

1. Contextual Understanding

LLMs capture deep semantic context, enabling interpretation of complex technical language and implicit meanings.

2. Cross-domain Generalization

Models trained on diverse data can transfer knowledge across domains, reducing need for domain-specific training.

3. Reduced Annotation Requirement

LLMs perform extraction using prompts or few-shot examples, minimizing labeled data needs.

4. Multi-document Reasoning

LLMs integrate information across multiple sources, enabling holistic knowledge synthesis.

5. Interactive Discovery

Users can query LLMs conversationally to explore knowledge spaces dynamically.

CHALLENGES AND LIMITATIONS

Despite advantages, LLM-based knowledge discovery faces several challenges.

1. Hallucination

LLMs may generate plausible but incorrect relationships or facts. This is critical in scientific and engineering domains.

2. Bias and Data Quality

Training data biases may influence discovered knowledge. Incorrect or incomplete source data leads to misleading insights.

3. Lack of Explainability

Neural knowledge representations are opaque, making it difficult to trace reasoning steps.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

4. Knowledge Validation

Automatically discovered knowledge must be validated against empirical evidence or expert review.

5. Computational Cost

Large models require high computational resources for training and inference.

EVALUATION OF KNOWLEDGE DISCOVERY SYSTEMS

Evaluating automated discovery is challenging because knowledge novelty and correctness must be assessed.

Table 2: Evaluation Metrics for LLM-based Knowledge Discovery

Metric	Description
Precision	Correctness of extracted knowledge
Recall	Coverage of relevant knowledge
Novelty	Newness of discovered relations
Consistency	Logical coherence
Explainability	Interpretability of reasoning
Human validation	Expert assessment

Evaluation often combines automated metrics and human expert review.

EMERGING RESEARCH DIRECTIONS

1. Autonomous Scientific Discovery

LLMs integrated with simulation and experimental systems may autonomously generate and test hypotheses.

2. Domain-adapted LLMs

Fine-tuning LLMs on domain corpora improves specialized knowledge discovery such as biomedical or materials science models.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

3. Hybrid Symbolic-Neural Systems

Combining LLMs with symbolic reasoning and knowledge graphs enhances explainability and logical consistency.

4. Trustworthy Knowledge Systems

Future research focuses on reducing hallucination and ensuring verifiable knowledge outputs.

5. Multimodal Knowledge Discovery

Integration of text, images, and sensor data enables richer knowledge extraction in engineering and science.

CONCEPTUAL FRAMEWORK FOR LLM-DRIVEN KNOWLEDGE DISCOVERY

A general framework for automated knowledge discovery using LLMs includes:

1. Data ingestion layer
2. LLM extraction layer
3. Knowledge structuring layer
4. Reasoning and synthesis layer
5. Validation layer
6. Application interface

This layered architecture supports scalable and trustworthy discovery systems.

CONCLUSION

Automated knowledge discovery using large language models represents a major shift in how knowledge is extracted from vast unstructured information sources. Traditional knowledge discovery methods depend heavily on structured data and manual feature engineering, limiting their ability to capture semantic relationships and cross-document insights. Large language models overcome these limitations through contextual understanding, semantic reasoning, and cross-domain learning.

This review examined the architecture and methodology of LLM-based knowledge discovery pipelines, including entity extraction, relation mapping, knowledge graph construction, and hypothesis generation. Applications across scientific research, healthcare, engineering, business intelligence, and cybersecurity demonstrate the transformative potential of automated

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved

discovery systems.

However, challenges such as hallucination, bias, explainability, and validation remain significant barriers to reliable deployment. Addressing these issues requires hybrid symbolic-neural approaches, domain-adapted models, and trustworthy evaluation frameworks.

Future research is expected to move toward autonomous discovery systems capable of generating and testing hypotheses, integrating multimodal data, and collaborating with human experts. Automated knowledge discovery using large language models has the potential to accelerate innovation, enhance decision-making, and transform knowledge-intensive disciplines across science and engineering.

REFERENCES

1. Vaswani, A., et al. (2017). Attention is all you need. *Neural Information Processing Systems*.
2. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers. *NAACL*.
3. Brown, T., et al. (2020). Language models are few-shot learners. *NeurIPS*.
4. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *Stanford CRFM Report*.
5. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Scientific text representation. *EMNLP*.
6. Lee, J., et al. (2020). BioBERT: Biomedical language representation. *Bioinformatics*.
7. Ji, S., et al. (2021). Survey on knowledge graphs. *IEEE TKDE*.
8. Paulheim, H. (2017). Knowledge graph refinement. *Semantic Web Journal*.
9. Wang, X., et al. (2023). Large language models for knowledge discovery. *AI Review*.
10. Zhou, D., et al. (2023). Foundation models in scientific discovery. *Nature Reviews AI*.

Page 26-41 © MANTECH PUBLICATIONS 2016. All Rights Reserved