2026 Volume: 02 Issue: 08 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-11

AN INTELLIGENT ONTOLOGY-BASED SYSTEM FOR MEDICAL DOCUMENT CATEGORIZATION

[*1]S. Narmatha, [2]Dr. V. Maniraj

[1]Research Scholar, Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli), Poondi, Thanjavur, Tamilnadu, India.

[2]Research Advisor, PG and Research Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli)., Poondi, Thanjavur, Tamilnadu, India.

Article Received: 28 June 2026
Article Revised: 17 July 2026
Published on: 05 August 2026

***Corresponding Author: S. Narmatha**
Research Scholar, Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli), Poondi, Thanjavur, Tamilnadu, India.

DOI: https://doi-doi.org/101555/ijrpa.5240

ABSTRACT

This study investigates the challenges and effectiveness of employing domain ontologies for the classification of online medical content. Specifically, the Medical Subject Headings (MeSH) thesaurus is utilized to enhance the categorization of medical documents. Based on the insights obtained, a novel document representation is proposed. The proposed approach is evaluated and compared with traditional stem-based representations using two widely adopted data mining algorithms, C4.5 and K-Nearest Neighbor (KNN). Experimental results demonstrate that the ontology-based method significantly outperforms conventional approaches. By integrating semantic concepts and hypernym relationships from the domain ontology, the document vectors used for classification are substantially enriched, leading to improved organization and accuracy. Validation using the Ohsumed biomedical benchmark dataset confirms the effectiveness of the proposed framework, achieving an improvement of approximately 30% over traditional stem-based representations.

INTRODUCTION

The rapid expansion of the Internet has greatly facilitated information dissemination; however, efficiently retrieving relevant and valuable information remains a major challenge. As a result, substantial research efforts are directed toward enhancing automated information retrieval systems. These advancements are particularly critical for the future evolution of the

Semantic Web, which relies on innovative data organization techniques and the incorporation of semantic annotations, often implemented through metadata. Ontologies have emerged as an effective mechanism for knowledge representation in this context. By providing a structured, domain-specific conceptual framework, ontologies define relationships among concepts and enable meaningful organization of information.

Within this framework, an important research question arises: to what extent can a conceptual model support the automated classification of medical documents? This paper seeks to address this question by investigating the role of ontology-based representations in medical document categorization. The remainder of this paper is organized as follows. Section 2 presents a review of related work. Section 3 details the proposed framework, including its architecture, methodology, and distinguishing features. Section 4 describes the Ohsumed benchmark dataset and the MeSH medical ontology, and presents a comparative evaluation against the traditional bag-of-stems approach. Section 5 discusses the experimental results and performance analysis, and Section 6 concludes the paper with final observations and directions for future research.

RELATED WORK

Despite advances in text mining and information retrieval, many document classification techniques continue to depend on the traditional bag-of-words model. Although widely used, this approach fails to capture semantic relationships between terms, limiting its effectiveness in representing the underlying meaning of textual content. Consequently, researchers have increasingly focused on developing concept-based representations, with ontologies emerging as a promising solution.

Notably, Amine's work highlights the benefits of enhancing text document clustering through the integration of ontological resources such as WordNet, demonstrating improved semantic understanding. More recently, G has shown that ontology-based text classification is both feasible and effective, particularly in the categorization of bilingual websites using multilingual ontologies. Their approach involves constructing topic-specific ontologies by combining linguistic and statistical techniques, providing an alternative strategy for extracting meaningful information from online documents.

For ontologies to accurately convey semantic meaning, web content must be classified with both precision and contextual relevance. This requirement emphasizes the need for automated classification algorithms that leverage domain-specific ontologies, especially in scenarios where predefined knowledge bases or supervised learning models are unavailable.

A METHOD FOR CONCEPTUAL REPRESENTATION

Effective text classification requires a document encoding strategy that enables efficient processing while preserving only the information relevant to the classification task. Although the bag-of-words model is widely adopted for this purpose, its inherent limitations have motivated researchers to explore alternative representation techniques. In this study, we propose a concept-based representation method designed to enhance semantic expressiveness while simultaneously reducing the dimensionality of the feature space. These improvements contribute to increased efficiency and performance in the proposed text classification system.

To facilitate a fair comparison with traditional approaches, our method is implemented in two sequential stages:

ConceptualTransformation

In this stage, selected textual terms are mapped to corresponding concepts through an appropriate matching and disambiguation process. This transformation enables the incorporation of contextual information and richer semantic relationships into the document representation.

Hyperonym-BasedEnhancement

Building upon the conceptual representation, the feature vector is further enriched by integrating hyperonym information. This step captures higher-level semantic abstractions, thereby improving the robustness and accuracy of the document representation.

Text Pre-processing Stage

At this stage, the raw textual data are refined to ensure that only content relevant for classification is retained. Non-textual elements commonly found in web documents, such as images and HTML tags, are removed to eliminate noise. Pre-processing is essential to filter out irrelevant information and to prevent multiple occurrences of identical terms or phrases from being treated as separate features. This step prepares the data for subsequent lexical analysis and stemming, ultimately improving both the quality of the representation and the efficiency of the classification process.

Web documents often contain content that does not directly contribute to their main subject. To address this issue, stop-word removal is applied to eliminate frequently occurring but semantically insignificant terms. In addition, stemming is employed to normalize word forms, thereby reducing feature redundancy and enriching the document representation vectors.

Stop-Word Removal

Stop words are high-frequency terms that carry minimal semantic value in text classification tasks. Eliminating these words during pre-processing reduces the dimensionality of the text representation, resulting in faster processing and improved classification performance. Such words are typically well documented in language-specific dictionaries.

Stemming

Stemming, introduced by Porter in 1980 for the English language, aims to group words by their common morphological roots. This approach mirrors human reading behavior, where different forms of a word are interpreted as expressing the same underlying concept. For instance, terms such as *walk*, *walker*, and *walking* are all associated with the concept of walking. Without stemming, these variations would be treated as distinct features. Porter's stemming algorithm consolidates related word forms into a single representative term by analyzing their morphological structure. Its successful application to literary texts, such as Shakespeare's works, has inspired adaptations for other languages, often requiring specialized linguistic knowledge.

Examples of Infections and Clinical Manifestations

In a clinical setting, an infection might present with the following signs in a patient with hemiplegia:

- **Urinary Tract Infections (UTIs)**
  - **Clinical Signs:** Fever, cloudy or strong-smelling urine, increased urinary frequency or incontinence, and sometimes altered mental status (confusion, agitation), which may be the only sign in a patient with a neurological deficit.
  - **Reason:** Bladder dysfunction and indwelling catheters are common in immobile patients.
- **Pneumonia**
  - **Clinical Signs:** Fever, cough, difficulty breathing, crackling sounds in the lungs upon examination, and reduced oxygen saturation.
  - **Reason:** Impaired swallowing (dysphagia) can lead to aspiration of food or fluid into the lungs, causing infection.
- **Skin and Soft Tissue Infections**
  - **Clinical Signs:** Redness, warmth, swelling, pain, or pus around pressure points (heels, lower back), or at injection/catheter sites.
  - **Reason:** Immobility can lead to pressure ulcers (bedsores), which are prone to infection.0

- **Sepsis**
  - **Clinical Signs:** A severe, systemic response to infection that can manifest as fever or hypothermia, rapid heart rate, low blood pressure, and significant changes in mental status.
  - **Reason:** The body's overwhelming response to an untreated or severe localized infection.

Relevant MeSH Terms

Medical Subject Headings (MeSH) are a standardized vocabulary for indexing biomedical information. Relevant terms for this topic include:

- **Hemiplegia**
- **Infection**
- **Central Nervous System Infections**
- **Brain Diseases**
- **Stroke** (a common cause of hemiplegia)
- **Pneumonia, Aspiration**
- **Urinary Tract Infections**
- **Sepsis**
- **Complications** (as a subheading under Hemiplegia in MeSH)

Figure 1 shows an example of a cleaned and stemmed representation vector. This diagram likely illustrates how a text, such as "Infect Clinic Hemiplegia," is transformed into a numerical or vector representation, which encodes the semantic meaning and relationships between words. This vector allows the text to be processed and analyzed by algorithms.

……	Infect	clinic	Hemiplegia	.....

Mapping Terms to Concepts

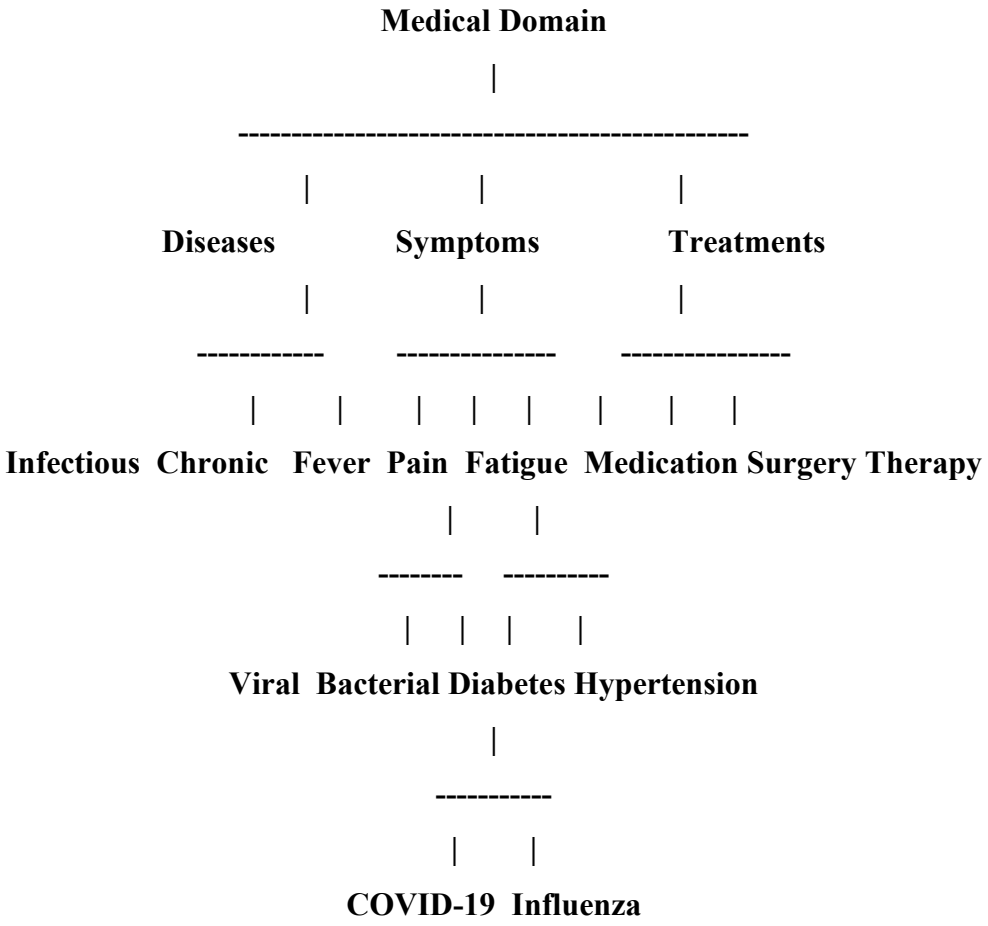
Relationship Between Terms and Concepts

Figure 2 illustrates the process of mapping individual terms to their corresponding conceptual representations. This mapping is a core component of natural language processing, as it enables textual data to be interpreted at a conceptual level by capturing abstract meanings and semantic relationships among words.

In this framework, the ontology serves as a semantic backbone by linking linguistic expressions to the concepts they denote. Different terms that refer to the same underlying

concept are unified within the ontology. For example, the terms *"appendicitis"* and *"appendiceal"* are both associated with the concept of appendicitis. Rather than treating these terms separately, their individual frequencies are combined to represent the occurrence frequency of the corresponding concept. This aggregation allows for a more accurate and semantically meaningful representation of document content.

Example Ontology Diagram: ***Medical Diagnosis Domain***



Methods for Handling Ambiguity

Because natural language is inherently subjective, assigning unambiguous labels to concepts is often a complex task. Many words possess multiple meanings depending on context, making sense disambiguation an essential component of semantic analysis. Word Sense Disambiguation (WSD) is widely recognized as one of the most challenging problems in artificial intelligence, comparable in difficulty to other fundamental AI tasks. Although sophisticated WSD techniques exist, their complexity can limit practical implementation. Therefore, this study concentrates on two relatively simple disambiguation strategies.

Irrespective of the specific concepts involved, the proposed approach is based on the assumption that identifying the dominant themes or core ideas within a document can substantially improve its representation. These themes tend to emerge naturally, even as the dimensionality of the representation space grows. Concept frequencies are then computed based on these identified themes, following the procedure described below.

Concept Frequency as Sum of Term Frequencies

This is the most common and intuitive alternative:

$$cf(d,c) = \sum_{t \in T(d),\ c \in ref(t)} tf(d,t)$$

Meaning

The frequency of concept $c$in document $d$is calculated as the sum of the frequencies of all terms $t$in the document that map to concept c.

Weighted Concept Frequency

This version assigns weights to terms based on relevance or confidence of mapping:

$$cf(d,c) = \sum_{t \in T(d)} w(t,c) \cdot tf(d,t)$$

Meaning

Each term contributes proportionally to the concept frequency using a weight $w(t,c)$, which reflects how strongly the term is associated with the concept.

Normalized Concept Frequency

Useful when document lengths vary significantly:

$$cf_{norm}(d,c) = \frac{cf(d,c)}{\sum_{c_i \in C} cf(d,c_i)}$$

Meaning

The concept frequency is normalized by the total number of concept occurrences in the document

Binary Concept Occurrence

A simplified representation focusing only on presence or absence:

$$cf(d,c) = \begin{cases} 1, & \text{if } \exists\, t \in T(d) \text{ such that } c \in ref(t) \\ 0, & \text{otherwise} \end{cases}$$

Meaning

The concept is marked as present if at least one related term appears in the document.

TF–IDF–Like Concept Weighting

A semantic extension of TF–IDF:

$$w(d,c) = cf(d,c) \times \log \frac{N}{df(c)}$$

Meaning

Concept importance increases if it appears frequently in a document but rarely across the corpus.

The FTIDF (Term Frequency-Inverse Document Frequency) method is more commonly applied and is a better weighting mechanism. It is used within the vector model framework, which involves:

«Frequency Term »* « Inverse document frequency »

FTIDF Formula

FTIDF(Ci, Wj) = TF(Ci, Wj) * log (nbr_category / DF(Wj))

Where "FT: Frequency Term" refers to the number of occurrences of a specific term within the relevant category, and "IDF: Inverse Document Frequency" is calculated as the total number of categories divided by the number of categories containing the term in question.

$$\mathbf{FTIDF(Ci,Wj)}\_TF(Ci,wj) * log \frac{\mathbf{nbr_{category}}}{\mathbf{DF(wj)}}$$

The dimensionality reduction selection process involves taking a set of features and creating a more focused subset that is effective for distinguishing between different categories. Managing the dimensions of the vector space is important for two reasons. First, when assessing the difficulty of a learning algorithm, both the volume of features and the number of learning examples must be considered. Reducing the number of index terms can improve the efficiency of these algorithms. Intuitively, having more features should lead to better classifiers, as...

EVALUATION METHODOLOGY

This section outlines experimental results using the **"F-measure"**, which is the harmonic

mean of recall and precision, defined as:

$$F= \frac{2*recall*precision}{recall+precision}$$

In text categorization, the two common metrics, recall and precision, are used to assess the algorithm's effectiveness within a particular category. These metrics are incorporated into the F-measure formula:

$$Precision= \frac{True\ positive}{True\ positive+false\ positive}$$

$$recall=\frac{True\ Positive}{True\ positive+false\ negative}$$

We refer to the **MeSH Ontology** in our research, using the biomedical thesaurus developed by the National Library of Medicine (NLM). The **Medical Subject Headings (MeSH)** thesaurus, established to help in indexing and retrieving medical information, has evolved over time, containing thousands of descriptors classified into categories like anatomy, diseases, organisms, and more, with up to 11 hierarchical levels.

For evaluating the proposed methodology, we used the OHSUMED dataset and two data mining algorithms: **C4.5** (a decision tree algorithm) and **KNN** (K-Nearest Neighbors). Both algorithms have proven effective for document classification. Our comparison is based on the representation of stems, and the methodology was tested across eight different groups from the OHSUMED corpus.

FINDINGS

Our study demonstrates that ontology-based representations significantly improve classification accuracy, achieving an impressive 30% performance increase compared to traditional methods. Additionally, the enhancement of the representation vectors through the inclusion of hypernyms and related semantic techniques has contributed to notable performance gains. These results reinforce the suitability of the K-Nearest Neighbor (KNN) algorithm for this classification task, particularly when paired with the χ^2 feature selection method.

The observed synergy between the KNN classifier and the χ^2dimensionality reduction technique appears to play a critical role in optimizing both classification effectiveness and feature representation.

CONCLUSION AND FUTURE WORK

The main objective of our study was to improve document classification by leveraging a domain-specific ontology. We focused on the medical domain due to its significance and the increasing application of data mining techniques in healthcare. Our approach was evaluated on a benchmark dataset using two well-known classification algorithms, KNN and C4.5. Initially, we compared results using the traditional stem-based representation and then incorporated MeSH concepts and hypernyms into the document representation.

The experimental outcomes confirm the effectiveness of our method, demonstrating a 30% improvement in classification accuracy over baseline models. These results underscore the value of subject-specific, ontology-driven document tagging as a powerful strategy for enhancing classification performance.

Looking forward, several avenues for further research are evident. One promising direction is to explore hypernymy relations in greater depth, extending the analysis across multiple levels of the MeSH ontology to potentially boost classification results. Our ultimate aim is to develop a more generalized framework that can be applied across diverse datasets and domains.

Furthermore, expanding this methodology to include a multilingual MeSH ontology could facilitate the classification of medical documents in various languages, thereby broadening the applicability and impact of our research.understand and leverage its potential.

REFERENCE

1. Hassan, Sahar, Franck Hétroy, and Olivier Palombi. "Ontology-guided mesh segmentation." *FOCUS K3D Conference on Semantic 3D Media and Content*. 2010.
2. Osborne, John D., et al. "Interpreting microarray results with gene ontology and MeSH." *Microarray Data Analysis: Methods and Applications* (2007): 223-241.
3. Trieschnigg, Dolf, et al. "MeSH Up: effective MeSH text classification for improved document retrieval." *Bioinformatics* 25.11 (2009): 1412-1418.
4. Elberrichi, Zakaria, Belaggoun Amel, and Taibi Malika. "Medical Documents Classification Based on the Domain Ontology MeSH." *arXiv preprint arXiv:1207.0446*

(2012).

5. Yoo, Illhoi, and Xiaohua Hu. "Biomedical ontology mesh improves document clustering qualify on medline articles: A comparison study." *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, 2006.
6. Ayeldeen, Heba, Aboul Ella Hassanien, and Ali Ali Fahmy. "Evaluation of semantic similarity across MeSH ontology: a Cairo University thesis mining case study." 2013 12th Mexican International Conference on Artificial Intelligence. IEEE, 2013.
7. Gope, Hira Lal, et al. "Medical document classification from OHSUMED dataset." IJCSN International Journal of Computer Science and Network 3.4 (2014): 215-219
8. Katris, Nikolaos, Richard Sutcliffe, and Theodore Kalamboukis. "Using a Cross-Language Information Retrieval System based on OHSUMED to Evaluate the Moses and KantanMT Statistical Machine Translation Systems." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016
9. Sayed, Mahmoud F., and Douglas W. Oard. "Jointly modeling relevance and sensitivity for search among sensitive content." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.
10. Gupta, Vishwa, and Sitesh Kumar Sinha. "Comparison of machine learning methods for OHSUMED-F Data Set: A Cardiovascular Diseases Simulation Study." NeuroQuantology 20.15 (2022): 385.