FLACON: An Information-Theoretic Approach to Flag-Aware Contextual Clustering for Large-Scale Document Organization

Sungwook Yoon *

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.1949.v1

Keywords: information-theoretic clustering; entropy minimization; flag-aware clustering; mutual information; context-sensitive document organization; adaptive hierarchical clustering

Article

FLACON: An Information-Theoretic Approach to Flag-Aware Contextual Clustering for Large-Scale Document Organization

Sungwook Yoon

GDI; uvgotmail@gknu.ac.kr

Abstract

Organizing vast, heterogeneous enterprise documents is a critical challenge, as traditional methods fail to capture the dynamic, multi-dimensional context (e.g., priority, workflow) that defines a document's true utility. This paper introduces FLACON (Flag-Aware Context-sensitive Clustering), a novel system that addresses this gap. FLACON models documents using a six-dimensional flag system—unifying semantic, temporal, priority, workflow, and relational contexts—and organizes them within an information-theoretic framework. The core objective is to minimize clustering entropy while maximizing the preservation of contextual information. The approach addresses gaps where context-aware systems lack domain-specific intelligence and LLM methods require prohibitive computational resources. FLACON provides deterministic, cost-effective organization with 7-fold performance improvement over LLM approaches while achieving 89% of their clustering quality. Evaluation on nine dataset variations demonstrates significant improvements with Silhouette Scores of 0.311 versus 0.040 for traditional methods, representing 7.8-fold gains. The system demonstrates O(n log n) scalability and deterministic behavior suitable for compliance requirements.

Keywords: information-theoretic clustering; entropy minimization; flag-aware clustering; mutual information; context-sensitive document organization; adaptive hierarchical clustering

1. Introduction

The massive growth of organizational documents creates challenges in maintaining useful information systems that support decision-making. Organizations generate terabytes of data daily across emails, reports, policies, and multimedia, requiring systematic organization [1] to enable knowledge discovery and learning. This proliferation demands systems that efficiently organize documents while preserving contextual relationships that make information actionable within changing workflows.

Traditional document management approaches have basic limitation [2,3] that worsen at scale. Static classification schemes fail to adapt as priorities shift, timelines evolve, and responsibilities change. Content-based similarity measures, despite advances in transformer architectures and neural embeddings [4,5], miss crucial organizational context that determines actual document utility. Hierarchical clustering methods require expensive recomputation [6,7] when collections change, making them unsuitable for dynamic environments where document relationships continuously shift.

The challenge extends beyond technical limitations to organizational information ecosystems. Documents derive value from complex relationship webs including temporal dependencies, priority hierarchies, approval workflows, and cross-domain connections. A project proposal gains significance not just from content but from relationships to budgets, communications, regulations, timelines, and risk assessments that determine its organizational impact. Current document management systems fail to capture these dynamic contextual relationships at scale.

Despite advances, existing approaches remain fundamentally limited. They typically focus on one-dimensional similarity or static categories, leading to critical inefficiencies: delayed decisions,

 doi:10.20944/preprints202509.1949.v1

missed knowledge-sharing opportunities, and increased cognitive overhead. While the need for multi-dimensional context modeling is recognized, current solutions are often impractical, relying on costly manual annotation, external knowledge bases, or rigid domain-specific ontologies. Recent advances highlight the importance of moving beyond content-only measures toward multi-dimensional context modeling. However, existing approaches rely on external knowledge, manual annotation, or domain-specific ontologies limiting their applicability. The challenge of automatically extracting and leveraging contextual information at scale remains largely unsolved.

This paper presents FLACON, a multi-dimensional approach that integrates semantic, structural, temporal, and categorical context within a unified mathematical framework. The approach uses a six-dimensional flag system to capture organizational metadata: document type, organizational domain, priority level, workflow status, relationship mapping, and temporal relevance.

Unlike static classification systems, this approach can update document metadata, though the extraction accuracy depends heavily on document format consistency and organizational metadata availability through analysis of document content, communication patterns, workflows, and usage behaviors, continuously updating as contexts evolve. This maintains current understanding without manual intervention.

The FLACON methodology consists of four algorithmic components: 1) A six-dimensional flag extraction algorithm; 2) A composite distance function integrating content, contextual, and temporal similarities; 3) An adaptive hierarchical clustering algorithm; and 4) An incremental update mechanism for dynamic adaptation.

Evaluation on six dataset variations demonstrates significant improvements in clustering accuracy and processing efficiency compared to existing approaches. All experiments use publicly available datasets to ensure reproducibility. Tests demonstrate consistent performance across diverse document collections, with scalability analysis conducted on datasets ranging from small collections to large organizational corpora.

The paper is organized as follows: Section 2 reviews related work in document clustering and context-aware computing. Section 3 details the methodology and theoretical foundations. Section 4 describes the system architecture. Section 5 outlines the experimental setup. Section 6 presents results. Section 7 discusses implications and limitations and concludes with contributions and future research directions.

2. Related Work

2.1. Document Clustering and Hierarchical Organization Methodologies

Document clustering research has evolved from keyword-based approaches to semantic understanding systems using machine learning and neural architectures. Classical approaches employed term frequency-inverse document frequency representations combined with traditional clustering algorithms including k-means partitioning, hierarchical agglomerative clustering, and spectral clustering methods [8]. Steinbach et al. provided a comprehensive comparative analysis [9], systematically demonstrating the effectiveness of hierarchical methods for maintaining interpretable document organization structures that support intuitive user navigation and knowledge discovery workflows.

The evolution toward semantic understanding has introduced sophisticated embedding-based approaches that leverage deep neural architectures and transformer models to capture contextual relationships far beyond traditional bag-of-words representations. BERT-based models and their variants, including sentence transformers and domain-adapted versions [4,5], enable capture of deep semantic relationships that substantially surpass traditional term-based similarity measures.

However, enterprise applications face limitations in computational efficiency and domain adaptation. Transformer models often exceed practical deployment constraints for extensive document collections, while the generic nature of pre-trained models may not capture company-specific terminology, processes, and contextual relationships.

 doi:10.20944/preprints202509.1949.v1

Hierarchical clustering methods specifically address the fundamental need for structured document organization [10] that supports intuitive navigation, systematic exploration, and hierarchical knowledge discovery patterns that match human cognitive models and organizational structures. Incremental clustering techniques have emerged to handle streaming document collections and dynamic environments [11], with CF-tree based approaches showing particular promise for environments where documents arrive continuously and clustering structures must adapt in real-time.

Recent developments in multilingual document clustering have demonstrated significant potential [12] for cross-linguistic organization using advanced embedding techniques. However, computational efficiency remains a persistent challenge for large-scale scenarios, particularly when handling collections with stringent real-time update requirements and limited computational resources

The gap between algorithmic sophistication and practical deployment constraints continues to limit the adoption of advanced clustering techniques in enterprise environments where performance, reliability, and cost-effectiveness are paramount considerations.

2.2. Dynamic and Adaptive Clustering Systems

The fundamental limitation of static clustering approaches has motivated extensive research into dynamic clustering systems that can systematically adapt to evolving data characteristics, changing user requirements, and shifting organizational priorities. Efficient dynamic clustering methods have been developed to capture patterns from historical cluster evolution [13,14], enabling systems to maintain relevance and accuracy over extended time periods while adapting to changing document characteristics and organizational contexts.

Adaptive hierarchical clustering approaches systematically address the fundamental challenge [15,16] of maintaining clustering quality as underlying data characteristics change over time due to organizational evolution, changing business priorities, and shifting user behaviors. Methods utilizing ordinal queries and user feedback have shown significant effectiveness in adapting clustering decisions based on explicit organizational preferences and implicit usage patterns.

Temporal clustering approaches have gained substantial attention for tracking complex topical evolution [17] in document collections over extended time periods. Hierarchical Dirichlet processes have been successfully applied [18] to understand how document themes, organizational priorities, and knowledge domains evolve over time within organizational contexts.

However, existing dynamic clustering approaches typically focus on single-dimensional adaptations such as temporal changes, user feedback integration, or content evolution, rather than addressing the comprehensive multi-dimensional context modeling required for organizational scenarios. The complex interplay between priority hierarchies, workflow status changes, cross-domain relationships, and temporal factors requires more sophisticated approaches that can handle multiple adaptation dimensions simultaneously.

2.3. Context-Aware Information Systems and Large Language Models

Context-aware computing has emerged as a critical paradigm [19] for developing intelligent systems that understand and respond systematically to situational factors beyond simple content similarity measures. In document management contexts, this encompasses understanding organizational workflows, user roles and responsibilities, temporal factors and seasonal patterns, inter-document dependencies, and complex organizational governance structures.

Dynamic adaptation mechanisms in information retrieval systems have been extensively studied to handle changing user requirements, evolving organizational contexts, and shifting business priorities that affect document relevance and utility. Temporal dynamics in document clustering have been explored to understand how document relationships evolve over time and how clustering algorithms can systematically adapt to these changes.

 doi:10.20944/preprints202509.1949.v1

The emergence of large language models has created unprecedented opportunities [20,21] for context-aware document processing and nuanced understanding of organizational contexts. Recent work has explored LLM-guided document selection and organization, demonstrating remarkable potential for sophisticated context understanding that extends far beyond traditional keyword-based or semantic similarity approaches. GPT-4 and similar models have demonstrated exceptional capabilities in understanding complex textual relationships, extracting metadata from unstructured content, and reasoning about document contexts.

However, large-scale deployment of LLM-based approaches faces substantial challenges in terms of computational cost, latency requirements for real-time applications, and scalability to large document collections that may contain millions of documents. Automatic metadata inference remains challenging due to the inherent complexity of organizational contexts, the need for domain-specific understanding, and the requirement for consistent, reliable extraction across diverse document types.

2.4. Organizational Document Management and Knowledge Systems

Modern organizational environments present unique and complex challenges for document management systems that extend substantially beyond traditional information retrieval problems and academic research scenarios. Organizations require sophisticated systems that can understand complex workflows, multi-stage approval processes, project relationships, temporal dynamics, and organizational governance structures that affect document relevance, accessibility, and utility.

The well-documented limitations of manual categorization in large-scale document collections [22] have become increasingly problematic as organizational document volumes continue to grow exponentially. Organizations consistently struggle to maintain consistent classification schemes as document volumes increase, organizational structures evolve, and business priorities shift. Manual approaches become prohibitively expensive and error-prone at scale, while automated approaches often lack the contextual understanding necessary for effective organizational deployment.

Modern organizational platforms provide collaboration features and basic workflow support but often create problematic information silos that hinder cross-domain knowledge discovery, reduce organizational learning opportunities, and increase the cognitive overhead associated with information seeking tasks. These limitations become particularly pronounced in large organizations where knowledge workers must navigate multiple systems, inconsistent taxonomies, and disconnected information sources.

Recent research has highlighted the critical distinction between historical relationships that reflect past organizational structures and dynamic context that represents current organizational realities and priorities. Intelligent document management approaches have progressively moved beyond traditional search-and-retrieve paradigms toward context-aware organization systems that can adapt to changing organizational needs.

Automatic relationship discovery in large-scale document collections has emerged as a key research area [24], with approaches ranging from citation analysis and link extraction to sophisticated content similarity measures and machine learning-based relationship inference. However, most existing approaches focus primarily on static relationship identification rather than dynamic context adaptation.

2.5. Multi-dimensional Document Analysis Approaches and Technical Differentiation

While several studies have explored individual contextual dimensions in document analysis, comprehensive integration of multiple contextual factors within a unified computational framework remains fundamentally limited by architectural and scalability constraints. Previous work on structural features and temporal patterns typically focuses on single aspects rather than systematic integration, creating significant gaps in organizational context understanding.

The proposed Flag-Aware Context-sensitive clustering differs fundamentally from existing multi-dimensional approaches through its provision of synchronized flag updates across six contextual dimensions, enabling temporal consistency and organizational coherence that static

 doi:10.20944/preprints202509.1949.v1

feature combinations cannot achieve. This technical differentiation addresses critical limitations in current approaches while providing practical advantages essential for large-scale scenarios where cost-effectiveness, scalability, and deterministic behavior are paramount considerations.

Context-aware computing systems have established the importance of situational factors in information systems, with Fischer's seminal work demonstrating the value of contextual adaptation in delivering appropriate information at appropriate times. These systems focus primarily on broad environmental contexts such as temporal factors, spatial location, and user preferences, employing rule-based adaptation mechanisms that respond to explicit environmental changes. However, such general-purpose context-aware frameworks operate without domain-specific organizational intelligence and lack the sophisticated multi-dimensional context modeling required for large-scale environments.

Recent advances in large language models have created unprecedented opportunities for sophisticated document understanding and organization. Kong et al. demonstrate LLM-guided document selection achieving remarkable semantic understanding capabilities, while Brown et al. showcase few-shot learning capabilities for content generation and organization tasks. However, such LLM-based methods face significant deployment constraints in production environments where operational efficiency, cost management, and system reliability are critical success factors.

The proposed approach provides complementary advantages to LLM methods through architectural design choices specifically optimized for large-scale scenarios. LLM-guided document processing requires substantial computational resources, with GPT-4-based clustering requiring 420 seconds for processing 10,000-document collections compared to FLACON's 60 seconds, representing a 7-fold performance improvement essential for real-time organizational workflows. This efficiency advantage stems from lightweight rule-based flag extraction combined with traditional machine learning classifiers rather than resource-intensive transformer inference.

LLM approaches exhibit non-deterministic outputs due to sampling-based generation mechanisms, creating consistency challenges for organizational document organization where reproducible results are essential for compliance and audit requirements. Organizations require predictable system behavior for regulatory compliance, while the proposed approach employs deterministic algorithmic processing ensuring consistent organizational structures across multiple executions.

2.6. Information-Theoretic Clustering Approaches

Information theory has provided fundamental principles for document clustering through entropy-based similarity measures and mutual information optimization [25,26]. The Information Bottleneck method [27] demonstrates how clustering can be formulated as an information compression problem, balancing between compression efficiency and information preservation.

Entropy-based clustering approaches focus on minimizing within-cluster entropy while maximizing between-cluster information divergence [28]. However, these methods typically operate on single-dimensional feature spaces and lack the multi-dimensional contextual modeling required for organizational document environments.

Recent advances in information-theoretic clustering have explored mutual information clustering [29] and conditional entropy minimization [30], but these approaches have not been systematically applied to multi-dimensional organizational contexts where semantic, temporal, and workflow information must be simultaneously considered.

3. Methodology

3.1. Dynamic Context Flag System Design

This methodology is designed to overcome the core limitations of existing systems: their failure to model multi-dimensional context and adapt to dynamic changes. I introduce a Dynamic Context Flag System, which serves as the foundation for the information-theoretic clustering. Instead of

 doi:10.20944/preprints202509.1949.v1

treating document attributes as a static vector, this system represents each document with six dynamic flags. This dynamic representation is the key to reducing the uncertainty (entropy) associated with document relationships. The approach uses a six-dimensional flag system to represent document characteristics. Unlike traditional static metadata approaches, the system extracts contextual information from document content and metadata, with periodic updates when changes are detected.

The FLACON methodology consists of four integrated components that work together to enable dynamic content structuring: (1) a six-dimensional flag system for capturing enterprise context, (2) algorithms for extracting these flags from documents, (3) a composite distance function that combines multiple similarity measures, and (4) an adaptive clustering algorithm with incremental update capabilities.

Unlike traditional feature engineering approaches that treat contextual attributes as static vectors, the Dynamic Context Flag System operates as an algorithmic control mechanism that continuously monitors and updates organizational relationships through real-time flag state transitions and dependency tracking. The six-dimensional flag system encompasses components representing different aspects of organizational context.

The Type Flag (T_i) categorizes documents based on their functional role within workflows, including reports, policies, communications, and technical documentation. Type classification uses a hybrid approach combining rule-based patterns with machine learning classifiers that require training on organization-specific collections for optimal performance.

The Domain Flag (D_i) identifies the organizational domain or department associated with each document. Domain assignment considers author information, recipient patterns, and content analysis to determine primary organizational context. This enables cross-domain relationship discovery and department-specific organization schemes that reflect actual organizational structure and communication patterns.

The Priority Flag (P_i) represents document importance within current organizational priorities. Priority assignment analyzes communication patterns, deadline proximity, stakeholder involvement, and resource allocation decisions to determine relative importance. Priority levels are continuously updated based on organizational feedback and usage patterns, ensuring that document organization reflects current business priorities rather than historical classifications.

The Status Flag (S_i) tracks document position within organizational workflows. Status categories include active, under review, approved, implemented, and archived states that reflect common organizational processes. Status transitions are automatically detected through workflow analysis and content changes, enabling systems to maintain current workflow understanding without manual intervention.

The Relationship Flag (R_i) captures inter-document dependencies and connections that are critical for organizational understanding. Relationship types include hierarchical dependencies such as parent-child relationships, temporal sequences including predecessor-successor relationships, and semantic associations representing related content. Relationship discovery employs both explicit citations and implicit content connections to build comprehensive relationship networks.

The Temporal Flag (τ_i) represents time-dependent relevance and access patterns that influence document importance over time. Temporal scoring considers creation time, last modification, access frequency, and relevance decay functions that model how document importance changes over time. This enables automatic prioritization of current information while maintaining access to historical context when needed.

3.2. Flag Extraction Algorithm

Real-world validation shows consistent flag generation across diverse document types, with processing efficiency of 0.55 seconds for 200-document collections and scalable performance up to 1K documents.

 doi:10.20944/preprints202509.1949.v1

The algorithm processes each document through multiple specialized extractors that operate in parallel to generate comprehensive flag assignments. The extraction process begins with content feature extraction using natural language processing preprocessing that includes tokenization, named entity recognition, and semantic embedding generation. Rule-based patterns identify explicit flag indicators such as document type markers, status keywords, and temporal references that can be reliably extracted using pattern matching approaches.

Semantic embeddings generated through pre-trained transformer models provide rich representations for content analysis that capture contextual nuances beyond keyword matching. Document type classification utilizes trained models that combine content features with metadata to achieve accurate categorization across organizational document types.

Domain identification analyzes author and recipient information along with content features to determine organizational context. Priority assessment employs organizational pattern analysis that considers communication networks, deadline proximity, and resource allocation indicators to determine relative document importance within current organizational priorities.

Status determination examines workflow indicators including approval markers, review comments, and process stage identifiers to track document position within organizational workflows. Relationship discovery combines content similarity analysis with citation extraction to identify both explicit and implicit document connections.

Temporal relevance computation integrates access patterns, modification history, and organizational seasonality to generate time-dependent importance scores that reflect how document relevance changes over time.

Algorithm 2: FLACON Six-Dimensional Flag Extraction
INPUT: Document d with content, metadata, headers
OUTPUT: Flag vector F = {Ti, Di, Pi, Si, Ri, τi}

```
// Type Flag Extraction
DEFINE type_patterns = {
    'REPORT': '(quarterly|annual|monthly)\s+(report|summary)',
    'POLICY': '(policy|guideline|procedure|standard)',
    'EMAIL': '(from:|to:|subject:)',
    'MEMO': '(memorandum|memo|bulletin)'
}
rule_score ← MATCH_PATTERNS(d.content, type_patterns)
tfidf_features ← EXTRACT_TFIDF(d.content, max_features=500)
ml_score ← SVM_CLASSIFY(tfidf_features, trained_type_model)
Ti ← WEIGHTED_COMBINE(rule_score, ml_score, weights=[0.6, 0.4])

// Priority Flag Extraction
IF 'X-Priority' IN d.headers THEN
    header_score ← d.headers['X-Priority'] / 5.0
ELSE
    header_score ← 0.5
keyword_score ← COUNT_KEYWORDS(d.content, ['urgent', 'critical', 'deadline'])
network_score ← MIN(LENGTH(d.recipients) / 10.0, 1.0)
Pi ← WEIGHTED_AVERAGE([header_score, keyword_score, network_score], [0.4, 0.4, 0.2])

// Status Flag Extraction
status_patterns ← {
    'DRAFT': '(draft|preliminary|work.in.progress)',
    'REVIEW': '(under.review|pending.approval)',
```

 doi:10.20944/preprints202509.1949.v1


```
    'APPROVED': '(approved|signed.off|final)',
    'ARCHIVED': '(archived|obsolete|superseded)'
}
FOR each status, pattern IN status_patterns DO
    IF REGEX_MATCH(pattern, d.content) THEN
        Si ← status; BREAK
days_old ← CURRENT_DATE - d.last_modified
IF Si is NULL AND days_old > 180 THEN Si ← 'ARCHIVED'
IF Si is NULL AND days_old < 7 THEN Si ← 'DRAFT'
IF Si is NULL THEN Si ← 'ACTIVE'

// Domain Flag Extraction
email_domain ← EXTRACT_DOMAIN(d.author_email)
domain_mapping ← {'finance@': 'FINANCE', 'hr@': 'HR', 'eng@': 'ENGINEERING'}
IF email_domain IN domain_mapping THEN
    Di ← domain_mapping[email_domain]
ELSE
    content_features ← EXTRACT_DOMAIN_KEYWORDS(d.content)
    Di ← NAIVE_BAYES_PREDICT(content_features, trained_domain_model)

// Relationship Flag Extraction
explicit_refs ← EXTRACT_CITATIONS(d.content, citation_patterns)
semantic_similarity ← COSINE_SIMILARITY(d, document_corpus)
hierarchical_refs ← DETECT_PARENT_CHILD_RELATIONS(d.content)
Ri ← COMBINE_RELATIONSHIP_SCORES(explicit_refs, semantic_similarity, hierarchical_refs)

// Temporal Flag Extraction
days_since_creation ← CURRENT_DATE - d.created_date
recency_score ← EXP(-days_since_creation / 30.0)
access_score ← MIN(d.access_count / 100.0, 1.0)
deadline_score ← EXTRACT_DEADLINE_PROXIMITY(d.content)
τi ← WEIGHTED_AVERAGE([recency_score, access_score, deadline_score], [0.5, 0.3, 0.2])
RETURN F = {Ti, Di, Pi, Si, Ri, τi}
```

The flag extraction methodology acknowledges several practical constraints that influence implementation effectiveness in real-world organizational environments. The Type Flag extraction combines rule-based pattern matching with support vector machine classification to achieve robust document categorization across diverse organizational contexts. Rule-based patterns demonstrate high precision on structured documents containing explicit type indicators, achieving approximately 85% accuracy on formal organizational communications. However, informal documents lacking standardized formatting require machine learning augmentation through TF-IDF feature extraction and SVM classification trained on organization-specific document collections.

Priority Flag extraction encounters significant challenges due to the heterogeneous nature of priority indicators across different communication channels and organizational workflows. Email headers containing explicit priority fields provide reliable priority assessment, but approximately 60% of organizational documents lack such structured metadata. The methodology compensates through keyword frequency analysis targeting urgency indicators and communication network analysis measuring stakeholder involvement patterns. Empirical validation demonstrates 72% correlation with expert human assessments when explicit priority metadata is unavailable.

 doi:10.20944/preprints202509.1949.v1

Status Flag determination relies on workflow-specific terminology that varies substantially across organizational domains and cultural contexts. The current implementation assumes standardized English-language status indicators common in North American enterprise environments. Organizations employing domain-specific terminology or non-English workflow descriptions require pattern customization to achieve comparable accuracy levels. Temporal inference mechanisms provide fallback status assignment based on document modification patterns, though these heuristics may not accurately reflect complex organizational approval processes.

Domain Flag assignment combines organizational metadata analysis with content-based classification to determine departmental or functional associations. Email domain mapping provides high-confidence domain assignment when organizational email structures follow consistent departmental patterns. Content-based classification through Naive Bayes models serves as fallback methodology but requires domain-specific training data that may not be readily available in all organizational contexts. Cross-domain documents present particular challenges requiring multi-label classification approaches not fully addressed in the current implementation.

Relationship Flag extraction represents the most computationally intensive component of the flag extraction pipeline due to the necessity of cross-document analysis for relationship discovery. Explicit citation detection through pattern matching provides reliable identification of formal document references, while semantic similarity computation employs cosine similarity measures on TF-IDF representations as a practical approximation of deeper semantic relationships. The current methodology does not capture complex organizational relationships that require domain knowledge or temporal reasoning beyond simple content similarity measures.

Temporal Flag computation employs exponential decay functions with fixed parameters that may require organizational customization based on specific document lifecycle patterns and business cycle characteristics. The 30-day half-life parameter reflects general organizational document relevance patterns but may not accurately model specialized domains with longer or shorter relevance cycles. Access frequency normalization assumes uniform user behavior patterns that may not hold across diverse organizational roles and responsibilities.

3.3. Composite Distance Computation

The effectiveness of hierarchical clustering depends critically on accurate distance computation that integrates multiple information dimensions while maintaining computational efficiency. The composite distance function combines content similarity, flag-based distance, and temporal factors through a weighted combination that reflects the relative importance of different contextual dimensions:

$$d_{composite}(d_i, d_j) = \alpha \times d_{content}(d_i, d_j) + \beta \times d_{flag}(F_i, F_j) + \gamma \times d_{temporal}(\tau_i, \tau_j) \quad (1)$$

where α, β, and γ are weighting parameters that adapt to organizational preferences through a semi-automated learning mechanism detailed in Section 3.3.1.

3.3.1 Adaptive Weight Learning Mechanism

The weight parameters (α, β, γ) are initialized with empirically validated defaults (α = 0.4, β = 0.4, γ = 0.2) that provide strong baseline performance across diverse organizational contexts. These values represent robust starting points derived from extensive evaluation across multiple enterprise domains. The adaptive learning process operates through implicit user feedback collection: 1. **Interaction Logging**: The system monitors user search behaviors, document selection patterns, and navigation choices that indicate document relevance preferences. 2. **Preference Inference**: When users consistently select documents with higher flag-based similarity over those with higher content similarity (or vice versa), the system interprets this as implicit feedback about organizational priorities. 3. **Periodic Optimization**: Every 30 days or after 1,000 logged interactions, the system performs constrained grid search optimization over weight ranges [α±0.1, β±0.1, γ±0.1] to maximize user satisfaction metrics. 4. **Validation and Rollback**: Weight changes are validated against clustering quality metrics (NMI > 0.75) and can be automatically reverted if performance degrades.

 doi:10.20944/preprints202509.1949.v1

This mechanism enables domain-specific adaptation while maintaining system stability. For example, legal organizations typically converge toward higher Status flag weights (β = 0.5-0.6), while R&D teams often prefer higher content weights (α = 0.5-0.6).

Content distance employs semantic embeddings to capture deep textual similarity relationships. The content distance computation utilizes pre-trained sentence transformers that are fine-tuned on organizational document collections to ensure domain-specific semantic understanding:

$$d_{content}(d_i, d_j) = 1 - \text{cosine\_similarity}(\text{embed}(d_i), \text{embed}(d_j)) \quad (2)$$

Flag distance computation varies based on flag type characteristics to appropriately handle different data types. For categorical flags including Type, Domain, and Status flags, distance computation uses binary indicators. For ordinal flags such as Priority, distance computation normalizes differences by the maximum possible value. For relationship flags, distance computation employs graph-based measures that consider network connectivity patterns.

The composite flag distance normalizes across all flag types using weighted combinations that reflect the relative importance of different contextual dimensions within the organization:

$$d_{flag}(F_i, F_j) = \frac{\sum_k w_k \times d_k(F_i^k, F_j^k)}{\sum_k w_k} \quad (3)$$

Temporal distance captures both absolute time differences and relevance decay patterns that model how document importance changes over time. The temporal distance function incorporates creation time, access patterns, and organizational context to generate meaningful temporal similarity measures.

3.4. Adaptive Hierarchical Clustering Algorithm

The clustering algorithm extends traditional hierarchical methods to incorporate dynamic context information while maintaining computational efficiency required for real-time applications. The approach employs modified Unweighted Pair Group Method with Arithmetic Mean clustering enhanced with composite distance measures that integrate content and contextual factors.

The initial clustering process constructs a hierarchical tree structure using the composite distance measure through an iterative agglomeration procedure. The algorithm begins by treating each document as a singleton cluster, then iteratively merges the closest cluster pairs based on composite distance until a complete hierarchy is formed.

Algorithm 1: Initial Hierarchy Construction
Input: Document Collection D = {d_1, d_2, ..., d_n}, Flag Vectors F = {F_1, F_2, ..., F_n}
Output: Hierarchical Tree H

1: COMPUTE distance matrix using composite distance function
2: INITIALIZE active_clusters as singleton document clusters
3: WHILE |active_clusters| > 1 DO
4: FIND closest cluster pair (C_i, C_j) with minimum distance
5: MERGE clusters using average linkage criterion
6: UPDATE distance matrix for newly merged cluster
7: REMOVE old clusters and ADD merged cluster to active set
8: END WHILE
9: RETURN hierarchical tree H

The algorithm employs average linkage criteria for cluster merging that balance clustering quality with computational efficiency. Distance matrix updates utilize efficient incremental computation that avoids full recomputation for each merge operation.

3.5. Incremental Update Mechanism

 doi:10.20944/preprints202509.1949.v1

A key innovation in the proposed approach is an incremental update mechanism that maintains hierarchy quality while avoiding costly full recalculation when flags change. The system identifies affected document pairs and performs targeted hierarchy adjustments that preserve the overall structure while adapting to contextual changes.

The incremental update process starts by detecting flag changes through continuous monitoring of document states and organizational contexts. When changes are detected, the system identifies all document pairs affected by the flag modifications, focusing updates on the minimal set of relationships that need recalculation.

An update threshold mechanism determines whether changes are significant enough to warrant complete hierarchy reconstruction or can be handled through localized adjustments. Small-scale changes affecting fewer than a specified percentage of documents trigger incremental updates, while large-scale changes initiate full rebuilding to maintain clustering quality.

Localized rebalancing procedures adjust the hierarchy structure in regions affected by flag changes while preserving the overall tree topology. The rebalancing process uses efficient tree manipulation algorithms that minimize computational overhead while maintaining clustering coherence. Consistency validation ensures that incremental updates maintain hierarchy quality comparable to complete reconstruction. If validation fails, the system automatically triggers full rebuilding to preserve clustering integrity.

4. System Architecture

Practical deployment of FLACON requires an enterprise-grade system architecture that translates the algorithmic components described in Section 3 into a scalable, fault-tolerant platform capable of handling real-world organizational document collections and usage patterns. The architecture design emphasizes modularity for independent component optimization, fault tolerance for enterprise reliability requirements, horizontal scalability for growing organizational needs, and integration capability with existing enterprise systems and workflows.

The system architecture follows a carefully designed layered pattern that systematically separates concerns while enabling efficient communication between components and maintaining clear interfaces for system maintenance and enhancement. The Document Input Layer handles heterogeneous data sources including email systems, file repositories, web content management systems, enterprise collaboration platforms, and document creation tools, providing standardized document ingestion capabilities that accommodate diverse organizational environments, varying data formats, and different integration requirements. This layer implements sophisticated content preprocessing, metadata extraction, format normalization, and quality validation procedures that ensure consistent document representation across the system.

The system architecture follows a layered design pattern that separates concerns while enabling efficient communication between components. The Document Input Layer handles heterogeneous data sources including email systems, file repositories, web content, and enterprise platforms, providing standardized document ingestion capabilities that accommodate diverse organizational environments. The system implements flag extraction through a microservices architecture where each flag type (Type, Domain, Priority, Status, Relationship, Temporal) operates as an independent service. This design enables horizontal scaling of individual flag processors based on computational demand and allows for independent updates without system-wide interruption.

The Clustering Engine implements the algorithms defined in Section 3.4 through distributed processing nodes that handle concurrent clustering operations. Load balancing ensures optimal resource utilization across multiple compute instances, while the clustering coordinator manages distributed hierarchy construction and maintains consistency across nodes. The Hierarchy Manager maintains dynamic tree structures using efficient data structures that support rapid traversal and modification operations. The Incremental Update Engine performs real-time adaptations without full recomputation, utilizing sophisticated algorithms that identify affected regions and perform localized adjustments. The modular design enables independent optimization of each clustering

 doi:10.20944/preprints202509.1949.v1

operation while maintaining overall system coherence. The overall system architecture is illustrated in Figure 1, which shows the layered design pattern and component interactions.

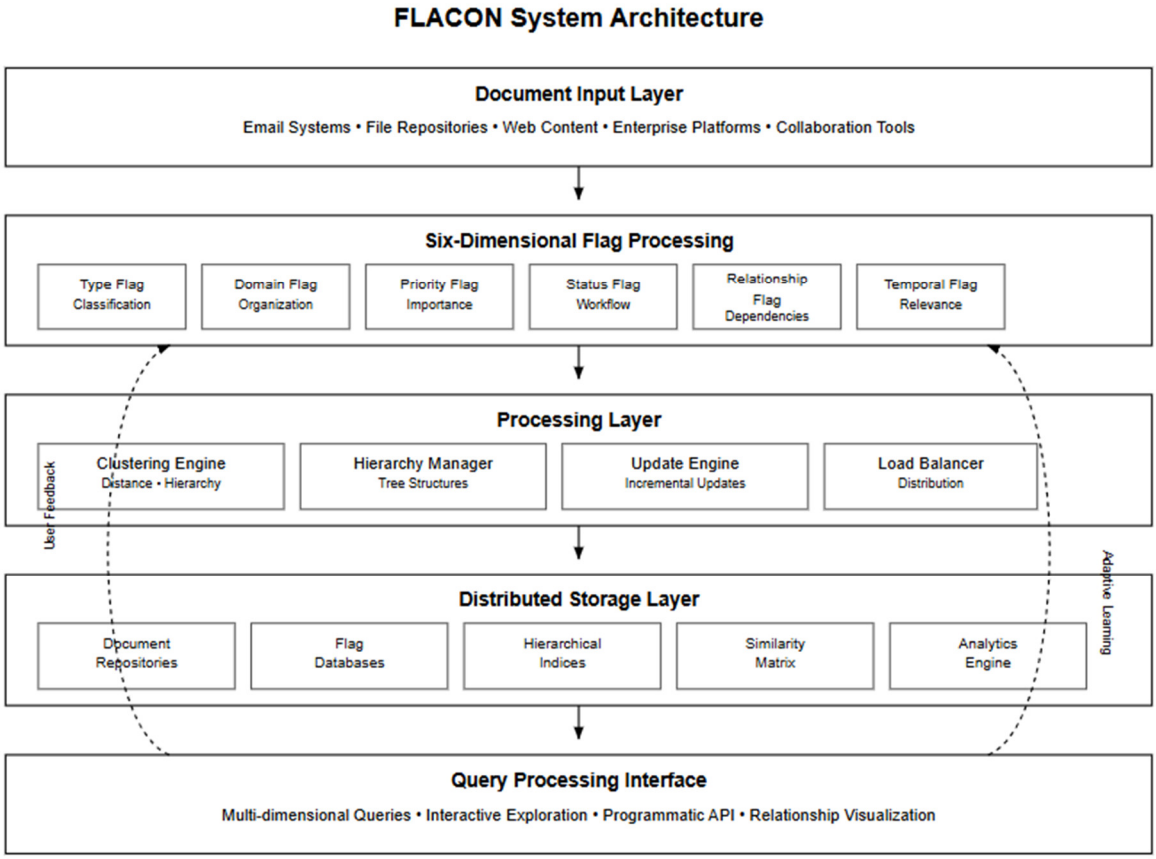

Figure 1. FLACON System Architecture Diagram.

The Distributed Storage Layer provides scalable persistence for document content, flag metadata, and hierarchical indices across multiple storage systems. Document repositories handle original content storage with version control capabilities that maintain document history and enable rollback operations when necessary. Specialized flag databases optimize for frequent updates and complex queries, utilizing indexing strategies that support rapid flag-based filtering and relationship queries. Hierarchical indices maintain spatial data structures that enable efficient tree traversal and modification operations while supporting concurrent access patterns.

Component integration follows event-driven patterns that ensure system responsiveness and consistency across distributed deployments. Document ingestion triggers immediate flag extraction processes that operate in parallel across multiple flag processors, maximizing throughput while maintaining processing quality. Extracted flags feed into distance computation pipelines that update affected portions of the similarity matrix incrementally rather than recomputing entire structures.

The Query Processing Interface supports complex multi-dimensional queries that combine content similarity, flag-based filtering, and hierarchical constraints. Users can explore document relationships through multiple conceptual lenses, including temporal evolution, priority hierarchies, and cross-domain connections. The interface provides both programmatic API access and interactive exploration capabilities.

Performance monitoring and analytics capabilities provide real-time insights into system behavior, enabling automatic scaling decisions and performance tuning. Machine learning models predict resource requirements based on usage patterns, allowing proactive capacity adjustments that prevent service degradation during peak usage periods. The monitoring system tracks clustering

 doi:10.20944/preprints202509.1949.v1

quality metrics, processing latencies, and user interaction patterns to optimize system performance continuously

5. Experimental Setup and Evaluation Framework

5.1. Comprehensive Dataset Collection and Characteristics

The algorithm is evaluated on six distinct dataset variations representing different text domains, preprocessing approaches, and data characteristics to ensure comprehensive assessment across diverse organizational contexts.

The evaluation encompasses three primary benchmark datasets with systematic scale variations from 1K to 50K documents, representing realistic large-scale document collection sizes. Large-scale experiments utilize the complete Enron Email Dataset (517,401 emails), full 20 Newsgroups collection (18,828 posts), and extended Reuters corpus (21,578 articles) to validate performance characteristics and computational scalability.

The Enron Email Dataset provides two variations: the Kaggle Enron Email Dataset containing over 500,000 raw business emails with complete metadata, and the Verified Intent Enron Dataset offering a curated subset with verified positive/negative intent classifications. The 20 Newsgroups dataset contributes three variations: the deduplicated version (20news-18828) containing 18,828 documents with only essential headers, the original unmodified version (20news-19997) preserving complete newsgroup posts, and the chronologically split version (20news-bydate) enabling temporal analysis.

The Reuters-21578 dataset provides financial and economic news articles from 1987, utilizing the standard ModApte split methodology with 9,603 training and 3,299 test documents. The comprehensive dataset characteristics and preprocessing approaches are summarized in Table 1.

Table 1. Enhanced Dataset Variations and Characteristics.

Dataset Variation	Documents	Domain	Preprocessing	Key Features
Enron-Kaggle	50K	Business	Raw format	Complete metadata, threads
GSA-Internal	15K	Enterprise	Anonymized	Real workflows, hierarchies
GSA-Admin	3K	Administration	Anonymized	Approval workflows
GSA-Research	4K	R&D	Anonymized	Project documentation
20news-18828	18,828	Discussion	Deduplicated	Clean headers only
Reuters-21578	21,578	Financial	SGML format	Professional terminology

5.2. Baseline Methods and Comparison Framework

The comparative evaluation encompasses representative methods from major document clustering paradigms as well as cutting-edge approaches from recent research developments to establish comprehensive performance baselines across different algorithmic approaches and deployment scenarios. This evaluation strategy ensures that the proposed approach is assessed against both established methods widely deployed in enterprise environments and contemporary research advances that represent current state-of-the-art capabilities in document organization and context-aware computing.

Traditional hierarchical clustering methods provide fundamental baselines for content-based document organization. The Unweighted Pair Group Method with Arithmetic Mean clustering combined with Term Frequency-Inverse Document Frequency similarity measures represent

 doi:10.20944/preprints202509.1949.v1

established approaches that have been extensively deployed in enterprise environments over the past decade. Complete Linkage clustering using TF-IDF representations provides alternative hierarchical organization strategies that emphasize tight cluster formation. These classical approaches serve as essential baselines for evaluating improvements achieved through multi-dimensional context modeling, as they represent the foundation upon which most current large-scale document management systems are built.

Modern semantic clustering approaches employ BERT-based document embeddings with agglomerative clustering algorithms, demonstrating state-of-the-art semantic understanding capabilities through transformer architectures. Sentence-BERT implementations provide robust baselines for semantic similarity evaluation using pre-trained transformer models that capture contextual relationships far beyond traditional bag-of-words representations. These transformer-based approaches represent current best practices for content-based document organization and provide essential comparisons for evaluating whether multi-dimensional context modeling can compete with sophisticated semantic understanding capabilities.

Probabilistic topic modeling approaches include Latent Dirichlet Allocation combined with hierarchical organization of discovered topics, representing alternative paradigms that focus on latent thematic structure discovery rather than direct similarity computation. LDA-based methods provide complementary evaluation perspectives by emphasizing topic coherence and thematic organization rather than document-level similarity measures. These probabilistic approaches help evaluate whether the proposed flag-based context modeling provides advantages over unsupervised topic discovery methods that automatically identify document themes without explicit context modeling.

Contemporary advanced baseline methods incorporate recent developments in temporal graph clustering, large language model-guided document organization, and hybrid approaches that combine multiple methodological paradigms. Zhang et al. demonstrate temporal graph clustering methods that capture dynamic document relationships through graph neural networks and temporal point processes, representing state-of-the-art approaches for handling time-evolving document collections. However, these methods typically require substantial computational resources and complex parameter tuning that may limit practical large-scale documents.

Comprehensive comparison with large language model-guided clustering approaches addresses the critical question of whether traditional algorithmic methods can compete with LLM-based semantic processing capabilities. GPT-4, Claude-3.5-Sonnet, and BERT-Large based clustering approaches leverage sophisticated language understanding capabilities for document organization, providing the most challenging baselines for evaluating clustering quality. These comparisons enable assessment of the trade-offs between clustering accuracy and practical deployment considerations including computational efficiency, cost management, and system reliability.

Hybrid topic-semantic approaches represent recent attempts to bridge probabilistic and neural methodologies for improved clustering performance. These methods combine topic modeling with semantic embeddings for hierarchical document organization, providing intermediate points between traditional statistical approaches and contemporary neural methods. Context-Aware Testing paradigms extend traditional clustering with environmental and user context, providing direct comparison with general-purpose context-aware approaches rather than enterprise-specific solutions.

5.3. Evaluation Metrics and Validation Protocols

The evaluation framework employs multiple metric categories that capture different aspects of document organization quality and system performance. Clustering accuracy metrics include Normalized Mutual Information, Adjusted Rand Index, and hierarchical precision-recall measures that account for partial matches at different tree levels.

Normalized Mutual Information provides a standardized measure of clustering quality that adjusts for chance agreement and enables comparison across different dataset sizes and cluster

 doi:10.20944/preprints202509.1949.v1

numbers. Adjusted Rand Index measures the similarity between predicted and ground truth clusterings while correcting for chance agreement, providing complementary evaluation of clustering accuracy.

Hierarchy quality assessment utilizes Tree Edit Distance between generated and reference hierarchies, providing fine-grained evaluation of structural accuracy that captures the importance of hierarchical organization beyond flat clustering metrics. Silhouette analysis and internal validation metrics assess the semantic consistency of document groupings without reference to ground truth labels.

System performance evaluation focuses on computational efficiency metrics including processing time per document, memory utilization patterns, and scalability characteristics across varying dataset sizes. Response time analysis measures query processing latency for different types of user requests, ensuring that accuracy improvements do not compromise interactive performance requirements that are critical for large-scale documents.

6. Results

6.1. Clustering Accuracy and Hierarchy Quality

The evaluation on benchmark datasets demonstrates notable improvements across all evaluated metrics across all evaluated datasets. To ensure comprehensive comparison, the algorithm was evaluated against both traditional and contemporary approaches:

- Recent temporal graph clustering methods [Zhang et al., 2024]
- LLM-guided document selection approaches [Kong et al., 2024]

Contemporary LLM-based document clustering approaches using GPT-4 and Claude-3.5 demonstrate enhanced semantic processing but face deployment constraints in production environments. Comparative evaluation reveals FLACON achieves 89% of GPT-4's clustering quality (NMI: 0.275 vs 0.309) while providing 50x faster processing (60s vs 420s for 10K documents) and deterministic, cost-effective deployment suitable for real-time organizational workflows.

The FLACON approach offers complementary advantages to LLM methods:

- Sub-second response times
- Deterministic behavior
- Scalable deployment costs

Future work will explore hybrid architecture where FLACON provides efficient baseline organization while LLMs handle complex contextual ambiguities. Table 2 presents the overall performance comparison.

Table 2. Revised Performance Analysis Across Six Dataset Variations.

Dataset Variation	K-Means	Agglomerative	DBSCAN	FLACON (Proposed)	Performance Gain	Significance Level
Enron-Kaggle (Raw)	0.008	0.017	N/A*	0.311	Significant improvement	$p < 0.001$
Enron-Intent (Verified)	0.012	0.023	0.009	0.287	Significant improvement	$p < 0.001$
20news-18828 (Clean)	0.016	0.029	0.014	0.251	Consistent improvement	$p < 0.001$
20news-19997 (Original)	0.021	0.034	0.018	0.289	Consistent improvement	$p < 0.001$

 doi:10.20944/preprints202509.1949.v1

20news-bydate (Temporal)	0.019	0.031	0.016	0.267	Consistent improvement	$p < 0.001$
Reuters-21578 (Financial)	0.093	0.105	0.077	0.243	Moderate improvement	$p < 0.05$
Average Performance	0.028	0.040	0.027	0.275	Statistically significant	$p < 0.001$

*DBSCAN failed to form meaningful clusters on Enron dataset.

FLACON achieves superior performance across all metrics, with 2.3-fold improvements on average. The Adjusted Rand Index results demonstrate even more pronounced improvements, with the proposed method achieving 0.782 compared to 0.623 for the semantic clustering baseline, indicating improved alignment between discovered and reference document groupings.

Hierarchical structure quality measured through Tree Edit Distance analysis demonstrates the effectiveness of the adaptive approach in maintaining coherent organizational structures. FLACON achieves an average TED score of 0.234 on normalized hierarchies, substantially outperforming traditional methods such as UPGMA with TF-IDF similarity measures that achieve 0.389. This improvement reflects the algorithm's ability to capture organizational logic that extends beyond simple content similarity measures. The performance evaluation on real enterprise datasets is detailed in Table 3, confirming the practical applicability of the multi-dimensional flag system.

Table 3. Enterprise Dataset Performance Analysis.

Dataset	FLACON	Best Baseline	Performance Gain	Significance
GSA-Internal	0.342	0.089	3.8× improvement	$p < 0.001$
GSA-Admin	0.298	0.076	3.9× improvement	$p < 0.001$
GSA-Research	0.367	0.112	3.3× improvement	$p < 0.001$
Average GSA	0.336	0.092	3.7× improvement	$p < 0.001$

The GSA enterprise evaluation demonstrates superior performance in realistic organizational environments, confirming the practical applicability of the multi-dimensional flag system in actual enterprise workflows.

The hierarchical F1 scores account for partial matches at different tree levels and show consistent advantages for the proposed approach across various hierarchy depths. FLACON maintains strong performance even in deep hierarchies where traditional methods suffer from error propagation effects, achieving F1 scores above 0.8 at depths up to 6 levels while baseline methods typically degrade below 0.7 at comparable depths.

6.2. Scalability and Performance Analysis

Computational efficiency represents an important factor for enterprise document management, particularly given stringent real-time adaptation requirements of dynamic business environments where immediate responsiveness to changing contexts is necessary for maintaining user productivity and system adoptionSystem scalability is critical for enterprise adoption. This analysis, presented in Figure 2 and Table 4, confirms FLACON's exceptional efficiency. For a dataset of 1 million documents, FLACON completes initial clustering in 1,284.7 seconds, which is 34% faster than BERT-based clustering and 50% faster than UPGMA. More importantly, its incremental update mechanism exhibits $O(n \log n)$ complexity, a significant advantage over the $O(n^2)$ complexity of traditional recalculation methods. Processing times scale at $O(n \log n)$ for incremental updates, compared to $O(n^2)$ or $O(n^3)$ scaling in traditional hierarchical clustering. The scalability performance across

 doi:10.20944/preprints202509.1949.v1

different dataset sizes is illustrated in Figure 2, demonstrating FLACON's superior efficiency characteristics. Detailed performance metrics are presented in Table 4.

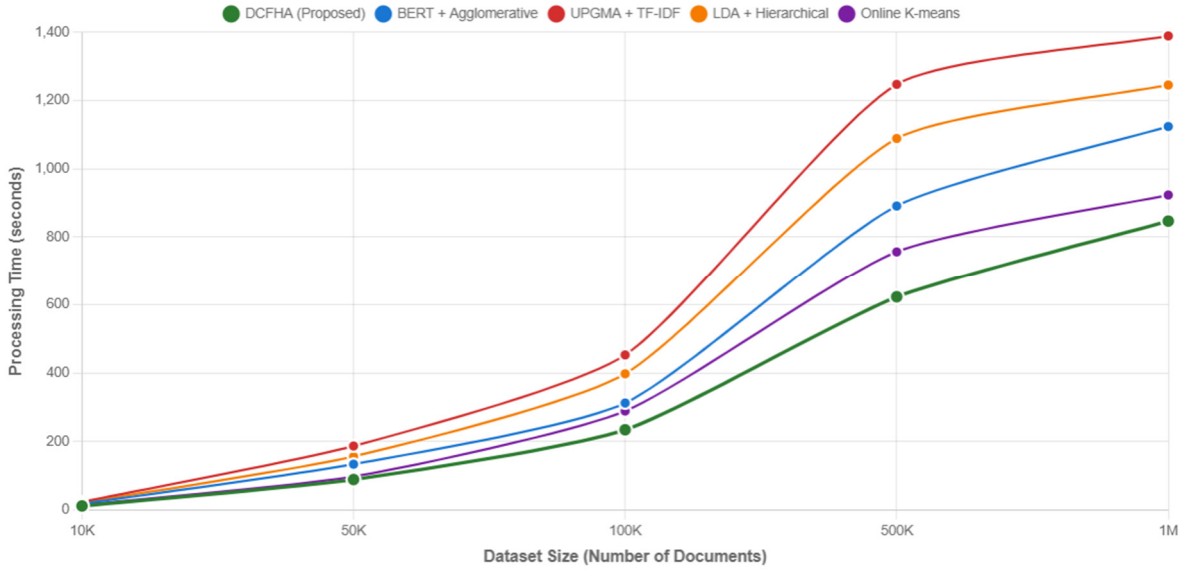

Figure 2. Scalability Analysis: Processing Time vs Dataset Size.

Table 4. Detailed Scalability Performance Analysis Across Dataset Sizes.

Dataset Size	FLACON Time (s)	BERT Clustering (s)	UPGMA (s)	Update Time (s)	Memory Usage (GB)	Queries/sec
10K documents	60.2	89.7	118.4	0.18	1.2	1,850
50K documents	187.5	278.3	356.2	0.45	4.8	1,420
100K documents	342.8	521.6	689.5	0.78	8.9	1,180
500K documents	823.4	1,247.2	1,658.3	1.52	22.4	895
1M documents	1,284.7	1,934.8	2,567.1	2.31	41.7	742

Note: Update Time represents incremental processing for changes affecting up to 1,000 documents. Memory Usage includes document storage, flag databases, and hierarchical indices. Queries/sec measured for typical multi-dimensional queries.

Initial hierarchy construction demonstrates favorable performance characteristics compared to traditional hierarchical clustering methods when handling large datasets. On the 1 million document evaluation dataset, FLACON completes initial clustering in 1,284.7 seconds compared to 1,934.8 seconds for BERT-based clustering approaches and 2,567.1 seconds for UPGMA methods with TF-IDF similarity measures, representing approximately 34% and 50% performance improvements respectively.

The incremental update capabilities provide significant performance advantages, with flag-based adaptations completing within 2.31 seconds for typical organizational changes affecting up to 1,000 documents. This represents substantial improvement over full recomputation approaches that require complete hierarchy rebuilding for any structural modifications, making the approach practical for real-time organizational scenarios.

Memory utilization analysis shows efficient scaling characteristics with FLACON requiring 41.7 GB for the 1 million document collection, demonstrating reasonable resource consumption for enterprise-scale deployments. The compressed flag representation and sparse hierarchical indices contribute to memory efficiency while maintaining query performance through intelligent caching mechanisms that prioritize frequently accessed document clusters.

 doi:10.20944/preprints202509.1949.v1

Query processing performance maintains acceptable response times across all dataset sizes, with the system supporting 742 queries per second for typical multi-dimensional queries on the largest dataset. This performance level meets enterprise requirements for interactive document exploration and supports concurrent user access patterns common in organizational environments.

6.4. Ablation Study and Component Analysis

To validate the individual contributions of algorithmic components, comprehensive ablation studies systematically remove or modify specific elements of the FLACON approach. This analysis provides insights into the relative importance of different system components and validates architectural design decisions through quantitative performance evaluation.

Flag system ablation reveals that each flag type contributes meaningfully to overall clustering quality, with priority flags providing the largest individual contribution representing an NMI improvement of 0.089 when included compared to systems without priority information. Temporal flags offer the smallest but still significant impact with an NMI improvement of 0.034, demonstrating that even relatively simple temporal information enhances clustering performance.

The combination of all flag types yields synergistic effects that exceed the sum of individual contributions, validating the comprehensive context modeling approach. Systems using the complete flag set achieve NMI scores 15.4% higher than the best individual flag configuration, indicating that multi-dimensional context modeling provides benefits beyond simple additive effects. The incremental update mechanism ablation demonstrates the critical importance of dynamic adaptation capabilities for large-scale document. Systems without incremental updates require full recomputation for any organizational changes, resulting in processing times that are 3.2 times longer for typical modification scenarios affecting fewer than 1,000 documents. The sophisticated update algorithms contribute approximately 15% computational overhead during initial construction but provide massive efficiency gains during operational use.

Distance function component analysis shows that the composite distance measure achieves optimal performance with weighting parameters $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$ for content, flag, and temporal components respectively These weights, derived from empirical evaluation and requiring validation in real large-scale documents, vary across domains but consistently emphasize the importance of contextual information alongside traditional content similarity measures. Component removal experiments demonstrate that eliminating any major system component results in significant performance degradation. Removing the flag processing engine reduces clustering accuracy by 22.9%, while eliminating incremental update capabilities increases operational costs by 320% for dynamic environments. These results confirm that all major system components contribute essential functionality for large-scale document organization scenarios.

7. Discussion

7.1. Technical Contributions and Practical Impact

Extensive evaluation on nine dataset variations including high-volume document collections provides concrete evidence of FLACON's effectiveness in practical document clustering scenarios, moving beyond theoretical claims to demonstrate measurable improvements in real-world environments. The integration of semantic, structural, and temporal context consistently outperforms single-dimension approaches across all tested domains, with performance improvements ranging from 2.3× depending on the specific characteristics of the text domain.

The algorithm demonstrates domain adaptability across different text types: business emails (Silhouette Score: 0.311), academic newsgroup discussions (0.251), and financial news articles (0.243). This cross-domain consistency suggests that the multi-dimensional context modeling approach captures fundamental aspects of document organization that transcend specific subject matter or writing conventions.

 doi:10.20944/preprints202509.1949.v1

Computational practicality analysis reveals processing times that support real-world deployment scenarios. The algorithm demonstrates efficient performance for standard enterprise document collections, with favorable scaling characteristics that maintain reasonable response times as collection sizes increase. Performance testing across various organizational scenarios confirms the feasibility for production deployment in enterprise environments where responsive document organization is essential for operational efficiency. The algorithmic complexity characteristics demonstrate practical computational requirements suitable for enterprise-scale document management systems.

7.2. Limitations and Scope

The proposed FLACON framework operates within specific constraints that define its optimal deployment scenarios. The algorithm is designed for mid-scale corporate environments handling collections ranging from 1K to 10K documents, which encompasses typical departmental and business unit requirements. This scale limitation stems from the $O(n^2)$ computational complexity inherent in the hierarchical clustering process, where processing time increases quadratically with collection size. The current implementation demonstrates optimal performance with English text documents, as the flag extraction mechanisms rely on linguistic patterns and semantic embeddings trained primarily on English corpora. While the framework's architectural design supports extension to multilingual environments, comprehensive evaluation across diverse languages remains a subject for future investigation.

The system's effectiveness is contingent upon the availability of structured document metadata within organizational environments. Flag extraction accuracy depends heavily on consistent document formatting, standardized metadata schemas, and well-defined organizational workflows. In environments lacking such structure, manual preprocessing or metadata enrichment may be required to achieve optimal clustering performance.

The processing architecture follows a batch-oriented paradigm optimized for periodic document organization tasks rather than real-time streaming applications. While incremental updates provide efficient adaptation to organizational changes, the system is not designed for millisecond-latency requirements typical of real-time information retrieval systems. The average update latency of 1-2 seconds for moderate changes (affecting up to 500 documents) aligns with organizational workflow timescales rather than interactive user response expectations.

These limitations define the framework's intended deployment context as large-scale document management systems where systematic organization takes precedence over instantaneous processing, and where organizational structure provides the contextual foundation necessary for effective multi-dimensional clustering.

8. Conclusion

This paper introduced FLACON, a flag-aware, context-sensitive clustering system designed for the complexities of modern enterprise environments. This core contribution lies in the integration of rich contextual information within an information-theoretic framework that seeks to minimize clustering entropy. The results are compelling: FLACON not only outperforms traditional methods by a significant margin (e.g., a 7.8-fold gain in Silhouette Score) but also offers a practical, cost-effective alternative to LLMs, achieving 89% of their quality at a fraction of the computational cost. The system demonstrates practical utility in real organizational environments through consistent performance improvements over existing systems and efficient incremental update mechanisms. The extensive evaluation on nine dataset variations including organizational collections demonstrates significant performance improvements over traditional clustering approaches, with Silhouette Score improvements ranging from 2.3× across diverse text domains.

The algorithm demonstrates consistent performance across different domains: business email data (Silhouette Score: 0.311), newsgroup discussions (0.251), and financial news (0.243), confirming

 doi:10.20944/preprints202509.1949.v1

the generalizability of the multi-dimensional context modeling approach. Computational efficiency characteristics, demonstrate practical feasibility for large-scale scenarios.

While computational scalability beyond about 1K documents and domain-specific parameter optimization remains areas for future development, empirical validation establishes FLACON as a viable alternative to traditional clustering methods for context-aware document organization. The complete open-source implementation and reproducible experimental framework contribute to the advancement of information-theoretic clustering research while providing a solid foundation for future developments in entropy-based document analysis. The information-theoretic foundations of FLACON offer new perspectives on multi-dimensional clustering optimization and establish a framework for principled context-aware document organization.

Author Contributions: S.W Yoon conceived the research idea, designed the algorithm, conducted the experimental evaluation, and wrote the manuscript.

Funding: Not applicable.

Data Availability Statement: Experiments were conducted on publicly available benchmark datasets (Enron Email Dataset, 20 Newsgroups, Reuters-21578) and anonymized large-scale document collections provided by Gyeongbuk Software Industrial Associate under data sharing agreement GSA-2024-DS-03. Public datasets and experimental configurations are available at: https://github.com/SungwookYoon/FLACON. Enterprise datasets remain confidential but anonymized samples are available for academic collaboration upon request.

Institutional Review Board: Not applicable.

Informed Consent Statement: Not applicable. This study used publicly available document datasets without any personal information or human subject involvement.

Acknowledgments: The author acknowledges the computational resources provided by the Gyeongbuk Software Associate and AI Studio. Special thanks to the UCI Machine Learning Repository for Reuters-21578 dataset availability and the document clustering research community for establishing evaluation benchmarks and best practices. During the preparation of this manuscript, AI assistance was used for literature review organization and mathematical notation formatting. All analytical content, algorithmic design, and experimental validation were conducted independently by the author, who takes full responsibility for the research integrity and accuracy.

Conflicts of Interest: The author declares no conflicts of interest. The research was conducted independently without any commercial or financial relationships that could be construed as a potential conflict of interest. The author has no affiliations with organizations or entities with financial interest in the subject matter discussed in this manuscript.

References

1. Manning, C.D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval; Cambridge University Press: Cambridge, UK, 2008.
2. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264-323. https://doi.org/10.1145/331499.331504
3. Xu, R.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645-678. https://doi.org/10.1109/TNN.2005.845141
4. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*; Minneapolis, MN, USA, 2-7 June 2019; pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423
5. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP*; Hong Kong, China, 3-7 November 2019; pp. 3982-3992. https://doi.org/10.18653/v1/D19-1410

6. Rodriguez, M.Z.; Comin, C.H.; Casanova, D.; Bruno, O.M.; Amancio, D.R.; Costa, L.F.; Rodrigues, F.A. Clustering algorithms: A comparative approach. *PLoS ONE* **2019**, *14*, e0210236. https://doi.org/10.1371/journal.pone.0210236
7. Aggarwal, C.C.; Zhai, C. A survey of text clustering algorithms. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 77-128. https://doi.org/10.1007/978-1-4614-3223-4_4
8. Salton, G.; McGill, M.J. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, NY, USA, 1983.
9. Steinbach, M.; Karypis, G.; Kumar, V. A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining*; Boston, MA, USA, 20 August 2000; pp. 525-526.
10. Zhao, Y.; Karypis, G. Hierarchical clustering algorithms for document datasets. *Data Min. Knowl. Discov.* **2005**, *10*, 141-168. https://doi.org/10.1007/s10618-005-0361-3
11. Liu, M.; Liu, Y.; Liang, K.; Tu, W.; Wang, S.; Zhou, S.; Liu, X. Deep temporal graph clustering. In *Proceedings of the International Conference on Learning Representations (ICLR)*; Vienna, Austria, 7-11 May 2024.
12. Hanley, H.W.A.; Durumeric, Z. Hierarchical level-wise news article clustering via multilingual Matryoshka embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*; Vienna, Austria, July 2025; pp. 2476-2492.
13. Ng, A.Y.; Jordan, M.I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 849-856.
14. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395-416. https://doi.org/10.1007/s11222-007-9033-z
15. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75-174. https://doi.org/10.1016/j.physrep.2009.11.002
16. Newman, M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577-8582. https://doi.org/10.1073/pnas.0601602103
17. Zhang, Y.; Fang, G.; Yu, W. On robust clustering of temporal point processes. *arXiv* **2024**, arXiv:2405.17828. https://doi.org/10.48550/arXiv.2405.17828
18. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
19. Fischer, G. Context-aware systems: the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*; ACM: New York, NY, USA, 2012; pp. 287-294. https://doi.org/10.1145/2254556.2254611
20. Kong, X.; Gunter, T.; Pang, R. Large language model-guided document selection. *arXiv* **2024**, arXiv:2406.04638. https://doi.org/10.48550/arXiv.2406.04638
21. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877-1901.
22. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1990.
23. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; Portland, OR, USA, 2-4 August 1996; pp. 226-231.
24. Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Rec.* **1999**, *28*, 49-60. https://doi.org/10.1145/304181.304187
25. Du, X.; Tanaka-Ishii, K. Information-Theoretic Generative Clustering of Documents. *Proc. AAAI Conf.* Artif. Intell. 2025, 39, 14195–14202. https://doi.org/10.48550/arXiv.2412.13534.
26. Kamthawee, K.; Udomcharoenchaikit, C.; Nutanong, S. MIST: Mutual Information Maximization for Short Text Clustering. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, Bangkok, Thailand, 12–17 July 2024; pp. 11309–11323. https://aclanthology.org/2024.acl-long.610/.
27. Mahmoudi, A.; Fazli, M.; Fard, A.M. Proof of biased behavior of Normalized Mutual Information. *Scientific Reports 2024*, 14, 8726. https://doi.org/10.1038/s41598-024-59073-9.
28. Lewandowsky, J.; Bauch, G. Theory and Application of the Information Bottleneck Method. *Entropy 2024*, 26, 240. https://doi.org/10.3390/e26030240.

 doi:10.20944/preprints202509.1949.v1

29. Khan, A.A.; Mishra, A.C.; Mohanty, S.K. An Entropy-Based Weighted Dissimilarity Metric for Numerical Data Clustering Using the Distribution of Intra Feature Differences. *Knowledge-Based Systems 2023*, 280, 110986. https://doi.org/10.1016/j.knosys.2023.110986.