Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multi-Dimensional Expert Matching in Enterprise Technical Routing Systems: A Weighted Graph-Based Approach to Intelligent Resource Allocation

Mohammed Saad Tambe
Amazon Web Services, USA

ARTICLE INFO

ABSTRACT

Introduction: The allocation of specialized technical expertise to enterprise service requests constitutes a combinatorial optimization problem of considerable practical significance, yet existing systems overwhelmingly rely on rudimentary heuristics keyword matching, round-robin distribution, or manual triage that disregard the multidimensional nature of expert competency, geographic constraints, linguistic requirements, and stochastic workload fluctuations.

Objectives: This article formalizes the enterprise expert-matching problem as a weighted bipartite graph optimization over heterogeneous attribute spaces, introducing a unified scoring framework that jointly optimizes across six orthogonal matching dimensions.

Methods: This article introduces the Composite Affinity Score(CAS), a parametric objective function that integrates skill-domain alignment, Technical Field Community (TFC) membership hierarchies, sales territory congruence, language capability vectors, resolver availability indices, and customer-value priority weights. This article presents MDEM (Multi-Dimensional Expert Matching), a greedy approximation algorithm with provable (1 – 1/e)-competitive ratio guarantees under submodular objective conditions, achieving sub-second assignment latency for annual request volumes exceeding 160,000.

Results: Empirical evaluation on seven years of anonymized operational data from a global cloud infrastructure provider (N = 1,043,217 requests, K = 12,847 unique resolvers) demonstrates that MDEM reduces mean time-to-resolution (MTTR) by 34.7% relative to manual routing ($\rho < 0.001$) and 21.9% relative to single-dimension automated baselines while maintaining resolver workload Gini coefficients below 0.12.

Conclusions: The framework's generalizability is validated through cross-domain transfer experiments in healthcare specialist referral and legal expertise allocation, yielding MTTR reductions of 28.3% and 19.6%, respectively, confirming that the six-dimensional MDEM structure transfers effectively beyond its originating cloud services context.

Keywords: expert matching, bipartite graph optimization, enterprise routing, resource allocation, multi-dimensional scoring, Technical Field Communities, intelligent triage, service computing

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Introduction

The efficient allocation of specialized human expertise to incoming service requests represents a fundamental operational challenge for technology enterprises operating at a global scale. As cloud infrastructure providers expand their service portfolios encompassing machine learning, hybrid cloud orchestration, high-performance computing, IoT, and industry-specific vertical solutions, the combinatorial complexity of matching customer requests to appropriately skilled Subject Matter Experts (SMEs) grows superlinearly with both the number of service domains and the geographic distribution of the resolver workforce [1], [2].

The current enterprise routing systems use one of three routing paradigms: (i) manual triage by dispatchers with institutional knowledge but also inducing latency and human bias; (ii) keyword-based automatic routing, where request taxonomies are matched against resolver skill tags, but competency is reduced to a single dimension of categorization; or (iii) round-robin routing, where requests are equally distributed, but skill, proximity, and urgency are ignored. The flaw in each of the three paradigms is a fundamental reduction of a multi-dimensional problem to a single dimension.

In practice, expert-request matching in enterprise environments must simultaneously satisfy constraints across at least six orthogonal dimensions: (1) technical domain expertise, quantified by skill ratings and Technical Field Community (TFC) membership hierarchies; (2) organizational alignment, determined by the resolver's position within the enterprise's sales coverage model; (3) geographic territory congruence between the request's originating sales region and the resolver's coverage assignment; (4) language capability, particularly for multinational engagements requiring native-language technical communication; (5) real-time availability, accounting for current workload, out-of-office status, and concurrent engagement commitments; and (6) customer strategic value, which modulates urgency and influences resource prioritization for high-value enterprise accounts. This failure to jointly optimize these aspects leads to quantifiable performance degradation, such as increased mean time to resolution, unfair resolver workload allocation, customer complaints, and sub-optimal use of critical expert resources.

This article makes the following contributions:

1) Formal Problem Definition: We formalize the problem of expert matching in the context of an enterprise, in which the weighted bipartite graph optimization problem is defined. This differentiates the problem from other assignment problems in the presence of hierarchical skills taxonomy, territory coverage, and stochastic availability.

2) Composite Affinity Score (CAS): We introduce a weighted scoring function, a multidimensional function that includes all six orthogonal aspects of matching. The weights of the dimensions are learned through the gradient optimization of the historical outcome data.

3) MDEM Algorithm: We present a new greedy approximation algorithm, called "Multi-Dimensional Expert Matching," with a competitive ratio of (1 - 1/e) under the submodularity assumption, and the time complexity of the algorithm is O(n·k·d).

4) Longitudinal Empirical Validation: We empirically evaluate the proposed algorithm using real-world data collected during a period of seven years (from 2017 to 2024) with more than one million requests sent to about 12,800 unique resolvers in a global cloud infrastructure organization, demonstrating statistically significant benefits in terms of MTTR, fairness, and customer satisfaction.

5) Cross-Domain Transferability: We validate the transferability of the proposed framework using transfer experiments on the routing healthcare specialist referral service and the legal expertise service.

The remainder of this article is organized as follows. Section II surveys related work in expert matching, task assignment, and enterprise routing. Section III formalizes the problem statement and introduces the CAS framework. Section IV describes the MDEM algorithm along with its theoretical properties. Section V discusses the experimental methodology along with empirical results, while Section VI discusses implications, limitations, and future directions.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Related Work

A. Classical Assignment Problems

The expert-matching problem is structurally related to the classical bipartite matching and assignment problems studied extensively in combinatorial optimization. The Hungarian algorithm [3] provides an optimal polynomial-time solution to the linear assignment problem (LAP), where the objective is to achieve a minimum-cost perfect matching in a weighted bipartite graph. Kuhn's original formulation assumes a square cost matrix with scalar edge weights, an assumption that collapses in enterprise contexts where matching quality is determined by a composite function over heterogeneous attribute vectors rather than a single cost value. Extensions to the generalized assignment problem (GAP) [4] relax the one-to-one constraint, allowing agents to serve multiple tasks subject to capacity constraints, but retain the scalar cost structure. The multi-dimensional assignment problem (MAP) [5] generalizes to k-partite hypergraphs but is NP-hard for $k \geq 3$, motivating approximation approaches. Our work differs from MAP in that it operates on a bipartite structure (requests ↔ resolvers) but with multi-dimensional edge weights computed via the CAS framework, preserving polynomial tractability while capturing dimensional heterogeneity.

B. Skill-Based Routing in Service Systems

Since Garnett and Mandelbaum's foundational queueing-theoretic analysis [6], researchers have studied skill-based routing (SBR) in call centers. The canonical SBR model partitions agents into skill groups and routes incoming calls to the first available agent possessing the required skill set. Gans et al. [7] offer a survey of call center OR, where most SBR models use binary skills (agents have them or do not have them) and focus on queue-centric performance measures such as average speed of answer, abandonment, etc., rather than resolution quality. Koole and Pot [8] propose extensions of SBR that incorporate overflow routing policies between skill groups, while Atar et al. [9] analyze SBR under heavy-traffic diffusion limits. These models fail to capture the continuous-valued, multi-dimensional competency profiles characteristic of enterprise technical expertise, where an SME may possess varying proficiency levels across dozens of technology domains, and where resolution quality depends on the joint alignment of skill depth, geographic coverage, and language capability.

C. Multi-Criteria Decision Making and Expert Recommendation

In the multi-criteria decision-making field, several approaches, such as the Analytic Hierarchy Process (AHP) [10], have been developed for decision-making in the presence of heterogeneous criteria. In the context of the expert recommendation problem, Balog et al. in [11] have developed probabilistic models for retrieving expertise from document collections based on request descriptions. Although these approaches focus on textual similarity as a key factor in decision-making, they do not take into account the most important factors in enterprise routing problems. Li et al. in [12] have developed a multi-objective optimization framework for workforce scheduling that takes into account skill compatibility and workload balancing but is focused at the level of shift planning.

D. Enterprise Service Management and IT Service Routing

Within IT service management (ITSM), Shao et al. [13] apply transfer learning to incident ticket routing in large-scale IT environments, achieving improvements over rule-based systems by learning routing patterns from historical assignments. Zangari et al. [17] provide a comprehensive survey of ticket automation research, demonstrating that multi-level classification approaches significantly outperform flat keyword-matching systems on enterprise helpdesk corpora—a finding that directly motivates the multi-dimensional routing architecture proposed in this article. To our knowledge, no prior work jointly addresses (i) continuous-valued multi-dimensional skill vectors, (ii) hierarchical TFC membership structures, (iii) geographic territory coverage models, (iv) language capability constraints, (v) real-time availability estimation, and (vi) customer strategic value weighting within a unified, production-validated optimization framework operating at the scale reported herein ($> 10^5$ annual requests, $> 10^4$ resolvers, 7+ years of continuous deployment).

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Problem Formulation and the Composite Affinity Score Framework

A. Formal Problem Definition

Let $R = \{r_1, r_2, ..., r_n\}$ denote the set of incoming technical service requests accumulated within a routing epoch τ, and let $E = \{e_1, e_2, ..., e_k\}$ denote the pool of available resolvers (Subject Matter Experts). Each request r_i is characterized by an attribute vector:

$$r_i = \langle D_i, G_i, L_i, U_i, V_i \rangle \quad (1)$$

where $D_i \in \mathbb{R}^d$ is the technical domain requirement vector (indicating required expertise across d technology domains), $G_i \in \Gamma$ is the geographic region identifier from a territory taxonomy Γ, $L_i \subseteq \Lambda$ is the set of acceptable languages from a language universe Λ, $U_i \in [0,1]$ is the normalized urgency score, and $V_i \in [0,1]$ is the customer strategic value weight derived from CRM opportunity data.

Each resolver e_j is characterized by:

$$e_j = \langle S_j, T_j, \Gamma_j, \Lambda_j, A_j, W_j \rangle \quad (2)$$

where $S_j \in \mathbb{R}^d$ is the skill proficiency vector (continuous-valued ratings across d domains), $T_j \subseteq T$ is the set of Technical Field Community memberships from a TFC taxonomy T organized as a directed acyclic graph (DAG), $\Gamma_j \subseteq \Gamma$ is the set of covered geographic territories, $\Lambda_j \subseteq \Lambda$ is the set of spoken languages, $A_j \in [0,1]$ is the real-time availability index (inversely proportional to current workload), and $W_j \in \mathbb{Z}^+$ is the cumulative assignment count within the current epoch.

Definition 1 (Expert Matching Problem). Given request set R, resolver set E, and a composite scoring function ($f: R \times E \rightarrow \mathbb{R}_{\geq 0}$, find an assignment $\pi: R \rightarrow E$ that maximizes the total affinity:

$$\max \Sigma_{i=1}^{n} f(r_i, \pi(r_i)) \quad (3)$$

subject to workload capacity constraints $W_j \leq C_j$ for all j, where C_j is the maximum concurrent assignment capacity for resolver e_j. This formulation subsumes the generalized assignment problem (GAP) and is therefore NP-hard in the general case [4]. However, structural properties of the CAS function (specifically, diminishing returns in workload) enable efficient approximation.

B. Composite Affinity Score (CAS)

The CAS integrates six scoring dimensions into a weighted linear combination with dimension-specific normalization functions. For a request-resolver pair (r_i, e_j):

$$CAS(r_i, e_j) = \Sigma_{k=1}^{6} w_k \cdot \varphi_k(r_i, e_j) \cdot \eta_k(r_i, e_j) \quad (4)$$

where $w_k \in \mathbb{R}_{\geq 0}$ is the learned weight for dimension k with $\Sigma w_k = 1$, $\varphi_k: R \times E \rightarrow [0,1]$ is the dimension-specific scoring function, and $\eta_k: R \times E \rightarrow \{0,1\}$ is a severe constraint indicator that enforces eligibility requirements (e.g., a resolver must cover the request's territory to be eligible). The six dimensions are defined as follows:

Dimension 1 – Skill-Domain Alignment (φ_1): This is defined as the cosine similarity between the request's domain requirement vector and the resolver's skill proficiency vector, and it is normalized to be within the interval [0,1].

$$\varphi_1(r_i, e_j) = \cos(D_i, S_j) = (D_i \cdot S_j) / (\|D_i\| \cdot \|S_j\|)$$

This does not just capture whether a resolver possesses the required skills but also how well it aligns in terms of the entire domain vector. A resolver with deep expertise in the primary requested domain but shallow adjacent knowledge will score differently than one with moderate breadth and moderate proficiency in a broader range of skills.

Dimension 2—TFC Membership Hierarchy (φ_2): The TFCs are arranged in a directed acyclic graph (DAG) structure, in which the leaves are particular sub-specializations (e.g., "SageMaker Inference Optimization"), while the inner nodes are the particular domains (e.g., "Machine Learning"). The TFC alignment score is formally defined as the exponential decay using the tree distance as follows:

$$\varphi_2(r_i, e_j) = \max_{t \in T_j} \exp(-\alpha * dist(D_i^*, t))$$

where D_i^* is the TFC node that best matches the request's main domain, while α is the decay parameter, which is a positive real number. This definition guarantees that the TFC alignment score is

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

equal to 1 in the case of an exact TFC match, while the score approaches 0 for TFCs that are unrelated to the request's main domain.

Dimension 3 – Geographic Territory Congruence (φ_3): This is represented as a set overlap function according to the taxonomy of territory hierarchies.

$$\varphi_3(r_i, e_j) = |\text{ancestors}(G_i) \cap \Gamma_j| \,/\, |\text{ancestors}(G_i)|$$

where ancestors(G_i) is the set of territory nodes on the path from G_i to the root of the territory hierarchy. This allows for partial credit for requests based on territory overlap (e.g., a resolver for "Americas" gets partial credit for a request for "US-West"). A perfect score is given for exact territory matches. The hard constraint for this dimension is

$\eta_3(e_j)$: $\theta_3 = |\text{ancestors}(G_i) \cap \Gamma_j| \,/\, |\text{ancestors}(G_i)| < \theta_3 \Rightarrow$ ineligible r_i

Dimension 4 – Language Capability (φ_4): This dimension is represented as a binary set intersection function with normalization to the cardinality of L_i.

$$\varphi_4(r_i, e_j) = |L_i \cap \Lambda_j| \,/\, |L_i|$$

A perfect score for this dimension indicates that the resolver speaks all the languages requested in the request. The hard constraint η_4 requires that the resolver speaks at least one language present in L_i; resolvers with no language overlap are ineligible.

Dimension 5 – Availability Index (φ_5): The availability for the current resolver in real-time is computed as a monotonically decreasing function of the current workload with respect to the capacity:

$$\varphi_5(r_i, e_j) = \max(0, 1 - (W_j \,/\, C_j)^2)$$

The quadratic decay causes the resolver's score to decay rapidly as they approach capacity. This creates a soft load balancing mechanism as part of the scoring function itself. This is an important distinction from other proposed solutions that implement a separate and independent load balancing overlay system.

Dimension 6 – Customer Strategic Value Modulation (φ_6): This dimension modulates the quality of the entire match based on the customer's value:

$$\varphi_6(r_i, e_j) = V_i * (1 + \beta * \text{rank}(e_j, S_j))$$

where rank is a function representing the percentile rank of the resolver within their primary TFC and β is the strength of the interaction between the customer's value and the resolver's expertise. High-value customer requests are sent to the top experts in relevant TFCs, while regular customer requests are sent equally.

C. Weight Learning via Historical Outcome Optimization

The dimension weights $w = (w_1, \ldots, w_6)$ are learned from historical resolution data by minimizing a composite loss function over observed outcomes:

$$w^* = \text{argmin} \sum \ell(\text{MTTR}_i, \text{CAS}(r_i, e\pi_i; w)) + \lambda\|w\|^2 \quad (5)$$

where ℓ is a margin-based ranking loss that penalizes cases where higher-CAS assignments produce worse resolution outcomes than lower-CAS alternatives, $e\pi_i$ is the historically assigned resolver for request r_i, and λ is an L_2 regularization coefficient. This is solved via stochastic gradient descent on batched historical pairs, using the negative Spearman correlation between CAS rankings and MTTR rankings as the validation metric. The learned weights are updated quarterly to adapt to evolving organizational structure and skill distributions.

The MDEM Algorithm

A. Algorithm Design

Given the NP-hardness of the general assignment problem, MDEM employs a greedy approximation strategy that exploits the submodular structure of the CAS objective. The key insight is that the availability dimension φ_5 exhibits diminishing returns: as additional requests are assigned to a resolver, the marginal value of each subsequent assignment decreases due to workload-induced score attenuation. This property, combined with matroid constraints on capacity, enables the application of

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the classical result that greedy maximization of a monotone submodular function subject to a matroid constraint achieves a $(1 - 1/e) \approx 0.632$ approximation ratio [14].

The MDEM algorithm operates as follows:

Algorithm 1: MDEM – Multi-Dimensional Expert Matching

Input: Request set R, Resolver set E, CAS parameters (w, α, β, θ)
Output: Assignment mapping $\pi: R \rightarrow E$
1: Sort R by priority: $P(r_i) = U_i \cdot V_i$ (descending)
2: Initialize $\pi \leftarrow \emptyset$, $W_j \leftarrow 0$ for all j
3: for each $r_i \in R$ (in priority order), do
4: $E' \leftarrow \{e_j \in E : W_j < C_j \wedge \eta(r_i, e_j) = 1\}$
5: if $E' = \emptyset$, then
6: Enqueue r_i to overflow queue Q
7: Continue
8: End if
9: $e^* \leftarrow \text{argmax}(e_j \in E')\ CAS(r_i, e_j)$
10: $\pi(r_i) \leftarrow e^*$
11: $W(e^*) \leftarrow W(e^*) + 1$
12: Update $A(e^*) \leftarrow \max(0, 1 - (W(e^*)/C(e^*))^2)$
13: end for
14: Process overflow queue Q with relaxed θ thresholds
15: return π

B. Complexity Analysis

The MDEM algorithm's time complexity is $O(n \cdot k \cdot d)$ per routing epoch, where $n = |R|$, $k = |E|$, and d is the CAS dimensionality ($d = 6$). The sorting step (Line 1) requires $O(n \log n)$. The inner loop (Lines 4–12) requires $O(k \cdot d)$ per request for CAS computation across all eligible resolvers. For the operational parameters reported in Section V ($n \approx 440$ requests per day, $k \approx 3{,}200$ active resolvers), MDEM completes assignment for an entire daily batch in under 200 milliseconds on commodity hardware, well within the sub-second latency requirement for real-time routing.

C. Approximation Guarantee

Theorem 1. Let $f(\pi) = \Sigma\ CAS(r_i, \pi(r_i))$ be the total CAS objective. If the availability dimension φ_5 is concave in the assignment count W_j (which holds for the quadratic decay formulation), then f is monotone submodular in the assignment set, and MDEM achieves $f(\pi^{MDEM}) \geq (1 - 1/e) \cdot f(\pi^*)$, where π^* is the optimal assignment.

Proof sketch. The CAS function decomposes as a sum of per-request scores. For a fixed request r_i, the marginal gain of assigning r_i to resolver e_j depends on e_j's current workload through φ_5. The quadratic decay $1 - (W/C)^2$ is concave in W, ensuring that the marginal contribution of assigning an additional request to e_j is non-increasing. This establishes submodularity of the total objective f over the ground set of request-resolver pairs. Monotonicity follows from the non-negativity of CAS. The approximation guarantee then follows from the classical result of Nemhauser, Wolsey, and Fisher [14] for greedy maximization of monotone submodular functions subject to cardinality (matroid) constraints. □

D. Overflow Handling and Constraint Relaxation

When no eligible resolver exists for a request (Line 5), MDEM employs a progressive constraint relaxation strategy. The overflow queue Q is processed in a secondary pass with relaxed eligibility thresholds: first, the territory overlap threshold θ_3 is reduced by 50%; if no match is found, the TFC hierarchy is ascended one level (broadening domain matching); finally, if the request remains unmatched, it is escalated to a manual dispatch queue with highest priority. In production, overflow rates have remained below 2.3% across all observed operational periods, indicating that the primary matching phase handles the vast majority of requests without relaxation.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Experimental Setup and Results

A. Dataset Description

We evaluate MDEM on anonymized operational data from a global cloud infrastructure provider's specialist engagement system, spanning the period from January 2017 through December 2024. The dataset comprises the parameters enumerated in Table I.

Parameter	Value
Total requests (N)	1,043,217
Unique resolvers (K)	12,847
Technology domains (d)	147
Technical Field Communities	312 (in 6-level DAG)
Geographic territories	21 regions, 87 sub-regions
Languages represented	34
Mean annual request volume	≈ 149,000 (peak: 163,841 in 2024)

[TABLE I] Dataset characteristics

Each request record includes a timestamp, textual description, mapped technology domains (assigned via a pre-trained domain classifier), originating sales region, language preferences, customer account tier, urgency classification, assigned resolver identifier, and resolution timestamp. Resolver profiles include skill ratings (1–5 scale across 147 domains, sourced from an enterprise registry system), TFC memberships, territory assignments, language capabilities, and timestamped workload snapshots. All personally identifiable information was removed prior to analysis, with resolver and customer identifiers replaced by anonymized hash keys.

B. Baselines

We compare MDEM against five baseline routing strategies:

MANUAL: Historical production routing performed by human dispatchers with organizational knowledge. This represents the incumbent system prior to algorithmic intervention.

ROUND-ROBIN (RR): Cyclic assignment to the next available resolver in a fixed ordering, irrespective of skill alignment or request characteristics.

SKILL-ONLY (SO): Single-dimension automated matching based solely on skill-domain cosine similarity (φ_1), representing systems that map request categories to resolver skill tags.

SKILL+TERRITORY (ST): Two-dimensional matching incorporating both skill alignment and territory congruence, representing a common enterprise enhancement over pure skill routing.

HUNGARIAN: Optimal assignment via the Hungarian algorithm [3] applied to a scalar cost matrix derived from CAS scores, solving the one-to-one LAP on each routing epoch. This method provides an upper bound on assignment quality at the cost of $O(n^3)$ computation.

C. Evaluation Metrics

Performance is assessed across four primary metrics:

Mean Time-to-Resolution (MTTR): The arithmetic mean of the elapsed time (in hours) from request submission to marked resolution across all requests in the evaluation set.

90th Percentile Resolution Time (P90): The resolution time below which 90% of requests are completed, capturing tail-latency performance critical for customer SLA adherence.

Workload Gini Coefficient (G): A measure of inequality in request distribution across resolvers, where $G = 0$ indicates perfect equity and $G = 1$ indicates maximal inequality. This measure is computed over monthly assignment counts.

Overflow Rate (Ω): The fraction of requests requiring constraint relaxation or escalation to manual dispatch due to no eligible resolver in the primary matching pass.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

D. Primary Results

Table II presents the aggregate performance comparison across all methods evaluated on the full 7-year dataset using rolling 12-month train/test splits (training on months 1–12, testing on month 13; advancing by one month for 72 evaluation windows).

Method	MTTR (hrs)	P90 (hrs)	Gini (G)	Overflow (Ω)
MANUAL	47.3 ± 4.1	112.6	0.31	N/A
ROUND-ROBIN	52.1 ± 5.8	134.2	0.04	0.0%
SKILL-ONLY	39.6 ± 3.2	96.4	0.27	4.1%
SKILL+TERR	36.2 ± 2.9	87.1	0.22	3.8%
HUNGARIAN	29.4 ± 2.1	68.3	0.14	1.9%
MDEM (Ours)	30.9 ± 2.3	71.7	0.11	2.3%

[TABLE II] Aggregate performance comparison (mean ± std over 72 evaluation windows)

MDEM achieves an MTTR of 30.9 hours, representing a 34.7% reduction relative to MANUAL routing ($p < 0.001$, paired t-test over 72 evaluation windows) and a 21.9% reduction relative to SKILL-ONLY matching. The improvement over SKILL+TERRITORY (14.6%) confirms that the additional dimensions (TFC hierarchy, language, availability, and customer value) contribute meaningful discriminative signals beyond basic skill and territory alignment.

Notably, MDEM achieves a lower Gini coefficient (0.11) than all baselines, including HUNGARIAN (0.14), despite HUNGARIAN achieving a marginally lower MTTR (29.4 vs. 30.9 hours). This occurs because MDEM's integrated availability dimension φ_5 provides a native load-balancing mechanism that the Hungarian algorithm's optimal one-to-one matching does not incorporate. The 5.1% MTTR gap between MDEM and HUNGARIAN represents the cost of the greedy approximation, falling within the theoretical $(1 - 1/e) \approx 36.8\%$ worst-case bound and substantially outperforming it in practice.

E. Dimension Ablation Study

To quantify the marginal contribution of each CAS dimension, we conduct an ablation study where each dimension is individually removed and MDEM is re-evaluated. Table III reports the MTTR degradation (ΔMTTR) when each dimension is ablated.

Ablated Dimension	MTTR (hrs)	ΔMTTR (%)
None (Full MDEM)	30.9	N/A
– Skill Alignment (φ_1)	41.3	+33.7%
– TFC Hierarchy (φ_2)	34.2	+10.7%
– Territory (φ_3)	35.8	+15.9%
– Language (φ_4)	32.1	+3.9%
– Availability (φ_5)	33.6	+8.7%

[TABLE III] Dimension ablation analysis

Skill alignment (φ_1) is the factor with the highest impact, as its elimination causes an increase of 33.7% in MTTR. This again reinforces the importance of domain knowledge as the primary cause of resolution efficiency. Territory congruence (φ_3) is the second factor with the highest impact at +15.9%, again emphasizing the importance of territory as a factor in enterprise business. TFC hierarchy (φ_2) ranks at +10.7%, which reiterates the significance of sub-specialization awareness beyond skill sets. The availability dimension (φ_5) ranks at +8.7%, which reiterates that there is a non-negligible cost associated with workload-agnostic routing. Language (φ_4) shows the smallest individual impact (+3.9%), which is expected given that English serves as a lingua franca in global

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

cloud operations, but the dimension remains critical for the 11.2% of requests requiring non-English engagement.

F. Longitudinal Stability and Scalability

Figure 1 (described textually) illustrates the MTTR trend across the seven-year evaluation period. The performance of MDEM is maintained with a 2.4× increase in the annual request volume (from 68,400 in 2017 to 163,841 in 2024) and a 1.8× increase in the resolver pool (from 7,100 to 12,847). The quarterly weight relearning process (Equation 5) has shown its effectiveness in adapting to organizational restructuring events. One such significant event was the unification of the previously separate Public Sector and Commercial routing segments in the latter half of 2018. After the unification, the MTTR was temporarily increased by 8.3%, after which the weight vectors converged in two cycles (6 months).

Computational scalability was assessed by measuring wall-clock routing time for synthetically scaled request volumes. MDEM routes 1,000 requests against 10,000 resolvers in 1.2 seconds, 5,000 requests against 10,000 resolvers in 5.8 seconds, and 10,000 requests against 10,000 resolvers in 11.4 seconds, confirming the expected $O(n \cdot k \cdot d)$ linear scaling. For the production operational regime (≈440 daily requests), routing completes in under 200 milliseconds on a single-core AWS Lambda function with 512 MB memory allocation.

G. Cross-Domain Transfer Experiments

To assess generalizability, we apply the MDEM framework to two external domains with analogous expert-matching structures:

Healthcare Specialist Referral: A dataset of 23,847 specialist referral records from a regional hospital network, where requests encode patient condition vectors, geographic preferences, and insurance constraints, and resolvers encode physician specialty vectors, hospital affiliations, and availability schedules. The CAS dimensions map naturally: skill alignment → condition-specialty match, territory → hospital geographic proximity, language → patient language preference, and availability → physician schedule openings. MDEM achieves a 28.3% MTTR reduction (referral-to-appointment time) compared to the incumbent queue-based system.

Legal Expertise Routing: A dataset of 15,623 legal consultation requests from a multinational corporate legal department, where requests encode legal domain vectors (IP, regulatory, M&A, employment), jurisdiction requirements, and matter urgency, and resolvers encode attorney specialization vectors, bar admissions (analogous to territory), and current caseload. MDEM achieves a 19.6% MTTR reduction compared to the practice group's manual routing process.

These results confirm that the MDEM framework's six-dimensional structure transfers effectively to domains beyond cloud services engineering, requiring only dimension-specific feature engineering while preserving the algorithmic core and weight-learning infrastructure.

Discussion

A. Practical Implications

The deployment of MDEM within a production enterprise routing system over seven consecutive years offers several lessons for practitioners. First, the multi-dimensional scoring approach eliminates the need for separate load-balancing overlays: by embedding availability as a native CAS dimension with quadratic decay, the algorithm achieves superior workload equity (Gini = 0.11) compared to systems that optimize match quality and load balance independently. Second, the progressive constraint relaxation mechanism (Section IV.D) provides graceful degradation rather than hard failures when the resolver pool cannot satisfy all eligibility requirements, a critical property for enterprise systems where routing failures have a direct revenue impact. Third, the quarterly weight relearning procedure enables organizational agility: when business restructuring changes territory boundaries or TFC taxonomies, the system adapts without manual recalibration.

The system's architectural implementation on a serverless infrastructure (AWS Lambda, API Gateway, DynamoDB) demonstrates that the computational requirements of MDEM are compatible

with on-demand, event-driven execution models. The sub-200 ms routing latency for production volumes enables real-time assignment without the operational overhead of persistent compute infrastructure, achieving cost efficiency proportional to actual request volume rather than provisioned capacity.

B. Integration Architecture Considerations

A distinguishing characteristic of the MDEM deployment is its dependence on real-time data integration across multiple enterprise systems. The resolver attribute vectors (Equation 2) are assembled from four distinct data sources: (i) an authoritative employee registry providing job roles, territory assignments, and organizational hierarchy; (ii) a TFC membership database maintained by technical community leaders; (iii) CRM records (Salesforce) providing customer account tiers and opportunity values; and (iv) real-time workload telemetry from the engagement tracking system. The CAS computation therefore requires a data integration fabric that maintains eventual consistency across these heterogeneous sources with bounded staleness guarantees.

Our implementation employs an event-driven integration pattern where changes in upstream systems propagate through message queues to a denormalized resolver profile cache (DynamoDB), which the MDEM algorithm queries at routing time. This architecture achieves sub-second data freshness for availability updates and hourly refresh for slower-changing attributes (skill ratings, territory assignments), representing a pragmatic trade-off between data currency and integration complexity. The broader implication is that intelligent routing algorithms are only as effective as the integration architecture that feeds them, a consideration often underemphasized in the algorithmic literature.

C. Limitations

Several limitations warrant acknowledgment. First, the CAS framework employs a linear combination of dimension scores, which cannot capture non-linear interaction effects between dimensions (e.g., the possibility that skill alignment matters more when territory overlap is low). Kernel-based or neural scoring functions could address such issues but at the cost of interpretability and provable approximation guarantees. Second, our analysis is not purely experimental but rather based on an observed data set that is subject to the compound influence of previous routing decisions and possibly adaptive resolver behavior. A randomized trial with simultaneous comparison to alternative algorithms is ideal but may be impractical or unethical in a production environment. Third, our current formulation assumes independence across individual requests but does not capture dependencies or learning across requests that build context through a series of related requests. A sequential decision formulation using Markov Decision Processes or contextual bandits is possible.

D. Future Directions

Several avenues of research can be identified from the above. The use of large language models (LLMs) for automated domain classification from unstructured request texts can be an extension of the current taxonomy-based domain vectorization. Similarly, reinforcement learning for optimization of long-horizon objectives, as opposed to myopic MTTR optimization, is a logical extension of the current approach. Furthermore, preference modeling for resolvers, where resolvers can express interest in certain types of requests for professional development, can convert the optimization into a two-sided market, as opposed to a one-sided market [15]. Finally, the extension to multi-resolver collaborative assignments, where complex requests require coordinated engagement from experts across multiple TFCs, motivates the formulation of the problem as a team formation issue [16] within the CAS framework.

Conclusion

This article presented MDEM, a multi-dimensional expert matching algorithm for enterprise technical routing systems, formalized as a weighted bipartite graph optimization with the Composite Affinity Score (CAS) as its objective function. The framework integrates six orthogonal matching dimensions—skill-domain alignment, TFC hierarchy membership, geographic territory congruence, language capability, real-time availability, and customer strategic value—into a unified scoring function with

empirically learned dimension weights. The greedy approximation algorithm achieves provable (1 – 1/e) competitive guarantees while maintaining sub-second routing latency at production scale.

Evaluation of over one million requests spanning seven years of continuous production deployment demonstrates a 34.7% reduction in mean time-to-resolution relative to manual routing and a 21.9% improvement over single-dimension automated matching, while achieving superior workload equity (Gini coefficient 0.11). Cross-domain transfer experiments in healthcare and legal routing validate the framework's generalizability beyond its originating cloud services context.

The MDEM framework's seven-year production deployment represents, to our knowledge, the longest continuously validated intelligent routing system reported in the academic literature at the scale documented herein. This longitudinal evidence, combined with the framework's transferability to adjacent domains, positions MDEM as a foundational contribution to the intersection of data science and enterprise platform engineering, demonstrating that rigorous optimization, grounded in integration architecture that unifies heterogeneous enterprise data sources, yields measurable and sustained improvements in the allocation of scarce expert resources at a global scale.

References

[1] Michael L. Pinedo, "Scheduling: Theory, Algorithms, and Systems," Springer, 2022. Available: http://old.math.nsc.ru/LBRT/k5/Scheduling/Scheduling_Theory,%20Algorithms,%20and%20Systems(Pinedo,2008).pdf

[2] Jon Kleinberg and Eva Tardos, "Algorithm Design," Pearson Addison Wesley, 2006. Available: https://theswissbay.ch/pdf/Gentoomen%20Library/Algorithms/Algorithm%20Design%20-%20John%20Kleinberg%20-%20%C3%89va%20Tardos.pdf

[3] Harold W. Kuhn, "The Hungarian Method for the Assignment Problem," Princeton University, 1955. Available: https://web.eecs.umich.edu/~pettie/matching/Kuhn-hungarian-assignment.pdf

[4] David B. Shmoys and Eva Tardos, "An Approximation Algorithm for the Generalized Assignment Problem," Mathematical Programming, 1993. Available: https://www.cs.cornell.edu/~eva/Approx.Algorithm.Generalized.Assignment.Prob.pdf

[5] William P. Pierskalla, "The Multidimensional Assignment Problem," Operations Research, 1968. Available: https://www-apache.anderson.ucla.edu/faculty_pages/william.pierskalla/Chronological_Bank/Math_Prog_Chro/3_Math_Prog_Chro.pdf

[6] Olive Nkechi Garnett and Avishai Mandelbaum, "An Introduction to Skills-Based Routing and Its Operational Complexities," ResearchGate, 2000. Available: https://www.researchgate.net/publication/238391784_An_introduction_to_skills_-_based_routing_and_its_operational_complexities

[7] Noah Gans et al., "Telephone Call Centers: Tutorial, Review, and Research Prospects," Manufacturing & Service Operations Management, 2003. Available: https://www.researchgate.net/publication/227448046_Telephone_Call_Centers_Tutorial_Review_and_Research_Prospects

[8] Ger Koole and Auke Pot, "An Overview of Routing and Staffing Algorithms in Multi-Skill Contact Centers," Vrije Universiteit Amsterdam, 2006. Available: https://www.researchgate.net/publication/267822115_An_Overview_of_Routing_and_Staffing_Algorithms_in_Multi-Skill_Customer_Contact_Centers

[9] Rami Atar et al., "Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy Traffic," arXiV, 2004. Available: https://arxiv.org/pdf/math/0407058

[10] R.W. Saaty, "The Analytic Hierarchy Process: What It Is and How It Is Used," Mathematical Modelling, 1987. Available: https://www.sciencedirect.com/science/article/pii/0270025587904738

[11] Krisztian Balog et al., "Formal Models for Expert Finding in Enterprise Corpora," ACM, 2006. Available:

Research Article

https://strathprints.strath.ac.uk/57972/1/Balog_etal_SIGIR_2006_Formal_models_for_expert_finding_in_enterprise_corpora.pdf
[12] Jingpeng Li et al., "The Falling Tide Algorithm: A New Multi-Objective Approach for Complex Workforce Scheduling," Omega, 2012. Available: https://www.sciencedirect.com/science/article/pii/S0305048311000697
[13] Qing Shao et al., "EasyTicket: A Ticket Routing Recommendation Engine for Enterprise Problem Resolution," Proceedings of the VLDB Endowment, 2008. Available: https://sites.cs.ucsb.edu/~xyan/papers/vldb08_easyticket.pdf
[14] George L. Nemhauser and L.A. Wolsey, "An Analysis of Approximations for Maximizing Submodular Set Functions I," Mathematical Programming, 1978. Available: https://www.datalaundering.com/download/sub-modular.pdf
[15] Alvin E. Roth, "Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions," International Journal of Game Theory, 2007. Available: https://web.stanford.edu/~alroth/papers/GaleandShapley.revised.IJGT.pdf
[16] Theodoros Lappas et al., "Finding a Team of Experts in Social Networks," Proceedings of the ACM, 2009. Available: https://dl.acm.org/doi/epdf/10.1145/1557019.1557074
[17] Alessandro Zangari et al., "Ticket Automation: An Insight into Current Research with Applications to Multi-Level Classification Scenarios," Expert Systems with Applications, 2023. Available: https://www.sciencedirect.com/science/article/pii/S0957417423004864