SEMANTIC DISTANCE INDEX (SDI) 1.0

A Multidimensional Framework for Measuring Semantic Alignment Between Search Intent and Content

SearchEngineZine Research Framework

Framework Version: SDI 1.0
Research Status: Proposed Methodology
Research Stage: Framework Construction
Publication Year: 2026
Empirical Validation: Pending
Framework Type: Original SEO / Information Retrieval Methodology

Abstract

Search queries provide only a partial representation of a searcher's underlying information need. A query may identify a topic while leaving the required task, entities, attributes, relationships, and expected outcome implicit.

Consequently, topical or semantic similarity between a query and a document does not necessarily demonstrate that the document adequately satisfies the underlying information need.

This paper proposes the Semantic Distance Index (SDI), a multidimensional framework for analyzing semantic separation between an information need and the content intended to satisfy it.

SDI decomposes semantic distance into five dimensions: Conceptual Distance, Intent Distance, Entity Distance, Attribute Distance, and Relational Distance. The framework converts an information need into a structured set of semantic requirements and evaluates a document against those requirements.

A provisional normalized scoring model is introduced to produce both a multidimensional SDI Profile and a composite SDI value ranging from 0 to 1. Equal weighting is used as the default SDI 1.0 baseline; these weights and interpretation bands are explicitly provisional and have not yet been empirically validated.

SDI is proposed as an analytical methodology for content research, auditing, optimization, and experimentation. It is not presented as a Google ranking factor, a proprietary search-engine metric, or a predictor of rankings.

Future research should test the framework against human information-need satisfaction, inter-rater reliability, established retrieval baselines, query types, content categories, and potentially search and AI retrieval outcomes.

SDI 1.0 at a Glance

Component	**SDI 1.0**
Purpose	Measure semantic separation between information needs and content
Dimensions	5
Dimensions	Conceptual, Intent, Entity, Attribute, Relational
Composite Scale	0-1
Lower SDI	Greater semantic alignment
Higher SDI	Greater semantic distance
Default Weighting	20% per dimension
Scoring Status	Provisional
Diagnostic Bands	Provisional
Research Status	Proposed methodology
Validation Status	Pending empirical research

Google Ranking Factor?	No claim
Primary Application	Content alignment diagnosis

Research Questions

The SDI framework is designed around the following research questions.

RQ1

Can semantic distance between an information need and a document be operationalized using multiple independent dimensions?

RQ2

Do Conceptual, Intent, Entity, Attribute, and Relational Distance provide distinct diagnostic information?

RQ3

Does a multidimensional SDI profile provide greater practical diagnostic value than a single semantic-similarity measurement?

RQ4

Is lower SDI associated with higher human-rated information-need satisfaction?

RQ5

Can SDI identify actionable content gaps without relying primarily on keyword frequency or lexical overlap?

RQ6

Do different query types require different relative weighting across SDI dimensions?

These questions are research questions, not established conclusions.

1. Introduction

Search queries are compressed representations of information needs.

A user may enter:

> how to improve LCP in WordPress

The query contains several useful signals:

- improve
- LCP
- WordPress

However, the complete information requirement is considerably broader.

The user may need to:

1. understand what is causing the problem,
2. diagnose the underlying issue,
3. identify relevant WordPress-specific causes,
4. implement appropriate fixes,
5. understand the consequences of those changes,
6. verify whether the improvement occurred.

A document may contain the words LCP and WordPress while failing to satisfy several of these requirements.

This creates a distinction between semantic relatedness and information-need satisfaction.

SDI is proposed to make that distinction measurable.

2. The Problem: Topic Relevance Is Not Enough

Traditional content analysis frequently evaluates whether a document is relevant to a query.

However, relevance can exist at different levels.

A page about LCP may be:

- lexically related to LCP,
- semantically related to LCP,
- conceptually related to LCP,
- entity-related to LCP,

while still failing to satisfy a procedural request such as:

> How do I improve LCP in WordPress?

The problem is therefore not necessarily that the document discusses the wrong topic.

The problem may be that the document is semantically incomplete relative to the information need.

SDI addresses this problem by asking:

> How far is the document from the semantic requirements necessary to satisfy the information need?

3. The Core SDI Principle

The central proposition of SDI is:

> Semantic relevance should be evaluated against the requirements of an information need, not merely against the words contained in a query.

A simplified representation is:

Search Query

↓

Information Need

↓

Semantic Requirements

↓

Document Representation

↓

Semantic Distance

↓

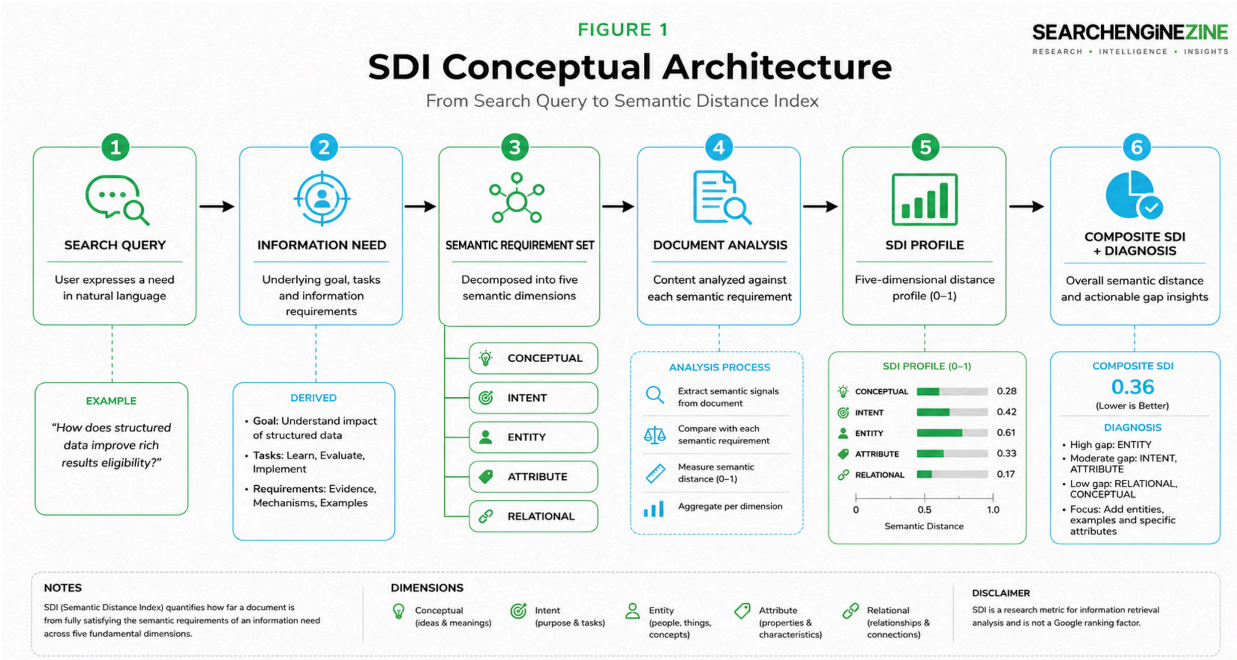
SDI Profile

This distinction is fundamental.

The query is the observable expression.

The information need is the underlying requirement.

The semantic requirement set is the structured representation used to evaluate whether the document adequately addresses that requirement. See information-need representation



The SDI framework models semantic separation between information requirements and document representations.

4. What SDI Is

The Semantic Distance Index is a proposed multidimensional methodology for evaluating the semantic separation between:

1. an information need, and
2. the content intended to satisfy that information need.

SDI uses five analytical dimensions:

- Conceptual Distance
- Intent Distance
- Entity Distance
- Attribute Distance
- Relational Distance

The framework produces two related outputs.

SDI Profile

A five-dimensional representation of semantic distance.

Composite SDI

A single normalized value derived from the five dimensions.

The profile should be treated as the more informative diagnostic object.

The composite score provides a convenient summary. Read more semantic SEO framework

5. What SDI Is Not

SDI is not:

- a confirmed Google ranking factor,
- a Google scoring system,
- Google's internal semantic metric,
- a ranking guarantee,
- a replacement for search-engine quality systems,
- a replacement for information retrieval models,
- a guarantee of AI-search citations,
- a guarantee of improved organic traffic.

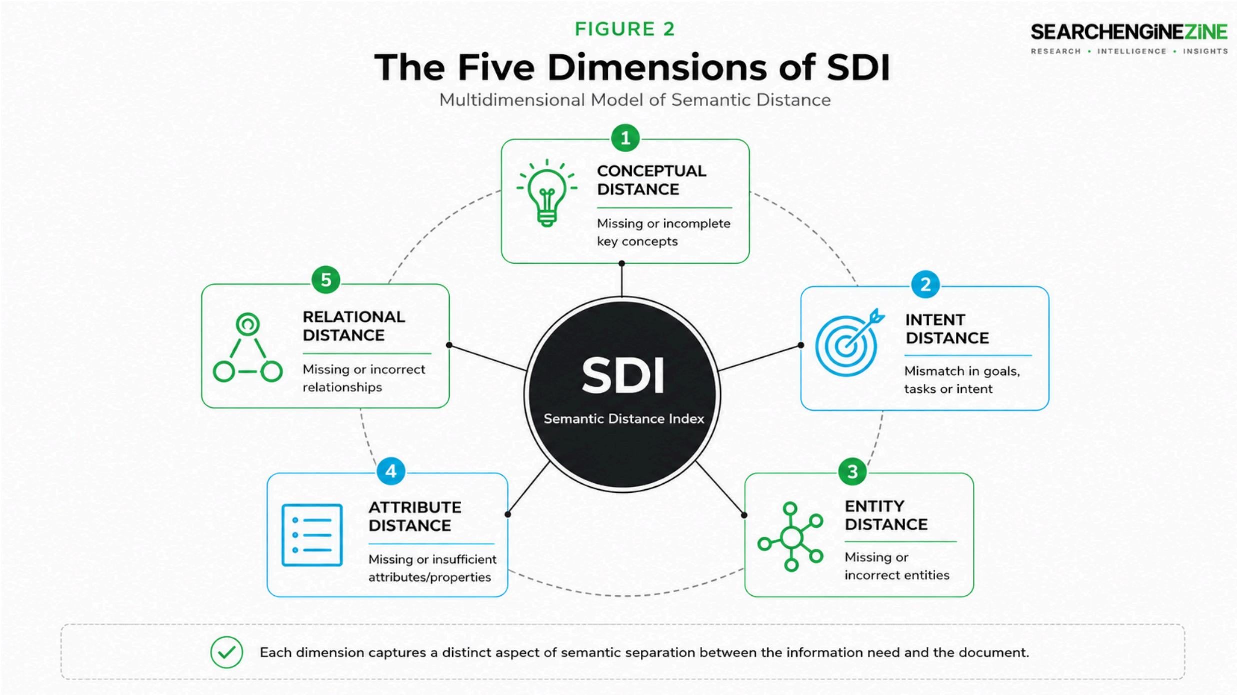
SDI is an independently proposed SearchEngineZine framework.

Its numerical weighting, scoring bands, reliability, predictive validity, and relationship with search performance remain subjects for empirical investigation.

6. The Five Dimensions of SDI

SDI decomposes semantic distance into five dimensions:

Dimension	Symbol	Core Question
Conceptual Distance	D^c	Are the required concepts represented adequately?
Intent Distance	D^I	Does the content satisfy the required task?
Entity Distance	D^E	Are the relevant entities represented adequately?
Attribute Distance	D^A	Are important properties represented?
Relational Distance	D^R	Are important relationships represented correctly?



7. Conceptual Distance

Definition

Conceptual Distance represents the degree to which the concepts required by an information need are absent, incomplete, or inadequately represented in a document.

Consider:

> how to improve LCP in WordPress

Potential concepts include:

- Largest Contentful Paint,
- page performance,
- rendering,
- loading,
- optimization,
- diagnosis.

A page that discusses only generic website speed may be topically adjacent while remaining conceptually distant from the specific information need.

Conceptual Distance increases when:

- essential concepts are absent,
- concepts are mentioned only superficially,

- concepts are replaced with overly broad terminology,
- important conceptual distinctions are missing.

8. Intent Distance

Definition

Intent Distance represents the degree to which the document fails to satisfy the tasks or goals implied by the information need.

This distinction is particularly important for procedural queries.

Consider:

> How to improve LCP in WordPress

A definition of LCP may be conceptually relevant.

But it does not necessarily satisfy the requested task.

A procedural information need may require:

Understand → Diagnose → Fix → Verify

If a document only defines LCP, its Intent Distance may remain high despite strong topical relevance.

Intent Distance therefore asks:

> Does the document actually perform the task the user is seeking?

9. Entity Distance

Definition

Entity Distance represents the degree to which relevant entities are absent, ambiguous, incorrect, or inadequately represented.

For:

> Googlebot crawl budget

Potential entities may include:

- Googlebot,
- crawler,
- website,
- URLs,
- crawl requests,

- server.

However, entity presence alone is insufficient.

A document can mention Googlebot without adequately explaining its role in crawling.

Therefore, entity analysis should consider:

- entity presence,
- entity specificity,
- entity context,
- entity adequacy,
- entity associations

10. Attribute Distance

Definition

Attribute Distance represents the degree to which important properties, characteristics, or attributes required by the information need are absent or insufficiently represented.

For an LCP-related information need, relevant attributes might include:

- measurement,
- thresholds,
- causes,
- optimization methods,
- diagnostic indicators,
- verification methods.

A document may contain the correct entity while failing to provide the attributes necessary to make the information useful.

Therefore:

> Entity presence does not guarantee attribute completeness.

11. Relational Distance

Definition

Relational Distance represents the degree to which required relationships between concepts and entities are absent, incomplete, or incorrectly represented.

Consider:

Large image

↓

Slow resource loading

↓

Delayed rendering

↓

Poor LCP

A document may contain all four concepts.

However, if the document does not explain their relationships, semantic representation remains incomplete.

Relational Distance therefore examines whether important relationships are:

- present,
- meaningful,
- sufficiently contextualized,
- correctly connected,
- internal relationships

12. Semantic Requirement Set

The most important operational step in SDI is converting the information need into a Semantic Requirement Set. Read more: document representation

Let:

$$R(Q) = \{\ C, I, E, A, L\ \}$$

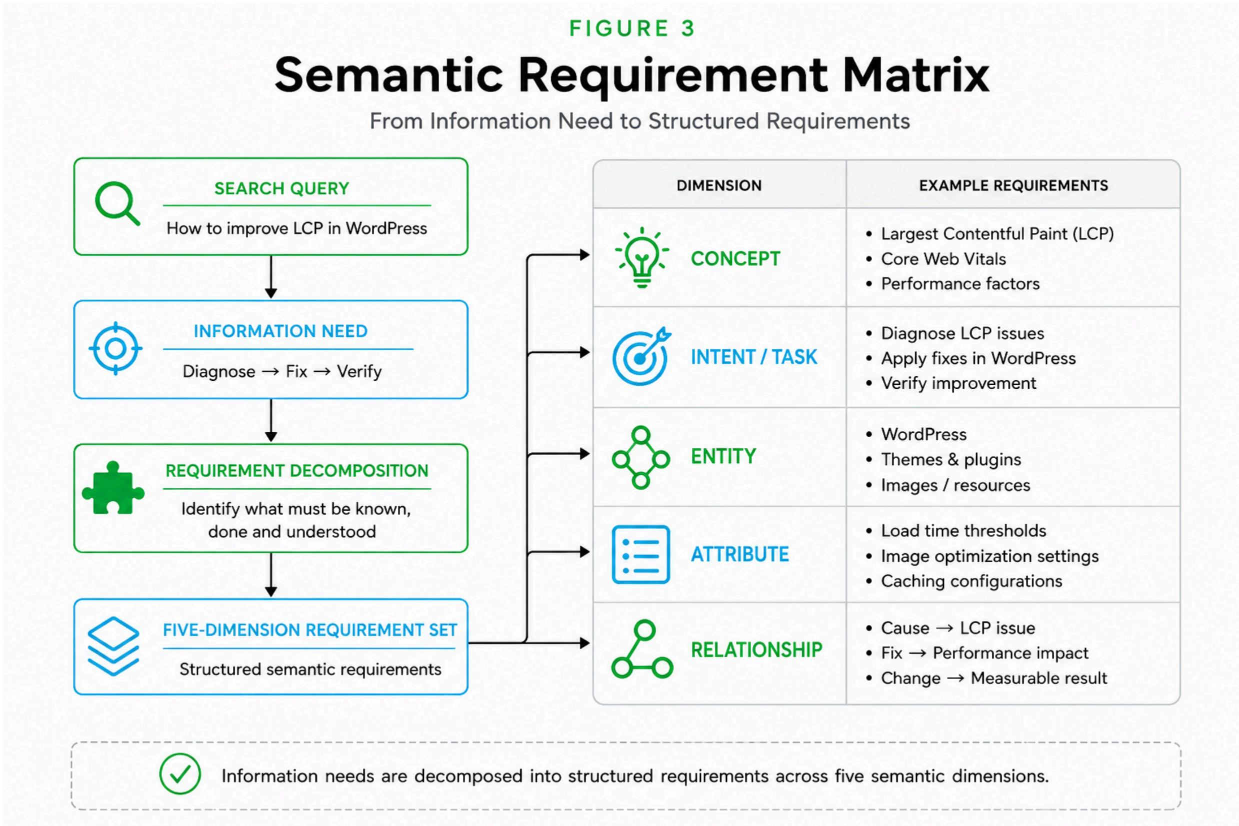
Where:

- C = required concepts,
- I = required intent or tasks,
- E = required entities,
- A = required attributes,
- L = required relationships.

The document can be represented as:

$$D = \{\ C', I', E', A', L'\ \}$$

> SDI evaluates the semantic distance between the requirement set and the document representation.



13. Operational Definitions

For SDI to become a measurable methodology, each dimension requires an operational definition.

Conceptual Distance

The degree to which required concepts are absent, incomplete, or inadequately represented.

Intent Distance

The degree to which required tasks or goals are absent or inadequately fulfilled.

Entity Distance

The degree to which required entities are absent, ambiguous, incorrect, or inadequately contextualized.

Attribute Distance

The degree to which important properties or characteristics required by the information need are absent or insufficiently represented.

Relational Distance

The degree to which required relationships between concepts and entities are absent, incomplete, or incorrectly represented.

These definitions are intended to make SDI independently scorable and testable.

14. Requirement Importance

Not every semantic requirement has equal importance.

SDI 1.0 proposes three importance levels.

3 — Essential

The information need would be substantially incomplete without the requirement.

2 — Important

The requirement strongly contributes to satisfying the information need.

1 — Supporting

The requirement provides useful supporting information but is not essential.

This allows the framework to distinguish between:

> missing one critical requirement

and:

> missing several minor supporting details.

15. Scoring Methodology

Each SDI dimension is normalized between:

$$0 \leq D^x \leq 1$$

where:

0 = minimum semantic distance

and:

1 = maximum semantic distance

The initial scoring protocol is:

Distance	Interpretation
0.00	Fully aligned
0.25	Minor gap
0.50	Partial / moderate gap
0.75	Major gap
1.00	Missing or substantially unsatisfied

These values are a provisional scoring protocol.

They should not be treated as empirically established measurement thresholds until tested.

16. Composite SDI

The initial SDI 1.0 composite model uses equal weighting.

$$SDI = w_C D^C + w_I D^I + w_E D^E + w_A D^A + w_R D^R$$

For the SDI 1.0 baseline:

$$w_C = w_I = w_E = w_A = w_R = 0.20$$

Therefore:

$$SDI = 0.20D^C + 0.20D^I + 0.20D^E + 0.20D^A + 0.20D^R$$

Subject to:

$$\sum w_i = 1$$

And:

$$0 \leq SDI \leq 1$$

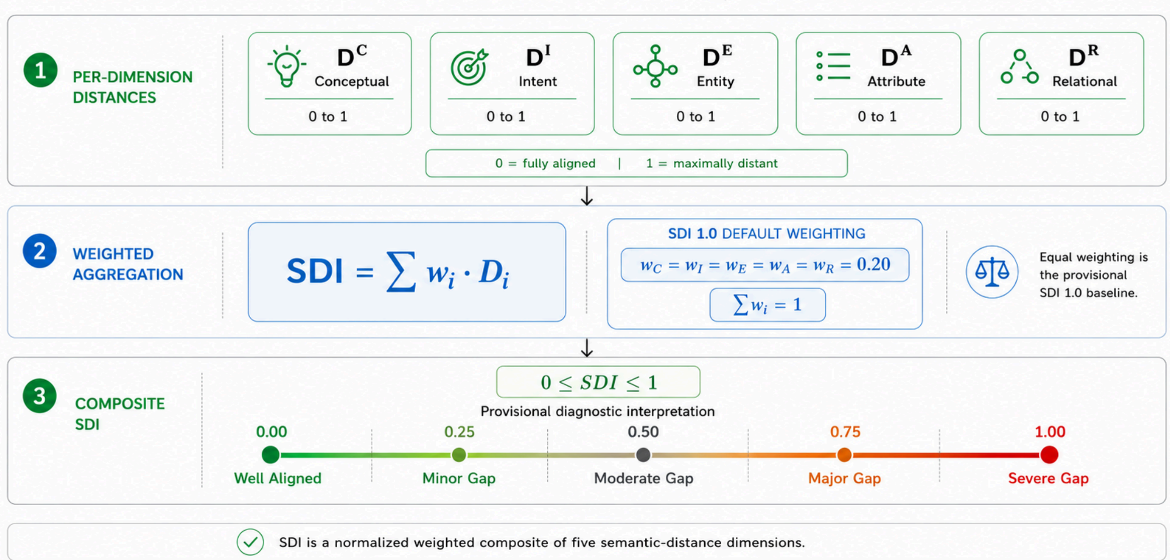
The equal weighting is deliberately described as a baseline configuration.

It is not claimed to be the optimal weighting.

FIGURE 4

SDI Scoring Model

From Per-Dimension Distances to Composite SDI



17. SDI Interpretation Bands

For practical analysis, SDI 1.0 proposes the following provisional bands:

Composite SDI	Interpretation
0.00–0.20	Very close
0.21–0.40	Close
0.41–0.60	Moderate
0.61–0.80	Distant
0.81–1.00	Very distant

These are diagnostic interpretation bands, not ranking thresholds.

They should be revised if empirical research demonstrates more appropriate boundaries.

18. Worked Example

Query

> **How to improve LCP in WordPress**

Information Need

> Diagnose, fix, and verify improvements to Largest Contentful Paint on a WordPress website.

Requirement Set

Type	Example Requirements
Concepts	LCP, rendering, loading performance
Intent	Diagnose, fix, verify
Entities	WordPress, LCP, image, CSS, JavaScript, server
Attributes	Causes, thresholds, optimization methods
Relationships	Resource loading → rendering → LCP

Suppose a hypothetical document receives:

Dimension	Distance
Conceptual	0.20
Intent	0.50
Entity	0.10
Attribute	0.30
Relational	0.60

The composite score becomes:

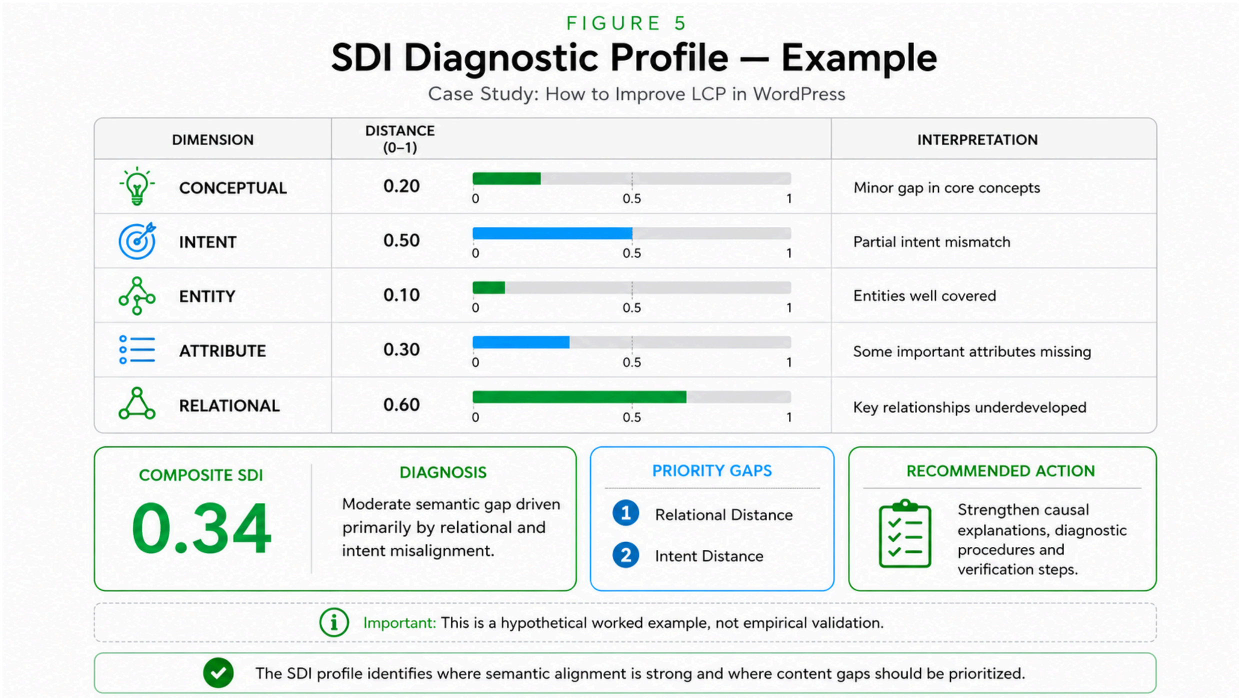
$$SDI = 0.20(0.20) + 0.20(0.50) + 0.20(0.10) + 0.20(0.30) + 0.20(0.60)$$

$$SDI = 0.34$$

Therefore:

$$\boxed{SDI = 0.34}$$

Under the provisional SDI 1.0 bands, this falls into the Close category.



19. Diagnosing the Example

The composite score alone is not sufficient.

The profile provides greater diagnostic value.

Conceptual Distance — 0.20

The document represents the core concepts relatively well.

Entity Distance — 0.10

The relevant entities are well represented.

Attribute Distance — 0.30

Some important attributes remain incomplete.

Intent Distance — 0.50

The document only partially satisfies the requested procedural task.

Relational Distance — 0.60

Important relationships between causes, interventions, and outcomes are insufficiently developed.

The resulting diagnosis is therefore:

> **The document is topically and entity relevant but has meaningful procedural and relational gaps.**

20. Content Intervention

The appropriate response is not automatically:

> **Add more keywords.**

Instead, the largest semantic gaps should be addressed.

To reduce Intent Distance

Add:

- diagnostic procedures,
- implementation steps,
- troubleshooting sequences,
- verification methods.

To reduce Relational Distance

Explain causal relationships such as:

Large LCP image

↓

Slow resource loading

↓

Delayed rendering

↓

Poor LCP

To reduce Attribute Distance

Add:

- relevant thresholds,
- causes,
- diagnostic signals,
- optimization methods,
- verification criteria.

The objective is not maximum content volume.

The objective is **greater semantic alignment with the information need**.

21. Presence Is Not Adequacy

One of the central rules of SDI is:

> **Presence is not adequacy.**

A keyword appearing on a page does not prove that the underlying concept is adequately represented.

An entity appearing does not prove that its role is explained.

A relationship being implied does not prove that it is sufficiently represented.

A definition does not prove that a procedural intent has been satisfied.

This principle separates SDI from simple keyword-coverage approaches.

22. SDI and Semantic Similarity

Semantic similarity and semantic distance are established concepts in information retrieval and natural-language processing.

However, SDI proposes a different analytical objective. Read more: NIST semantic matching definition

Semantic Similarity asks:

> How similar are two representations?

SDI asks:

> How far is a document from the semantic requirements necessary to satisfy an information need?

Semantic similarity may therefore contribute to SDI.

It does not constitute the entire framework.

SDI additionally considers:

- task requirements,
- entities,
- attributes,
- relationships,
- requirement importance.

23. SDI and Information Gain

SDI and Information Gain measure different properties.

SDI

Alignment

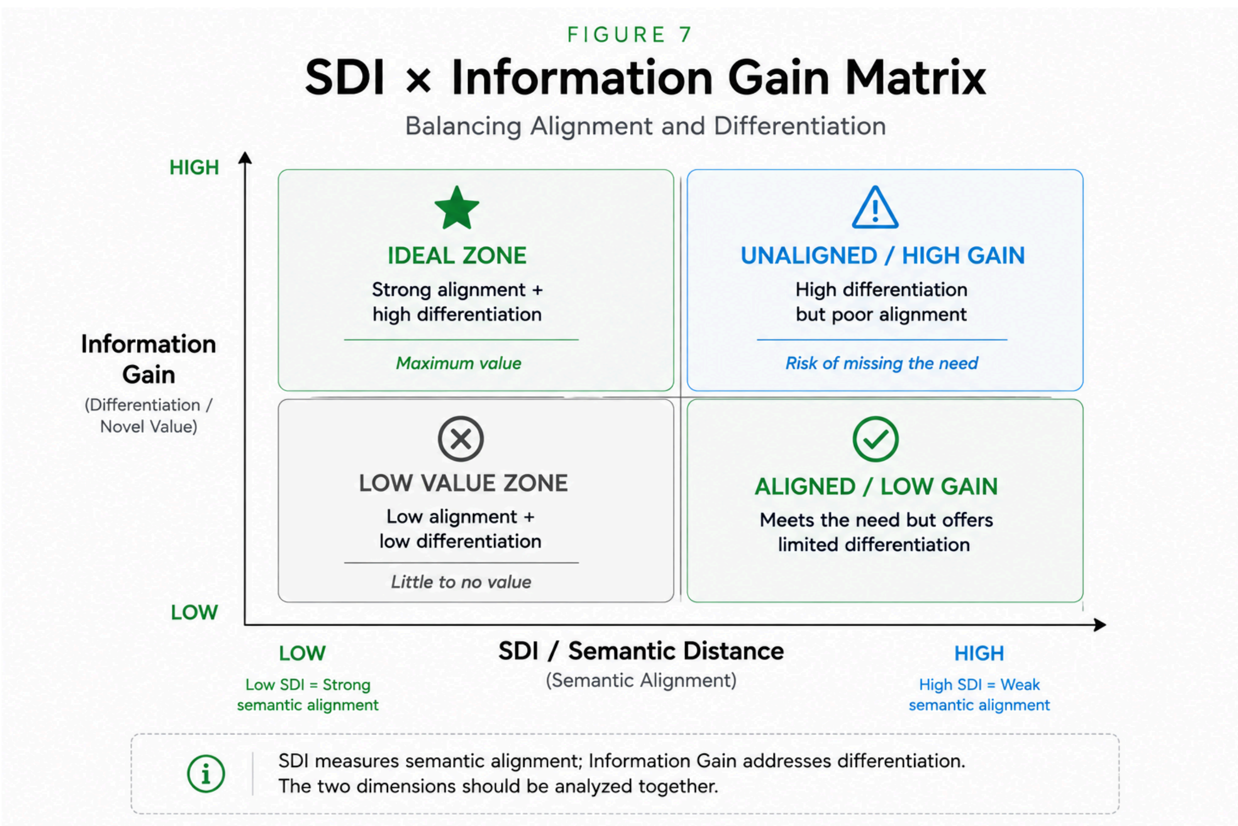
> Does the content address the information need?

Information Gain

Differentiation

> Does the content provide additional useful information beyond what is already represented? Read more: Information Gain

The two dimensions can therefore be considered together.



24. SDI and Semantic Drift

Semantic Drift and SDI address different analytical questions.

Semantic Drift

> How does semantic meaning move away from its intended direction?

SDI

> How far is the current document from a defined information need?

A simplified distinction is:

Semantic Drift = movement

SDI = distance

They can therefore be used together.

For example:

$$SDI_{initial} \longrightarrow Content\ Change \longrightarrow SDI_{final}$$

25. SDI and Vector Shift

Vector Shift can measure semantic movement between two document states.

SDI measures separation from a target.

A future experiment could therefore investigate:

$$SDI_{before}$$

versus:

$$SDI_{after}$$

while simultaneously measuring:

$$VectorShift_{before \rightarrow after}$$

> This could help determine whether a content modification actually moved the document closer to its intended semantic target.

26. SDI and Query Fan-Out

Query Fan-Out can be used to identify multiple information requirements associated with a search query.

SDI could then evaluate each requirement separately.

For example:

Main Query

↓

Requirement 1

↓

SDI

Requirement 2

↓

SDI

Requirement 3

↓

SDI

This produces a broader semantic coverage profile.

The potential benefit is that content can be evaluated against an information space, rather than against only one query string.

27. SDI for Content Briefs

SDI can be applied before writing.

Target Query

> How to optimize crawl budget

Information Need

Diagnose and improve crawl efficiency.

Required Concepts

- crawl budget,
- crawl demand,
- crawl rate,
- server capacity.

Required Entities

- Googlebot,
- robots.txt,
- XML sitemap,
- URLs,
- server.

Required Tasks

- diagnose,
- prioritize,
- optimize,
- verify.

Required Relationships

Server capacity → crawl rate

URL quality → crawl demand

robots.txt → crawler access

The resulting requirement set becomes the semantic foundation of the content brief.

The writer is therefore not starting with:

> "What keywords should I include?"

The writer is starting with:

> **"What must this document represent to satisfy the information need?"** Read more: keyword-to-content mapping

28. SDI for Content Auditing

The SDI audit workflow is:

Step 1 — Identify the target query

Step 2 — Define the information need

Step 3 — Build the requirement set

Step 4 — Assign requirement importance

Step 5 — Evaluate the document

Step 6 — Calculate the SDI profile

Step 7 — Identify the largest semantic gaps

Step 8 — Improve the document

Step 9 — Recalculate SDI

This creates a repeatable optimization process.

29. SDI for Content Pruning

SDI may also be useful for content consolidation and pruning.

A page originally targeting:

> SEO audit checklist

may gradually evolve into:

> General SEO guide

The page may still contain many related terms.

However, its current semantic representation may have moved away from the original information need.

An SDI analysis could therefore help determine whether a page should be:

- refocused,
- rewritten,
- consolidated,
- redirected,
- retargeted.

This application remains a proposed use case requiring testing.

30. SDI for Internal Linking

Two pages can be highly related while satisfying different information needs.

For example:

Page A

> What is crawl budget?

Page B

> How to optimize crawl budget.

Both may share:

- crawl budget,
- Googlebot,
- crawling,
- URLs.

But their intended tasks differ.

SDI can therefore potentially help distinguish:

- complementary content,
- duplicate intent,
- supporting content,
- semantic gaps,
- hub/spoke relationships.

31. SDI for Entity SEO

Entity optimization should not simply become:

> Add more entities.

The more useful question is:

> Which entities are required by the information need, and are they represented in the appropriate context?

Entity Distance and Relational Distance are particularly relevant here.

A document can contain many entities while still failing to represent the relationships that make those entities useful.

32. SDI for AI Search Research

SDI may eventually be tested against generative search systems.

A future research question could be:

> Do documents with lower SDI have a higher probability of retrieval or citation in AI-mediated search environments?

This is a hypothesis.

It should not be presented as a fact.

SDI does not claim that AI systems calculate SDI. Google's guidance on AI features in Search

33. Manual SDI

The first implementation of SDI can be performed manually.

An analyst identifies:

- information need,
- concepts,
- entities,
- attributes,
- relationships,
- tasks.

The analyst then evaluates the document against each requirement.

Advantages

- transparent,
- explainable,
- auditable,
- suitable for early research.

Limitations

- slower,
- potentially subjective,
- dependent on evaluator expertise.

Manual analysis is therefore particularly appropriate for early pilot research.

34. Model-Assisted SDI

A future automated implementation could use NLP and embedding-based systems.

A potential pipeline is:

Document

↓

Passage Extraction

↓

Concept Extraction

↓

Entity Extraction

↓

Attribute Extraction

↓

Relationship Extraction

↓

Semantic Comparison

↓

SDI Profile

The automated approach could increase scale.

However, it introduces additional methodological dependencies.

Different models may identify different concepts, entities, relationships, or semantic distances.

Therefore, model-assisted SDI should eventually be calibrated against human judgments.

35. Passage-Level SDI

SDI can potentially operate at two levels.

Document-Level SDI

Measures semantic distance across the document as a whole.

Passage-Level SDI

Measures individual passages against individual information requirements.

Passage-level analysis could be valuable because a document may contain:

- one highly aligned section,
- several moderately aligned sections,
- large amounts of unrelated material.

Future research should determine whether document-level SDI is best calculated directly or derived from passage-level measurements.

36. The SDI Optimization Cycle

Once semantic distance has been measured, the framework becomes iterative.

The process is:

Define

↓

Decompose

↓

Inspect

↓

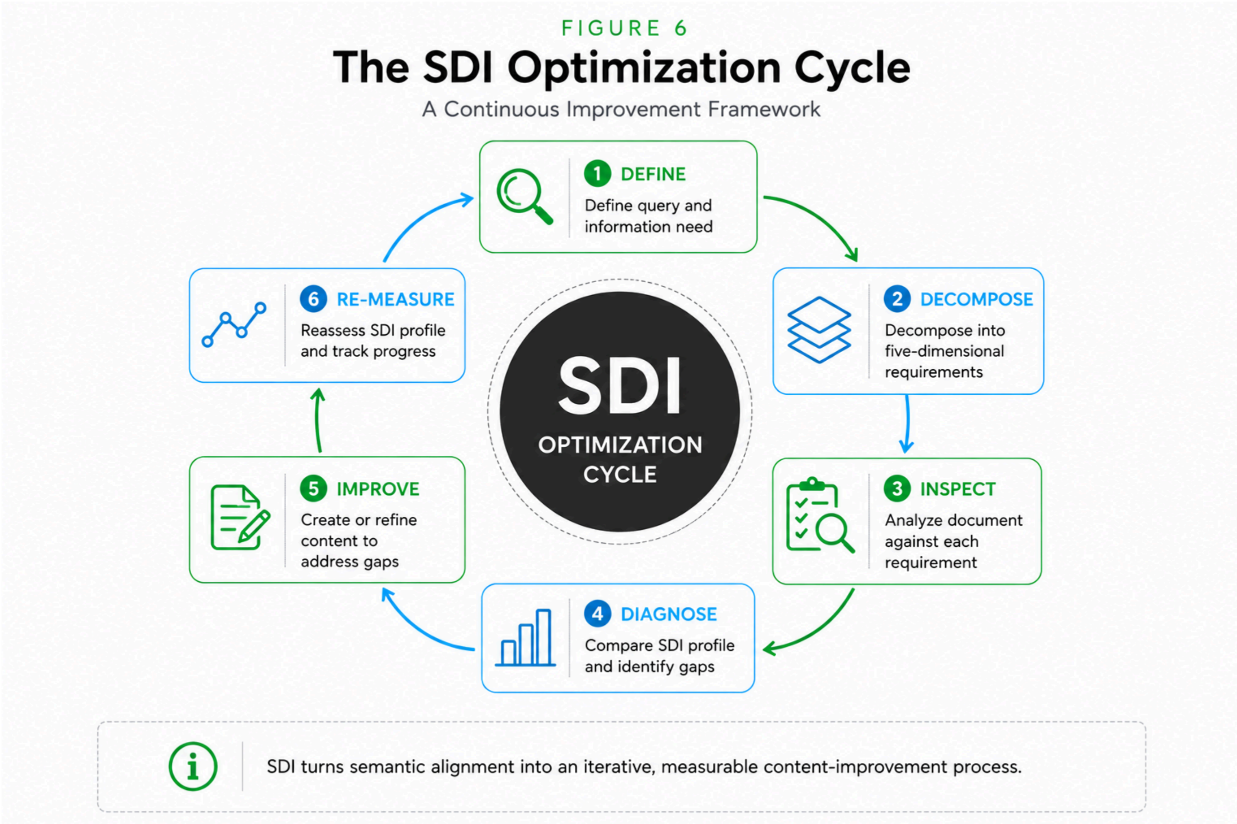
Diagnose

↓

Improve

↓

Re-measure



37. SDI Research Model

The long-term objective is not merely to create a score.

The objective is to determine whether SDI provides useful explanatory and diagnostic information.

The proposed research pathway is:

Framework

↓

Operationalization

↓

Human Evaluation

↓

Reliability Testing

↓

Baseline Comparison

↓

Ablation Testing

↓

Weight Optimization

↓

External Validation

38. Proposed Hypotheses

H1 — Information-Need Satisfaction

Lower SDI will be positively associated with human-rated information-need satisfaction.

H2 — Multidimensional Utility

The five-dimensional SDI model will provide greater diagnostic utility than a single-dimensional semantic similarity measurement.

H3 — Intent Effects

Intent Distance will contribute more strongly to information-need satisfaction for procedural queries than for definition-oriented queries.

H4 — Relational Contribution

Relational Distance will provide additional explanatory value beyond entity presence alone.

H5 — Intervention Effect

Targeted content modifications that reduce SDI will increase human-rated semantic alignment.

H6 — Query-Type Variation

The relative importance of SDI dimensions will vary across query types.

These are **proposed hypotheses**.

They are not findings.

39. Empirical Validation Framework

A serious validation program should compare SDI against simpler approaches.

Potential baselines include:

Baseline A

Lexical overlap.

Baseline B

BM25-style retrieval similarity.

Baseline C

Embedding-based semantic similarity.

Baseline D

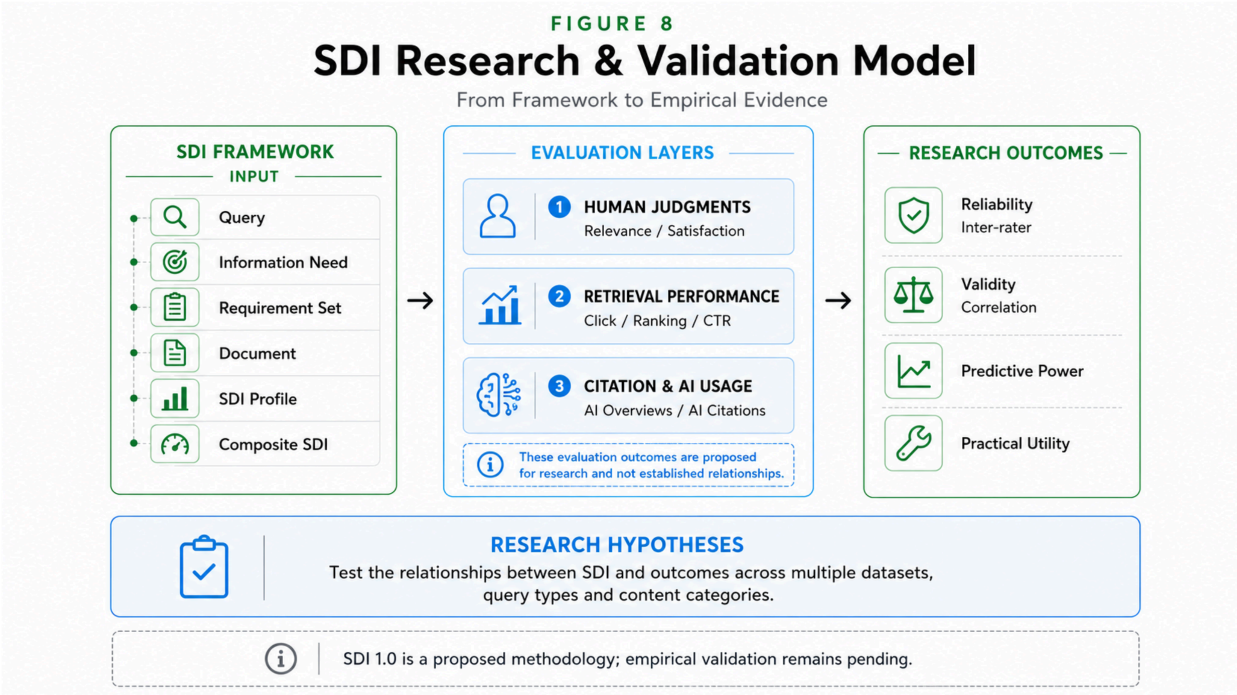
Entity overlap.

Proposed Model

Multidimensional SDI.

The central research question becomes:

> Does SDI explain human judgments of information-need satisfaction better than simpler semantic or lexical approaches?



40. Proposed Pilot Study

A reasonable initial pilot could use:

50 queries × 5 documents = 250 query-document pairs

The dataset could include multiple query categories, such as:

- informational,
- procedural,
- comparison,
- troubleshooting,
- transactional.

Each document would receive:

1. an SDI profile,
2. a composite SDI,
3. human information-need satisfaction ratings. See human relevance judgments

The study could then compare SDI with simpler baselines.

The purpose of the pilot would not be to prove that SDI predicts rankings.

It would first determine whether SDI is reliable, distinguishable, and practically useful as a measurement framework.

41. Inter-Rater Reliability

Because SDI involves interpretation, independent raters should eventually score the same query-document pairs.

For example:

Rater A → SDI 0.31

Rater B → SDI 0.48

A substantial difference would indicate a measurement problem.

Future research should therefore investigate appropriate inter-rater reliability statistics, potentially including:

- Cohen's kappa,
- weighted kappa,
- intraclass correlation,
- Krippendorff's alpha.

The appropriate statistic should depend on the final scoring design and unit of analysis.

No reliability claim is made for SDI 1.0 at this stage.

42. Ablation Study

One of the most important proposed experiments is an ablation study.

Compare progressively larger models:

Model 1

Concept only.

Model 2

Concept + Intent.

Model 3

Concept + Intent + Entity.

Model 4

Concept + Intent + Entity + Attribute.

Model 5

Concept + Intent + Entity + Attribute + Relationship.

The purpose is to determine whether each additional dimension contributes useful information.

If a dimension does not improve reliability or explanatory performance, the framework should be reconsidered.

This is preferable to assuming that all five dimensions are automatically necessary.

43. Query-Type Weighting

The equal weighting of SDI 1.0 is deliberately simple.

Future research should test whether different query types require different weights.

For example, a definition query may place greater importance on:

- concepts,
- entities.

A troubleshooting query may potentially place greater importance on:

- intent,
- relationships,
- attributes.

A comparison query may potentially place greater importance on:

- attributes,
- relationships.

These are research hypotheses, not established weighting rules.

44. SDI and Human Information-Need Satisfaction

The strongest early validation target is likely to be human evaluation.

A human evaluator could answer questions such as:

- Did the document address the requested task?
- Did it provide the necessary information?
- Did it address important entities?
- Did it explain important relationships?
- Did it provide enough detail to satisfy the stated need?

The resulting human judgments could then be compared against SDI.

This creates a measurable relationship:

SDI↔Human Satisfaction

The research objective would be to determine whether the relationship is sufficiently strong and consistent to justify SDI as a useful measurement instrument.

45. SDI and Search Performance

Search performance could be examined later.

Potential variables could include:

- retrieval,
- ranking,
- click-through rate,
- impressions,
- clicks,
- query coverage.

However, search performance is influenced by many factors beyond semantic alignment.

Therefore:

> A relationship between SDI and search performance would not by itself demonstrate that SDI is a ranking factor.

This distinction must remain explicit.

46. SDI and AI-Mediated Search

A later research phase could investigate:

- AI retrieval,
- AI-generated answers,
- citation inclusion,
- answer-source selection.

Potential questions include:

> Are lower-SDI documents more frequently retrieved?

and:

> Are lower-SDI documents more frequently cited?

Again, these are empirical questions.

No claim is made that current AI systems explicitly calculate SDI.

47. Limitations

SDI 1.0 has significant limitations.

47.1 Information-Need Ambiguity

A short query can support multiple interpretations.

Different analysts may therefore construct different information requirements.

47.2 Requirement-Construction Subjectivity

The quality of SDI depends partly on how accurately the information need is decomposed.

47.3 Scoring Subjectivity

Human evaluators may assign different distance values to the same content.

This makes inter-rater reliability essential.

47.4 Model Dependence

Automated implementations may produce different outputs depending on:

- language model,
- embedding model,
- extraction method,
- context window,
- relationship-detection method.

47.5 Multiple Valid Answers

Different documents may satisfy the same information need through different structures.

There may therefore be no single correct semantic representation.

47.6 Coverage Is Not Quality

Reducing semantic distance does not automatically make content:

- accurate,
- trustworthy,
- authoritative,
- useful,
- well-written.

47.7 Low SDI Is Not a Ranking Guarantee

A document can have low semantic distance and still perform poorly in search for many reasons.

47.8 High Information Coverage Can Become Excessive

Adding information merely to reduce semantic distance may produce unnecessary verbosity.

Semantic alignment should therefore be evaluated together with usefulness and differentiation.

48. Research Boundaries

The following boundaries are essential to SDI 1.0.

SDI does not claim:

- Google uses SDI.
- Google calculates an SDI score.
- Search engines assign SDI values to pages.
- Lower SDI guarantees higher rankings. Google's people-first content guidance
- Lower SDI guarantees AI citations.
- The proposed weights are optimal.
- The proposed diagnostic bands are empirically validated.

SDI does propose:

- a structured way to represent information needs,
- five dimensions of semantic distance,
- a provisional scoring system,
- a diagnostic methodology,
- a framework for empirical testing.

49. The SDI Research Roadmap

The framework can evolve through successive versions.

Phase 1 — Framework Construction

SDI 1.0

Define:

- dimensions,
- notation,
- scoring protocol,
- applications,
- research questions.

Phase 2 — Pilot Dataset

Develop a controlled query-document dataset.

Phase 3 — Human Evaluation

Collect independent information-need satisfaction judgments.

Phase 4 — Reliability Testing

Measure agreement between evaluators.

Phase 5 — Baseline Comparison

Compare SDI with lexical and semantic baselines.

Phase 6 — Ablation Study

Test the contribution of each dimension.

Phase 7 — Weight Optimization

Determine whether equal weighting should be replaced by empirically derived weights.

Phase 8 — External Validation

Test SDI across different query types, topics, domains, and document formats.

Phase 9 — Search Validation

Investigate relationships with retrieval and search performance.

Phase 10 — AI Search Validation

Investigate potential relationships with AI retrieval and citation behavior.

50. Practical SDI Analysis Workflow

A practitioner can apply SDI using the following sequence.

Step 1 — Define

Identify the target query.

Step 2 — Interpret

Define the likely information need.

Step 3 — Decompose

Create the five-dimensional requirement set.

Step 4 — Prioritize

Assign requirement importance.

Step 5 — Inspect

Evaluate the document.

Step 6 — Score

Assign dimension-level distances.

Step 7 — Calculate

Generate the composite SDI.

Step 8 — Diagnose

Identify the largest semantic gaps.

Step 9 — Improve

Modify the content.

Step 10 — Re-measure

Calculate SDI again.

This produces an iterative methodology:

> **Define → Decompose → Inspect → Diagnose → Improve → Re-measure**

51. SDI Diagnostic Worksheet

Target Query

Information Need

Required Concepts

1. ____
2. ____
3. ____

Required Intent / Tasks

1. ____
2. ____
3. ____

Required Entities

1. ____
2. ____
3. ____

Required Attributes

1. ____
2. ____
3. ____

Required Relationships

1. ____
2. ____
3. ____

Dimension Scores

Dimension	Score
Conceptual Distance	____
Intent Distance	____

Entity Distance ____

Attribute Distance ____

Relational Distance ____

Composite SDI ____

Primary Gap

Secondary Gap

Recommended Intervention

SDI After Revision

52. SDI in One Formal Model

The complete framework can be summarized as:

INFORMATION NEED

$N(Q)$

SEMANTIC REQUIREMENT SET

$R(Q) = \{ C, I, E, A, L \}$

DOCUMENT REPRESENTATION

$D = \{ C', I', E', A', L' \}$

DIMENSION-LEVEL DISTANCES

D^C, D^I, D^E, D^A, D^R

WEIGHTED COMPOSITE

$SDI = \sum_i w_i \cdot D_i$

DIAGNOSTIC PROFILE

$SDI_{Profile} = (D^C, D^I, D^E, D^A, D^R)$

The profile provides the multidimensional diagnosis.

The composite provides the summary measurement.

53. The Central SDI Proposition

The framework can ultimately be reduced to one proposition:

> Content should be evaluated not only by how closely it resembles a query, but by how completely it represents the semantic requirements of the information need behind that query.

SDI proposes a method for making that proposition operational.

54. Conclusion

Search queries are compressed expressions of information needs.

A document can therefore be semantically related to a query without fully satisfying what the searcher actually needs.

The Semantic Distance Index (SDI) proposes a multidimensional approach to this problem.

Rather than treating semantic alignment as a single similarity value, SDI decomposes it into five dimensions:

1. Conceptual Distance
2. Intent Distance
3. Entity Distance
4. Attribute Distance
5. Relational Distance

The resulting SDI Profile can reveal where content is aligned and where important semantic gaps remain.

The provisional composite score provides a concise summary:

$$0 \leq SDI \leq 1$$

with lower values representing greater proposed semantic alignment.

However, SDI 1.0 should be understood as a research framework, not an established search-engine metric.

Its weighting system, scoring protocol, interpretation bands, reliability, validity, and relationship with search performance remain open empirical questions.

That uncertainty is not a weakness of the framework.

It is the reason for the research program.

The next stage is therefore not to assume that SDI works.

It is to test whether it works.

Citation

SearchEngineZine. (2026). *Semantic Distance Index (SDI) 1.0: A Multidimensional Framework for Measuring Semantic Alignment Between Search Intent and Content.* SearchEngineZine Research Frameworks.

Research Note

> SDI is an independently developed SearchEngineZine research framework. It is presented as a proposed methodology for analyzing semantic alignment

between information needs and content. The framework does not claim to represent Google's internal ranking systems or any proprietary search-engine scoring mechanism. Numerical weights, scoring bands, reliability, validity, and relationships with search or AI-search outcomes remain subject to empirical validation.

About This Framework

The **Semantic Distance Index** is part of SearchEngineZine's research framework series exploring measurable approaches to modern search, semantic SEO, information retrieval, content architecture, and search-intent analysis.

The framework is designed to evolve through experimentation and empirical validation rather than remain a fixed theoretical model.