УДК 004.8:004.832

Жвакина Анна Васильевна, к.т.н., доцент
Белорусский государственный университет
информатики и радиоэлектроники
Zhvakina Anna Vasilevna
Парамонов Антон Иванович, к.т.н., доцент
Белорусский государственный университет
информатики и радиоэлектроники
Paramonov Anton Ivanovich
Круговой Владислав Николаевич, магистрант
Белорусский государственный университет
информатики и радиоэлектроники
Krugovoy Vladislav Nikolaevich

АНАЛИЗ МЕТРИК СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ДЛЯ ГИБРИДНОЙ МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОГО ПРОСТРАНСТВА
ANALYSIS OF SEMANTIC SIMILARITY METRICS FOR A HYBRID SMART SPACE MODEL

Аннотация. *Рассмотрены онтологические, информационные, дистрибутивно-векторные метрики оценки степени семантической близости и оценена возможность их использования в гибридной модели интеллектуального пространства*

Abstract. *Ontological, informational, and distribution-vector metrics for assessing the degree of semantic closeness are considered, and the possibility of their use in a hybrid model of intelligent space is assessed*

Ключевые слова: *Интеллектуальное пространство, метрики семантической близости, гибридная модель, распознавание образов, семантические сети, исследование операций*

Keywords: *Intelligent space, semantic similarity metrics, hybrid model, pattern recognition, semantic networks, operations research*

Подходы, используемые для обработки данных и знаний, отличаются математическим аппаратом, способами представления информации. Каждый из них эффективен в определенных случаях и может дополнять другой. Поэтому целесообразно использовать их сильные стороны, объединив в гибридной модели интеллектуального пространства, где будет иметь место комплексирование методов распознавания образов, семантических сетей и исследования операций.

Важную роль при этом играет количественная оценка степени близости смыслового сходства между объектами, терминами, ситуациями. Для этого применяются метрики семантической близости, то есть формализованные способы вычисления расстояния или сходства между элементами знаний.

Выделяют следующие классы метрик в зависимости от типа информации:

онтологические, когда выполняется анализ иерархии и расстояний между узлами понятий;

информационные, опирающиеся на частоту встречаемости терминов и их информационную ценность;

дистрибутивно-векторные, использующие эмбеддинги и геометрические меры близости.

В каждом из данных классов могут быть использованы различные варианты метрик.

Так для определения семантической близости на символьном уровне (**онтологические** метрики) применяется иерархия понятий, где любое понятие можно представить как вершину графа, а его рёбра задают отношения «is-a» (является подклассом). Семантическая близость определяется расстоянием между узлами (минимальное количество ребер) или положением наименьшего общего предка.

В метрике Рады (Rada metric) сходство понятий c_1 и c_2 обратно пропорционально кратчайшему расстоянию между ними

$$\mathrm{Sim}_{\mathrm{Rada}}\,(c_1, c_2) = \frac{1}{\mathrm{len}\,(c_1, c_2)} \qquad (1)$$

где $\mathrm{len}\,(c_1, c_2)$ – длина кратчайшего пути между понятиями в иерархическом графе.

Однако здесь не учитывается глубина залегания узлов, и все рёбра считаются равнозначными, что ограничивает применение данной метрики.

Метрика By – Палмера (Wu & Palmer) учитывает глубину наименьшего общего предка и глубину самих понятий:

$$\mathrm{Sim}_{\mathrm{WP}}\,(c_1, c_2) = \frac{2 \cdot \mathrm{depth}\,(\mathrm{LCS}\,(c_1, c_2))}{\mathrm{depth}\,(c_1) + \mathrm{depth}\,(c_2)} \qquad (2)$$

где $\mathrm{LCS}\,(c_1, c_2)$ – наименьший общий предок понятий c_1 и c_2, $\mathrm{depth}\,(\cdot)$ – глубина узла в иерархии понятий (в примере глубины отсчитываются от 1 для корневого узла).

Для глубоких онтологий, например, биомедицинских, применяется метрика Лечкака – Чодороу (Leacock & Chodorow), использующая логарифмическую шкалу:

$$\mathrm{Sim}_{\mathrm{LC}}\ (c_1, c_2) = -\log\left(\frac{\mathrm{len}\ (c_1, c_2)}{2D}\right) \qquad (3)$$

где D – максимальная глубина иерархии. Здесь мелкие различия на нижних уровнях считаются более важными, чем на верхних.

Метрика Лечкака – Чодороу, метрика Ву – Палмера и метрика Рада реализованы в пакете UMLS-Similarity [5].

Метрика Ли – Чжоу (Lee & Zhou) в сравнении с метрикой Рады использует веса на ребрах. Связи на верхних уровнях более абстрактны, поэтому должны весить меньше, а конкретные связи внизу – больше.

$$\mathrm{Sim}_{\mathrm{LZ}}\ (c_1, c_2) = \frac{1}{\sum_{e \in \mathrm{path}\ (c_1, c_2)} w\ (e)} \qquad (4)$$

где path (c_1, c_2) – множество рёбер на кратчайшем пути между понятиями, $w\ (e)$ – вес ребра e, который уменьшается с уменьшением глубины залегания (связи верхних уровней имеют меньший вес). Конкретная реализация весовой функции зависит от предметной области.

Структурное сходство двух фрагментов онтологии оценивается метрикой на основе максимального общего подграфа [3], где используются графовые нейронные сети, эмбеддинги узлов и косинусное сходство:

$$\mathrm{Sim}_{\mathrm{MCS}}\ (G_1, G_2) = \frac{|\ \mathrm{MCS}\ (G_1, G_2)\ |}{\max\ (|\ G_1\ |, |\ G_2\ |)} \qquad (5)$$

где MCS (G_1, G_2) – наибольший общий подграф фрагментов G_1 и G_2, $|\ G\ |$ – число узлов во фрагменте.

Например, при сравнении фрагмента «Компьютер → Ноутбук → Ультрабук» (3 узла) и фрагмента «ЭВМ → Портативный компьютер → Ноутбук» (3 узла) наибольший общий подграф включает два узла («Ноутбук» и его связь с вышестоящим понятием). Подставляя значения в (5), получаем сходство $= 2/3 \approx 0{,}67$ [3] В результате эффективность поиска MCS улучшается на более чем 90% для крупных графов по сравнению с существующими алгоритмами, что достигается за счёт применения best-first search вместо depth-first search.

Для выравнивания онтологий и оценки структурного сходства может использоваться модель Duan & Lee [2], которая комбинирует word embeddings и BERT для вычисления текстового сходства, а также алгоритм SimRank для структурного сходства между графами онтологий.

Метрика таксономического перекрытия (Taxonomic Overlap) анализирует множества, включающие само понятие, его предков (более общие понятия) и потомков (более конкретные понятия):

$$\mathrm{Sim}_{\mathrm{TaxOver}}(c_1, c_2) = \frac{|\, S(c_1) \cap S(c_2) \,|}{|\, S(c_1) \cup S(c_2) \,|} \qquad (6)$$

где $S(c)$ – семантическое множество, включающее понятие c, всех его предков и всех потомков в иерархии понятий. Для автоматического построения таксономий и преобразования таксономических данных в формальные OWL-онтологии может применяться инструмент Taxonomy OWLizer (TOWLizer) [5], который автоматизирует обработку синонимов и генерацию иерархий.

Общим ограничением всех графовых метрик является их неприменимость при отсутствии явной иерархии понятий или при наличии синонимии вне иерархии. Например, если в онтологии понятие «Легковой автомобиль» не связано с понятием «Седан» отношением «is-a», указанные выше метрики не смогут определить семантическую близость этих понятий, хотя в естественном языке они являются близкими.

Комплексный обзор современных методов семантической близости представлен в [4], где рассмотрены трансформерные модели (FarSSiBERT, DeBERTa-v3), контрастное обучение (AspectCSE), доменно-специфичные решения (CXR-BERT для медицины, Financial-STS для финансов), мультимодальные и графовые подходы. Особое внимание уделено методам, интегрирующим внешние знания (онтологии, графы знаний), что подтверждает актуальность гибридного подхода.

Информационные метрики (статистический уровень) опираются на теорию информации. Каждому понятию ставится в соответствие информационная ценность (Information Content – IC), которая измеряет, насколько понятие является специфичным. Чем реже понятие встречается в корпусе текстов (или чем больше потомков оно имеет в иерархии), тем выше его информационная ценность, которую можно вычислить по формуле:

$$IC\ (c) = -\log\ P\ (c) \qquad (7)$$

где $P\ (c)$ – вероятность встретить понятие c в корпусе текстов (оценивается по частоте) или доля потомков понятия c в иерархии понятий.

Чем больше информации несёт понятие, тем выше его IC. Корневые узлы (например, «Транспорт») имеют низкую IC, а листовые узлы (например, «Легковой автомобиль») – высокую.

Существует несколько метрик для данного класса. При использовании метрики Резника (Resnik) сходство между понятиями определяется исключительно информационной ценностью их наименьшего общего предка (LCS) [6]:

$$\mathrm{Sim}_{\mathrm{Resnik}}\ (c_1, c_2) = IC\ (\mathrm{LCS}\ (c_1, c_2)) \qquad (8)$$

Например, для понятий «Легковой автомобиль» и «Мотоцикл» наименьший общий предок – «Наземный транспорт». Если IC («Наземный транспорт») $= 0{,}5$, то и сходство = 0,5. При этом IC («Легковой автомобиль») и IC («Мотоцикл») не учитываются.

Однако два различных понятия, имеющих одного и того же предка, получают одинаковую оценку сходства, даже если сами понятия сильно различаются по информационной ценности.

Метрика Лина (Lin) учитывает как информационную ценность наименьшего общего предка, так и самих понятий [7]:

$$\mathrm{Sim}_{\mathrm{Lin}}\ (c_1, c_2) = \frac{2 \cdot IC\ (\mathrm{LCS}\ (c_1, c_2))}{IC\ (c_1) + IC\ (c_2)} \qquad (9)$$

Значение лежит в интервале [0,1]. Если понятия идентичны, сходство равно единице.

Например, пусть

IC («Наземный транспорт») $= 0{,}5$,

IC («Легковой автомобиль») $= 0{,}9$,

IC («Мотоцикл») $= 0{,}8$.

Подставляя значения в (9), получаем сходство = $(2 \cdot 0{,}5) / (0{,}9 + 0{,}8) = 1/1{,}7 \approx 0{,}59$. По сравнению с метрикой Резника (0,5) данная оценка учитывает специфичность сравниваемых понятий. Однако требуется корректная оценка IC для всех понятий; при отсутствии статистических данных результат может быть нестабильным.

При использовании метрики Цзян – Конрата (Jiang & Conrath) вычисляется семантическое расстояние, а не сходство [8]:

$$\mathrm{Dist}_{\mathrm{JC}}\ (c_1, c_2) = IC\ (c_1) + IC\ (c_2) - 2 \cdot IC\ (\mathrm{LCS}\ (c_1, c_2)) \qquad (10)$$

Сходство может быть получено как $\mathrm{Sim}_{\mathrm{JC}} = 1 / (1 + \mathrm{Dist}_{\mathrm{JC}})$.

В метрике Санчеса (Sánchez) информационная ценность вычисляется на основе структуры иерархии понятий [9]. В данном случае $IC\ (c)$ вычисляется по числу дочерних узлов:

$$IC_{\text{Sanchez}}\ (c) = -\log\left(\frac{|\ \text{leaves}\ (c)\ |}{|\ \text{leaves (root)}\ |}\right) \qquad (11)$$

где $|\ \text{leaves}\ (c)\ |$ – количество листовых узлов в поддереве с корнем c, а $|\ \text{leaves (root)}\ |$ – общее количество листьев во всей иерархии понятий.

Однако метрика не учитывает частоту реального употребления понятий: специфичное понятие может быть редко используемым, но формально иметь высокую IC.

Общим ограничением информационных метрик является зависимость от качества оценки информационной ценности IC. Если IC вычисляется на основе внешнего текстового корпуса, результаты чувствительны к его репрезентативности. Если IC вычисляется интринсивно (структурно), метрики не учитывают реальную частотность употребления терминов. Кроме того, информационные метрики, как и графовые, не работают при отсутствии явной иерархии понятий.

Лучшее значение средней степени достоверности составляет 0,86, а средней корреляции Пирсона 0,69 [1], что подтверждает эффективность IC-подходов для оценки семантической близости в иерархических структурах.

Дистрибутивно-статистические метрики (векторный уровень) опираются на гипотезу дистрибутивной семантики, согласно которой смысл слова определяется его контекстом. Понятия представляются как векторы в многомерном пространстве (эмбеддинги), получаемые с помощью нейросетевых моделей (Word2Vec, GloVe, BERT, LaBSE). Близость между понятиями вычисляется геометрически.

Простейшим вариантом является косинусное сходство (Cosine Similarity) – стандартная мера для сравнения векторных представлений, которая вычисляется по формуле:

$$\text{Sim}_{\text{Cosine}}\ (\text{a}, \text{b}) = \cos\ (\theta) = \frac{\text{a} \cdot \text{b}}{\|\ \text{a}\ \|\|\ \text{b}\ \|} \qquad (12)$$

где a и b – векторные представления сравниваемых понятий, $\text{a} \cdot \text{b}$ – скалярное произведение векторов, $\|\ \text{a}\ \|$ и $\|\ \text{b}\ \|$ – их евклидовы нормы.

Значение косинусного сходства лежит в интервале $[-1,1]$ (для неотрицательных векторов – $[0,1]$), где единица соответствует идентичным векторам, ноль – ортогональным (несвязанным) векторам.

Например, в модели Word2Vec вектор понятия «Автомобиль» может быть близок к вектору «Мотоцикл» (косинусное сходство ≈ 0,7), тогда как вектор «Автомобиль» будет далёк от вектора «Компьютер» (косинусное сходство ≈ 0,1). Данная метрика независима от длины векторов и широко применяется в информационном поиске, кластеризации текстов и выравнивании эмбеддингов сущностей.

Другим распространённым подходом является евклидово расстояние (Euclidean Distance), используемое в функциях потерь для плотных пространств:

$$\mathrm{Dist}_{\mathrm{Eucl}}(\mathrm{a},\mathrm{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \qquad (13)$$

где n – размерность векторного пространства, a_i и b_i – координаты векторов a и b по i-му измерению.

Чем меньше евклидово расстояние, тем ближе объекты друг к другу. Например, для векторов «Легковой автомобиль» и «Седан» евклидово расстояние будет меньше, чем для «Легковой автомобиль» и «Самолёт». Данная метрика чувствительна к масштабу признаков, поэтому требует предварительной нормализации векторов.

Общим достоинством векторных метрик является способность улавливать скрытые смысловые связи, синонимию и парафразы. Например, понятия «Легковой автомобиль» и «Седан» могут иметь высокое косинусное сходство, даже если в иерархии понятий они не связаны явным отношением «is-a». Однако результат нельзя объяснить логически, метрики чувствительны к шуму в обучающих данных и демонстрируют низкое качество для редких терминов, не представленных при обучении.

Общим ограничением всех трёх классов метрик является отсутствие универсального подхода, одинаково эффективного для всех типов данных и задач.

С помощью фреймворка OntoAligner [10] семнадцать моделей встраивания графов знаний (KGE), включая ConvE, TransF, DistMult и SE, оценивались на семи наборах данных OAEI-2024. Сходство

вычислялось как косинусная мера между нормализованными эмбеддингами сущностей. Применялась пороговая фильтрация для отсечения ложных срабатываний.

В результате модель DistMult (билинейная метрика, близкая к информационному классу) достигла точности 97,9% на задаче Mouse-Human (таксономия анатомии). Полнота при этом составила 69,0%, что типично для консервативных мер, отдающих приоритет правильности, а не полноте.

Модель TransF (гибкая трансляционная метрика, относящаяся к графовому классу) показала идеальную точность 100% на задаче FISH-ZOOPLANKTON (биоразнообразие). Однако полнота не превысила 60%.

Модель ConvE (гибридная сверточная метрика) тестировалась на базах знаний для ENVO-SWEET (окружающая среда и явления природы), CEON-BiOnto (переработка и биоэкономика) и OMIM-ORDO (биомедицина и редкие болезни). Полученные точность (в диапазоне 58–89%) и полнота (40–55%) делает её наиболее универсальной среди рассмотренных.

Интегральная метрика F1 (учитывает точность и полноту) для модели SE (структурная нейросетевая метрика) для онтологии NCIT-DOID (рак и все болезни) равна 60,2% (интегральная метрика, сочетающая точность и полноту). Такой же результат и для методов, не использующих нейросети.

Во всех результатах значение точности выше полноты, то есть метрики предпочитают найти меньше совпадений, но без ошибок.

Вариантом улучшения качества может быть применение разных метрик. Так в SUMEX использованы эмбеддинги ClinicalBERT (медицинские тексты) с фильтром медицинской онтологии. Это позволило достичь точности на 7% больше и уменьшить ложные срабатывания на 23% [11]. В CIDER-LM понятия из онтологии, представленные в виде обычных предложений, были обработаны моделью LaBSE (многоязычная). При этом F1 увеличился на 14–18% по сравнению с вариантом, когда понятия не переводились в текст [12].

Предлагаем гибридную модель, где семантическая близость вычисляется в несколько уровней: выполняется проверка по онтологии, затем по смыслу с помощью нейросетей и в заключении оба результата объединяются.

Уровень 1 (онтология) – для предметных областей (медицина, инженерия), где важна точность, проверка осуществляется по

классификации. Применяется метрика Цзян – Конрата или Лина [8] [7], которые строго контролируют иерархию. На вход подается онтология в формате OWL/RDF, а на выходе получаем оценку близости понятий в диапазоне [0, 1].

Уровень 2 (векторный) – для неструктурированных данных, пользовательских запросов и связывания сущностей. Используется косинусное сходство эмбеддингов (LaBSE, Sentence-BERT, BERT). На входе текст или понятие, на выходе – оценка близости в диапазоне [0, 1].

Уровень 3 (механизм интеграции). Итоговая близость считается как взвешенная сумма двух предыдущих оценок:

Общая близость = α × (оценка онтологии) + (1-α) × (векторная оценка)

Коэффициент α (от 0 до 1) настраивается под конкретную задачу. Если α близок к 1 – приоритет у онтологии, например, при постановке медицинского диагноза, в юриспруденции. Если α близок к 0 – у векторной близости, например, при поиске по текстам, в чат-боте. При α ≈ 0,5 оба подхода работают на равных. Важно, чтобы обе оценки (онтологическая и векторная) были предварительно приведены к единому диапазону [0, 1].

Дополнительно система может динамически выбирать α на основе метаданных. Если оба понятия присутствуют в онтологии и путь между ними определён, значение α увеличивается. Если одно или оба понятия отсутствуют или являются редкими (низкая информационная ценность IC), значение α снижается, передавая управление векторному уровню. Такой подход позволяет гибридной модели использовать логический вывод там, где он возможен, и переключаться на вероятностную семантику в условиях неполноты данных или при работе с естественным языком.

Предложенная архитектура может быть реализована в различных сценариях. Для выравнивания медицинских онтологий рекомендуется $\alpha = 0{,}7–0{,}9$ с применением метрик Цзян – Конрата и Лечкака – Чодороу. Для поиска по научным текстам рекомендуется $\alpha = 0{,}3–0{,}5$ с комбинацией косинусного сходства (BERT/LaBSE) и метрики Лина. Для чат-бота с поддержкой предметной области рекомендуется $\alpha = 0{,}2–0{,}4$ с применением косинусного сходства (Sentence-BERT) и фильтрации по онтологии на основе метрики Рады. Для интеграции разнородных баз знаний рекомендуется $\alpha = 0{,}5–0{,}6$ с применением метрики Ву – Палмера и косинусного сходства эмбеддингов сущностей.

Механизм динамической настройки коэффициента α позволяет переключаться между логическим выводом (при полноте онтологии) и вероятностной семантикой (при работе с неструктурированными данными или естественным языком). Предложенная модель носит теоретический характер и требует экспериментального подтверждения. В случае успешной валидации она закладывает основу для создания унифицированного аппарата обработки данных и знаний в системах искусственного интеллекта.

Список литературы:

1. Formica A., et al. Information Content Methods for Semantic Similarity: An Experimental Assessment // IEEE Access. 2025. Vol. 13. P. 113953-113966. DOI: 10.1109/ACCESS.2025.3584192.
2. Duan H., Lee Y. A Novel Ontology Matching Model to Address Ontology Heterogeneity // International Journal of Internet, Broadcasting and Communication. 2025. Vol. 17. No. 1. P. 151-162. DOI: 10.7236/IJIBC.2025.17.1.151
3. Quer S., Madeo T., Calabrese A., Squillero G., Carraro E. Node Embedding and Cosine Similarity for Efficient Maximum Common Subgraph Discovery // Applied Sciences. 2025. Vol. 15. No. 16. P. 8920. DOI: 10.3390/app15168920.
4. Kumar L., et al. Advances and Challenges in Semantic Textual Similarity: A Comprehensive Survey // arXiv. 2025.
5. Soares F., et al. Taxonomy OWLizer: A Tool for Converting Taxonomic Data into OWL // Biodiversity Information Science and Standards. 2025. Vol. 9. P. e176413. DOI: 10.3897/biss.9.176413.
6. Resnik P. Using Information Content to Evaluate Semantic Similarity // Proceedings of IJCAI 1995. Vol. 1. P. 448-453.
7. Lin D. An Information-Theoretic Definition of Similarity // Proceedings of ICML 1998. P. 296-304.
8. Jiang J., Conrath D. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy // Proceedings of ROCLING 1997. P. 19-33.
9. Sánchez D., Batet M., Isern D. Ontology-based information content computation // Knowledge-Based Systems. 2011. Vol. 24. No. 2. P. 297-303. DOI: 10.1016/j.knosys.2010.10.001.
10. Babaei Giglou H., D'Souza J., Sanaei M., Auer S. OntoAligner Meets Knowledge Graph Embedding Aligners // Proceedings of the 20th International Workshop on Ontology Matching (OM 2025) co-located with ISWC 2025. CEUR Workshop Proceedings. 2025. Vol. 4085.

11. SUMEX: A Hybrid Framework for Semantic Textual Similarity and Explanation Generation // Information Processing & Management. 2024. Vol. 61. No. 4. DOI: 10.1016/j.ipm.2024.103748.

12. CIDER-LM: Semantic Alignment of Multilingual Knowledge Graphs via Contextualized Vector Projections // arXiv. 2025. arXiv:2601.00814.