А

УДК 004.048

Н. К. Андриевская
Государственное образовательное учреждение высшего профессионального образования
«Донецкий национальный технический университет», г. Донецк
283001, г. Донецк, ул. Артема, 58

ГИБРИДНАЯ ИНТЕЛЛЕКТУАЛЬНАЯ МЕРА ОЦЕНКИ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ

Natalia Andrievskaya
State Educational Institution of Higher Education “Donetsk national technical University”, Donetsk
283001, Donetsk, Artyoma str., 58

HYBRID INTELLIGENT MEASURE OF SEMANTIC SIMILARITY EVALUATION

Н. К. Андрієвська
Державна освітня установа вищої професійної освіти
«Донецький національний технічний університет», м. Донецьк
283001, м. Донецьк, вул. Артема, 58

ГІБРИДНА ІНТЕЛЕКТУАЛЬНА МІРА ОЦІНКИ СЕМАНТИЧНОЇ БЛИЗЬКОСТІ

В статье рассматривается задача определения оценки семантической близости между двумя объектами на базе гибридной меры. В результате исследований была разработана гибридная интеллектуальная мера оценки семантической близости между двумя объектами по семантическому трехмерному тензору графа знаний. Полученная мера включает вычисление семантической близости по онтологии, по семантике, по частотным характеристикам терминов и косинусному сходству контекстных векторов.
Ключевые слова: семантическая близость, онтология, контекстный вектор, меры сходства.

The article deals with the problem of determining the assessment of semantic proximity between two objects on the basis of a hybrid measure. As a result of the research, a hybrid intelligent method was developed for evaluating the semantic similarity between two objects using the semantic three-dimensional tensor of the knowledge graph. The resulting hybrid measure includes the calculation of semantic similarity by ontology, by semantics, from Natural Language, using Term Frequency and Cosine Similarity of context vectors.
Key words: semantic similarity, ontology, context vector, similarity measures.

У статті розглядається задача визначення оцінки семантичної близькості між двома об'єктами на базі гібридної міри. У результаті досліджень було розроблено гібридну інтелектуальну міру оцінки семантичної близькості між двома об'єктами на базі семантичного тривимірного тензору графа знань. Отримана міра включає обчислення семантичної близькості за онтологією, за семантикою, за частотними характеристиками термінів і подібністю за косинусною схожістю контекстних векторів.
Ключові слова:семантична близькість, онтологія, контекстний вектор, міри подібності.

Общая постановка проблемы

Современный мир характеризуется резким ростом объёмов информации, особенно в сети Интернет. Это привело к стремительному развитию Веб-технологий, переходу к концепции Semantic Web, о чем свидетельствует ряд разработанных стандартов WWW, например, стандарт на язык представления знаний (RDF) и язык описания онтологий (OWL) [1], [2]. Онтологии становятся базой для построения систем управления знаниями (СУЗ). В связи с этим в работах [3], [4] был описан онтологический подход к построению СУЗ научно-образовательных организаций, каркасом которого является прикладная онтология в области профессиональной деятельности сотрудников научно-образовательных организаций. Использование современных семантических технологий и онтологии позволили перейти на новый уровень обработки данных – семантический, когда появляется возможность поиска и извлечения знаний по заданной теме, а не просто конкретного файла [5].

Ключевым моментом при построении СУЗ является разработка алгоритмов расчета количественных оценок семантической близости (СБ) онтологических термов. Функция *F(X, Y)*, ставящая в соответствие каждой паре термов *X* и *Y* некоторый вещественный коэффициент, называется функцией, определяющей семантическую близость между двумя термами.

Для *F (X, Y)* действительны следующие свойства:

- $0 \le F(X, Y) \le 1$;
- $F(X, Y) = 1 \Leftrightarrow X = Y$ (объекты *X*, *Y* идентичны);
- $F(X, Y) = 0 \Leftrightarrow X \ne Y$ (объекты *X*, *Y* совершенно различны и не имеют схожих характеристик);
- $F(X, Y) = F(Y, X)$, (свойство симметричности функции подобия).

В свою очередь, каждый терм представляет собой некоторое размытое множество, куда попадают и другие подобные термы со значением семантической близости выше заданного порога. Принадлежность к нечеткому множеству задается с помощью значения семантической близости. Подобие сущностей *X*, *Y* означает, что $F(X, Y) >= t$, где *t* – пороговая величина (уровень подобия, уровень отсечения). В свою очередь, все термы со значением менее некоторого порогового значения *t*1 считаются различными. Другими словами, функция принадлежности к нечеткому множеству следующая:

$$\mu_A(v) = \begin{cases} 1, & \text{если } F(X, Y) >= t \\ F(X, Y), & \text{если } t1 < F(X, Y) < t \\ 0, & \text{если } F(X, Y) < t1, \end{cases} \quad (1)$$

где $F(X, Y)$ – функция, определяющая семантическую близость между двумя термами, t, $t1$ – пороговые значения.

Таким образом, необходимо разработать способ определения СБ для дальнейшего использования в алгоритмах поиска и классификации.

Обзор существующих методов оценки близости

Мера семантической близости между концептами – это числовая оценка их смысловой связанности. По тематике, посвященной методам оценки СБ терминов написано немало работ. В работе [6] сделан обширный обзор различных методов вычислений мер семантической близости термов внутри онтологий.

Одной из первых моделей оценки СБ является геометрическая модель, при которой каждая ось представляет собой некоторое свойство, а близость объектов интерпретируется как расстояние между объектами. Недостатком этой модели является то, что геометрическая модель никак не учитывает добавление общих свойств к объектам, что по идее должно увеличивать их близость.

Другим устоявшимся подходом является подход Тверски, когда через сопоставление одинаковых и различных свойств определяется близость двух объектов. В большинстве мер при вычислении СБ используется нормализованная модель (*ratio model*), которая является развитием подхода Тверски:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \quad (2)$$

где A и B – множества свойств этих объектов,

f – некоторая функция,

α и β – веса общих и различных свойств объектов.

Для получения данных большей точности и непротиворечивости используются гибридные модели, которые являются свертками различных мер оценки критериев подобия концептов онтологий, при этом чаще всего в гибридных мерах используется аддитивная свертка:

$$S(c_1,c_2) = \sum_{i=1}^{n} w_i S^i(c_1,c_2), \quad (3)$$

где S^i – i-ая мера близости по определенному критерию, c_1,c_2 – два концепта, w_i – вес i-й меры, n – число мер.

Использование сигмоидальной функции позволяет повысить веса мер с большими значениями и практически пренебречь мерами с малыми весами:

$$S(c_1,c_2) = \sum_{i=1}^{n} w_i Sig(S^i(c_1,c_2)) \quad (4)$$

$Sig(x) = 1/((1 + e^{\wedge}(-ax)))$, где a>0

Веса рассчитываются вручную экспертами или автоматически с помощью интеллектуальных методов – нейронной сети или генетического алгоритма.

Выделяют следующие типы мер семантической близости [6]:

1) таксономические – на основе иерархических (родовидовых, таксономических) связей;

2) реляционные – на основе неиерархических (ассоциативных, проблемно специфических, «горизонтальных») связей между терминами онтологии;

3) атрибутивные.

В свою очередь таксономические подразделяются на:

1) на определении кратчайшего пути;

2) основанные на определении глубины иерархии;

3) учитывающие глубину ближайшего общего родителя (LCS);

4) основанные на понятии общей специфичности двух вершин.

Недостатками большинства мер, основанных на иерархических структурах, является симметричность. Кроме этого, эти меры не зависят от контекста и зависят от качества разработки онтологии.

Среди методов оценки СБ используются методы, основанные на вычислении подобия определенных характеристик. Среди мер, учитывающих значения атрибутов, известна атрибутивная мера близости, основанная на близости значений общих атрибутов понятий. Пусть A – множество атрибутов; $A(i)$ – множество атрибутов экземпляра i; $A_{CO} = A(i_1) \cap A(i_2)$ – множество общих атрибутов экземпляров i_1 и i_2, S (i_1, i_2, a) близость двух экземпляров i_1 и i_2 в отношении одного атрибута a. Тогда атрибутивная мера близости экземпляров i_1, i_2 с учетом всех атрибутов из $A_{co}=A(i_1)\cap A(i_2)$вычисляется по формуле [6]:

$$S(i_1,i_2) = \frac{1}{A_{co}} \sum_{a \in A_{co}} S(i_1,i_2,a)\,. \tag{5}$$

Также существуют и меры, основанные на обработке лексико-семантических ресурсов и основанные на обработке неразмеченных документов.

Традиционно в информационно-поисковых системах близость текстовых документов вычисляется как угол между векторами документов, образуемыми весами ключевых слов документов по схеме TF [7], [8]:

$$TF = \frac{\text{частота слова в документе}}{\text{общее количество слов в документе}}, \tag{6}$$

$$cos(\theta) = \frac{A * B}{\|A\| * \|B\|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}, \tag{7}$$

где A_i и B_i – компоненты векторов.

Многие подходы, основанные на обработке текстовых документов, используют вычисление сходства между словами. Так, коэффициент подобия Жаккара вычисляет количество уникальных терминов, совместно используемых между двумя текстами, а коэффициент Левенштейна состоит из минимального количества необходимых операций по трансформации одной строки в другую. Семантическое сходство между двумя текстами как максимальное значение, полученное между парами слов, определяется в мерах Леска, Резника и др.

Последнее семейство мер, Дайса и Жаккара, коэффициент Танимото – это простые, нерекурсивные меры близости, которые традиционно применялись в области информационного поиска. Эти меры определяют схожесть двух множеств на основе общих символов и удобно выражаются в множественно-теоретической форме. Несомненным их плюсом является низкая вычислительная сложность.

Пусть N_a – количество элементов в первом множестве, N_b – количество элементов во втором множестве, N_c – количество общих элементов в обоих множествах. Тогда коэффициент Танимото принимает значения от 0 до 1 (чем ближе к 1, тем больше сходство между множествами) [9]:

$$T = \frac{N_c}{N_a + N_b - N_c}, \tag{8}$$

Мера Дайса:

$$D = \frac{|N_a \cap N_b|}{|N_a \cup N_b|} \tag{9}$$

Коэффициент Жаккара [8]:

$$G = \frac{|N_a \cap N_b|}{|N_a \cup N_b| - |N_a \cap N_b|} \quad (10)$$

Среди примеров простейшей гибридной меры можно привести формулу «мягкого косинуса», который кроме векторных характеристик текста учитывает еще и семантику [8]:

$$soft\_cos(\theta) = \frac{\sum_{i,j=1}^{n} S_{ij} * A_i * B_i}{\sqrt{\sum_{i,j=1}^{n} S_{ij} * A_i * A_j} * \sqrt{\sum_{i,j=1}^{n} S_{ij} * B_i * B_j}} \quad (11)$$

A_i и B_i – компоненты векторов,

S_{ij} – матрица семантических связей.

В работе [10] предложен подход, когда гибридная мера близости содержит оценку критериев подобия понятий онтологии, состоящую из трех частей: атрибутивная мера, таксономическая мера и реляционная мера.

$$S(c_1, c_2) = tS^t(c_1, c_1) + pS^p(c_1, c_2) + aS^a(c_1, c_2), \quad (12)$$

где S^t, S^p, S^a – таксономическая, реляционная и атрибутивная составляющие коэффициента близости, соответственно, t, p и a – веса составляющих близости объектов в интервале [0, 1], их сумма равна единице.

Для решения задачи нахождения весовых коэффициентов в этой работе было предложено использование генетического алгоритма, который наиболее эффективно обеспечивает поиск решения для функций, имеющих несколько экстремумов. В качестве общей структуры алгоритма использовался модифицированный генетический алгоритм.

В работе [11] предложена методика оценки СБ, отличительной особенностью которой является автоматическое определение весовых коэффициентов с использованием метода роя частиц.

Таким образом, при исследовании методов были рассмотрены разные походы: теоретико-множественный подход, базирующийся на определении пересечения одинаковых и различных свойств; информационный подход на базе статистики (частоты встречаемости терминов); структурный подход, основанный на определении различных характеристик онтологий (длина пути, глубина иерархии и др.). Очевидно, что чем полнее учитываются различные характеристики двух сущностей, тем качественнее рассчитываются меры близости и гибридные меры близости, сочетающие несколько подходов и методов, являются наиболее перспективными.

Трехмерный тензор семантических связей на базе RDF-графа онтологии

Современным подходом при построении баз знаний является использование стека семантических технологий. Стандартизованный консорциумом W3C RDF специфицирует архитектуру, синтаксис и семантику, а также базовый словарь RDF Schema (RDFS) для построения моделей предметных областей [1]. В концепции Semantic Web модель данных в виде RDF-графа представляется следующим образом: Subject – Predicate –> Object. Каждая сущность в свою очередь имеет свой универсальный и уникальный идентификатор ресурса – URI (*Uniform Resource Identifier*).

Для моделирования бинарных отношений на RDF-графе удобно использовать трёхмерный тензор [12]. Наилучшим определением тензора будет цитата *Tamara G. Kolda: «A tensor is a multidimensional array»* [13]. У тензора бывают верхние, нижние и смешанные индексы. Верхние меняют значение при переходе от одной строки к другой, а нижние при переходе от одного столбца к другому.

В работе [14] авторами было предложено тензорное представление графа знаний в виде тензорного разложения RESCAL.

Для представления многомерных семантических данных авторы работы [15] использовали формализм RDF семантической сети, где отношения моделируются как тройки (субъект, предикат, объект) и где предикат моделирует либо отношения между двумя сущностями, либо между сущностью и значением атрибута. Тензорная запись $S_{ijk} \neq 1$ обозначала тот факт, что существует RDF отношение (*i*-я сущность, *j*-й субъект, *k*-й предикат). В противном случае для несуществующих и неизвестных отношений запись устанавливается равной нулю.

В данной работе впервые предлагается использование трехмерного тензора семантических связей, значения которого определяются различным образом для отношений графа знаний и содержат значения в диапазоне [0,1].

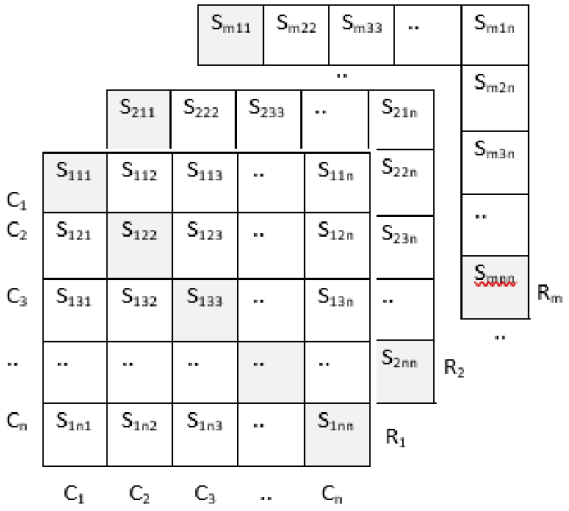

Рисунок 1 – Трехмерный тензор семантических связей

Рассмотрим меры СБ, которые были использованы в модели трехмерного тензора семантических связей. Для анализа трехмерного тензора рационально разложить трехмерный куб на отдельные матрицы, представляющие собой матрицы семантических связей для отдельно взятых отношений *Rk* (рис. 2). Количество полученных матриц соответствует числу используемых в модели мер.

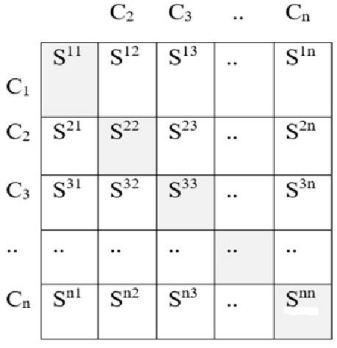

Рисунок 2 – След трехмерного тензора семантических связей S^{ij}_1 при k=1

Элемент матрицы S^{ij}_1 равен числу w, если существует ребро между вершинами C_i и C_i с весом w. Элемент S^{ij}_1 равен нулю, если рёбер между вершинами C_i и C_j не существует.

Меры для оценки СБ по иерархии онтологии

СБ двух термов может быть оценена по положению вершин в иерархической структуре данных – онтологии. Для получения иерархии классов предметной области в виде графа следует выделить скелет (каркас) онтологии. При этом ребра дерева отражают иерархические отношения типа:

1) *R1* – таксономические (IS-A, KIND-OF, has, имеет);
2) *R2* – партономические (PART-WHOLE, part-of, часть-целое);
3) *R3* – родовидные (PARENT-CHIEILD, TOPIC–SUBTOPIC, родитель-ребенок).

Зафиксировав индекс k=1 в трехмерном тензоре семантических связей, мы получили след матрицы S^{ij}_1 – матрицу семантических связей для первого отношения *R1*. Соответственно при k=2 и k=3 мы получили подобные матрицы для отношений *R2* и *R3*. За расстояние между концептами приняли длину минимального пути между ребрами, так как требуется построить в максимальной степени сбалансированное дерево. При расчете использовали путь от ближайшего общего предка обоих термов и лишь в случае отсутствия идем от корня.

Оценка семантических отношений, выраженных в OWL свойствами концептов

Важным свойством графовых моделей данных является возможность возникновения различных корреляций между множеством взаимосвязанных узлов. Подобные корреляции могут быть вычислены за счёт включения обработки атрибутов, связей и классов связанных сущностей. Мера СБ показывает высокие значения для концептов, которые находятся в семантических отношениях (синоним, гипоним, ассоциативность) и нулевые значения для всех остальных пар.

Свойства OWL типа *Data Propery* определяют связь между объектами и данными. Выделим отдельно отношение синонимизации ввиду его особой семантической важности. Зафиксировав индекс k=4, мы получим матрицу семантических связей для отношения R4 – синонимизации (эквивалентности, подобия). Если концепты связаны, то СБ=1, иначе 0.

Свойство типа *Object Property* определяет отношения между индивидуальными объектами. Зафиксировав индекс k=5, мы получим матрицу семантических связей для отношения R5 – *Individal.* Если два концепта связаны отношением *Object Property*, то их мера связанности также равна 1.

Между классами и экземплярами различных классов наряду с иерархическими связями могут быть и другие «горизонтальные» семантические связи, представляемые объектными бинарными отношениями, характерными для описываемой предметной области. Зафиксировав индекс k=6, мы получим матрицу семантических связей для отношения R6 – свойства ассоциации (семантических отношений). Поскольку данная матрица не является матрицей смежности, то с помощью эксперта рассчитываются значения всех семантически связанных ячеек, остальные обнуляются.

Рассмотрим процедуру определения СБ на примере фрагмента онтологии “Документ”, которая описывает различные виды документов в деятельности организаций. Часть документов, например, научные статьи, имеют достаточно шаблонную

структуру (если соответствуют ГОСТ) и содержат заголовок, аннотацию, список ключевых слов, тему документа, выводы. Эти элементы присутствуют почти во всех документах, несмотря на различные требования к оформлению научных статей. Логично предположить, что наиболее важная информация заключается в заголовке, аннотации и списке ключевых слов и эти термины должны иметь более высокий вес по сравнению с другими понятиями, встречающимися в основном тексте. Соответственно, веса могут быть откорректированы экспертным путем в любой момент времени.

Поскольку в OWL стандарте предусмотрено хранение свойств-атрибутов, то целесообразно будет и вычисление атрибутивной меры близости концептов. Таким образом, зафиксировав индекс k=7, мы получим матрицу семантических связей для отношения R7 – атрибутивную меру сходства.

Меры СБ, основанные на SVM и контекстных множествах

При решении задач поиска используются векторные меры сходства при решении некоторых задач, особенно когда онтология только наполняется данными или не имеет концептов по определённой теме. Сущности графа знаний могут быть эффективно представлены векторами их латентных свойств. Данные свойства называют латентными, т.к. они напрямую не описаны в данных, но могут быть выведены из имеющихся данных в процессе математической обработки. В работе [14] предложена модель, представляющая тройки посредством парного взаимодействия латентных свойств.

Следующей группой мер являются меры, построенные на предположении, что семантически близкие концепты встречаются в тексте в одинаковых контекстах, т.е. обладают похожим набором ключевых (контекстных) слов, который называется контекстным множеством. Для определения меры сходства между двумя документами можно использовать меры на базе косинусного сходства (см. форм. 7). Одна из причин популярности этой меры состоит в том, что она эффективна в качестве оценочной меры, особенно для разреженных векторов, когда необходимо учитывать только ненулевые измерения. Мы использовали «мягкую» косинусную меру сходства между двумя векторами. При расчете «мягкой» косинусной меры используется матрица S сходства между признаками. При отсутствии сходства между признаками (S_{ij} = 0 для $i \neq j$), данное уравнение эквивалентно общепринятой формуле косинусного сходства (см. формулу 11).

Следовательно, зафиксировав индекс k=8, мы получим матрицу семантических связей для отношения R8, рассчитанную по векторной мере сходства.

Таким образом, получаем таблицу используемых в гибридной интеллектуальной мере мер СБ.

Таблица 1 – Виды оценок СБ, используемых в гибридной модели

№	Описание	Вид оценки	Способ определения
1	R1 – таксономические R2 – партономические R3 – родовидовые	по иерархии онтологии	Длина кратчайшего пути
2	R4 – синонимы, R5 – individual	OWL свойства концептов	Длина кратчайшего пути
3	R6 – атрибутивная мера близости, основанная на близости значений общих атрибутов понятий	OWL по общим атрибутам	Атрибутивная мера (см. формулу 6)

Продолж. табл. 1

4	R6 – мера близости, основанная на ассоциативных семантических связях	по горизонтальным отношениям онтологии	Заполняется экспертом значениями в диапазоне [0;1]
5	R7 – мера Жаккара	Мера, основанная на контекстных множествах	Коэффициент Жаккара (см. формулу 10)
6	R8 – близость – угол между векторами терминов, образуемыми весами терминов множеств	Мера, основанная на SVM	TF (см. формулу 6) Косинусная мера (см. формулу 7) Мягкая косинусная мера (см. формулу 11)

Генетический алгоритм взвешивания коэффициентов гибридной модели

Тензорное представление графа семантических отношений позволяет эффективным образом вычислить оценки семантической близости между двумя концептами через факторизацию срезов тензора. Выполнив аддитивную свертку тензора $S^{ij}{}_k$ с вектором коэффициентов значимости для каждого типа отношения W^k получаем:

$$R^{ij} = \sum_{k=1}^{p} W^k S_k^{ij}, \quad (13)$$

где W^k – вес, который определяет относительную важность каждого типа отношения,
p – число отношений, R – матрица семантических связей ключевых концептов.

Для оценки весов W^k использовался генетический алгоритм. В результате сформировалось близкое к оптимальному распределение весов для конкретного набора отношений. Существует также возможность экспертного ввода весовых коэффициентов в настройках прикладной программы.

Эксперименты

С целью тестирования разработанных моделей и алгоритмов была разработана программа на языке PHP с использованием целого ряда библиотек и фреймворков. Тестирование проводилось на бюджетном ноутбуке со следующими параметрами: Процессор Intel(R) Core(TM) i3-4010U CPU @ 1.70GHz, 1700 МГц, ядер: 2, логических процессоров: 4, Установленная оперативная память (RAM) 4,00 ГБ.

Интерфейс программы приведен на рис. 3.

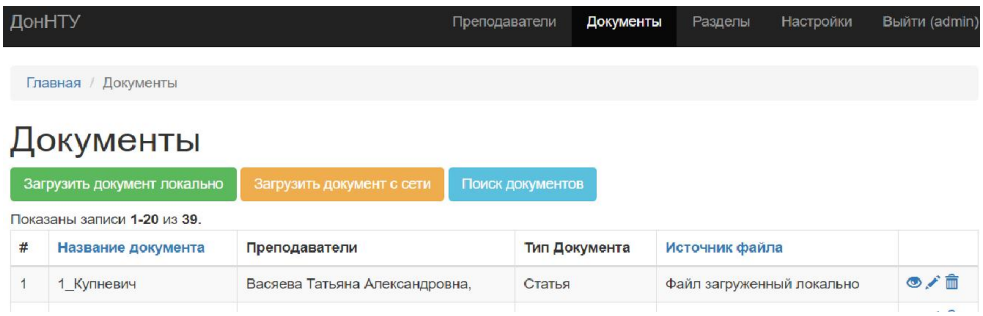

Рисунок 3 – Интерфейс программного модуля

Программа снабжена целым рядом пользовательских настроек (рис. 4).

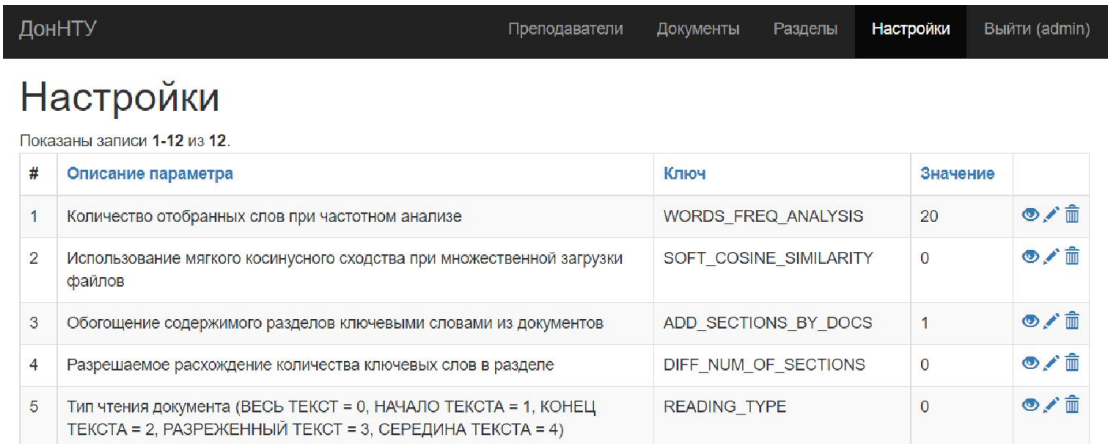

Рисунок 4 – Окно настройки параметров обработки файлов

Для определения оптимальных значений некоторых параметров были выполнены дополнительные исследования. Так, результаты выбора способа обработки документа сведены в табл. 2. Видно, что лучшим вариантом по сочетанию двух параметров – скорости обработки (t – время в секундах) и достигнутого результата (n – количество найденных ключевых слов) является разреженная обработка текста.

Таблица 2 – Выбор оптимального способа обработки файла

Документ	Начальный фрагмент		Конечный фрагмент		Средний фрагмент		Разреженный текст		Весь текст	
	t,c	n	t,c	n	t,c	n	t,c	n	t,c	n
6_Давыденко	5.54	10	6.16	10	6.56	5	6.24	14	8.42	14
1_Главацкий	10.62	10	10.99	12	9.78	7	11.4	17	13.8	17
..	..	..	..	..	..	..	..	..	..	..
1_Купневич	22,9	15	23,8	14	24	17	24	17	34	17
Итоги	132	120	136	118	134	96	138	160	188	160

Результаты экспериментов по определению оптимальной длины контекстного вектора приведены на рис. 5.

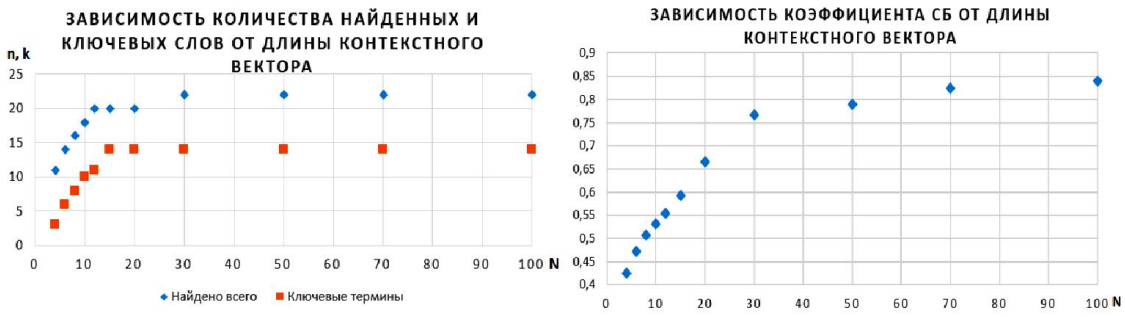

Рисунок 5 – Результаты экспериментов по определению оптимальной длины контекстного вектора

В результате обработки документа происходит извлечение полезных знаний и расчет отдельных мер семантической близости к тематическим разделам, а также взвешенной гибридной меры (рис. 6).

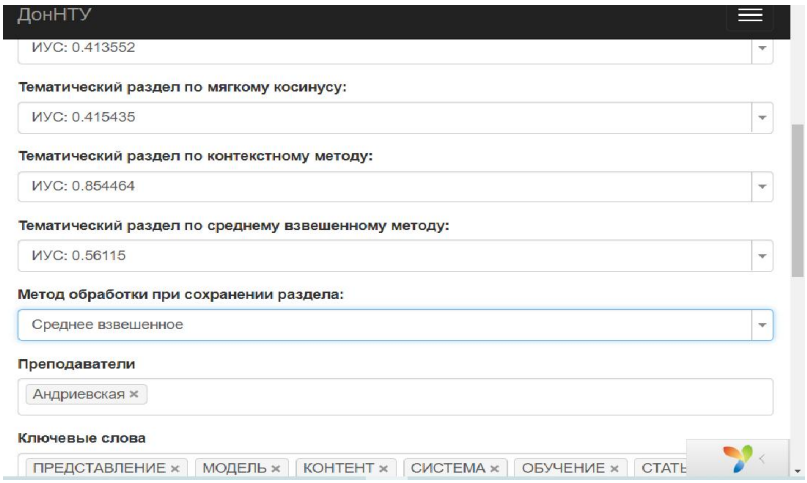

Рисунок 6 – Результаты обработки отдельного файла

В качестве базы для проведения экспериментов были использованы электронные документы – материалы конференций факультета за последние 10 лет. На выбор тестового набора повлиял тот факт, что документы в свое время уже были рецензированы и разделены на секции рецензентами-экспертами, то есть набор уже был размечен.

Результаты обработки показали, что использование гибридной взвешенной меры дает существенное преимущество в следующих случаях:

– когда онтология не наполнена терминами;
– когда документ отличается в значительной степени от других;
– когда возникают семантические неоднозначности.

Первоначальная близость тематических разделов друг к другу сильно снижает точность классификации. Несмотря на это, результаты оценки качества классификации документов с использованием гибридной взвешенной меры оценки семантической близости оказались достаточно высокими (табл. 3).

Таблица 3 – Результаты оценки работы системы

Раздел	Correct	Total	Precision (%)	AVG Precision%
ИУС (информационные системы)	237	332	71.3	72,1
КМ (Компьютерное моделирование)	118	167	70,6	
ИИ (Искусственный интеллект)	216	290	74,4	

Выводы

В результате исследований на базе тензорного представления графа знаний был получен многомерный тензор семантических связей, в котором учитываются различные меры СБ для разных типов отношений. После программной обработки документов оказалось, что для разных документов значения различных мер отличались между собой по значению, но только в отдельных случаях возникали конфликты в рекомендациях по выбору разделов. Наиболее корректные результаты были получены, как и ожидалось, для гибридной меры.

Таким образом, разработанная гибридная интеллектуальная мера оценки семантической близости концептов, полученная на базе многомерного семантического тензора, дает возможность более точно определять их сходство с учетом семантики, частотных характеристик, контекста, структуры онтологии и может быть использована в различных системах управления знаниями, а также при работе с онтологиями.

Список литературы

1. RDF - Semantic Web Standards [Электронный ресурс]: w3.org. – Режим доступа: https://www.w3.org/RDF / (дата обращения 28 декабря 2020).
2. OWL Web Ontology Language Guide [Электронный ресурс]:w3.org. – Режим доступа: https://www.w3.org / (дата обращения 28 декабря 2020).
3. Андриевская Н. К. Основные принципы и подходы при разработке системы управления профессиональными знаниями ВУЗа [Текст] / Н. К. Андриевская // Информатика и кибернетика. –2019. – № 4 (18).
4. Андриевская Н.К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций [Текст] / Н. К. Андриевская // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта». – 2020. – № 1 (18).
5. Андриевская Н.К. Разработка прикладной онтологии в системах обработки данных научных и научно-образовательных организаций [Текст] / Н.К. Андриевская // Вестник Донецкого национального университета. – Серия Г: Технические науки. – 2020. –№ 3. – С. 43–51.
6. Меры семантической близости в онтологии [Текст] / К. В. Крюкова , Л. А. Панкова, В. А. Пронина, В. С. Суховеров, Л. Б. Шипилина // Пробл. управл. – 2010. – Выпуск 5. – С. 2–14.
7. TF-IDF [Электронный ресурс]: Википедия. Свободная энциклопедия. – Режим доступа: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения 28 декабря 2020).
8. «Мягкая»_косинусная_мера [Электронный ресурс]: Википедия. Свободная энциклопедия. – Режим доступа: https://ru.wikipedia.org/wiki/Векторная_модель#«Мягкая»_косинусная_мера (дата обращения 28 декабря 2020).
9. Индекс Жаккара [Электронный ресурс]: Википедия. Свободная энциклопедия. – Режим доступа: https://wikichi.ru/wiki/Jaccard_index# Similarity_of_asymmetric_binary_attributes (дата обращения 28 декабря 2020).
10. Бова В. В. Эволюционный подход к решению задачи интеграции онтологий [Текст] / В. В. Бова, Д. В. Заруба, В. В. Курейчик // Известия ЮФУ. Технические науки. – 2015. – № 6 (167).– С. 41–56.
11. Семенова А. В. Оптимизация отображения онтологий методом роя частиц [Текст] / А. В. Семенова, В. М. Курейчик // Онтология проектирования. – 2018. – Т. 8, № 2(28). – С. 285–295.
12. Nickel M. A Three-Way Model for Collective Learning on Multi-Relational Data [Текст] / Nickel M., Tresp V., Kriegel H. P. // ICML. – 2011. – Vol. 11.
13. Kolda Tamara G. Tensor Decompositions and Applications [Текст] / Kolda Tamara G., Bader Brett W. // SIAM Rev. – 51(3). – P. 455–500.
14. Nickel, M. A review of relational machine learning for knowledge graphs [Текст] / M. Nickel et al. // Proceedings of the IEEE. – 2015. – Vol. 104, No. 1. – P. 11–33.
15. Муромцев, Д. И. Модели и методы индивидуализации электронного обучения в контексте онтологического подхода [Текст] / Д. И. Муромцев // Онтология проектирования. – 2020. – Т. 10, № 1(35). – С. 34–49.

References

1. RDF –Semantic Web Standards site: w3.org URL: https://www.w3.org/RDF (last accessed on 28 December 2020)
2. OWL Web Ontology Language Guide site: w3.org URL:https://www.w3.org/OWL (last accessed on 28 December 2020)
3. Andriyevskaya N.K. Osnovnyye printsipy i podkhody pri razrabotke sistemy upravleniya professional'nymi znaniyami VUZa [Basic principles and approaches in the development of a professional knowledge management system of a university]. *Informatika i kibernetika* [Informatics and Cybernetics], 2019, No. 4 (18).

4. Andriyevskaya N.K. Ontologicheskiy podkhod v sistemakh obrabotki dannykh nauchnykh i nauchno-obrazovatel'nykh organizatsiy [Ontological approach in data processing systems of scientific and scientific and educational organizations] *Mezhdunarodnyy retsenziruyemyy nauchno-teoreticheskiy zhurnal «Problemy iskusstvennogo intellekta»* [International peer-reviewed scientific and theoretical journal "Problems of Artificial Intelligence], 2020, No. 1 (18).
5. Andriyevskaya N.K. Razrabotka prikladnoy ontologii v sistemakh obrabotki dannykh nauchnykh i nauchno-obrazovatel'nykh organizatsiy [Development of applied ontology in data processing systems of scientific and scientific-educational organizations]. *Vestnik Donetskogo natsional'nogo universiteta* [Bulletin of Donetsk National University] Seriya D: Tekhnicheskiye nauki, 2020, No. 3, S. 43-51.
6. Kryukova K. V., L. A. Pankova, V. A. Pronina, V. S. Sukhoverov, L. B. Shipilina, Mery semanticheskoy blizosti v ontologii [Measures of semantic proximity in ontology] *Probl. upravl*. [Management problems], 2010, vypusk 5, S. 2–14
7. TF-IDF – site: Wikipedia (wikipedia.org) URL: https://ru.wikipedia.org/wiki/TF-IDF (last accessed on 28 December 2020)
8. Vector space model site: Wikipedia (wikipedia.org) URL: https://ru.wikipedia.org/wiki/vector_space_model (last accessed on 28 December 2020)
9. Jaccard index site: wikichi.ru URL: https://wikichi.ru/wiki/Jaccard_index#Similarity_of_asymmetric_binary_attributes / (last accessed on 28 December 2020)
10. Bova, V.V., Zaruba D.V., Kureychik V.V. Evolyutsionnyy podkhod k resheniyu zadachi integratsii ontologiy [An evolutionary approach to solving the problem of ontology integration]. *Izvestiya YUFU. Tekhnicheskiye nauki* [Izvestia SFedU. Technical science], No. 6 (167), 2015, S. 41-56
11. Semenova, A.V., Kureychik V.M. Optimizatsiya otobrazheniya ontologiy metodom roya chastits [Optimizing the display of ontologies by the particle swarm method] *Ontologiya proyektirovaniya* [Design Ontology], 2018, T. 8, No. 2 (28), S. 285-295.
12. Nickel M., Tresp V., Kriegel H. P. A Three-Way Model for Collective Learning on Multi-Relational Data // ICML. 2011. Vol. 11.
13. Kolda Tamara G., Bader Brett W. Tensor Decompositions and Applications // SIAM Rev., 51(3), 455–500.
14. Nickel, M. et al. A review of relational machine learning for knowledge graphs / M. Nickel et al. // Proceedings of the IEEE. – 2015. – Vol.104. No. 1. - P.11-33
15. Muromtsev, D.I. Modeli i metody individual'nogo elektronnogo obucheniya v kontekste ontologicheskogo podkhoda [Models and methods of individualization of e-learning in the context of an ontological approach] *Ontologiya proyektirovaniya* [Design Ontology], 2020, T. 10, No.1 (35)., S. 34-49.

RESUME

Natalia Andrievskaya

Hybrid intelligent measure of semantic similarity evaluation

The most frequently solved problem is the problem of finding semantically similar objects due to the transition to modern and semantic information processing technologies. Obviously, to obtain a qualitative assessment of the semantic similarity between two objects, it is necessary to use hybrid measures.

At the moment, many different measures have been developed to determine the semantic similarity between concepts. However, there was a need to develop a measure that would use different measures in the calculation: based on ontology, using semantics, frequency characteristics of the text, calculated on the basis of context vectors.

As a result of the research, a hybrid intelligent measure was developed for evaluating the semantic similarity between two objects using the semantic three-dimensional tensor of the knowledge graph. The resulting hybrid measure includes the calculation of semantic similarity by ontology, by semantics, from Natural Language, using Term Frequency and Cosine Similarity of context vectors.

Experiments have shown that the developed hybrid measure more accurately determines the similarity of objects than the measures separately. The developed algorithm is supposed to be used in knowledge management systems of organizations, as well as in intelligent data retrieval systems.

РЕЗЮМЕ

Н. К. Андриевская

Гибридная интеллектуальная мера оценки семантической близости

В связи с переходом на современные и семантические технологии обработки информации наиболее часто решаемой задачей стала задача поиска семантически близких объектов. Очевидно, что для получения качественной оценки семантической близости между двумя объектами необходимо использовать гибридные меры. В этой статье была поставлена цель – разработать гибридный интеллектуальный способ оценки семантической близости между двумя объектами.

На текущий момент разработано много различных мер определения семантической близости между концептами. Однако возникла необходимость разработки такой меры, которая бы использовала при вычислении различные меры: базирующуюся на онтологии, использующую семантику, частотные характеристики текста, рассчитанные на базе контекстных векторов.

В результате исследований была разработана гибридная интеллектуальная мера оценки семантической близости между двумя объектами по семантическому трехмерному тензору графа знаний. Полученная гибридная мера включает вычисление нескольких мер семантической близости и затем интеллектуальное взвешивание с помощью аппарата генетического программирования.

Эксперименты показали, что разработанная гибридная мера более точно определяет сходство объектов, чем меры по отдельности. Разработанный алгоритм предполагается использовать в системах управления знаниями организаций, а также в системах интеллектуального поиска данных.

Статья поступила в редакцию 16.12.2020.