Подход к грануляции и измерению близости онтологий для интеллектуальной поддержки принятия решений в слабоструктурированных предметных областях

А. В. Чернов[1], М. А. Бутакова[2], О. В. Дейнеко, А. М. Мирошников
Ростовский государственный университет путей сообщения
[1]a.v.chernov@ieee.org, [2]butakova@rgups.ru

***Аннотация*. Слабоструктурированные предметные области являются одним из основных объектов изучения в теории искусственных интеллектуальных систем. В качестве крупнейшей интеллектуальной системы управления в слабоструктурированной области рассмотрена интеллектуальная система управления железнодорожным транспортом. Рассмотрена задача построения гранулированных онтологий для слабоструктурированной предметной области, отличающаяся от известных устройством и синхронизацией данных между гранулами. На основе выполненного анализа мер семантической близости контекстов онтологий предлагается подход к измерению степени их схожести. Предложенный подход основан на графовом представлении онтологий, является универсальным и применимым для интеллектуальной поддержки принятия решений в слабоструктурированных предметных областях.**

Ключевые слова: системы поддержки принятия решений; интеллектуальное принятие решений; слабо структурированные данные; семантическая близость онтологий

I. Введение

Информационные системы современного поколения должны оперировать, за некоторыми исключениями слабоструктурированными данными. В качестве исключений можно отнести лишь системы обработки строгой отчетности, например, экономической, налоговой и финансовой. Все системы, которые можно отнести к *Big Data* занимаются обработкой данных из разнородный источников. Примером системы, обрабатывающей весьма разнородные данные, а также служащей для принятия решений на железнодорожном транспорте является интеллектуальная система управления железнодорожным транспортом. Одной из основных подсистем интеллектуальной системы управления железнодорожным транспортом является автоматизированная система управления инфраструктурой, которая предназначена для обработки широкого класса ситуаций и инцидентов, возникающих на инфраструктуре. Авторы ранее рассматривали специфику обработки инцидентов на инфраструктуре железнодорожного транспорта в работах [1,2], а также связанные с ней понятия об информационной грануляции [3] и средствах представления и извлечения знаний в слабоструктурированных предметных областях рассматривались в работе [4].

Многозначное понятие «онтология» для искусственных интеллектуальных систем наиболее удачно определено в работе Т. Грубера [5], как формальная явная спецификация концептуализации. Под концептуализацией в данном случае понимается процесс построения модели для выбранной предметной области, а также погружение в нее с целью решения назревших задач. Таким образом, в рамках данного доклада можно ограничить весьма обширное понятие об онтологиях некоторым формальным процессом построения модели предметной области искусственных интеллектуальных систем в виде достаточно распространенного и наглядного графового представления. Конкретизируем, также, что дальнейшее рассмотрение семантического подобия онтологий, связано с семантическим расстоянием и справедливо для декларативного описания знаний. При этом необходимо иметь также прототипы понятий, имеющие смысл типичных представителей понятий и измерения выполнять путем определения расстояния конкретного экземпляра до типичного представителя.

Задача измерения расстояния между онтологиями выделяется многими исследователями, как весьма важная задача, имеющая вполне конкретные приложения, как в искусственном интеллекте, так и в общеинженерном значении. Особое значение измерениям сходства онтологий уделяется в биомедицинских информационных системах, например, в работе [6]. В этой области внимание уделяется также разработке программного обеспечения [7]. В области промышленности и производства рассматривались задачи разработки семантических мер схожести для поиска информации о производимом продукте, например, работа [8]. Для информационно-поисковых систем, расположенных в Semantic Web, задача

Работа выполнена при финансовой поддержке РФФИ, проекты № 17-07-00620-а, 18-01-00402-а, 19-01-00246-а

установления мер сходства между понятиями, терминами, включенными в базы знаний, является весьма важной. Такая задача рассматривается в работе [9]. В следующем разделе рассматриваются предварительные исследований, касающиеся, в основном, информационной грануляции онтологий, как предварительного этапа грануляции данных. В третьем разделе доклада предлагается гибридная мера семантической схожести онтологий, имеющая высокую степень универсальности своего применения. В заключении делаются выводы и указываются преимущества и недостатки предложенного подхода.

II. Предварительные исследования

Слабоструктурированные данные, как указано во введении в настоящее время составляют значительную часть информационного обмена в вычислительной инфраструктуре железнодорожного транспорта. Доменная онтология для этой системы, а также многоуровневая модель данных для нечеткой и слабоструктурированной информации была предложена в работе [10]. , заметим также, что в области практического и формализованного описания слабоструктурированной информации наибольшее использование имеют следующие форматы [4]: 1) *XML* (*eXtensible Markup Language*); 2) *RDF* (*Resource Description Framework*), 3) *OWL* (*Ontology Web Language*); 4) *OEM* (*Object Exchange Model*); 4) *JSON* (*JavaScript Object Notation*). Каждый из перечисленных форматов, предназначенных для описания слабоструктурированными данными, имеет свои преимущества и недостатки, а также область использования.

Методы информационной грануляции, которые рассматривались ранее в работах [11, 12] можно считать одними из наиболее перспективных методов, дающих достаточно полное и точное решение в условиях неопределенности поступающей информации. Хорошо известно также, что гранулярные структуры обобщают классические теоретико-множественные подходы, в том числе подходы, основанные на теории нечетких множеств. Одной из проблем информационной грануляции является невозможность динамически изменять степень гранулирования. В своих предыдущей работе [13] авторы предложили развитие информационного гранулирования на случай динамического интерактивного взаимодействия гранул, а также изменения степени грануляции посредством аппарата теории грубых множеств [1].

В следующем разделе представлен новый подход к измерению сходства онтологий.

III. Предлагаемый подход

Измерение сходства онтологий обычно понимается, определение степени похожести некоторых концептов из одной онтологии с некоторыми концептами из другой онтологии. Любая онтология, в отношении ее концептов может иметь структуру множества. Одной из самых известных мер подобия можно считать информационно-теоретическую меру, предложенную А. Тверским [14]:

$$S_{ratio}(X,Y)=\frac{f(X\cap Y)}{f(X\cap Y)+\alpha f(X-Y)+\beta f(Y-X)}. \qquad (1)$$

В мере (1) X и Y представляют собой множества, которые описывают объекты x и y соответственно. $(X\cap Y)$ представляет собой набор свойств, которые присущи одновременно и объекту x и объекту y. $(X-Y)$ описывает свойства, которые имеются у объекта x, но которых не имеет объект y, а, $(Y-X)$ в свою очередь описывает свойства, имеющиеся у объекта y, но не у объекта x. Скаляры α и β служат в (1) для обозначения увеличения, либо уменьшения степени влияния свойств объектов x и y. Заметим, что мера (1) является нормализованной $0\le S_{ratio}(X,Y)\le 1$.

Для любой онтологии важным аспектом является ее информационный контекст, в котором она рассматривается. Одной из первых известных информационно-теоретических мер, учитывающих контекстную схожесть онтологий является мера, предложенная П. Резником [15]:

$$IC(c)=-\log p(c). \qquad (2)$$

Нетрудно заметить некоторое сходство с известной формулой энтропии Хартли, а в (2) $p(c)$ является вероятностью концепта c в некоторой рассматриваемой онтологии, рассчитываемую по частотному принципу расчета вероятностей. Например, если онтология представлена в описательной текстовой форме, то рассчитывается вероятность появления слова-концепта в описании. Для более распространенного графового описания онтологий, где принимается, что концепт размещается в вершине графа, может быть рассчитана вероятность по отношению к общему числу вершин.

Не менее важным для концептов онтологий, которые представлены графовым образом является наличие дочерних вершин у концепта. В связи с эти, имеется доработанный вариант информационно-теоретической меры, учитывающий наличие потомков некоторого концепта:

$$IC_{ont}=\log\frac{(n_d(c)+1)}{\max_{ont}}/\log\frac{1}{\max_{ont}}=1-\frac{\log(n_d(c)+1)}{\log(\max_{ont})}, \qquad (3)$$

где $n_d(c)$ – количество потомком у концепта c; $\max_{ont}$ – максимальное число концептов в рассматриваемой онтологии.

Прежде чем перейти к изложению меры, предложенной автором диссертации, выполним краткий обзор имеющихся мер контекстно-зависимой семантической схожести онтологий. Это даст возможность установить отличия предлагаемой меры от ранее известных. Ввиду значительного числа ранее предложенных мер семантической близости онтологий ограничимся рассмотрением только тех мер, которые используют представление онтологии в виде ориентированного

ациклического графа и имеют информационно-теоретическое происхождение. Расстояние между двумя концептами онтологий обычно понимается как кратчайший путь на графе между двумя вершинами a,b и далее обозначается $dist(a,b)$, а глубина концепта c, также имеет сходной значение с глубиной вершины графа и обозначается $depth(c)$.

Для составления меры пользуются также свойствами самих графов, описывающих онтологии. Например, можно задать веса для дуг графа, либо принять веса дуг в графе одинаковыми или равными единице. В случае взвешенных и агрегированных по некоторому типу t дуг имеется мера, определяемая, как $S_{weight}(a,b)=\alpha_t \prod_{i=1}^{dist(c_1,c_2)} \beta_{t_i}$, где α_t – это весовой коэффициент группы дуг, относящихся к типу t; β_{t_i} – это весовой коэффициент i-той дуги, относящейся к типу t.

Следующая мера использует значение направления при прохождении пути по ориентированному графу от одного концепта a к другому концепту b. Если обозначить $turns(a,b)$ как число изменений направления, то можно использовать, например, меру, предложенную Г. Хирстом [16]:

$$S_{Hirst}(a,b)=C-dist(a,b)-k\cdot turns(a,b),$$

где C и k – это некоторые константы, причем у Г. Хирста $C=8$, $k=1$.

При определении мер, которые основаны на (3) часто рассматривается числовая величина, называемая *Least Common Subsumer* (LCS). Величина *LCS* определяется как минимальная дистанция до ближайшего родительского концепта, который является одновременно родителем концепта a и концепта b. Используя $LCS(a,b)$ определяется мера $S_{LCS}(a,b)=IC(LCS(a,b))$, а также нормализованный её вариант

$$S_{NormLCS}(a,b)=\frac{2S_{LCS}(a,b)}{(IC(a)+IC(b))}. \quad (4)$$

Далее, используя идеи мер из (3) и (4), представлена мера, которая учитывает имеющиеся подходы к агрегации слабоструктурированных данных.

Этапы расчета состоят в следующем.

Этап 1. Обозначим x,y – концепты онтологий, $IC(x)$, $IC(y)$ – рассчитываются по (3).

Этап 2. Вводим весовые коэффициенты W_t и рассчитываем значения

$$SV_x(x)=\sum_{t\in A_x} W_t IC(x);$$
$$SV_y(y)=\sum_{t\in A_y} W_t IC(y),$$

где A_x, A_y – это множество агрегированных данных, описывающих свойства концептов x,y соответственно.

Этап 3. Рассчитываем нормализованные функции подобия для концептов x и y, аналогично выражению (4):

$$C_a(x,y)=\frac{2\sum_{a\in(A_x\cap A_y)} IC(a)}{SV_x(x)+SV_y(y)};$$
$$C_b(x,y)=\frac{2\sum_{b\in(A_x\cap A_y)} IC(b)}{SV_x(x)+SV_y(y)}.$$

Этап 4. Рассчитываем нормализованную меру контекстной семантической схожести онтологий:

$$S_{sim}=\frac{1}{|A_x|\times|A_y|}\sum_{t_1\in A_x}\sum_{t_2\in A_y} CC(t_1,t_2), \quad (5)$$

где $CC(t_1,t_2)=C_a(x,y)\times C_b(x,y)$.

Меру (5) можно использовать для принятия решений о близости онтологий в различных системах интеллектуальной поддержки принятия решений в случае, если онтология построена для слабоструктурированной предметной области.

IV. Заключение

В докладе рассмотрен подход к измерению близости онтологий для принятия интеллектуальной поддержки принятия решений. Предварительные исследования проводились для обсуждения возможности интерактивной грануляции. Можно сказать, что в нашем случае интерактивная грануляция выполняется для отсечения решений на некотором заданном уровне при неких обстоятельствах. В дальнейшем это ускоряет процессы принятия решений. Центральный подход к измерению близости основан на достаточно универсальной мере схожести онтологий для графового их представления. Преимуществами данной меры является ее высокая практическая применимость, так как графовое представление слабоструктурированных данных хорошо задокументировано различными распространенными форматами и описаниями. Недостатком данной меры является использование минимальной дистанции до ближайшего родительского концепта, так как для многих графовых представлений, данная величина не является вполне информативной.

Список литературы

[1] Chernov A.V., Bogachev V.A., Karpenko E.V., Butakova M.A., Davidov Y.V. Rough and fuzzy sets approach for incident identification in railway infrastructure management system // Proceedings of 2016 19th IEEE International Conference on Soft Computing and Measurements, SCM 2016, pp. 228-230.

[2] Chernov A.V., Kartashov O.O., Butakova M.A., Karpenko E.V. Incident data preprocessing in railway control systems using a rough-set-based approach // Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017, pp. 248-251.

[3] Butakova M.A., Chernov A.V., Guda A.N., Vereskun V.D., Kartashov O.O. Knowledge representation method for intelligent situation awareness system design // Advances in Intelligent Systems and Computing, 875, 2019. pp. 225-235.

[4] Карташов О.О., Бутакова М.А., Чернов А.В., Костюков А.В., Жарков Ю.И. Средства представления знаний и извлечения данных для интеллектуального анализа ситуаций // Инженерный вестник Дона, №4, 2018.

URL: http://ivdon.ru/ru/magazine/archive/n4y2018/5421

[5] Gruber T.R. A translation approach to portable ontology specifications // Knowledge Acquisition, 5(2), 1993. Pp 199-220.

[6] Cross V. Ontological Similarity // In: M. Popescu, Dong. Xu (Eds.), Data Mining in Biomedicine Using Ontologies, Artech House, MA, Norwood, 2009, pp. 23–43.

[7] Li L.J., et. al. CellSim: a novel software to calculate cell similarity and identify their co-regulation networks // BMC Bioinformatics, 20:111, 2019.

[8] Akmal S., Shih L.H., Batres R. Ontology-based similarity for product information retrieval // Computers in Industry, 65:1, 2014. pp. 91-107.

[9] Slimani T. Description and evaluation of semantic similarity measures approaches // International Journal of Computer Applications, 80(10). pp. 25-33.

[10] Yants V.I.; Chernov A.V.; Butakova M.A.; Klimanskaya E.V.: Multilevel data storage model of fuzzy semi-structured data. In: Soft Computing and Measurements (SCM), 2015 XVIII International Conference, vol. 1, 2015., pp.112-114.

[11] Бутакова М.А., Гуда А.Н., Иванченко О.В., Карпенко Е.В. Элементы теории гранулярных вычислений с нечеткими приближенными информационными гранулами // Вестник Ростовского государственного университета путей сообщения. 2015. № 4 (60). С. 27-33.

[12] Бутакова М.А., Иванченко О.В. Методы информационного гранулирования для решения задач редукции условных атрибутов в системах поддержки принятия решений // Вестник Ростовского государственного университета путей сообщения. 2016. № 4. С. 136-144.

[13] Чернов А.В., Дейнеко О.В. Интерактивные гранулярные вычисления как инструмент принятия решений в сложных киберфизических системах // Материалы 8 международной научно-технической конференции «Технологии разработки информационных систем», 2017. С. 119-122.

[14] Tversky A., Features of Similarity // Psychological Rev. 84. 1977. pp. 327–352.

[15] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res. 11. 1999. pp. 95–130.

[16] Hirst G., St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms // In: C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 305-332.