

Література

1. Клинковштейн, Г. И. Организация дорожного движения [Текст] : учебное пособие / Г. И. Клинковштейн, В. И. Коноплянко. – М. : Транспорт, 1977. – 58 с.
2. Приборы и методики психофизиологического обследования водителей автомобилей [Текст] / Н. А. Игнатов, В. М. Мишурин, Р. Т. Мушегян, В. А. Сергеев. – М. : Транспорт, 1978. – 88 с.
3. Лобанов, Е. М. Проектирование дорог и организация движения с учетом психофизиологии водителя [Текст] / Е. М. Лобанов. – М. : Транспорт, 1980. – 311 с.
4. Баевский, Р. М. Прогнозирование состояний на грани нормы и патологии [Текст] / Р. М. Баевский. – М. : Медицина, 1979. – 298 с.
5. Доля, В. К. Пасажирські перевезення : підручник / В. К. Доля. – Харків: Видавництво «Форт», 2011. – 504 с.
6. Коноплянко, В. И. Основы управления автомобилем и безопасность движения [Текст] / В. И. Коноплянко, С. В. Рыжков, Ю. В. Воробьев. – М. : ДОСААФ, 1989. – 224 с.
7. Клеббельсберг, Д. Транспортная психология [Текст] : пер. с нем. / под. ред. В. Б. Мазуркевича. – М. : Транспорт, 1989. – 367 с.

Розглядається задача усунення різноманітності онтологій. Дана загальна характеристика основних підходів до співставлення онтологій. Представлено оптимальний за швидкодією алгоритм автоматичного співставлення онтологій, заснований на принципах навчання штучної нейронної мережі

Ключові слова: онтологія, співставлення, адаптивне машинне навчання

Рассматривается задача устранения разнородности онтологий. Дана общая характеристика основных подходов к сопоставлению онтологий. Представлен оптимальный по быстродействию алгоритм автоматического согласования онтологий, основанный на принципах обучения искусственной нейронной сети

Ключевые слова: онтология, сопоставление, адаптивное машинное обучение

The problem of elimination of heterogeneity among different ontologies is considered. General description of the main ontology matching approaches is given. The optimal by performance algorithm for automatic ontology matching, based on principles of artificial neural network learning, is presented

Key words: ontology, matching, adaptive machine learning

УДК 519.68:007.5

ОБ ОДНОМ ПОДХОДЕ К СОПОСТАВЛЕНИЮ ОНТОЛОГИЙ НА ОСНОВЕ АДАПТИВНОГО МАШИННОГО ОБУЧЕНИЯ

Е. В. Бодянский

Доктор технических наук, профессор*
E-mail: natavoloshina@gmail.com

Н. А. Волошина

Старший преподаватель*
Контактный тел.: 095-56-46-203
E-mail: natavoloshina@gmail.com

Н. В. Рябова

Кандидат технических наук, доцент, заведующий
кафедрой*

*кафедра Искусственного интеллекта
Харьковский национальный университет
радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
Контактный тел.: (057)70-21-337, 050-958-05-72
E-mail: ryabova.nv@gmail.com

1. Введение

Обеспечивая формальную и декларативную модель представления знаний, онтологии приобретают все большую значимость при создании Web-систем различного назначения. Они формируют семантическую основу для многих областей приложений, таких как Web-сервисы, электронная коммерция, Semantic Web, где программные системы должны совместно использовать и понимать информационный контент друг друга. Тем не менее, поскольку онтологии отражают концептуальный взгляд их разработчиков на «модель мира», неотъемлемой их характеристикой является неоднородность. Таким образом, проблема сопоставления и согласования онтологий (ontology matching), т.е. нахождение семантического соответствия между элементами двух независимо разработанных онтологий, является весьма актуальной как в развитии новых направлений онтологического инжиниринга, так и в практических разработках интеллектуальных Web-систем.

Для управления проблемой разнородности в онтологиях предлагались различные подходы. Однако все они в общем могут быть отнесены к одному из двух классов решений. Первый – централизованное решение, при котором основная онтология уже построена. Такая онтология должна быть согласованной, глобальной и единой, содержащей минимально необходимое множество концептов, удовлетворяющее потребностям всех пользователей. Типичным примером такого централизованного подхода является широко известная онтология верхнего уровня (Top Level) СУС. Недостатки данного класса решений очевидны. Централизованная онтология не может быть настолько большой и совместимой, чтобы включить все концепты, представляющие интересы каждого коллектива разработчиков онтологий. Если же допустить возможности для модифицирования и расширения такой онтологии, то весьма вероятно, что, каждое новое ее расширение нарушит свойства однородности и совместимости.

Таким образом, более применимым на практике является второй класс решений, распределенный. Основным допущением здесь является возможность сопоставления и согласования независимо разработанных онтологий друг с другом, и возможности их совместного использования. Сопоставление онтологий изначально выполнялось специалистами-онтологами вручную. Однако, данный подход требует больших временных затрат, привлечения высококвалифицированных специалистов, является весьма трудоемким, дорогостоящим, предрасположенным к ошибкам в силу человеческого фактора, и не масштабируемым. Следовательно, крайне необходимым является разработка автоматического или полуавтоматического инструментария для сопоставления онтологий.

Большинство известных технологий сопоставления и согласования онтологий либо базируются на правилах (rule-based), либо на методах машинного обучения (learning-based). Оба подхода имеют недостатки, например, игнорирование разработчиками информации, содержащейся в экземплярах классов онтологий. Более серьезным недостатком является то, каким образом эти технологии трактуют семантические составляющие онтологий во время процесса

сопоставления. В общем случае онтологии характеризуются такими семантическими составляющими как имена концептов, свойства концептов и отношения на множестве концептов. Эти составляющие вносят различный вклад в понимание онтологической семантики. Представляется целесообразным назначить (или установить каким-либо образом) различные веса этим семантическим составляющим онтологии для получения более точного и значимого результата сопоставления. Для определения этих весов можно использовать, например, знания экспертов и/или предварительные знания о предметной области (ПрО). Основными проблемами алгоритмов сопоставления на основе обучения являются относительно продолжительное время их выполнения и сложность получения выборок данных достаточного объема и качества.

В данной работе рассматривается подход к сопоставлению и согласованию онтологий, комбинирующий в себе как алгоритмы, основанные на правилах, так и основанные на обучении. Предлагаемый подход интегрирует нейросетевую технологию в разработанный алгоритм сопоставления онтологий. Таким образом, веса семантических составляющих онтологии, упоминавшиеся выше, могут быть получены путем обучения искусственной нейронной сети (ИНС) в отличие от определенных заранее экспертами.

2. Общее описание предлагаемого подхода к сопоставлению онтологий

В работах [1,2] предложен новый комбинированный подход к согласованию онтологий, основанный одновременно и на правилах, и на машинном обучении. Использование нейросетевых алгоритмов обучения позволяет автоматически определять веса семантических составляющих онтологии вместо их априорного задания экспертом в соответствующей ПрО. Основным преимуществом такого подхода является минимизация элемента субъективизма, вносимого человеком. Авторы [1,2] постулируют, что семантика онтологического концепта формируется тремя его составляющими: именем, свойствами, и системой отношений с другими концептами. Именно эти три основных признака лежат в основе математической модели любого концепта. Тогда для решения задачи сопоставления концептов двух онтологий каждый концепт в результате может быть описан (3×1) -вектором $\vec{S} = (S_1, S_2, S_3)^T$, который описывает соответственно

степень схожести имен концептов S_1 , их свойств S_2 и отношений S_3 .

Заметим, что значения величины S_1 (посимвольного совпадения имен концептов) лежат в диапазоне $[0, 1]$. Понятно также, что сравниваемые имена концептов (строки символов) проходят этап предобработки, а именно, удаление дефисов, символов подчеркивания, при необходимости преобразование имени существительного (имени концепта) из множественного числа в единственное. Если имена сравниваемых концептов посимвольно совпадают или являются синонимами, то S_1 принимает значение 1, в противном случае вычисляется по формуле

$$S_1 = 1 - \frac{d}{l} \quad 0 \leq S_1 \leq 1,$$

где d - редактируемое расстояние между именами концептов, (количество несовпавших символов в двух строках), l - количество символов более длинной строки в сравниваемой паре.

Свойства сравниваемых концептов задаются двумя множествами P_1 и P_2 соответственно. Мера схожести свойств двух сравниваемых концептов S_2 также находится в диапазоне $[0, 1]$ и определяется по формуле

$$S_2 = \frac{n}{m} \quad 0 \leq S_2 \leq 1,$$

где n - количество согласованных (сопоставленных) пар свойств, m - мощность меньшего из множеств P_1 и P_2 . Заметим также, что для пары согласованных свойств их типы данных должны быть одинаковы или сопоставимы, мера схожести их имен, определяемая аналогично S_1 , должна быть выше установленного порогового значения.

В данной работе, как и в [1,2], при решении задачи сопоставления онтологий анализируются только иерархические концептуальные отношения суперкласс/подкласс (`superClassOf/subClassOf`), принимая во внимание, что в дальнейшем типы сравниваемых онтологических отношений будут расширены. Кроме того, вводит ограничение, что все анализируемые онтологии, имеют суперконцепт верхнего уровня «Thing» («Сущность»). Однако, даже при таком ограниченном подходе задача не является тривиальной. Не редкой является ситуация, когда для пары эквивалентных концептов двух сравниваемых онтологий количество промежуточных концептов (между Thing и анализируемым концептом) будет различно. Это означает, что одна из онтологий разработана более детально. Чтобы избежать потери информации, рассматривается не только прямой «родитель» концепта, но и все его «предшественники», т.е. концепты, лежащие на пути от данного концепта вверх до вершины Thing. «Потомки» (прямые и не прямые дочерние вершины для данного концепта) при этом не принимаются во внимание. Для двух множеств «предков» сравниваемых концептов a_1 и a_2 мера их схожести S_3 находится в интервале $[0,1]$ и определяется следующим образом. Сначала вычисляется мера подобия для пар концептов (первый из множества a_1 , второй - из a_2 , рассматривая все комбинации), затем величине S_3 приписывается максимальное значение. Заметим, что это рекурсивная, но конечная процедура, поскольку число рассматриваемых концептов конечно, онтологии просматриваются снизу-вверх и имеют одинаковую корневую вершину Thing.

После расчета величин S_1, S_2, S_3 , связывающих два концепта C_1 и C_2 , общая мера подобия может быть рассчитана в виде взвешенной линейной комбинации

$$S = \sum_{i=1}^3 w_i s_i = \vec{w}^T \vec{s} \quad (1)$$

при ограничении

$$\sum_{i=1}^3 w_i = \vec{w}^T \vec{I}_3 = 1, \quad \vec{I}_3 = (1,1,1)^T \quad (2)$$

Исходной информацией для обучения согласованию двух онтологий O_1 и O_2 является $(n_1 \times n_2)$ - матрица M , i, j -тый элемент которой описывает меру подобия i -го концепта в O_1 и j -го концепта в O_2 .

В [1,2] нахождение вектора параметров \vec{w} рассматривается как процесс обучения трехходового адаптивного линейного ассоциатора (адалина) [3] с помощью рекуррентной градиентной процедуры с постоянным параметром шага. Целью настоящей работы является синтез адаптивных алгоритмов обучения с улучшенными динамическими свойствами, позволяющих находить эффективные решения в условиях малых обучающих выборок.

3. Оптимальный по быстродействию алгоритм обучения

Процесс обучения рассматривается как оптимизация стандартного квадратичного критерия

$$E(\hat{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 \quad (3)$$

где D - множество обучающих примеров; t_d - внешний обучающий сигнал для d -го примера; выходной сигнал адальны для d -го входного образа \vec{s}_d вычисляется по формуле

$$O_d = \hat{w}_{d-1}^T \vec{s}_d \quad (4)$$

\hat{w}_{d-1}^T - оценка искомого вектора \vec{w} , полученная по предыдущим $d-1$ примерам. С целью содержательной интерпретации полученных результатов, в [1] предлагается вместо единого обучающего сигнала t_d , использовать строки t_r и столбцы t_c матрицы M . При этом вместо критерия (3) естественно использовать целевую функцию

$$E(\hat{w}) = \frac{1}{2} \sum_{d \in D} ((t_{rd} - o_d)^2 + (t_{cd} - o_d)^2) \quad (5)$$

градиентный алгоритм оптимизации которой может быть записан в виде

$$\hat{w}_d = \hat{w}_{d-1} + \eta_d ((t_{rd} - \hat{w}_{d-1}^T \vec{s}_d) + (t_{cd} - \hat{w}_{d-1}^T \vec{s}_d)) \vec{s}_d \quad (6)$$

где η_d - параметр шага, выбираемый из эмпирических соображений.

С целью оптимизации по скорости процесса обучения, вместо (6) целесообразно использовать либо оптимальный по быстродействию алгоритм Качмаша-Уидроу-Хоффа [3,4] в модифицированной версии:

$$\hat{w}_d = \hat{w}_{d-1} + \eta \frac{(t_{rd} - \hat{w}_{d-1}^T \vec{s}_d) + (t_{cd} - \hat{w}_{d-1}^T \vec{s}_d)}{\|\vec{s}_d\|^2} \vec{s}_d, \quad 0 \leq \eta \leq 2, \quad (7)$$

либо его аддитивную форму [5]

$$\hat{w}_d = \hat{w}_{d-1} + \frac{(t_{rd} - \hat{w}_{d-1}^T \vec{s}_d) + (t_{cd} - \hat{w}_{d-1}^T \vec{s}_d)}{\gamma + \|\vec{s}_d\|^2} \vec{s}_d, \quad (8)$$

где $0 \leq \gamma \leq 3$ - параметр регуляризации.

4. Оптимальный по быстродействию алгоритм обучения при наличии ограничений

Основным недостатком алгоритмов обучения (6) - (8) является то, что в процессе настройки не выполняется условие (2), имеющее принципиальное значение при согласовании онтологий. Преодолеть указанное затруднение можно, воспользовавшись принципом адаптивного обобщенного прогнозирования [6,7], сводящимся в данном случае к поиску седловой точки функции Лагранжа

$$L(\hat{w}, \lambda) = E(\hat{w}) + \lambda(\hat{w}^T I_3 - 1) = \frac{1}{2} \sum_{d \in D} (t_{rd}^2 - 2t_{rd} \hat{w}^T \bar{s}_d + \hat{w}^T \bar{s}_d \bar{s}_d^T \hat{w} + t_{cd}^2 - 2t_{cd} \hat{w}^T \bar{s}_d + \hat{w}^T \bar{s}_d \bar{s}_d^T \hat{w}) + \lambda(\hat{w}^T I_3 - 1) \tag{9}$$

где λ – неопределенный множитель Лагранжа. Решение системы уравнений Куна-Таккера

$$\begin{cases} \nabla_w L(\hat{w}, \lambda) = -\sum_{d \in D} t_{rd} \bar{s}_d + \left(\sum_{d \in D} \bar{s}_d \bar{s}_d^T \right) \hat{w} - \sum_{d \in D} t_{cd} \bar{s}_d + \left(\sum_{d \in D} \bar{s}_d \bar{s}_d^T \right) \hat{w} + \lambda I_3 = 0 \\ \frac{\partial L(\hat{w}, \lambda)}{\partial \lambda} = \hat{w}^T I_3 - 1 = 0, \end{cases} \tag{10}$$

или

$$\begin{cases} -q_r + Q_d \hat{w} - q_c + Q_d \hat{w} + \lambda I_3 = 0, \\ \hat{w}^T I_3 - 1 = 0 \end{cases} \tag{11}$$

(здесь $q_r = \sum_{d \in D} t_{rd} \bar{s}_d$, $q_c = \sum_{d \in D} t_{cd} \bar{s}_d$, $Q_d = \sum_{d \in D} \bar{s}_d \bar{s}_d^T$) ведет к оценке

$$\begin{aligned} \hat{w} &= 0,5 Q_d^{-1} \left(q_r + q_c - \frac{0,5 I_3^T Q_d^{-1} (q_r + q_c) - 1}{0,5 I_3^T Q_d^{-1} I_3} I_3 \right) = \\ &= w^* - Q_d^{-1} \frac{0,5 I_3^T Q_d^{-1} (q_r + q_c) - 1}{I_3^T Q_d^{-1} I_3} I_3 = w^* - Q_d^{-1} \frac{I_3^T w^* - 1}{I_3^T Q_d^{-1} I_3} I_3 \end{aligned} \tag{12}$$

где

$$w^* = \frac{Q_d^{-1} q_r + Q_d^{-1} q_c}{2} \tag{13}$$

оценка, полученная с помощью стандартного метода наименьших квадратов.

С целью получения адаптивного алгоритма обучения, воспользуемся градиентной оптимизацией Лагранжа (9) на основе процедуры Эрроу-Гурвица:

$$\begin{cases} \hat{w}_d = \hat{w}_{d-1} - \eta_d \nabla_w L(\hat{w}, \lambda), \\ \lambda_d = \lambda_{d-1} + \eta_{\lambda d} \frac{\partial L(\hat{w}, \lambda)}{\partial \lambda}, \end{cases} \tag{14}$$

или, что то же самое,

$$\begin{cases} \hat{w}_d = \hat{w}_{d-1} + \eta_d (((t_{rd} - \hat{w}_{d-1}^T \bar{s}_d) + (t_{cd} - \hat{w}_{d-1}^T \bar{s}_d)) \bar{s}_d - \lambda_{d-1} I_3), \\ \lambda_d = \lambda_{d-1} + \eta_{\lambda d} (\hat{w}_d^T I_3 - 1), \end{cases} \tag{15}$$

где $\eta_d, \eta_{\lambda d}$ - скалярные параметры шага поиска.

Используя далее технику оптимизации по быстродействию алгоритмов обучения [8], приходим к окончательному результату

$$\begin{cases} \hat{w}_d = \hat{w}_{d-1} + \frac{((t_{rd} - \hat{w}_{d-1}^T \bar{s}_d) + (t_{cd} - \hat{w}_{d-1}^T \bar{s}_d)) \bar{s}_d - \lambda_{d-1} I_3}{\eta_{wd}}, \\ \eta_{wd} = \alpha \eta_{wd-1} + \|\bar{s}_d\|^2, 0 \leq \alpha \leq 1, \\ \lambda_d = \lambda_{d-1} + \eta_{\lambda d} (\hat{w}_d^T I_3 - 1). \end{cases} \tag{16}$$

Адаптивный алгоритм обучения согласованию онтологий, обладая высоким быстродействием и дополнительными сглаживающими свойствами, обеспечивает устойчивое поддержание ограничений, порождаемых сущностью рассматриваемой проблемы.

5. Выводы

Предложена процедура машинного адаптивного обучения согласованию онтологий на основе нейросетевого подхода. Введенный алгоритм обучения характеризуется вычислительной простотой, высоким быстродействием, возможностью устойчивой работы в нестационарных условиях.

Литература

1. Huang J. Ontology matching using an artificial neural network to learn weights / J. Huang, J. Dang, J.M. Vidal, M.N. Huhns // Proc. IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition (SWECKA-07). – Hyderabad, India, January 2007. – P. 80-85.
2. Huang J. Use artificial neural network to align biological ontologies / J. Huang, J. Dang, M.N. Huhns, W.J. Zheng // BMC Genomics. – 2008. – Vol. 9 (Suppl. 2): S16.
3. Widrow B. 30 years of adaptive neural networks: perceptron, adaline and backpropagation / B. Widrow, M. Lee // Proc. IEEE. – 1990. – 78. – P. 1415-1442.
4. Kaczmarz S. Approximate solution of systems of linear equations / S. Kaczmarz // Intern. J. Control. – 1993. – Vol. 57, No 6. – P. 1269-1271.
5. Райбман Н.С. Построение моделей процессов производства [Текст] / Н.С. Райбман, В.М. Чадаев. – М.: Энергия, 1975. – 376 с.
6. Бодянский Е.В. Адаптивное обобщенное прогнозирование многомерных случайных последовательностей [Текст] / Е.В. Бодянский, И.П. Плисс, Т.В. Соловьева // Доклады АН УССР. – 1989. – А., № 9. – С.73-75.
7. Bodyanskiy Ye. Adaptive generalized forecasting of multivariate stochastic signals / Ye. Bodyanskiy, I. Pliss // Proc. of Latvian Sign. Process. Int. Conf. – Riga, 1990. – Vol.2. – P. 80-83.
8. Bodyanskiy Ye. An adaptive learning algorithm for a neuro-fuzzy network / Ye. Bodyanskiy, V. Kolodyazhnyi, A. Stephan // Ed. by B. Reusch. Computational Intelligence. Theory and Application. – Berlin-Heidelberg-New York: Springer, 2001. – P. 68-75.