OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.89:004.4

АГЕНТНЫЙ ПОДХОД К РАЗРЕШЕНИЮ НЕ-ФАКТОРОВ В ЗАДАЧЕ СЕМАНТИЧЕСКОГО СРАВНЕНИЯ ИМЕН СУЩНОСТИ ПРИКЛАДНОГО ПРОГРАММНОГО РЕШЕНИЯ

Бердник В.Л., Заболеева-Зотова А.В.

Волгоградский государственный технический университет,
г. Волгоград, Россия

bwlg@inbox.ru

zabzot@gmail.com

В работе рассматривается вопросы автоматизации семантического сравнения символьных имен на эквивалентность. При эксплуатации широкого класса программ, как простейших учетных систем, так и систем управления жизненным циклом изделия (PLM), остается актуальной задача поддержания перечней материалов, полуфабрикатов и т.п. в актуальном, внутренне непротиворечивом состоянии, а также, нахождение их отражения в прейскурантах поставщиков для автоматизации деятельности служб снабжения.
Ключевые слова: моделирование принципа интуиции; многоагентные системы; семантическое сравнение символьных имен; НЕ-факторы.

Введение

Под сравнением символьных имен в статье будем понимать определение эквивалентности денотатов (кореферентности) имен для последующего прикладного использования в различных системах автоматизации производства, торговли и учета. Пусть S – множество строк символов (символьных названий), D – множество документов d. Каждый документ:

$$d = \langle e, S^d \rangle, \quad (1)$$

где e – денотат, соответствующий сущности реального мира (товар, физическое лицо, услуга, подразделение организации и т.п.);

S^d – подмножество строк S, для которых известен денотат *e*.

Таким образом

$$S = S^i \cup \left(\bigcup S^d \right), \quad (2)$$

где S^i – подмножество строк S, для которых денотат не определен.

Пример символьного имени:

«Растворитель технический 646 0,5л»

Определение денотата *e* для строк S^i является практически востребованной задачей в системах автоматизации учета предприятий. Разработанные авторами модели и методы позволяют существенно снизить затраты предприятий в долгосрочном плане. В неавтоматизированном виде решение задачи (ввод в систему приходных накладных, заказов покупателей, созданию прайс-листов, поддержания в актуальном состоянии электронных справочников и т.п.) требует существенных затрат труда оператора ЭВМ. Практически всегда имеется возможность получения электронной версии вводимого документа по электронной почте или сканированием печатной формы, но последующий автоматический ввод в систему учета предприятия не возможен в силу разных способов именования объектов учета (разных строк S^i денотата *e)*.

Рассмотрим пример НЕ-фактора в задаче семантического сравнения символьных названий. Пусть имеется некий документ описывающий товар «Материнская плата 3PE-A». В какой-то момент времени производитель материнской платы модифицирует изделие и появляется в продаже «Материнская плата 3PE-A Green». На уровне модели знаний программной системы неизвестно, какую семантику несет терм «Green» так как этот терм встретился впервые. Возможно это добавление новой подсистемы энергосбережения в изделие, или другой цвет окраски корпуса. Фактор не полноты знаний приводит к нечеткости отношения между

символьным обозначением и денотатом документа. Более подробный анализ вопросов семантического сравнения был представлен в работе [Бердник, 2012a].

Разрешение очередного НЕ-фактора это добавление нового знания о терме, его семантике, семантических взаимосвязях. Каждый новый терм несет новое семантическое значение, которое не выводимо на основе правил из семантики БЗ. Учитывая рассуждения, приведенные выше, применение какого – либо логического аппарата представляется не эффективным.

Как известно, интуиция — непосредственное постижение истины без логического анализа, основанное на воображении, эмпатии и предшествующем опыте. В нашей задаче, интуиция – это совокупность гипотез, построенных на системе закономерностей достоверных знаний системы, а также взаимосвязей внутри текущего фактически существующего комплекса НЕ-факторов.

Интуиция оперирует латентными (скрытыми) знаниями и ассоциациями. Как известно [Чанышев. 2011], слова одного предложения ассоциативно связаны. В нашем случае в качестве предложения выступает строка символьного обозначения, совокупность термов отражает закономерность строения изделия или маркетинга. В символьных обозначениях термы имеет скрытые ассоциации между собой, которые могут быть экстраполирована на новые символьные обозначения. Авторам представляется наиболее удачным моделирование интуиции описанным ранее методом "семантического пятна" [Бердник, 2012b].

Критерии эффективности сравнения

Методы семантического сравнения, основаны на разбиении множества D на подмножества D^r на основе род-видовых отношений таким образом, что каждый род сущностей $r \in R$ отличается присущим только ему набором типов существенных признаков. Для каждого типа существенного признака (например, цвет) задается семантическое поле — совокупность семантических признаков взаимно отделяющие виды сущности друг от друга в пределах рода [Кобозева, 2007] (например, белый, черный, серебристый). Соотнесение строки s^i и рода сущностей r является простой задачей, так как признак рода в строке указывается в явном виде.

Таким образом, задача семантического сравнения включает в себя уточнение цепочки НЕ-факторов [Нариньяни, 1994] за счет сокращения не полноты информации о семантике термов, к разрешению всего комплекса НЕ-факторов, например, между «Материнская плата 3PE-A» и «Материнская плата 3PE-A Green», и выбора сингулярного значения оператором системы.

Эффективность автоматизации семантического сравнения зависит от качества базы знаний, ее полноты, непротиворечивости, нечеткости и т.д. Количественным показателем эффективности автоматизации является значение обратно зависимое от мощности множества D^i.

$$D^i(s^i) = \{d \mid \mu(d, s^i) > 0\}, \quad \forall s^i \in S^i \qquad (3)$$

Ключевыми источниками знаний в нашей задаче являются операторы ЭВМ. Помимо вопросов к оператору, гипотезы для НЕ-факторов можно извлекать из структуры коллекции документов D, из корпусов текстов сети Интернет и т.п. Эффективность автоматизации задачи тем выше чем меньше количество вопросов от системы к оператора. Количество вопросов к оператору ЭВМ должно быть значительно меньше количества вопросов при непосредственном сравнении символьных обозначений за длительный период времени эксплуатации системы.

Требуется построить такую систему автоматизации, которая при ограниченном объеме получения знаний от пользователя ЭВМ обеспечивала фиксацию не менее заданного количества пар $<s^i,d>$ в одном сеансе взаимодействия с пользователем, и минимизировала среднюю за длительный период времени ее эксплуатации мощность множества:

$$D^i(s^j) = \{d \mid \mu(d, s^j) > 0\}, \quad \forall s^j \in S^j, \quad (4)$$

где S^j – наиболее востребованные в будущем операторами ЭВМ символьные обозначения. Под $D^i(s^j)$ будем понимать активные в производственном процессе (например, в торговле - список продаваемых товаров) документы $\forall d \in D^i$ в ограниченный период времени предшествующий текущему, (например, 3 месяца).

Агентный метод разрешения НЕ-факторов

Как правило, в беседе двух людей каждый задаваемый вопрос имеет интуитивную предпосылку. Собеседники могли бы задавать большое число вопросов, восполняя каждую мелочь своего незнания. Однако, каждая из сторон имеет своѐ виденье ситуации, это видение является гипотезой, дающей ответы на множество примитивных вопросов. В начале беседы собеседники ненавязчиво сверяют свои взгляды – идет корректировка гипотезы. Затем собеседники вопросами и ответами взаимно дополняют свое представление о предмете разговора. Такой способ получения информации продуктивнее списка простых вопросов, позволяет существенно меньше обращаться к собеседнику и более комфортный для человека. Сверенная точка зрения собеседников позволяет экстраполировать ответы на вновь возникающие вопросы. Именно такая концепция

положена в основу разрабатываемого интеллектуального интерфейса.

Как свойственно любой гипотезе, ее достоверность окончательно не известна. Целесообразно в ходе беседы иметь несколько различных гипотез, из которой выбирается более правдоподобная. Кроме того, мы имеем дело с множеством сравниваемых пар <s^i,d>. В данной постановке задачи отслеживается наличие конкуренции между множеством гипотез для каждой из сравниваемых пар <s^i ,d>. Ограничивающим ресурсом задачи является право задать вопрос пользователю ЭВМ. Средой конкуренции является лабиринт (граф) из семантических элементов (комнат) и взаимосвязями между ними (проходов). НЕ-факторы в таком пространстве являются потайными проходами и комнатами, а также ложными связями (проходами).

Столь нетрадиционная постановка задачи дана здесь для пояснения применения многоагентного подхода в разрешении НЕ-факторов при построении интеллектуального пользовательского интерфейса на основе интуиции.

Архитектура агента

Как известно, агенты — это программы акторы в проблемной области, которые имеют взаимные обязательства, определяемые в процессе диалога, ведут переговоры и координируют передачу информации. [Рассел, 2006]

В разрабатываемой многоагентной среде для каждого d^j создадим виртуальный агент -гуманоид, который в зависимости от цели может создавать реактивные агенты-животные для решения локальной задачи [Braspenning, 1997]. Полученная информация пополняет базу знаний, а агент информирует других агентов. Если агенты выполняют свою цель, то они получают дополнительное право на ошибку. Агенты являются дружественными, если разрешение НЕ-фактора одного агента с некоторой вероятностью может разрешить НЕ-фактор другого агента, даже если они созданы разными агентами-гуманоидами. Сила такой группы увеличивается в том агенте, от которого зависит разрешение всей цепочки НЕ-факторов.

Агент-гуманоид создается для каждого $d \in D^j$. При создании, ему предоставляется релевантная выборка

$$S^r(d) = \{s | \mu(d,s) > L\}, \quad \forall s \in S^i, \qquad (5)$$

где L – пороговое значение, заданное пользователем в политике интеллектуальной системы, μ – рассчитывается по методу TF-IDF. Учитывая, что $d = <e,S^d>$, в распоряжении агента имеется множество S^d – строки, для которых семантическая эквивалентность задана $\mu(d^j, s^d) = 1$, $\forall s^d \in S^d$.

Учитывая что строка есть множество термов, то представляется правдоподобным $\mu(d, s^d) = 1$, если

$$\bigcup (s^r / s^d) = \varnothing, \quad \forall s^d \in S^d, \qquad (6)$$

где S^d - строки документа d, s^r - множество термов сравниваемой строки, исходя из вырожденной синтаксической структуры символьного обозначения.

Выбор цели агента гуманоида происходит на графе методом «семантическое пятно». Граф предварительно настраивается агентами-животными. Вершинами графа являются термы, семантические поля, документы, роды сущности. Направленным дугам графа назначен вес, который рассчитывается как количество ассоциаций меду вершинами.

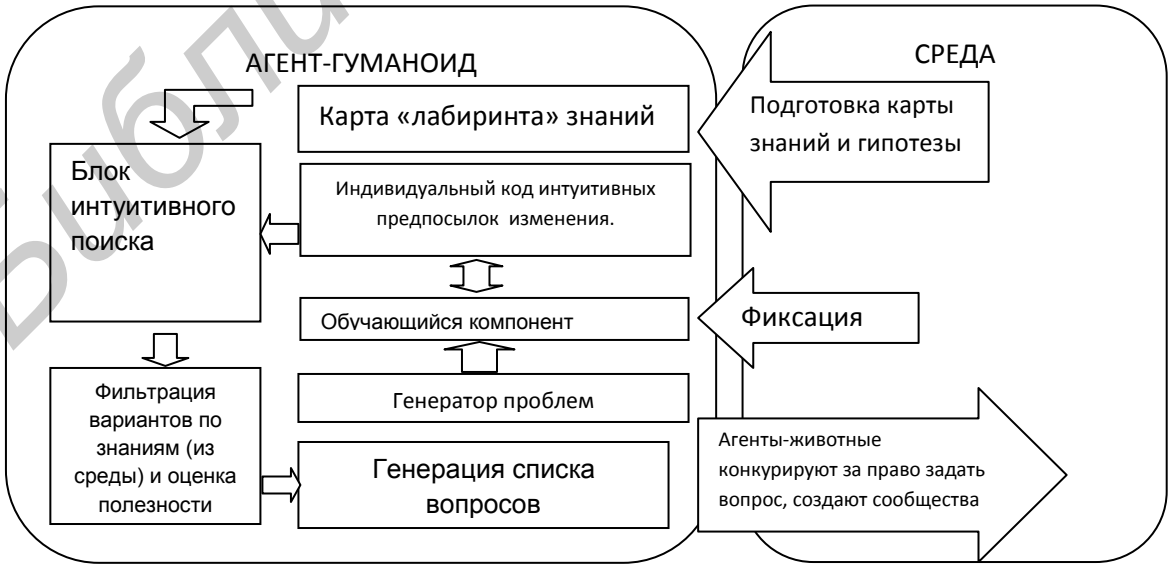

Рисунок 1 - Архитектура интеллектуального агента.

Отношение терм-терм складывается из частоты встречаемости термов в одном символьном обозначении, для терм – документ определяется как количество строк документа, содержащего терм, и т.д.

Далее веса графа нормируются, таким образом, что бы сумма весов всех исходящих вершин равнялась 1. Таким образом, вес ребра графа находится в интервале [0..1]. При анализе s^r - множество термов сравниваемой строки находится в вершинах графа. Для найденной вершины задается возбуждение = 1. Затем волновым методом возбуждение распространяется по графу. При прохождении ребра значение возбуждения умножается на вес ребра. Возбуждения, поступившие по разным маршрутам в одну вершину, складываются. Вершины, получившие наибольшее вторичное возбуждение воспринимаются агентом как гипотезы. Так, если не задана фрейм-модель, в качестве гипотезы выбирают наиболее возбужденную вершину рода сущности, для термов, семантика которых неизвестна, выдвигаются гипотезы об их семантических полях из соответствующих возбужденных вершин, относящихся к роду сущности, и т.д.

Полученные таким образом гипотезы сверяются с доской объявлений. Если какая либо гипотеза возникла у другого агента, то они обмениваются адресами. Если для гипотезы не найдена на доске объявлений аналогичная – эта гипотеза помещается на доску объявлений.

Для наиболее возбужденных вершин – документов отыскиваются агенты. Если такие агенты есть, то с ними так же происходит обмен адресов.

Каждая группа выбирает вопрос пользователю ЭВМ. Если гипотеза оказалась верной, каждый элемент группы получает дополнительный бал. Полученное знание фиксируется в БЗ. Если гипотеза неверна, очко снимается со всех агентов группы, и если у агента не было в запасе бала, агент умирает. Выжившие агенты запоминают гипотезы умерших агентов как ошибочные.

Заключение

Работа выполнена при поддержке гранта РФФИ № 13-07-00461

Библиографический список

[Бердник, 2012a] Бердник, В.Л. Семантический анализ символьных обозначений в коллекции документов: Монография/ В.Л. Бердник, А.В. Заболеева-Зотова, Ю.А. Орлова; ВолгГТУ. – Волгоград, 2012. – 124 с.

[Бердник, 2012b] Бердник В.Л.. Модель «семантическое пятно» в сложноформализуемых задачах интеллектуальной обработки информации/ В.Л. Бердник, А.В, Заболеева-Зотова// Известия ЮФУ. Технические науки.-2012.-№ 1.-С. 116-121.

[Кобозева, 2007] Кобозева И.М. Лингвистическая семантика: Учебник. Изд.3-е, стереотипное. М.:КомКнига, 2007.-352с.

[Нариньяни, 1994] Нариньяни А.С. НЕ-факторы и инженерия знаний: от наивной формализации к естественной прагматике В сб. Труды IV Национальной конференции Искусственный Интеллект94.v.1, Рыбинск 1994

[Рассел, 2006] Рассел, Стюарт, Норвиг, Питер Искусственный интеллект: современный подход, 2-е изд.: Пер. с англ.- М.: Издательский дом «Вильямс», 2006 .- 1408с.

[Чанышев, 2011] Чанышев О.Г. «Ассоциативные поля доминант и анализ текста», Материалы Всероссийской конференции с международным участием ~~–~~Знания-Онтологии-Теории" (ЗОНТ-11), 3-5 октября 2011 г.,Т.2, стр. 126-135, Новосибирск 2011.

[Braspenning, 1997] Braspenning, P. Plant-like, Animal-like and Humanoid Agents and Corresponding Multi-Agent Systems / P. Braspenning // Proceedings of the International Workshop "Distributed Artificial Intelligence and Multi-Agent Systems" (DAIMAS'97, St. Petersburg, Russia, June 15-18, 1997). - P. 64-77.

AN AGENT-BASED APPROACH FOR RESOLVING NON-FACTORS IN THE PROBLEM OF SEMANTIC COMPARISON OF ESSENCE NOTIONS IN APPLIED PROGRAM SOLUTION

Berdnik V.L., Zaboleeva-Zotova A.V.

Volgograd State Technical University

bwlg@inbox.ru

zabzot@gmail.com

Automation of semantic comparison of symbol names is an imminent task of practical application in automation accounting systems of companies. It is basically always possible to get an electronic copy of incoming document by, but subsequent incorporation into a company's accounting system may be difficult due to different ways of naming an accounted item. Architecture of intellectual agent and intuition modeling inference engine by the method of "semantic spot" for that task are reported here. Resolution of non-factor is adding new knowledge about the term of symbolic name, its semantics, semantic inter links. Each new term has a new semantic meaning which cannot be derived on rules of knowledge base. In view of experiments, conducted by the authors and above considerations, the use of any logical device is not deemed efficient. Intuition operates latent knowledge and associations is used here. Normally, agent's behavior is based on knowledge base and an inference engine. Intelligent agents might observable environments to achieve their goals. Possibility to build an agent behavior by intuitive are reported here.

Keywords: modeling of the principle of intuition; multiagent systems; semantic comparison of symbolic names; non-factors.