OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

МЕТОД ЦЕЛЕНАПРАВЛЕННОГО ПОИСКА СВЯЗАННЫХ ДАННЫХ В УСЛОВИЯХ СТРУКТУРНОЙ НЕОПРЕДЕЛЕННОСТИ

Тертышный В.А.[*], Шаповал И.С. [*], Щербак С.С.[*]

[*]*Кременчугский национальный университет имени Михаила Остроградского, г. Кременчуг, Украина*

vladislafus@gmail.com

sergey.shcherbak@gmail.com

В работе рассмотрены вопросы организации поиска связанных данных в условиях отсутствия структурно-логических схем распределенных источников. Разработан метод поиска на основе взаимных отображений объектов и связанных данных, и предложены рекомендации по созданию пользовательских интерфейсов поиска.

Ключевые слова: поиск; связанные данные; паттерны реализации; информационное пространство

Введение

Распределенность информационных ресурсов (ИР), содержимое которых создано на основе концепции связанных данных (англ. Linked Data) приводит к необходимости переосмысления понятия информационного поиска как такового. В рамках концепции связанных данных рассмотрим ИР (IR), как составляющие некоторого информационного пространства I. В рамках такого подхода будем рассматривать ИР как совокупность объектов, объединенных единым контекстом и принадлежащих некоторой предметной области, и связанных между собой отношениями «быть подклассом» и другими, характерными для объектно-ориентированного подхода.

Информационное пространство (ИП) обеспечивает распределенное хранение объектов - экземпляров классов, описанных в ИР с наборами действий в виде присоединенных к объектам процедур, применяемых для организации поиска. Таким образом, целью поиска связанных данных в информационном пространстве будет нахождение наиболее релевантного запросу объекта в виде набора характеристик объекта или объектов и их состояний. Так как хранилища, соответствующие источникам связанных данных, из которых состоят информационные пространства, позволяют для обеспечения синтаксической гибкости представлять одни и те же синтаксические конструкции по-разному, то пользователь перед взаимодействием с источником информационного пространства должен указать, на основе какого отображения интерпретировать связанные данные в терминах объектно-ориентированного подхода, поэтому определим два отображения (1), (2).

Выражение δ_1 определяет, что контексты G^c отображаются на классы C, предикаты P на свойства F, объекты O – на значения свойств V, а субъекты S – на экземпляры класса C', которые связаны с C' отношением «быть экземпляром класса» R^i:

$$\delta_1 : G^c \to C, P \to F, O \to V, S \to C' R^i C, \quad (1)$$

Выражение δ_2 определяет, что контексты G^c отображаются на экземпляры класса C', который связан с C отношением «быть экземпляром класса» R^i, предикаты P – на свойства F, а объекты O – на значения свойств V:

$$\delta_2 : G^c \to C' R^i C, P \to F, O \to V, \quad (2)$$

С учетом выражений δ_1, δ_2 и необходимостью при поиске наиболее релевантных объектов анализа не только составляющих структурно-логических схем классов, но и состояний их экземпляров, релевантность запроса экземпляра класса будем определять на основе метрики подобия. Таким образом, наиболее релевантным запросу экземпляром класса будет экземпляр класса, степень сходства которого согласно выбранной метрике будет наибольшей.

Метод целенаправленного поиска связанных данных

Для разработки метода целенаправленного поиска связанных данных уточним определения используемых в работе понятий.

Определение 1. Запрос представляет собой набор характеристик и их состояний, описывающий в терминах концепции связанных данных искомого в информационном пространстве объекта.

Определение 2. Наиболее релевантный запросу класс – это класс, у которого число совпадающих свойств класса и запроса наибольшее.

Замечание 1. Осознавая необходимость учитывать не только составляющие схемы класса, но и состояния его экземпляров, введем также понятие наиболее релевантного запросу экземпляра класса, соответствующего определению 2, но с дополнением в виде «метрики сходства». Таким образом, наиболее релевантным запросу экземпляром класса будет являться экземпляр класса, мера сходства которого согласно выбранной метрике является наибольшей.

Определение 3. Наиболее релевантная иерархия классов – это часть иерархии классов информационного пространства, базовый класс которой наиболее релевантный запросу.

Определение 4. Базовый класс – это класс источника связанных данных информационного пространства, с которым другие классы находятся в отношении «быть подклассом» (англ. «isSubclass»).

Для распознавания состояния объекта будем использовать следующую последовательность действий, а именно распознавание класса объекта, соответствующего запросу, на основе источников связанных данных ИП в соответствии с правилами интерпретации, представленных в виде присоединенных процедур, каждая из которых может использоваться экземпляром класса в соответствии с его структурно-логической схемой.

Замечание 3.2. В дальнейшем, с целью упрощения математической реализации предлагаемого метода задачу поиска решений будем рассматривать как задачу поиска на основе одного запроса, что ни в коей мере не сужает общности суждений.

Учитывая вышесказанное, последовательность метода поиска решений определим как совокупность действий, представленных следующими этапами и направленных на нахождение экземпляра класса информационного пространства, соответствующего состоянии объекта информационного пространства информирования процедуры визуализации состояния этого объекта:

Этап 1. Формирование запроса к ИП через пользовательский интерфейс, анализ его на противоречивость и построение соответствующего запроса экземпляра класса, поиск которого будет осуществляться.

Этап 2. Выбор метрики сходства, на основе которой будет осуществляться поиск.

Этап 3. Передача запроса к источнику СД ИП, а именно к их экземплярам базового класса, по подклассам которого будет проводиться дальнейший поиск.

Этап 4. Целенаправленный поиск наиболее релевантного запросу класса объектов для сокращения пространства поиска путем установления соответствия между схемами классов запроса и ИП.

Этап 5. Определение наиболее релевантных объектов по состоянию их характеристик на основе выбранной метрики сходства путем анализа ограничений назначения свойств класса, соответствующих объектов с целью выявления шаблонов визуализации в виде присоединенных процедур, формирующих ответ на запрос в виде визуализации в определенной форме содержимого объекта и синтез при необходимости нового шаблона визуализации.

Этап 6. Агрегация результатов поиска.

Этап 7. Оценка и формирование результатов поиска, упорядочивание на основе их рейтинга и представление в виде перечня визуализированных объектов по запросу пользователю.

Особенностью подхода, изложенного в методе, является то, что среда WWW, используемая в работе в качестве коммуникационной среды, позволяет параллельно осуществлять запросы к различным источниками связанных данных ИП. За счет этого одновременно осуществляется поиск в различных источниках. Кроме того, за счет анализа схем классов объектов и запроса, осуществляется целенаправленный поиск объектов и соответствующих им шаблонов визуализации путем сокращения пространства поиска благодаря выбору наиболее релевантных данному запросу объектных иерархий источника связанных данных, по которому осуществляется поиск.

Согласно вышеизложенному, реализацию математического обеспечения метода осуществим путем добавления к ИП присоединенных процедур, соответствующих положениям этого метода для визуализации содержимого объектов, что позволит находить объекты и формировать или использовать существующие шаблоны визуализации объектов.

Математическое обеспечение метода ценаправленного поиска связанных данных

Пусть Pr^{ds} – множество присоединенных процедур модели специализированной поисковой системы M1 (СПС), которые реализуют операционную составляющую этой системы в виде подмножеств присоединенных процедур $\mathrm{Int}, \mathrm{Viz}$, реализующих интерпретацию и визуализацию

содержимого объектов, которые соответствуют запросу, представленному в виде экземпляра класса $E \subset G_t^S$, где G_t^S – это структурно-логическая схема связанных данных; $A_1,...,A_k$ – источники ИП, хранящие данные об объектах, A_r – хранилище специализированной поисковой системы, причем $\{A_1,...,A_k\} \bigcup A_r = Src$, тогда определим присоединенную процедуру ρ^{cpn}, обеспечивающую взаимодействие пользователя специализированной поисковой системы с ИП как средство формирования управляющего воздействия в виде запроса к ИП через специализированную поисковую систему, выполнение соответствующей входному запросу q процедуры объекта СПС $\rho^{act} \in Pr^{ds}$ (реакция ИП на входное воздействие) и визуализация результатов выполнения запроса (ρ^{viz}):

$$\rho^{cpn}(q) = \left\{ \frac{\rho^{err}(q) \mid \rho^{act}(q) = \varnothing}{\rho^{viz}(\rho^{act}(q) : E) \mid else} \right\}, \quad (3)$$

где E – множество экземпляров классов, полученных в результате выполнения запроса процедуре ρ^{act}, ρ^{err} – реакция ИП в случае ошибки процедуры ρ^{cpn}.

Пусть $\rho^{act}, \rho^{search}, \rho^{int}, \rho^{dec} \in Pr^{ds}$ – процедуры обработки управляющих воздействий пользователя, поиска, добавления критериев интерпретации значений показателей объекта, формирования шаблонов визуализации пользователя соответственно, тогда ρ^{act} формально определим так:

$$\rho^{cpn}(q, in) = \left\{ \begin{array}{c} \rho^{search}(q)|_{in=1} \\ \rho^{int}(\min^{prm}, \max^{prm})|_{in=2} \\ \rho^{dec}(q)|_{in=3} \end{array} \right\}, \quad (4)$$

Рассмотрим математическую реализацию ρ^{search} с учетом (3) – (4) как средства представления основных положений метода целенаправленного поиска объектов в ИП (рис. 1):

$$\rho^{search}(q) = [\rho^{get}(q) : \{E^{A_1},...,E^{A_k}\}, \quad (5)$$

$$\rho^{rat}(E^{A_1},...,E^{A_k}) : E^A], \quad (6)$$

где ρ^{get}, ρ^{rat} – процедуры получения данных из источника, определения наиболее релевантного класса объектов и соответствующих экземпляров прецедентов ПС E, $A_1,...,A_k$ – источники ИП, которые хранят данные об объектах.

Учитывая то, что схема источника связанных данных ИП иерархически организована на основе отношения «быть подклассом», то доступ к источнику осуществляется через базовый класс этого источника, то есть запрос необходимо подавать на базовый класс источника ИП.

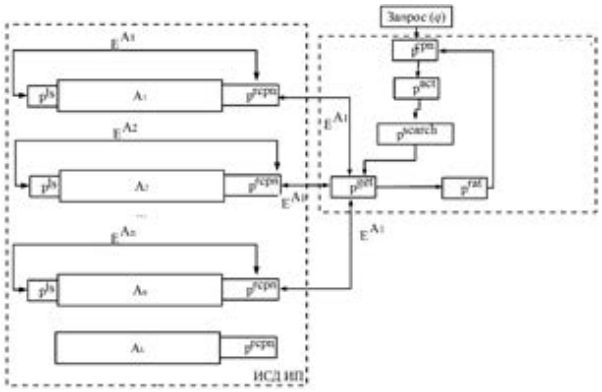

Рисунок 1 – Схема взаимодействия процедур при поиске в ИП ПП

Процедура ρ^{get} формально определена через ρ_A^{rcpn} – присоединенные процедуры взаимодействия с удаленными источниками $A_1,...,A_k$ – следующим образом:

$$\rho^{get}(q) = [\rho_{A_1}^{rcpn}(q) : E^{A_1},...,\rho_{A_k}^{rcpn}(q) : E^{A_k}], \quad (7)$$

где $\forall_i \rho_{A_i}^{rcpn}(q) : [\rho_i^{ls}(q) : E^{A_i}]$,

E^{Ai} – экземпляры наиболее релевантного класса ПС (прецеденты), q – запрос, ρ_i^{ls} – i-я ниже определенная процедура поиска прецедентов источников СД на основе метрики, определенной в $\rho_{A_i}^{metric}$:

$$\rho_i^{ls}(q, b^{A_i}) = [\rho_{A_j}^{metric}(q, E_b^{A_j}) \to Max], \quad (8)$$

где $E_b^{A_j}$ –экземпляры E класса b j– источника A.

Процедура ρ^{rat} предназначена для определения среди источников $A_1,...,A_k$ наиболее релевантного результата поиска в виде класса объектов и соответствующих экземпляров объектов на основе их оценки с помощью процедуры $\rho_{A_j}^{rat}$:

$$\rho^{rat}(E^{A_1},...,E^{A_m}) = [\forall_j \rho_{A_j}^{maxrat}(q, E^{A_j}) \to Max], \quad (9)$$

Визуализацию содержимого объектов определим с помощью ρ^{dec}:

$$\rho^{dec}(q) = [\rho_1^d(q),...,\rho_r^d(q)], \quad (10)$$

где $\rho_1^d(q),...,\rho_r^d(q)$ – это процедуры, которые последовательно применяются к экземпляру класса соответствующим запросом q для визуализации его содержимого или для формирования соответствующей запросу реакции системы.

Таким образом, решение в виде ρ^{dec}, применяемое к экземплярам класса, определенного выражением (10), и есть искомое содержание, которое затребовал пользователь.

Для улучшения процедуры поиска объектов в связанных данных пользователь может добавить новые показатели и соответствующие им

ограничения, которые будут отслеживаться СПС и учитываться при поиске объектов. Для этого пользователь может вызвать процедуру ρ^{cpn}, определенную выражением (4) с установленным параметром $in = 2$. В этом случае запрос q к СПС будет представлен парой с минимальным ($\min^{prm}$) и максимальным ($\max^{prm}$) ограничением значения показателя, идентифицируемый порядковым номером в множестве показателей Crt^c, а присоединенная процедура добавления нового показателя и его ограничений представлена так:

$$\rho^{int}(\min^{prm}, \max^{prm}) = [Crt^c \cup [\min^{prm}, \max^{prm}]] \quad (11)$$

Особое место в интерактивном взаимодействии пользователя через СПС с ИП занимает процедура формирования запроса. Рассмотрим ее более подробно.

Так как объект – экземпляр класса объектов, тогда его структурно-логическая схема может быть задана следующим образом:

$$x = << x\ instanceOf\ X^{'} >, f_1 ... f_k \in F, p^{dec} >, \quad (12)$$

где *x* – экземпляр класса объектов $X^{'}$, свойства $f_1 ... f_k \in F$, p^{dec} – процедура визуализации содержания объектов.

Класс объектов $X^{'}$, подкласс базового класса c^{base} ИСД ИП с установленным отношением isSubclass и определенным множеством *F*, зададим с помощью следующего выражения:

$$X^{'} = << X^{'}\ isSubclass\ c^{base} >, F, \varnothing >, \quad (13)$$

Запрос к ИП, задаваемый пользователем через СПС определим как шаблон объектахкласса $X^{'}$ в виде набора свойств Fc установленными значениями v_1, v_k следующим образом:

$$q = << x\ instanceOf\ X^{'} >, < t_1, l_1, v_1 > ,..., < t_k, l_k, v_k > \in F, \varnothing >, \quad (14)$$

где t_1, t_k – типы данных, l_1, l_k – языки представления, *k* – номер свойства.

Учитывая возможность построения противоречивых запросов к ИП, определим понятие противоречивости запроса и разработаем процедуру проверки запроса на противоречивость.

Определение 5. Под противоречивостью запроса будем понимать появление в запросе двух и более одинаковых свойств, как по структуре, так и по содержанию.

Для проверки запроса на противоречивость, согласно определению (5), определим следующее выражение, результат которого 1 означает наличие противоречия в запросе:

$$pq(q) = \begin{Bmatrix} 1 \mid \exists f \in F^q \mid f = f^{'} \\ 0 \mid else \end{Bmatrix}, \quad (15)$$

где $f, f^{'} \in F$ – свойства, определенные в запросе q.

С учетом вышесказанного, математическое обеспечение метода поиска решений в информационном пространстве может быть определено с помощью формул (1) – (15).

В качестве рекомендаций по проектированию поисковых интерфейсов можно выделить следующие положения, а именно подобные интерфейсы должны:

- обеспечивать своевременное информирование пользователя о текущем состоянии системы, смене состояния в результате выполнения поисковых запросов;

- формировать команды по принципу модели "объект - действие" вместо модели «действие-объект»;

- поддерживать монотонность интерфейса и интерфейсы на различных устройствах и платформах, включая режим сенсорного ввода.

Заключение

В работе рассмотрены вопросы организации целенаправленного поиска связанных данных в условиях отсутствия структурно-логических схем распределенных источников. Разработан метод поиска на основе взаимных отображений объектов и связанных данных, и предложены рекомендации по созданию пользовательских интерфейсов поиска.

Библиографический список

[Майер-Шенбергер В., 2014] Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Майер-Шенбергер В., Кукьер К. – М.: Манн, Иванов и Фербер, 2014. – 240 с.:ил.

[Гудсон, 2013] Гудсон Дж. Практическое руководство по доступу к данным / Д. Гудсон, Р.Стюарт. – СПб.:БВХ-Петербург, 2013. – 304с.:ил.

METHOD OF LINKED DATA SEARCH IN CONDITIONS OF STRUCTURAL UNCERTAINTY

Tertishniy V.A., Shapoval I.S., Shcherbak S.S.*

**Kremenchuk Mykhailo Ostrohraskyi National University, Kremenchuk, Ukraine*
vladislafus@gmail.com
sergey.shcherbak@gmail.com

In this paper linked data search issues have been examined in conditions of structural and logical schema absence for distributed resources. A search method has been developed, which is based on mutual mappings of objects and lined data. Recommendations concerning search user interface development have been proposed.