Конструктор запросов интеллектуального поиска

© О.Л. Обухова, Т.К. Бирюкова, М.М. Гершкович, И.В. Соловьев, А.П. Чочиа

Институт проблем информатики РАН, г. Москва

obuhova@amsd.com

Аннотация

Для описания и анализа семантической структуры коллекции научных материалов авторы предлагают метод создания специальных формул, при построении которых используются правила теории исчисления предикатов и алгоритм «решающих правил» из Data Mining. Полученные формулы составляют основу запросов интеллектуального поиска, а также служат для решения задач исследования научных идей определенной тематики, представленных в коллекции публикаций.

1 Введение

В данной работе рассматривается метод решения задачи интеллектуального поиска в коллекции научных материалов с помощью разработанной авторами модели онтологии [1], в которую встроены сервисы анализа семантического содержания коллекции.

Содержание коллекции публикаций трактуется в зависимости от того, как понимаются и систематизируются научные тексты. «Понимание научного текста – это формирование смысловой структуры: выделение «смысловых вех» и связывание их в единую семантическую структуру» [2]. Для формирования смысловой структуры научной публикации авторы используют подход, базирующейся на фасетной классификации [3]. Родственные «смысловые вехи» или, иными словами, семантические единицы объединяются в фасеты, где фасетный признак является обобщающим понятием для входящих в данный фасет значений. Для каждой научной публикации при занесении в коллекцию создается описание смысловой структуры в форме *фасетной формулы объекта*, представленной в виде множества совокупностей: <фасетный признак: список значений> для данного объекта [4]. Набор фасетных признаков диктуется направленностью предметной области – к примеру, для коллекции электронных документов сайта научного института набор фасетных признаков определяется видами и характером научной деятельности. Количество фасетных признаков выбирается в соответствии с принципом, сформулированным американским психологом Миллером [5]: для того чтобы выбор был эффективным, количество элементов в нем не должно быть больше семи – девяти. Научная публикация совместно с *фасетной формулой объекта* называется информационным объектом (ИО) [6]. Объединение всех фасетных формул определяет семантическое содержание коллекции научных публикаций. Цель работы – построение модели онтологии предметной области со встроенными сервисами семантического анализа для исследования коллекции научных материалов, представления её структурной организации и формирования запросов интеллектуального поиска.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

2 Онтологическая модель предметной области

Концептуальная модель предметной области включает классы, представленные на рис. 1.

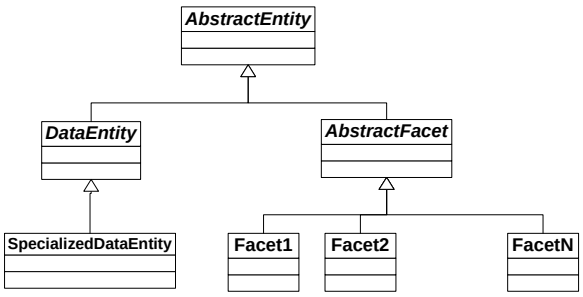

Рис. 1

Класс **AbstractEntity** – абстрактный класс, который имеет два наследуемых абстрактных класса **DataEntity** и **AbstractFacet.**

Класс **SpecializedDataEntity** является наследником класса **DataEntity**, экземплярами класса являются информационные объекты.

Классы $\mathbf{Facet_f}$, f=1, …, n, являются наследниками класса **AbstractFacet**, экземплярами класса являются объекты, представляющие допустимые значения определенного фасетного признака для всех ИО, формирующих актуальный состав коллекции. Объекты классов $\mathbf{Facet_f}$, f=1, …, n, определяют набор семантических понятий предметной области.

Для построения онтологической модели необходимо определить множества свойств и отношений объектов [7].

Поскольку понятие «бинарное отношение» и «признак (свойство)» идентичны в случае, если значением свойства является объект другого класса, то для представленной модели множество свойств является подмножеством множества отношений.

Фасетные формулы объектов (ФФО) определяют отношения между объектами класса **SpecializedDataEntity** и классов **Facet$_f$**, f=1, …, n, и являются аксиомами онтологической модели.

Множество отношений между объектами классов **Facet$_f$**, f=1, …, n, определим следующим образом.

Определение 1. *Двуместный предикат*

$$R^{2}_{f_1 f_2}(X^{f_1}_{i_1}, X^{f_2}_{i_2})$$

обозначает бинарное отношение двух объектов $X^{f_1}_{i_1}$ *и* $X^{f_2}_{i_2}$, *где* $X^{f_1}_{i_1}$, $X^{f_2}_{i_2}$ *– объекты классов* **Facet** $_{f_1}$ *и* **Facet** $_{f_2}$, $f_1 \neq f_2$, $i_1 \in \{1, \ldots, k_{f_1}\}$, $i_2 \in \{1, \ldots, k_{f_2}\}$, k_{f_1}, k_{f_2} *– количество объектов соответствующего класса, и принимает значение из множества* {true, false} *в зависимости от актуального набора ФФО.*

N-арные отношения (предикаты), связывающие три и более объектов разных классов **Facet$_f$**, f=1,…,n, определяются рекурсивно.

Определение 2. *Трехместный предикат*

$$R^{3}_{f_1 f_2 f_3}(X^{f_1}_{i_1}, X^{f_2}_{i_2}, X^{f_3}_{i_3}),$$

где $f_1, f_2, f_3 \in \{1, \ldots, n\}$, $i_1 \in \{1, \ldots, k_{f_1}\}$,

$$i_2 \in \{1, \ldots, k_{f_2}\}, i_3 \in \{1, \ldots, k_{f_3}\},$$

k_{f_1}, k_{f_2}, k_{f_3} *– количество объектов соответствующего класса, выражающий отношение трех объектов, принимает значение* true, *если* $R^{2}_{f_1 f_2}(X^{f_1}_{i_1}, X^{f_2}_{i_2})$ = true *и* $\exists$ *объект* $X^{f_3}_{i_3}$, *такой, что* $R^{2}_{f_1 f_3}(X^{f_1}_{i_1}, X^{f_3}_{i_3})$=true *и* $R^{2}_{f_2 f_3}(X^{f_2}_{i_2}, X^{f_3}_{i_3})$ = true, *иначе* $R^{3}_{f_1 f_2 f_3}(X^{f_1}_{i_1}, X^{f_2}_{i_2}, X^{f_3}_{i_3})$ *принимает значение* false.

Определение 3. N-*местный предикат*

$R^{N}_{f_1 f_2 \ldots f_N}(X^{f_1}_{i_1}, \ldots, X^{f_N}_{i_N})$, *где* $f_1, f_2, \ldots, f_N \in \{1, \ldots, n\}$, $i_1 \in \{1, \ldots, k_{f_1}\}$, $i_2 \in \{1, \ldots, k_{f_2}\}, \ldots, i_N \in \{1, \ldots, k_{f_N}\}$, k_{f_1}, k_{f_2}, …, k_{f_N} *– количество объектов соответствующего класса, выражающий отношение* N *объектов, принимает значение* true, *если* $R^{N-1}_{f_1 \ldots f_{N-1}}(X^{f_1}_{i_1}, \ldots, X^{f_{N-1}}_{i_{N-1}})$=true *и* $\exists$ объект $X^{f_N}_{i_N}$, *такой, что для* $\forall X^{f_j}_{i_j}$, $j \in \{1, \ldots N\text{-}1\}$, *справедливо*: $R^{2}_{f_j f_N}(X^{f_j}_{i_j}, X^{f_N}_{i_N})$=true, *иначе* $R^{N}_{f_1 f_2 \ldots f_N}(X^{f_1}_{i_1}, \ldots, X^{f_N}_{i_N})$ *принимает значение* false.

Таким образом, отношение существует, если соответствующий предикат принимает значение true.

В основу интеллектуального поиска заложено выявление синтагматических (ситуационных) связей между объектами, попадающими в область интереса пользователя и составляющими часть предметной области, ограниченной многомерным параллелепипедом, выбор граней которого определяет пользователь на уровне графического интерфейса. В представленной модели элементы пользовательского интерфейса являются визуальным отображением каждого онтологического класса. В силу сложившихся традиций и акцентируя тот факт, что для построения онтологической структуры использовалась фасетная классификация, графические элементы называют фасетами [8].

Конструирование запросов интеллектуального поиска заключается в выполнении итерационного и интерактивного процесса выбора тех объектов классов **Facet$_f$**, f=1,…,n, между которыми существуют отношения. Существование отношений выявляется и доказывается с помощью правил теории исчисления предикатов и использования алгоритма «решающее правило» из Data Mining [9], т. е. в соответствии с набором правил, где очередное правило строится путем последовательного добавления к нему условий. Выбранные объекты и отношения составляют формулу специального вида, которая является расширением и обобщением ФФО, отражает состояние коллекции в определенный момент времени и используется для реализации интеллектуального поиска и анализа коллекции публикаций.

3 Заключение

В структуру предлагаемой онтологической модели предметной области авторы встроили механизмы формирования синтагматических связей объектов, записываемых формулами специального вида в форме многоместного предиката. Данный подход обеспечивает проверку существования связей и количественную оценку числа публикаций, семантическая структура которых, выраженная в виде *фасетной формулы объекта*, отвечает построенной формуле. Изложенный метод позволяет пользователям строить запросы на языке, приближенном к естественному языку. Поскольку отношения – это есть поименованная связь, то формула специального вида автоматически превращается в запрос, к примеру, такого содержания: «Какое количество работ по направлению «Инженерия знаний» в виде Диссертации, или Научного доклада, или Статьи было опубликовано за последние 25 лет?».

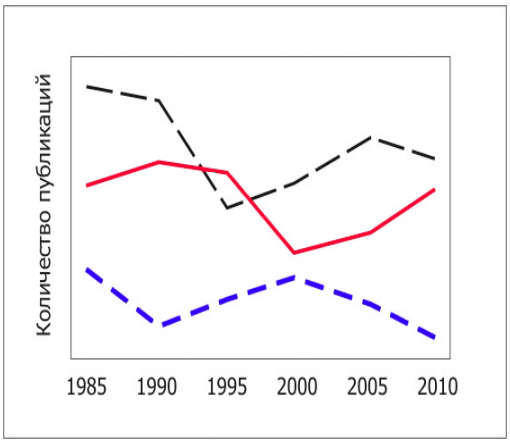

Рис. 2

Ответ можно предоставить в графическом виде, что проиллюстрировано на рис. 2, где линия __ __ обозначает количество публикаций в виде статьи, линия _ _ обозначает количество публикаций в виде доклада, линия _____ обозначает количество публикаций в виде диссертации.

Литература

[1] Gruber T.R. A translation approach to portable ontologies// Knowledge Acquisition. – 1993. – V. 5, No 2. – P. 199-220.

[2] Сиротко-Сибирский С.А. О проблеме понимания текста в лингвистике и психолингвистике // Слово отзовется: памяти А.С. Штерн и Л.В. Сахарного. – Пермь, 2006.

[3] Ранганатан Ш.Р. Классификация двоеточием. Основная классификация: Пер. с англ. / Под ред. Т.С. Гомолицкой и др. – М., 1970. – 422 с.

[4] Чочиа А.П., Соловьев И.В., Обухова О.Л., Бирюкова Т.К., Гершкович М.М. Модель адаптивной фасетной навигации в открытых электронных коллекциях // Системы и средства информатики. – М.: Наука, 2008. – Вып. 18. – С. 294-309.

[5] Миллер Дж. Магическое число семь, плюс или минус два. – В кн.: Инженерная психология. – М. 1964.

[6] Обухова О.Л., Бирюкова Т.К., Гершкович М.М., Соловьев И.В., Чочиа А.П. Метод динамического создания связей между информационными объектами базы знаний//Труды 11-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2009, Петрозаводск, Россия, 2009. – С. 39-45.

[7] Ручкин В.Н., Фулин В.А. Универсальный искусственный интеллект и экспертные системы. – С-Петербург: «БХВ - Петербург», 2009.

[8] Обухова О.Л., Соловьев И.В., Бирюкова Т.К., Гершкович М.М., Чочиа А.П. Модель фасетного информационного поиска в коллекции научных материалов// Системы и средства информатики, доп. выпуск. – М.: Наука, 2009. – С. 163-174.

[9] Piatetsky-Shapiro G. A comprehensive microarray data generator to map the space of classification and clustering methods// Tech. Report No 2004-016, U. Massachusetts Lowell, 2004.

Designer for requests of intelligent search

Olga Obuhova, Tatiana Biryukova,
Maxim Gershkovich, Ivan Soloviev, Anton Chochia

We propose the method to describe and analyze the semantic structure of the scientific publications. The method suggests design of specific formulas using the predicate's calculations theory and Data Mining's algorithm of Decision Rules. Formulas form a structure of the intelligent search's request and are used to investigate scientific ideas from publication's data base.