УДК 681.5.015

РЕГУЛЯРИЗАЦИЯ ПРОЦЕДУР ОБРАБОТКИ ДАННЫХ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В.П. Авраменко[1], Н.А. Валенда[2]

[1] ХНУРЭ, г. Харьков, Украина

[2] ХНУРЭ, г. Харьков, Украина, valenda@ukr.net

Исследованы причины некорректности задач обработки данных в системах контроля знаний дистанционного обучения. Предложены процедуры регуляризации (исправления) некорректно поставленных задач обработки данных на основании принципа внешнего дополнения. Внедрение регуляризованных процедур обработки данных позволило повысить эффективность поиска данных благодаря отбору информации, более релевантной запросу.

КОРРЕКТНОСТЬ ЗАДАЧ ОБРАБОТКИ ДАННЫХ, ПРОЦЕДУРЫ РЕГУЛЯРИЗАЦИИ

Введение

В основу системы контроля знаний дистанционного обучения, как правило, закладывается принцип текстового диалога, когда вопрос задается обучающей стороной, а ответ формируется обучаемым по определенным правилам. В системах контроля знаний важное место занимает проблема однозначного смыслового толкования и восприятия задаваемого вопроса и получаемого ответа. Для решения этой проблемы в формализованных системах используются процедуры регуляризации [1, 2].

Процедура регуляризации представляет собой вычислительный процесс, который сначала исправляет исходную некорректно поставленную задачу на корректно поставленную за счет привлечения дополнительной информации, а затем отыскивает приближенное решение. Термин «регуляризация» восходит от латинского слова *regula*, которое означает *правило* или *сделать правильным*. Процедуры обработки данных не всегда удовлетворяют желаемым *правилам*, а поэтому по природе своей принадлежат к некорректно поставленным задачам.

В алгебраических системах при обработке данных важное место занимают регулярные выражения и языковые среды. Под выражением принято подразумевать совокупность действий, которые выполняются в заданной последовательности, для того чтобы получить вполне определенное значение некоторого алгебраического объекта. Среди алгебраических выражений важное место занимают регулярные выражения, составленные по определенным правилам на специальном алгебраическом языке.

Каждому алгебраическому объекту ставится в соответствие набор правил «хорошего поведения», при соблюдении которых «регуляризованные» вычислительные процедуры ведут себя «наилучшим образом». Регулярным объектом может выступать регулярный оператор, в частном случае, хорошо обусловленная (невырожденная) матрица, а нерегулярным объектом может выступать нерегулярная (вырожденная) матрица. Для вычисления определителя нерегулярной (плохо обусловленной) матрицы применяются процедуры «регуляризации».

Целью данной статьи является анализ причин возникновения некорректности в задачах обработки данных и разработка процедур регуляризации (исправления) некорректно поставленных задач с целью повышения эффективности функционирования систем искусственного интеллекта на примере системы контроля знаний дистанционного обучения.

1. Проблематика некорректности задач обработки данных

Качество принимаемых решений широкого класса математических задач тесно связано с корректностью их исходной постановки. Большинство оптимизационных задач обработки данных принадлежит к классу обратных некорректно поставленных задач. Первые этапы исследования некорректно поставленных задач направлены на уточнение вопроса о разрешимости и устойчивости задачи [1, с. 16; 2, с. 26].

Задача математического программирования

$$Q^* = Q(x^*) = \min_{x \in D} Q(x), \quad D = \{x : x \in E_+^n, G(x) \geq 0\} \quad (1)$$

является корректно поставленной при одновременном выполнении следующих условий:

– решение x^* существует, принадлежит допустимому множеству D и обеспечивает минимум функционалу

$$Q^* = Q(x^*) = \min_{x \in d} Q(x); \quad (2)$$

– решение единственно, то есть если $x_1^* \in D$ и $x_2^* \in D$, то из соотношения

$$Q^* = Q(x_1^*) = Q(x_2^*) = \min_{x \in d} Q(x) \quad (3)$$

следует, что $x_1^* \equiv x_2^* = x^*$;

– решение устойчиво по отношению к отклонениям исходных данных, то есть достаточно малым погрешностям элементов матрицы условий, величин выделенных ресурсов и коэффициентов

целевой функции соответствуют погрешности того же порядка в определении решения.

2. Регуляризация процедур обработки данных

Некорректность задач обработки данных можно трактовать в узком и широком смысле относительно класса аппроксимируемых моделей (линейного и нелинейного программирования) и количества аппроксимируемых критериев (однокритериальных и многокритериальных). Примером однокритериальной модели скалярной оптимизации обработки данных может служить модель математического программирования (1) – (3).

Большинство задач обработки данных в лучшем случае являются слабо корректными в силу их слабой структурированности. В них отсутствует достоверная информация о непротиворечивости ограничений, характере возмущающих воздействий и погрешностях вычислений. В качестве регуляризованного решения задачи можно принять нормальное решение $x_н$, наименее уклоняющееся от некоторого заданного вектора x_0.

Чем ближе искомое решение $x_н$ к некоторому желаемому вектору x_0, тем эффективнее полученное решение. Меру уклонения нового решения от старого (критерий близости) можно задать квадратом нормы:

$$\Omega[x_н - x_0] = \| x_н - x_0 \|_n^2 . \qquad (4)$$

Предположим, что задан некоторый вектор $x_0 \in X$. Искомый вектор $x_н$ представляет нормальное решение задачи линейного программирования (по отношению к x_0), если справедливо соотношение

$$\| x_н - x_0 \|^2 = \min_{x^* \in X} \| x^* - x_0 \|^2 , \qquad (5)$$

где x^* – любое решение этой задачи. Из совокупности квазиоптимальных решений с помощью интеллектуальной технологии (способа задания критерия близости) выделяется нормальное решение, наилучшее в смысле выбранной функции уклонения.

Многокритериальные технологии эффективной обработки данных принято относить к классу многократно некорректных задач. Их некорректность возникает, во-первых, из-за некорректности задач локальной оптимизации, во-вторых — из-за процедур принятия многокритериальных решений, в основу которых положен принцип неединственности. Множественность принимаемых решений является скорее достоинством, а не недостатком, поскольку «жесткие» схемы получения единственного решения неадекватны сущности задач многокритериальной оптимизации, а имеющаяся интеллектуальная «свобода» выбора предпочтительного решения из множества эффективных позволяет учесть неопределенность целей и критериев.

3. Применение регулярных выражений для обработки данных

В роли математического описания последовательности выполняемых операций в системе контроля знаний предложено использовать регулярные выражения формальных языков и грамматик. Регулярные выражения служат удобным описанием программных компонентов типа программ текстового поиска и программ текстового перевода. Регулярные выражения строятся на основе алгебраических законов, которые определяют структуру данных с помощью текстовых цепочек.

Регулярные выражения позволяют создать регулярные языки с заданными свойствами, которые определяют допустимые цепочки декларативным способом. Поэтому регулярные выражения используются в качестве входного языка во многих системах обработки текстовых цепочек. Различные поисковые системы преобразуют регулярные выражения в конечные автоматы, а такие автоматы используются для поиска текстовых цепочек в файле.

Регулярные выражения определяются в специальной алгебраической языковой среде регулярных событий, взаимосвязанных набором операций. Множество всех событий представляет собой некоторую универсальную алгебру, то есть над событиями определяются алгебраические операции. В качестве регулярных событий выступает формальный язык (произвольное множество слов или текстовых цепочек), выражения которого задают события над некоторым алфавитом. В качестве операций обычно выступают три оператора регулярных выражений: объединение, конкатенация и итерация.

Эффективность обработки текстовых данных определяется набором основных требований соответствующих интеллектуальных систем к инструментальным средствам перевода, компиляции или иной обработки данных. Проблема оценки качества перевода волнует теоретиков и практиков в связи с возросшими запросами в этом направлении, развитием машинного перевода и необходимостью создания общей теории перевода.

Одним из основных требований к тексту перевода является необходимость единства формы и содержания текста оригинала средствами языка перевода. Адекватность переводов обычно определяется категорией семантической полноты и точности, которая дополняется стилистической эффективностью, включающей в себя принцип соответствия текста перевода стилистическим нормам языка перевода. По этим параметрам оценивается качество перевода. На сегодняшний день проблема теоретических оценок качества перевода еще не решена. Практические оценки качества перевода определяются по эмпирико-интуитивным соотношениям, базирующимся на личном профессиональном опыте [3, с. 124–142].

Регуляризация является важным аспектом при обработке языков. Она может распространяться на несколько уровней. Наиболее распространенным является использование регулярных выражений в качестве средства описания языка. Существует трехсторонняя эквивалентность между регулярными выражениями, конечными автоматами и регулярными грамматиками при задании языка.

Это свойство широко используется в компиляторах языков программирования. Для описания конструкций языка используются регулярные выражения, на основании которых автоматический генератор может построить распознающие конечные автоматы. В результате можно получить лексический анализатор языка на основе его описания средствами регулярных выражений.

Разработанные генераторы лексического анализа, такие как lex или flex, позволяют автоматизировать программы обработки данных [4, с. 236–239]. Кроме программирования эти средства могут применяться для обработки естественных языков, например, русского. Такой лексический анализатор преобразовывает последовательность символов исходного файла во множество элементов, типы которых определяются видом регулярных выражений.

Ниже приведен пример программы на flex для разбиения текста русского языка на слова, аббревиатуры, числа, классы знаков пунктуации:

```
digit        [0-9]
intconst     [+\-]?{digit}+
realconst [+\-]?{digit}+\.{digit}+(e[+\-]?{digit}+)?
letter       [а-я]
capital      [А-Я]
word  ({letter}|{ capital}){letter}*|{letter}+-{letter}+
abbreviation {capital}*
punctuation {,|:|;|-| (|)|{|}| [|]}
stops {.|!|?}
%%
```

{intconst} Процедура формирования лексемы – целое число;

{realconst} Процедура формирования лексемы – дробное число;

{word} Процедура поиска слова в словаре и формирования лексемы слово;

{abbreviation} Процедура формирования лексемы аббревиатура;

{punctuation} Процедура формирование лексемы – знак препинания;

{stops} Процедура формирование лексемы – признак конца предложения;

%%

Таким образом, регулярное описание языка позволяет строить для него автоматические распознаватели. Дополнив конечные автоматы процедурами формирования лексем, получаем конечные распознаватели (КР) языковых конструкций. По результатам автоматизированной работы flex построен анализатор, имеющий структуру представленную на (рис. 1).

Другим уровнем применения регуляризации к языкам является преобразование конструкций естественных языков к регулярному виду. Естественные языки из-за их многозначности представляют сложный объект для описания – каждому слову мо-

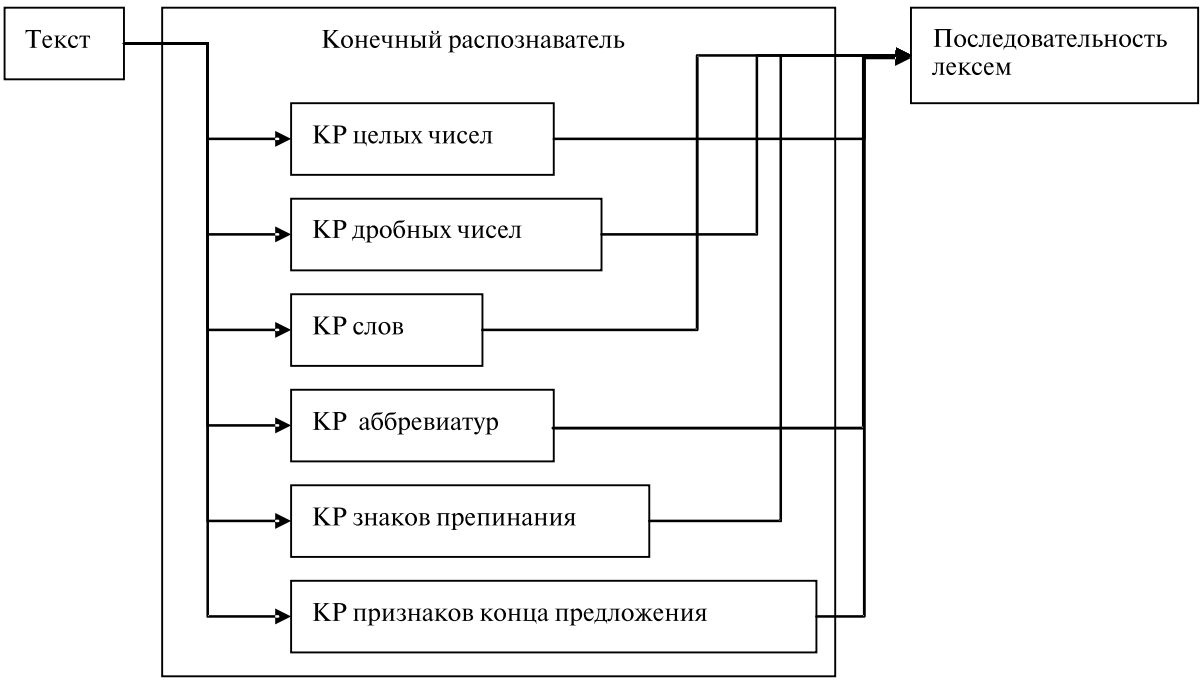

Рис. 1. Структура лексического анализатора на базе конечных распознавателей

жет соответствовать несколько значений и множеств морфологических признаков, предложению может соответствовать несколько структур разбора.

Назовем регулярным объектом языковую конструкцию, где каждому слову сопоставлено единственное значение в данном контексте и единственная морфологическая интерпретация; элементам предложения однозначно приписаны их роли, что дает единственную структуру разбора. Построение такого регулярного представления происходит в процессе поэтапного анализа языковых конструкций.

В качестве инструментов обработки данных могут использоваться различные формальные представления, удовлетворяющие ограничениям, налагаемым на данное представление. Ниже приведен пример регулярного представления языковой конструкции «Программный продукт строит график функции» на основе семантических функций:

$$f_1(V_3^2(y),V_1^1(x),f_3^4(V_1^4(z),V_1^4(u)))\,,$$

где f_i — функции, определяющие отношения в предложении; $V_j(x)$ — функция, определяющая j значение аргумента x.

Важным вопросом является эквивалентность формального представления исходной языковой конструкции. Пусть исходная языковая конструкция A преобразуется процедурой регуляризации R в формальное представление B:

$$A \xrightarrow{R} B\,.$$

Пусть существует формальная процедура R', позволяющая получать из формального представления языковую конструкцию

$$B \xrightarrow{R'} A'\,.$$

Важным при этом является вопрос об эквивалентности языковых конструкций A и A'. Ответ на этот вопрос зависит от симметричности преобразований R и R'. Симметричность преобразований могут обеспечить эквивалентные структуры анализа и синтеза на уровне синтаксического и семантического анализа. Главная проблема состоит в адекватности отражения семантико-синтаксических связей предложения на формальном уровне, их использование на этапах синтеза и анализа обработки данных.

Выводы

Усовершенствованы инструментальные средства обработки данных в системах искусственного интеллекта за счет применения семантических функций для анализа языковых конструкций.

Разработаны процедуры регуляризации семантического анализа данных на основе применения логического вывода на семантических функциях, что позволило выбирать единственное значение для многозначных слов и получать единственное формальное представление языковых конструкций.

Список литературы: 1. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. – М.: Наука, 1979. – 288 с. 2. *Авраменко В.П.* Управление производством в условиях неопределенности. – Киев: УМК ВО, 1992.– 48 с. 3. *Волошин В.Г.* Комп'ютерна лінгвістика. – Суми: Університетська книга, 2004. – 382 с. 4. *Валенда Н.А.* Применение методов анализа естественного языка для поисковых систем // Вестник Херсонского государственного политехнического университета. – 2002. – № 1 (14). – С. 236 – 239.

Поступила в редколлегию 08.10.2007