Московский государственный университет
имени М. В. Ломоносова

Научно-исследовательский институт механики

На правах рукописи

Хазова Елена Евгеньевна

АЛГОРИТМЫ ПОСТРОЕНИЯ ОДНОСЛОВНОЙ ПЕРЕЗАПИСИ РЕГУЛЯРНЫХ ПУТЕВЫХ ЗАПРОСОВ

Специальность 05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2009

Работа выполнена в Научно-исследовательском институте механики Московского государственного университета имени М. В. Ломоносова.

Научный руководитель:	доктор физико-математических наук, профессор *Васенин Валерий Александрович*
Официальные оппоненты:	доктор физико-математических наук, профессор *Алешин Станислав Владимирович* (Московский государственный университет имени М. В. Ломоносова)
	доктор технических наук, профессор *Кузнецов Сергей Дмитриевич* (Институт системного программирования Российской Академии Наук)
Ведущая организация:	Институт математики имени С. Л. Соболева Сибирского отделения РАН

Защита диссертации состоится 25 февраля 2009 г. в 16 ч. 45 мин. на заседании диссертационного совета Д.501.002.16 при Московском государственном университете имени М.В.Ломоносова по адресу: Российская Федерация, 119991, Москва, ГСП-1, Ленинские горы, д.1, Московский государственный университет имени М.В.Ломоносова, механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в библиотеке механико-математического факультета МГУ (Главное здание, 14 этаж).

Автореферат разослан 22 января 2009 г.

Ученый секретарь диссертационного
совета Д.501.002.16 при МГУ
доктор физ.-мат. наук А. А. Корнев

Общая характеристика работы

Актуальность темы. Современное общество, которое принято именовать информационным, характеризуется тем, что эффективные технологии обработки данных во многом определяют успехи в области материального производства, в научно-технической, социально-экономической и в политической сферах деятельности человека. В условиях быстрого развития сетевой инфраструктуры и объемов, сосредоточенных в ней данных, методы их поиска, систематизации и анализа на сегодня отстают от возрастающих еще более высокими темпами потребностей общества. Одна из причин сложностей, возникающих на пути их удовлетворения, связана с существенно неоднородным характером подлежащей обработке информации, как с позиции поддерживающей ее аппаратно-программной платформы, так и в плане модели представления этих данных. На сегодня сложно представить себе случай, когда интересующие пользователя практически значимые данные могут находиться под управлением традиционной реляционной СУБД, какой либо другой модели их описания (модели данных) или системы. Для того, чтобы удовлетворять такие потребности, необходимо *интегрировать* данные из автономных, независимо администрируемых информационных источников, как правило, слабо связанных между собой или обладающих разной моделью описания данных. Как следствие, задача управления данными с неоднородной, перманентно изменяющейся или частично неизвестной структурой становится практически важной и актуальной. Такие данные принято называть слабоструктурированными или полуструктурированными. Далее чаще используется второе из этих определений.

Одним из результатов в задаче поиска информации по разнородным источникам данных является новая абстракция управления информацией, именуемая *пространством данных*[1,2,3]. Цель поддержки пространства данных состоит в обеспечении базового набора функций над всеми источниками данных. Понятие пространства данных предполагает поддержку нескольких уровней доступа к информации, отличающихся степенью детальности описания данных. На самом нижнем уровне поддерживается поиск по ключевым словам. При обработке более сложных поисковых запросов, учиты-

[1] M. Franklin, A. Halevy, D. Maier, *From databases to dataspaces: A new abstraction for information management* // SIGMOD Rec., **34**(4), 27–33, ACM Press, 2005

[2] M. Franklin, A. Halevy, D. Maier, *Principles of dataspace systems* // Proc. of the 25th SIGMOD symp. on Principles of database systems, 1–9, ACM Press, 2006

[3] M.A. Vaz Salles, J.-P. Dittrich, S. Karakashian et al, *iTrails: pay-as-you-go information integration in dataspaces* // Proc. of the 33rd int. conf. on Very large data bases, 663–674, VLDB Endowment, 2007

вающих структуру документов, могут применяться модели как полуструктурированных данных, рассматриваемых в настоящей диссертации, так и модели реляционных данных. При этом следует принимать во внимание, чем выше уровень, тем сложнее организация доступа к данным.

Формальной моделью полуструктурированных данных является модель регулярных путевых запросов, предполагающая описание данных в виде графовых структур. Вычисление регулярных путевых запросов является трудоемкой операцией. Однако, применение ряда механизмов оптимизации, в том числе рассматриваемых в качестве основных в настоящей работе материализованных представлений, позволяет во многих случаях значительно сократить время вычисления запросов. Исследованиям на этом направлении посвящены работы большого числа зарубежных ученых, отмеченные в тексте диссертации и в списке литературы. Актуальность данной работы определяется тем обстоятельством, что существующие алгоритмы построения перезаписи регулярных путевых запросов по заданной системе представлений позволяют построить максимальную перезапись, которая обладает значительной избыточностью. Вычисление максимальной перезаписи может быть эквивалентно вычислению исходного запроса даже в том случае, когда представления в явном виде содержат результат вычисления запроса.

Цель и задачи работы. Целью данной диссертации является разработка алгоритмов построения однословной перезаписи регулярных путевых запросов по заданной системе представлений при наличии дополнительных ограничений на структуру перезаписи. Для достижения этой цели решаются следующие задачи.

1. Формализуется понятие ограничения на структуру перезаписей и разрабатывается алгоритм построения однословной перезаписи при наличии дополнительного ограничения.

2. Разрабатывается механизм оценки выразительной силы набора представлений, а также выразительной силы набора представлений при наличии ограничений.

3. Разрабатываются алгоритмы построения всех возможных однословных перезаписей запроса.

4. Алгоритмы реализуются в виде программ и их работоспособность демонстрируется в ходе тестовых испытаний.

Научная новизна. С учетом введенного в диссертации понятия однословной перезаписи регулярных путевых запросов и формулировки основной задачи в терминах полугрупп регулярных языков, научной новизной обладает следующие ее результаты. Решена задача принадлежности регулярного языка рациональному подмножеству полугруппы регулярных языков, что является обобщением результатов Хашигучи[4]. Сформулировано условие конечности полугруппы и рационального множества регулярных языков, которое существенно сильнее критериев конечности[5,6], справедливых для произвольных полугрупп. Исследована структура полугрупп регулярных языков, доказана регулярность класса эквивалентности. Доказано, что полугруппы регулярных языков над однобуквенным алфавитом являются рациональными, автоматными и полугруппами Клини.

Практическая ценность. Результаты работы могут применяться при разработке систем управления полуструктурированными данными, в основе которых лежит графовая модель их описания, в частности, в системах интеграции данных и в информационно-поисковых системах, учитывающих логическую структуру документов. Предложенные алгоритмы могут быть использованы для повышения эффективности вычисления регулярных путевых запросов.

Апробация работы и публикации. Основные положения работы докладывались на международной конференции «Developments in Language Theory», Палермо, Италия (2005 г.), на международной конференции «Semigroups and Languages», Лиссабон, Португалия (2005 г.), на международной конференции «Automata and Formal Languages», Добогоко, Венгрия (2005 г.), на международной конференции «Актуальные проблемы вычислительной математики» посвященной памяти академика Н.С. Бахвалова (2006 г.), на механико-математическом факультете МГУ им. М. В. Ломоносова на семинаре «Проблемы современных информационно-вычислительных систем» под руководством проф. В. А. Васенина (2005 г., 2007 г.), на международной конференции «Automata and Formal Languages», Балатонфюред, Венгрия (2008 г.), на международной конференции «Мальцевские чтения», Новосибирск (2008 г.), на кафедре Математической теории интеллектуальных систем механико-математического факультета МГУ им. М. В. Ломоносова

[4]K. Hashiguchi, *Representation theorems on regular languages* // Journal of computer and system sciences, **27**, 101–115, 1983

[5]A. de Luca, A. Restivo,*A finiteness condition for finitely generated semigroups* // Semigroup forum **28**, 123–134, 1984

[6]S. Varricchio, *A finiteness condition for finitely generated semigroups* // Semigroup Forum, **38**(1), 331–335, Springer New York, 1989

на семинаре «Дискретный анализ» под руководством проф. С. В. Алешина, проф. В. А. Буевича, с.н.с. М. В. Носова (2008 г.).

По теме диссертации опубликовано 6 печатных работ, в том числе две из них [3,5] в журналах, внесенных в список ВАК.

Основные результаты работы

В ходе выполнения настоящей работы автором получены следующие основные результаты.

1. Предложен метод описания ограничения на структуру перезаписи в виде рационального множества регулярных языков и разработан алгоритм проверки их принадлежности рациональному множеству регулярных языков.

2. Сформулированы необходимое и достаточное условия конечности полугруппы и рационального множества регулярных языков.

3. Доказана регулярность класса эквивалентности конечно порожденной полугруппы регулярных языков. Разработан алгоритм, позволяющий построить регулярный язык, который соответствует классу эквивалентности полугруппы.

4. Исследованы свойства полугрупп регулярных языков над однобуквенным алфавитом, доказана их рациональность. Разработан алгоритм, выписывающий автоматную структуру для данных полугрупп.

5. Программно реализованы алгоритмы построения однословной перезаписи регулярного путевого запроса. В ходе их тестовых испытаний продемонстрировано, что вычислительная сложность алгоритмов может быть существенно снижена.

Структура и объем работы. Работа состоит из введения, четырех глав, заключения, приложения и списка литературы. Общий объем диссертации 135 страниц. Список литературы содержит 68 наименований.

Краткое содержание работы

Во **введении** описываются цели и задачи работы, обосновывается ее актуальность. Представлены основные результаты диссертации, ее научная новизна и практическая значимость.

Первая глава является вводной, в ней рассматриваются задачи, которые возникают при вычислении запросов с использованием представлений. К их числу относятся:

- выразимость запроса через набор представлений;
- эквивалентность перезаписей запросов;
- пустота пересечения множеств запросов;
- конечность множества запросов, выводимых из набора представлений;
- эквивалентность множеств запросов.

Представление (от английского слова view) - это запрос к базе данных, написанный на некотором заранее определенном языке, именуемом языком представлений, который может использоваться при формировании других запросов к базе данных.

В главе также рассматриваются возможные области практического применения перечисленных выше задач. К их числу относятся: виртуальная интеграция в единую базу данных из источников с разными моделями их описания; семантическое кеширование, которое широко используется в клиент-серверных, OLAP-системах и мобильных вычислениях; дополнительные механизмы логического разграничения доступа к данным с целью обеспечения их конфиденциальности. Вводится модель полуструктурированных данных, описывается формальная математическая постановка задач, которые составляют содержание диссертации.

После введения основных определений и понятия матриализованного представления в виде пары (запрос; результат его вычисления), решаемые задачи формулируются в общей постановке, вне зависимости от конкретного языка запросов и модели данных. Далее приводится используемая в работе модель описания полуструктурированных данных. В соответствии с этой моделью данные представляются коллекцией объектов. Формальной математической моделью полуструктурированных данных является помеченный ориентированный граф вида:

$$\mathcal{B} = (V, E, v),$$

где V — вершины графа, соответствующие объектам, v — помечающая функция, которая сопоставляет каждому объекту его значение, $E \subseteq V \times \Sigma \times V$ — отношение инцидентности, где Σ — множество меток.

В такой модели задача поиска по базе полуструктурированных данных состоит в нахождении вершин ориентированного графа, связанных определенными условиями. В качестве базового механизма для языков запросов к полуструктурированным данным используются путевые запросы. Основная идея, положенная в его основу, состоит в том, что структура данных и часть самих данных описываются путями между вершинами.

Каждому ребру, соединяющему вершины x, y, в графе $\mathcal{B}$ сопоставлен некоторый символ $a \in \Sigma$, при этом $(x, a, y) \in E$. Таким образом, каждому пути

$$(x_1, a_1, x_2), (x_2, a_2, x_3), \ldots, (x_n, a_n, x_{n+1}) \in E, \tag{1}$$

соединяющему вершины x_1, x_{n+1}, приписано слово $a_1 a_2 \ldots a_n$ в алфавите Σ. Следовательно, в основу запросов к полуструктурированным данным можно положить ограничения, которые накладывются на слова, приписанные путям, соединяющим вершины графа, формализующего полуструктурированные данные.

В работе рассматриваются *регулярные путевые запросы*, в которых ограничения задаются регулярными выражениями. Пусть $L \subseteq \Sigma^*$ — регулярный язык. *Результатом* данного запроса будут все возможные пары вершин x, y графа $\mathcal{B}$, для которых существует путь вида (1), где $x = x_1, y = x_{n+1}$ и слово $a_1 a_2 \ldots a_n$ принадлежит регулярному языку L. Регулярные выражения позволяют формулировать запросы, не имея полной информации о структуре данных. В ответ на запрос, представленный регулярным выражением вида $a\Sigma^* b$, будут найдены пары вершин, которые соединены путями, начинающимися с буквы a и заканчивающиеся буквой b.

Для ответов на запрос с использованием набора представлений в работе используется метод построения перезаписи. *Перезапись* — это разложение запроса по системе представлений. Моделью представления данных в работе принимается модель полуструктурированных данных с регулярными языками в качестве путевых запросов. В этой связи, если для построения перезаписи использовать только одну операцию, то множество запросов, которое покрывается конечным множеством материализованных преставлений, является полугруппой регулярных языков.

Введем определения. Пусть Δ и Σ два непересекающихся конечных алфавита, которые будем называть алфавитом представлений и алфавитом базы данных, соответственно. Через $Reg(\Sigma)$ обозначим множество регулярных языков над алфавитом Σ. Регулярная подстановка φ — это гомоморфизм $\Delta^+ \to Reg(\Sigma)$. Таким образом, регулярная подстановка определяет

полугруппу $\mathcal{S}_\varphi = \langle\{\varphi(\delta) \mid \delta \in \Delta)\}\rangle$.

Если породающие элементы конечно порожденной полугруппы запросов рассматривать как символы алфавита Δ, то любой элемент полугруппы может быть представлен в виде слова над этим алфавитом. Полугруппа задается языком Δ^* и регулярной подстановкой. Определим рациональное множество запросов, как множество регулярных языков, для которого существует регулярный язык над Δ и регулярная подстановка, которые его задают.

Определение (Рациональное множество). Множество $\mathcal{R}$ регулярных языков над Σ называется *рациональным*, если существует конечный алфавит Δ, регулярный язык $K \subseteq \Delta^+$, и регулярная подстановка $\varphi : \Delta^+ \to Reg(\Sigma)$ такая, что

$$\mathcal{R} = \{\varphi(w) \mid w \in K\}.$$

Таким образом, рациональные множества являются рациональными подмножествами конечно порожденных полугрупп регулярных языков. Будем считать, что пара (K, φ) является *представлением рационального множества* и обозначать его через $\mathcal{R} = (K, \varphi)$.

В первой главе описываются существующие методы построения перезаписей (максимальная, частичная, точная). Рассматриваются случаи, в которых максимальная перезапись оказывается избыточной. По этой причине в работе вводится понятие однословной перезаписи и перечисленные ранее задачи решаются с их использованием. В контексте полуструктурированных баз данных перезапись запроса по конечному набору представлений является однословной, если набор операций над представлениями содержит лишь одну операцию (условно названную композицией). Последовательность букв этого слова определяет композицию представлений.

Определение. Пусть $\varphi : \Delta^+ \to Reg(\Sigma)$ регулярная подстановка. *Однословной перезаписью* регулярного языка $R \in Reg(\Sigma)$ по подстановке φ называется слово $w \in \Delta^+$ такое, что $\varphi(w) = R$.

Далее приведены формальные постановки задач, решаемых в диссертации, а также описываются области приложений, в которых они возникают.

Перезапись запросов как задача принадлежности для полугруппы регулярных языков. В области интеграции данных основной задачей является виртуальное объединение в единую базу(хранилище) данных из разных источников с целью предоставления пользователю унифици-

рованных представлений[7]. Обобщенная модель системы интеграции данных включает: глобальную схему данных, схемы источников данных и отображение между ними. Локальные схемы описывают структуру источников, где расположены реальные данные. Глобальная схема предоставляет согласованное виртуальное описание всех данных на ее основе.

Один из методов интеграции носит название LAV (local as view). Особенностью этого метода является тот факт, что локальные источники данных L_i описываются как представления глобальной базы данных G. Если система интеграции получает запрос к глобальной базе данных, то необходимо его представить в терминах представлений локальных источников. Такое задание можно описать следующим образом: для каждого локального источника L_i существует некоторый запрос R_i к глобальной базе данных такой, что верно $L_i = R_i(G)$. Для того, чтобы вычислить произвольный запрос M, поступивший извне, необходимо разложить его через запросы $\{R_i\}$. Если предположить, что доступна только одна операция над запросами, которую условно назовем композицией ($\circ$), то необходимо найти такой набор $l_1, ...l_k$, что $M = R_{l_1} \circ R_{l_1} \ldots \circ R_{l_k}$, либо констатировать, что такого разложения не существует.

Вопрос перезаписи запросов возникает также в семантическом кешировании. Семантическое кеширование широко используется в клиент-серверных системах[8], OLAP системах[9], мобильных вычислениях[10] и гетерогенных системах[11]. Семантическое кеширование отличается тем обстоятельством, что кешируются результаты запросов, а не строки базы данных или страницы. В данном случае кеш-память состоит из множества элементов, связанных семантическим описанием. Одним из этапов семантического кеширования является проверка того, можно ли поступивший запрос M выразить через те запросы $\{R_i\}$, результаты вычисления которых уже известны и хранятся в кеш-памяти системы.

С помощью представлений обеспечиваются дополнительные механизмы защиты данных. Пользователю могут, например, предоставляться права на

[7]Lenzerini M., *Data Integration: A Theoretical Perspective* // Proc. of the 21th SIGMOD Conf., 233–246, ACM, 2002

[8]Shaul Dar, Michael J. Franklin, Björn T. Jónsson, Divesh Srivastava, Michael Tan, *Semantic Data Caching and Replacement* // VLDB '96: Proceedings of the 22th International Conference on Very, 330–341, 1996

[9]Prasad M. Deshpande, Karthikeyan Ramasamy, Amit Shukla, Jeffrey F. Naughton, *Caching multidimensional queries using chunks*, 259–270, 1998

[10]Ken. C. K. Lee, H. V. Leong, Antonio Si, *Semantic query caching in a mobile environment* // SIGMOBILE Mob. Comput. Commun. Rev., **3**(2), 28–36, 1999

[11]Parke Godfrey, Jarek Gryz, *Semantic Query Caching for Hetereogeneous Databases* // Knowledge Representation Meets Databases, 6.1-6.6, 1997

доступ к данным только через представления, благодаря чему он не будет иметь доступа к данным, хранящимся в основной базе, которые, согласно принятой политике безопасности, не предназначены для него. Ответы на некоторые другие запросы, согласно положений той же политики, могут быть конфиденциальными и для этой категории пользователей. В таком случае описанные механизмы могут потребовать обоснования того, что раскрытые представления не дают информации, позволяющей вычислять такие конфиденциальные запросы. Здесь возникает задача по постановке и технике ее решения близкая к задачам интеграции и кеширования. Ее суть в том, чтобы уметь проверять, принадлежит ли какой-либо запрещенный запрос M множеству запросов, вычислимых на основании разрешенных представлений $\{R_i\}$[12].

Перечисленные три задачи сводятся к одной, которую можно формализовать следующим образом. Как отмечалось ранее, конечный набор регулярных путевых запросов можно представлять конечным множеством регулярных языков, а для эффективности вычисления перезаписей область поиска можно ограничивать только однословными перезаписями. Если предполагать, что композиция запросов — это конкатенация регулярных языков, то описанная задача сводится к проверке принадлежности регулярного языка полугруппе.

Задача. Пусть $\varphi : \Delta^+ \to Reg(\Sigma)$ — регулярная подстановка, $R \in Reg(\Sigma)$ — произвольный регулярный язык. Существует ли алгоритм проверки принадлежности языка R полугруппе $\mathcal{S}_\varphi$, порожденной регулярной подстановкой φ? Можно ли найти слово $w \in \Delta^+$ такое, что верно $\varphi(w) = R$?

В данной задаче вопрос существования алгоритма проверки принадлежности является более существенным, чем вопрос поиска однословной перезаписи. В случае существования алгоритма проверки, перезапись может быть найдена конечным перебором. Однако в данном случае значительную роль играет и сложность решения задачи. По этой причине следует рассматривать оба вопроса. Данная задача решена Хашигучи[13] и описана в первой главе.

Поиск перезаписи с дополнительными ограничениями. В задачах перезаписи запроса через набор представлений может появиться необхо-

[12] В более общей форме задачу можно ставить как проверку пустоты пересечения множества запрещенных запросов с множеством запросов, вычислимых на основании набора разрешенных представлений. Однако в контексте настоящей работы эта задача не рассматривается

[13] K. Hashiguchi, *Representation theorems on regular languages* // Journal of computer and system sciences, **27**, 101–115, 1983

димость вводить некоторые ограничения на структуру искомой перезаписи. Например, в задачах интеграции и кеширования такая необходимость может быть связана с временными ограничениями на выполнение запросов, в области информационной безопасности — непосредственно с моделями логического разграничения доступа к защищаемым данным.

В данной работе предлагается использовать рациональные множества регулярных языков в качестве ограничений на структуру перезаписей. В таком случае описанные задачи могут быть сформулированы следующим образом.

Задача. Пусть $\mathcal{R} = (K, \varphi)$ — рациональное множество регулярных языков над Σ и $R \subseteq \Sigma^*$ — произвольный регулярный язык. Существует ли алгоритм проверки принадлежности языка R множеству $\mathcal{R}$? Можно ли найти слово $w \in \Delta^*$ такое, что оно принадлежит языку K и верно $\varphi(w) = R$?

Решение данной задачи представлено во второй главе диссертации.

Применение задачи равенства слов при перезаписи запросов. Два пользователя системы хранения данных могут задать один и тот же запрос по-разному. В тоже время, перезаписи запросов через представления могут отличаться по виду. Если система позволяет сравнивать перезаписи запросов, то в случае поступления запроса, идентичного поступившему ранее, новых вычислений не потребуется. Таким образом, с использованием идеи проверки эквивалентности перезаписей при кешировании данных возможно сокращение времени обработки запросов. В модели полуструктурированных баз данных вопрос сравнения однословных перезаписей запросов формально звучит как задача равенства слов в полугруппе, где полугруппа порождена конечным числом регулярных языков.

В общем случае, когда полугруппа задана определяющими соотношениями на порождающие элементы, а именно —

$$\mathcal{S} = \langle \Delta \mid v_i = u_i, v_i, u_i \in \Delta^*, i \in [1, \ldots, k] \rangle,$$

задача равенства слов является неразрешимой[14]. В рассматриваемом случае, когда порождающие элементы являются регулярными языками, задача равенства слов разрешима. Для сравнения слов полугруппы, достаточно построить соответствующие автоматы и проверить их на равенство. Однако, и это необходимо отметить, задача эквивалентности недетерминированных конечных автоматов является NP-полной относительно размера автоматов.

[14] А. А. Марков, Н. М. Нагорный, *Теория алгорифмов*, Наука, 1984

Вместе с тем, существуют полугруппы, которые допускают решение задачи равенства слов за полиномиальное время. Например, автоматные полугруппы допускают ее решения за квадратичное время относительно длины слов. Для того, чтобы дать определение автоматности, введем отображение $\theta : \Delta^+ \times \Delta^+ \to \Delta(2,\$)^*$, где $\Delta(2,\$) = ((\Delta \cup \{\$\}) \times (\Delta \cup \{\$\})) - \{(\$,\$)\}$, следующим образом:

$$\theta(u,v) \rightleftharpoons \begin{cases} (\delta_1,\delta'_1)\dots(\delta_n,\delta'_n), & \text{если } m = n; \\ (\delta_1,\delta'_1)\dots(\delta_n,\delta'_n)(\$,\delta'_{n+1})\dots(\$,\delta'_m), & \text{если } n < m; \\ (\delta_1,\delta'_1)\dots(\delta_m,\delta'_m)(\delta_{m+1},\$)\dots(\delta_n,\$), & \text{если } m < n. \end{cases}$$

Определение. Пусть Δ конечный алфавит, $\mathbb{S}$ конечно порожденная полугруппа. Пусть $L \subseteq \Delta^*$ регулярный язык, $\varphi : \Delta^+ \to 2^{\Sigma^*}$ регулярная подстановка такие, что $\varphi(L) = \mathbb{S}$. Пара (Δ, L) является *автоматной структурой* для $\mathbb{S}$, если:

$L_= = \{\theta(u,v) \mid u,v \in L, \varphi(u) = \varphi(v)\}$ регулярный язык в $\Delta(2,\$)^*$;

$L_\delta = \{\theta(u,v) \mid u,v \in L, \varphi(u\delta) = \varphi(v)\}$ регулярный язык в $\Delta(2,\$)^*$ для любого $\delta \in \Delta^*$.

Полугруппа называется *автоматной*, если она обладает автоматной структурой.

С помощью автоматной структуры происходит сравнение слов. Поскольку автоматная структура может быть построена заранее, то ее можно считать своего рода индексом для задачи равенства слов. В настоящей работе делается предположение, что полугруппы регулярных языков являются автоматными, однако доказывается оно только для полугрупп регулярных языков над однобуквенным алфавитом.

Оценка выразительной силы набора представлений. Возвращаясь к вопросу о кешировании данных, следует отметить, что одна из возникающих в данной области задач — это минимизация объема содержимого кеш-памяти с сохранением его выразительной силы. Потребность уменьшать размер кеш-памяти появляется в целом ряде практически значимых задач. Как следствие, возникает необходимость проверки эквивалентности наборов представлений. Переходя от абстрактных запросов и данных к регулярным путевым запросам, к полуструктурированной модели данных, задача об эквивалентности наборов представлений формально может быть сформулирована следующим образом.

Задача. Пусть $\mathfrak{S}_\varphi$ и $\mathfrak{S}_\psi$ — две полугруппы, заданные регулярными подстановками $\varphi : \Delta_1 \to Reg(\Sigma)$ и $\psi : \Delta_2 \to Reg(\Sigma)$. Можно ли проверить вложенность полугрупп?

В случае, когда верно вложение полугрупп $\mathfrak{S}_\varphi \subseteq \mathfrak{S}_\psi$, система представлений заданная подстановкой ψ обладает большей выразительной силой, чем система представлений заданная φ.

При существовании ограничений на выполнение запросов в процессе интеграции данных, задача проверки эквивалентности множества разрешенных запросов формально может быть представлена следующим образом.

Задача. Пусть (K_1, φ) и (K_2, ψ) — рациональные множества, заданные регулярными подстановками $\varphi : \Delta_1 \to Reg(\Sigma)$, $\psi : \Delta_2 \to Reg(\Sigma)$ и регулярными языками $K_1, K_2 \in Reg(\Delta)$. Можно ли проверить вложенность рациональных множеств (K_1, φ) и (K_2, ψ)?

В случае полугрупп задача решается путем проверки выразимости порождающих элементов одной полугруппы через порождающие элементы другой. В случае рациональных множеств, данная задача не разрешима, что доказывается во второй главе.

Оценка выразительной силы набора представлений может быть получена с помощью множества запросов, которое покрывается данным набором представлений. Возникает следующая формальная задача.

Задача. Пусть $\mathcal{R} = (K, \varphi)$ — рациональное множество регулярных языков над Σ, заданное регулярным языком $K \in Reg(\Delta)$ и регулярной подстановкой $\varphi : \Delta \to Reg(\Sigma)$. Существует ли алгоритм проверки конечности рационального множества $\mathcal{R} = (K, \varphi)$?

Разрешимость данной задачи доказывается во второй главе настоящей работы.

В заключительном разделе первой главы перечислены некоторые типы полугрупп с эффективно решаемой задачей равенства слов, такие как автоматные, рациональные. Приводится также определение полугрупп Клини.

Во **второй главе** решена задача выразимости запроса через набор представлений при наличии ограничений на структуру перезаписи, а также задача оценки выразительной силы набора представлений. Доказано, что множество всех однословных перезаписей может быть эффективно построено.

В первом разделе представлено предложенное автором доказательсво того факта, что задача принадлежности регулярного языка рациональному

множеству разрешима. В контексте полуструктурированных баз данных, когда запрос представляется регулярным языком, а множество материализованных представлений конечно, вопрос о вычислимости запроса через представления эквивалентен вопросу о принадлежности регулярного языка рациональному множеству. Как уже отмечалось ранее, задача проверки принадлежности регулярного языка полугруппе разрешима[15]. В данной главе доказывается следующая теорема.

Теорема 1. Задача принадлежности регулярного языка $R \subseteq \Sigma^*$ рациональному множеству $\mathcal{R} = (K, \varphi)$ регулярных языков над Σ разрешима.

Доказательство данной теоремы является конструктивным. В первом разделе главы приводится алгоритм проверки принадлежности регулярного языка рациональному множеству и анализируется сложность данного алгоритма.

Во втором разделе доказывается, что для любой полугруппы $\mathcal{S}$ регулярных языков и любого слова $w \in \mathcal{S}$ данной полугруппы верно утверждение: множество слов, равных w, является регулярным.

Теорема 2. Пусть $\varphi : \Delta^+ \to Reg(\Sigma)$ — регулярная подстановка, $\mathcal{S}_\varphi$ — конечно порожденная полугруппа регулярных языков, w — слово в алфавите Δ. Множество

$$[w] = \{u \in \Delta^+ \mid \varphi(u) = \varphi(w)\}$$

является регулярным языком над Δ.

Множество $F = [w]$ может быть построено с помощью следующего алгоритма (на вход подается слово $w \in \Delta^+$).

- Положить $K = \Delta^+$, $F = \varnothing$.
- Повторять следующие шаги пока $\varphi(w) \in (K, \varphi)$:
  - найти самое короткое слово $v \in K$ такое, что $\varphi(v) = \varphi(w)$;
  - добавить слово v в F;
  - положить $K = K \setminus E(v, \Delta_0)$.

В работе доказывается корректность данного алгоритма и анализируется его сложность. Регулярность класса эквивалентности позволяет выбирать

[15] K. Hashiguchi, *Representation theorems on regular languages* // Journal of computer and system sciences, **27**, 101–115, 1983

оптимальную перезапись регулярного путевого запроса через материализованные представления. Сформулируем алгоритм поиска оптимальной перезаписи. Первый шаг — проверка принадлежности рациональному множеству и поиск произвольной однословной перезаписи. Второй шаг — построение класса эквивалентности найденной перезаписи. Третий шаг — при использовании *функции оценки сложности* f поиск слова (в классе эквивалентности найденной перезаписи), на котором f имеет минимальное значение.

При последовательном поступлении эквивалентных запросов, являющихся словами над алфавитом представлений, полезно иметь возможность вычислять только один из них, а на второй уже выдавать результат не производя вычислений. Предложенный алгоритм поиска оптимальной перезаписи позволяет решать данную задачу. Он требует построения класса эквивалентности для каждого нового запроса и основывается на проверке ограниченности автомата расстояния. Эта задача, в свою очередь, является PSPACE-полной. Более эффективный метод решения задачи эквивалентности запросов может основываться на решении задачи о равенстве слов в том случае, если она не является вычислительно сложной. Данный вопрос рассматривается в следующей главе.

В третьем разделе решается задача проверки конечности полугруппы и конечности рационального множества. Для оценки выразительной силы набора представлений можно оценивать множество запросов, которое покрывается набором представлений. Таким образом, становится актуальной задача определения конечности рационального множества регулярных языков.

Регулярный язык L обладает *свойством конечной степени*, если существует натуральное число $k \in \mathbb{N}$ такое, что $L^* = L^k$. В 1966 Брозовский поставил вопрос о разрешимости данной проблемы. Независимо друг от друга положительные ответы дали Хашигучи[16] и Саймон[17]. Будем писать $fpp(L) = p$ или $fpp(L) < \infty$ если $L^* = L^p$, и $fpp(L) = \infty$, если такого числа не существует.

Теорема 3. Пусть $\varphi : \Delta^+ \to Reg(\Sigma)$ — регулярная подстановка. Конечно порожденная полугруппа $\mathcal{S}_\varphi$ конечна тогда и только тогда, когда для любого

[16] K. Hashiguchi, *Limitedness Theorem on Finite Automata with Distance Functions* // Journal of computer and system sciences, **24**, 233–244, 1982

[17] I. Simon, *Limited subsets of a free monoid* // Proceedings of the 19st Annual Symposium on Foundations of Computer Science, 143–150, 1978

m-подмножества $\{\delta_1, \ldots, \delta_m\} \subseteq \Delta$ $(m = 1, \ldots, |\Delta|)$:

$$fpp(\varphi(\delta_1\delta_2 \ldots \delta_m)) < \infty.$$

Следует отметить, что полученное условие накладывает существенное ограничение на структуру соотношений в полугруппе. Так, например, полугруппа $\mathcal{S}$, заданная представлением $S = \langle \Delta \mid x^3 = x^2 \text{ для всех } x \in \Delta^* \rangle$ является бесконечной, если алфавит содержит по крайней мере две буквы[18]. Однако, полугруппа регулярных языков, удовлетворяющих соотношению $x^3 = x^2$ конечна по Теореме 3 (из соотношения $x^2 = x^3$ следует, что $fpp(\varphi(x)) = 2$). Этот факт означает, что полугруппа регулярных языков должна удовлетворять дополнительным соотношениям, которые являются следствием структуры регулярных языков.

В третьем разделе данной главы также доказывается следующая теорема.

Теорема 4. Задача конечности рационального множества $\mathcal{R} = (K, \varphi)$ разрешима.

В третьем разделе анализируется сложность алгоритмов проверки принадлежности регулярного языка полугруппе и рациональному множеству регулярных языков.

В четвертом разделе второй главы доказывается следующая теорема.

Теорема 5. Задача вложенности рациональных множеств регулярных языков не разрешима.

Третья глава посвящена решению задачи о равенстве слов для полугрупп регулярных языков.

В модели полуструктурированных баз данных вопрос сравнения однословных перезаписей запросов формально может быть представлен как задача о равенстве слов в полугруппе, которая порождена конечным числом регулярных языков. Автоматные полугруппы допускают решения задачи за квадратичное время относительно длины слов. В данной главе автором делается предположение, что полугруппы регулярных языков являются автоматными. Приводится пример полугруппы регулярных языков над многобуквенным алфавитом, которая является автоматной. Доказывается, что в случае однобуквенного алфавита полугруппа всегда является автоматной и

[18] J.A Brzozowski, K. Culik, A. Gabrielian, *Classification of noncounting events* // Journal of computer and system sciences, **5**, 41–53, 1971

рациональной. Данный факт и существование автоматных полугрупп в общем случае позволяет сделать предположение, что полугруппы регулярных языков являются автоматными и в общем случае, однако данный вопрос остается открытым.

Рассмотрим полугруппу $\mathcal{S}$ порожденную языками

$$x = (a+b)^*a, \;\; y = \varepsilon + a + b, \text{ и } z = b^*$$

над $\Sigma = \{a, b\}$. В работе доказывается, что полугруппа $\mathcal{S}$ является автоматной. Для этого строятся представление $\mathcal{S}$

$$\langle x, y, z \mid yx = x, zx = x, z^2 = z, xzy = xz, xy^k z = xz \; (k \geqslant 1) \rangle$$

и автоматы для $L_=$ и L_x,L_y,L_z.

Во второй главе доказывается, что ядро регулярной подстановки не обязательно регулярно. По этой причине в общем случае (A, Δ^+) не является автоматной структурой для полугруппы регулярных языком. Тем не менее, выдвигается следующее предположение.

Предположение. Каждая полугруппа регулярных языков является автоматной.

Несмотря на то, что автоматные структуры с помощью несложных рассуждений могут быть построены для некоторых полугрупп регулярных языков, это предположение не является тривиальным. Например, пусть $\mathcal{S}$ полугурппа конечных языков. Каждый конечный язык может быть разложен в конкатенацию простых языков, однако это разложение не единственно ($\{\varepsilon, a, a^2, a^3\} = \{\varepsilon, a\}^3 = \{\varepsilon, a\}\{\varepsilon, a^2\}$). Таким образом, простые факторы могут удовлетворять нетривиальным соотношениям. В общем случае, ситуация еще более сложна.

В первом разделе данной главы доказывается, что рациональные полугруппы являются автоматными.

Теорема 6. Рациональная полугруппа является автоматной.

Таким образом, возможным направлением доказательства автоматности полугруппы могло бы быть доказательство утверждения о ее рациональности. Однако в работе доказывается, что полугруппы регулярных языков рациональными не являются.

Теорема 7. Не все полугруппы, порожденные конечным множеством регулярных языков, являются рациональными.

Следует заметить, что в частном случае, когда порождающие элементы полугруппы являются регулярными языками над однобуквенным алфавитом, полугруппы являются автоматными, полугурппами Клини и рациональными. Данные факты доказываются в третей главе диссертационной работы.

Доказательство основано на том, что коммутативные группы Клини являются рациональными[19], а полугруппа является конечно порожденной тогда и только тогда, когда распознаваемое множество является рациональным[20].

Определение. Для любого языка $L \subseteq \Delta^*$, любой регулярной подстановки φ язык $[L]_\varphi = \{w \in \Delta^* \mid \exists u \in L : \varphi(u) = \varphi(w)\}$ называется *замыканием* языка L.

В работе доказываются следующие теоремы.

Теорема 8. Для конечно порожденной полугруппы $\mathcal{S}$ и регулярной подстановки $\varphi : \Delta^+ \to \mathcal{S}$ верно: если для любого регулярного языка L над алфавитом Δ замыкание $[L]_\varphi$ является регулярным языком, то полугруппа $\mathcal{S}$ является полугруппой Клини.

Теорема 9. Пусть $|\Sigma| = 1$. Для конечно порожденной полугруппы $\mathcal{S}$, для любого регулярного языка $L \subseteq \Delta^*$ и регулярной подстановки $\varphi : \Delta^+ \to \mathcal{S}$ замыкание $[L]_\varphi$ является регулярным языком.

Таким образом доказывается, что полугруппы регулярных языков над однобуквенным алфавитом являются полугруппами Клини, а поскольку они являются коммутативными, то доказывается, что они также являются рациональными.

Ранее предполагалось, что произвольные коммутативные полугруппы являются автоматными[21]. Однако оказалось, что существует коммутативная полугруппа не являющаяся автоматной[22]. Существование иных соотношений, кроме соотношений коммутативности, диктуется структурой порождающих элементов полугруппы, которые в нашем случае задаются автоматами. В многобуквенном случае регулярных языков полугруппы коммутативными не являются. Вместе с тем, специфика строения порождающих

[19] C. P. Rupert, *On commutative Kleene monoids* // Springer New York, Semigroup Forum, **43**(2), 163-177, 1991

[20] R. Gilmer, *Commutative Semigroup Rings* // University of Chicago, Chicago, 1984

[21] Colin M. Campbell, Edmund F. Robertson, Nikola Ruškuc, Richard M. Thomas, *Automatic semigroups* // Journal of Theoretical Computer Science, **250** (1–2), 365–391, 2001

[22] Michael Hoffmann, Richard M. Thomas, *Automaticity and commutative semigroups* // Glasgow Mathematical Journal, **44**, 167–176, 2002

элементов позволяет сделать предположение, что полугруппы регулярных языков являются автоматными и в общем случае.

В **четвертой главе** описывается созданная автором на языке Mathematica программная реализация алгоритмов, представленных в работе. Программные средства использовались в ходе тестовых испытаний разработанных алгоритмов.

Программный комплекс включает в себя следующие подсистемы:

- подсистема вычисления запроса;
- подсистема построения максимальной перезаписи запроса;
- подсистема построения однословной перезаписи запроса.

Сложность построения максимальной перезаписи запроса через представления является экспоненциальной относительно размера автомата, представляющего язык запроса. В предложенном в работе алгоритме для нахождения однословной перезаписи требуется построить автомат расстояния из максимальной перезаписи и проверить его ограниченность. Для того, чтобы проверить его ограниченность, необходимо воспользоваться константой Хашигучи

$$2^{4*n^3+n*\log(n+2)+n},$$

где n — это число состояний автомата расстояния. Константа Хашигучи ограничивает длину слов, которые могут являться однословной перезаписью запроса.

Ограничение Хашигучи на длину перезаписи не позволяет эффективно решать задачу построения однословной перезаписи. Однако, анализ проведенных экспериментов позволяет сделать следующий вывод: длину перезаписей можно ограничивать меньшей константой, нежели константа Хашигучи; алгоритмы построения однословных перезаписей таким образом можно реализовать на практике.

В **заключении** формулируются результаты, полученные в рамках настоящей диссертационной работы.

Благодарности

Автор глубоко благодарен научному руководителю — доктору физико-математических наук, профессору Васенину Валерию Александровичу и кандидату физико-математических наук Афонину Сергею Александровичу за постановку задач, обсуждение результатов и постоянное внимание к работе.

Список опубликованных работ по теме диссертации

1. «О представлении регулярных языков в виде конкатенации заданных», Е.Е.Хазова // Информационные технологии и программирование: Межвузовский сборник статей. Вып. 3(8), 23–38, - М.:МГИУ, 2003.

2. «Membership and Finiteness Problems for Rational Sets of Regular Languages», Afonin S., Khazova E., Lecture Notes in Computer Science // Springer-Verlag GmbH, 88–99, 2005.

   (Е.Е.Хазовой принадлежат доказательство теоремы 5 и техническая реализация теоремы 3)

3. «Membership and Finiteness Problems for Rational Sets of Regular Languages», Afonin S., Khazova E., International Journal of Foundations of Computer Science // World Scientific, 17(3), 493–506, 2006.

4. «A note on finitely generated semigroups of regular languages», Afonin S, Khazova E, International Conference «Semigroups and Languages» // World Scientific, 1–8, 2007.

   (Результаты данной статьи являются плодом совместной работы С.А.Афонина и Е.Е.Хазовой, эти результаты не могут быть разделены.)

5. «К вопросу об автоматности полугрупп регулярных языков», Хазова Е., Вестник Московского университета, сер.1, математика. механика, 6, 55–59, 2007.

6. «Semigroups of regular languages over one letter alphabet are rational», Afonin S., Khazova E., Proceedings of 12th International Conference on Automata and Formal Languages, AFL'08, 61–73, 2008.

   (Е.Е.Хазовой принадлежит доказательство основных теорем 2 и 3, представленных в данной статье.)