

УДК 004.08:004.6

РАСПРЕДЕЛЕННЫЕ СИСТЕМЫ ОБРАБОТКИ ДАННЫХ**Попов С.В.***ООО «Научно-внедренческая фирма БП+», Москва, e-mail: s-v-popov@yandex.ru*

Под распределенной информационной системой понимается множество узлов, с каждым из которых ассоциированы пользователь и база данных. Все узлы разбиты на (в общем случае пересекающиеся) подмножества – так называемые кластеры. Пользователь любого узла имеет доступ к базам данных других узлов лишь в пределах кластера, которому все они принадлежат. Следовательно, пользователь обменивается информацией с другими пользователями в пределах одного кластера и может получать информацию сразу из нескольких баз данных также в пределах одного кластера. Так как в общем случае кластеры пересекаются, то один пользователь может принадлежать нескольким кластерам. В статье предлагается способ разбиения множества узлов на кластеры, который позволяет программными средствами реализовать упомянутую методику обмена и переработки данных. Способ состоит в использовании следующих меток узлов информационной системы: каждый узел метится некоторым интервалом единичного n -мерного куба таким образом, что пересечение меток произвольного множества узлов не пусто, тогда и только тогда, когда это множество включено в один кластер. Существует простая процедура порождения меток и проверки множества меток на непустоту. В результате сравнительно просто можно на заданном множестве узлов динамически формировать кластеры, реализуя обмен данными в пределах одного кластера. Приводятся примеры представительно-го множества распределенных информационных систем, которые описываются как совокупность кластеров с информационным обменом в пределах одного кластера.

Ключевые слова: распределенная информационная система, кластеры, несовместимость узлов, семантика совместимости/несовместимости, граф несовместимости, интервал единичного куба

DISTRIBUTED DATA PROCESSING SYSTEM**Popov S.V.***LLC «Nauchno-vnedrencheskaya firma BP+», Moscow, e-mail: s-v-popov@yandex.ru*

Under the distributed information systems the set of nodes, each of which has an associated user and the database. All nodes are divided into (generally overlapping) subsets, called clusters, any node has access to the databases of other nodes within the cluster to which they belong. Therefore, the user exchanges information with other users within the same cluster and can retrieve information from several databases within the same cluster. As in the General case, the clusters intersect, then one user can belong to multiple clusters. The paper proposes a method of partitioning the multiple nodes into clusters, which allows software to implement the said method the exchange and processing of data. Method is to use the following labels of the nodes of the information system: each node is marked a certain interval of a unit n -dimensional cube in such a way that the intersection of the labels of an arbitrary set of nodes is not empty, then and only then, when this set is included in one cluster. There is a fairly simple procedure of generating the labels and check the labels on many not the void. The result is relatively easy on a given set of nodes to dynamically determine clusters and to implement the exchange of data within the same cluster. Given examples are representative of many distributed information systems, which are described as a set of clusters with the above-mentioned method of information exchange within the same cluster.

Keywords: distributed information system, clusters, the incompatibility of nodes, the semantics of compatibility/incompatibility, graph the incompatibility, the interval of a single cube

В настоящее время явно выраженный тренд развития информационных систем состоит в использовании распределенных систем обработки информации [1, 2]. Их преимущество в том, что они дают возможность каждому пользователю для решения собственных задач подключать дополнительные ресурсы, которыми он в данный момент не обладает. Современные сетевые технологии позволяют реализовать практически любые распределенные системы, не ограничивая возможностей передачи данных между узлами сетей [3, 4]. Еще один аспект, который следует учитывать при разработке и реализации распределенных информационных систем, состоит в обеспечении безопасности хранимых данных и исключении противоречий при извлечении новых данных из имеющихся. В совре-

менных сетях это достигается различными способами: аутентификацией, распределением прав доступа, разбиением сети на подсети, регулированием доступа к данным, механизмами обеспечения целостности и непротиворечивости, используемыми при формировании запросов к разным узлам, и пр. Поэтому при их разработке требуется уделять значительное внимание вопросам безопасности хранения и обработки данных. Здесь безопасность понимается в широком смысле – это не только исключение несанкционированного доступа к данным, но и исключение противоречий при использовании данных из разных узлов информационной системы.

Имеется еще один факт, который необходимо учитывать, когда мы говорим о распределенных информационных системах,

так как уже сейчас наблюдается их большая разновидность. Каждый вид распределенных информационных систем характеризуется своей методикой хранения и обработки данных, не применимой для иных типов. В связи с этим актуален вопрос построения такого способа реализации распределенных информационных систем, который учитывал бы их большое разнообразие в плане адекватного удовлетворения требований к информационному обмену и обработке данных.

Цель работы

В связи с этим цель статьи состоит в описании методики проектирования распределенных информационных систем переработки информации, которая обеспечивает распределение прав доступа пользователей к имеющимся данным по заранее заданным содержательным законам. При этом исключаются технические средства разбиения на подсети или рабочие группы. Все методы распределения доступа реализуются программно, тем самым обеспечивая возможность оперативного изменения параметров доступа к данным. То есть при различных распределениях прав доступа в распределенной информационной системе, возникающих в связи с различными обстоятельствами, происходит лишь перенастройка параметров программного обеспечения.

Распределенные информационные системы

Распределенной информационной системой (РИС) называется множество $N = \{1, 2, \dots, n\}$ узлов, i -й узел, $i \in N$, представляет собой пару, первым компонентом которой есть база данных (БД) B_i , а вторым – пользователь u_i . Под базой данных здесь подразумевается как хранилище фактических данных (например, реляционная БД), так и база знаний, которая вместе с фактическими данными содержит утверждения, истинные на имеющихся данных БД B_i (декларативные данные). Эти утверждения используются для получения новых данных, используя средства логического вывода. Тем самым декларативные данные используются в процедурах извлечения новых данных. Помимо этого имеется q подмножеств $N_1 \subseteq N$, $N_2 \subseteq N$, ..., $N_q \subseteq N$, которые назовем *кластерами*. Информационное взаимодействие в такой РИС состоит в том, что пользователь кластера имеет неограниченный доступ к каждой его БД и не имеет доступа к БД тех кластеров, в которые он не входит.

Примеры. Приведем несколько примеров РИС, отметив, что каждая из них ха-

рактеризуется своей методикой обработки информации.

1. *Базы знаний учебного заведения.* Пусть $N = \{1, 2, \dots, n\}$ есть множество узлов, в каждой БД которых хранятся учебно-методические и обучающие материалы по определенной компетенции. А все множество $\{B_1, B_2, \dots, B_n\}$ этих БД образует учебно-методическую базу учебного заведения, в котором имеется q специальностей. В рамках одной специальности доступны учебные материалы строго определенных компетенций, и не доступны материалы тех компетенций, которые не предполагаются к использованию их учебными планами. Тогда i -я специальность, $i = 1, 2, \dots, q$, представляет собой совокупность N_i узлов множества N , базы знаний которых определяют учебно-методические материалы, которые необходимо освоить для получения i -й специальности.

2. *Схема документооборота.* Пусть РИС имеет множество узлов $N = \{1, 2, \dots, n\}$, каждый узел $i \in N$ ассоциирован с определенным работником u_i . В РИС имеется q кластеров $N_1 \subseteq N$, $N_2 \subseteq N$, ..., $N_q \subseteq N$. Все работники обмениваются данными: каждый из них имеет возможность создать документ и отослать его другим работникам. Документ метится уникальной подписью работника, который его создал. Документ, сгенерированный работником u_i , рассылается всем работникам РИС, но открывается работниками лишь того кластера, в который входит i -й узел. Тем самым обмен информацией между работниками одного кластера происходит свободно, в то время как обмен данными между разными кластерами не разрешен. Тем самым реализуется распределение доступа к информации в зависимости от специализации пользователей РИС.

3. *Распределенный банк данных.* Пусть $N = \{1, 2, \dots, n\}$ есть множество узлов, с каждым из которых ассоциирована пара: база данных и пользователь, и $N_1 \subseteq N$, $N_2 \subseteq N$, ..., $N_q \subseteq N$ суть множество кластеров. Данные разных кластеров могут быть противоречивыми, поэтому на информационные процессы накладываются определенные ограничения. В рамках такого распределенного банка данных явно выделяются два процесса: помещение информации в БД и извлечение новых данных, исходя из хранящейся в БД информации. При этом как размещение информации, так и извлечение может происходить с использованием нескольких БД, ассоциированных с разными узлами. Что касается помещения новой информации, то пользователь u_i , $i \in N$, может помещать информацию в любую базу кластера, которой

принадлежит узел i . Но при формировании пользователем u_i запросов к распределенной базе данных разрешено использовать данные из любого подмножества БД в точности того кластера, которому принадлежит узел i . Тем самым исключается появление противоречия, так как для БД одного кластера соблюдается непротиворечивость данных. В то же время использование БД из разных кластеров не разрешено, так как в этом случае использование данных из таких БД может привести к противоречию.

4. *Частичный порядок кластеров.* На множестве кластеров одной РИС можно ввести отношение смежности кластеров, сделав ее, в определенном смысле, двухуровневой. Для этого надо ввести дополнительное ребро, по которому будут общаться между собой два кластера K_1 и K_2 . В исходной РИС это ребро отсутствует. Оно вводится лишь для передачи определенной информации между кластерами K_1 и K_2 , фильтруя ее с помощью определенного фильтра. В разных направлениях ребро фильтрует информацию по-разному. Тем самым лишь определенная информация из кластера K_1 будет доступной в кластере K_2 , и наоборот. В разных направлениях передаваемая информация имеет разные характеристики. В итоге реализуется вполне удобный метод защиты информации, циркулирующей между кластерами, делая РИС менее статичной по сравнению с описанными выше.

Итак, задача разработки РИС на заданном множестве узлов состоит в том, чтобы:

1) пользователь мог общаться с другими пользователями только в рамках тех кластеров, которым он принадлежит. Общение с пользователями других кластеров невозможно, только если это не предусмотрено дополнительными средствами РИС в виде дополнительных ребер;

2) пользователю, ассоциированному с узлом $n \in N$, была доступна информация, ассоциированная с узлами $n_1, n_2, \dots, n_m \in N$ только, если последние принадлежат тому же кластеру, что и он. Исключение составляют РИС с дополнительными ребрами, обладающими свойством фильтрации информации.

Тем самым в РИС осуществляется разделение прав доступа пользователей, в зависимости от разбиения ее узлов на кластеры. В предлагаемой далее методике проектирования РИС предполагается, что выделение кластеров происходит путем присвоения меток узлам и описания критерия, позволяющего по любому множеству меток определять, принадлежат ли они одному кластеру. Подобная методика позволяет осуществлять динамическое разбиение

узлов на кластеры, так как присвоение меток узлам и проверка критерия принадлежности узлов одному кластеру осуществляется программными средствами.

Реально РИС задаются путем определения ее узлов с соответствующими БД и пользователями, а затем перечислением всех кластеров. Таким образом, РИС мыслится как пара (N, N^*) , где N есть конечное множество узлов, а $N^* \subseteq 2^N$ – совокупность кластеров, которые суть подмножества множества N . Один узел может входить в несколько кластеров в качестве составной части. Если два узла одновременно не принадлежат никакому кластеру, то они называются *несовместимыми*, в противном случае – *совместимыми*. Содержательно несовместимость узлов может интерпретироваться по-разному: это может быть несовместимость соответствующих баз данных в распределенной системе обработки данных, невозможность пользователей иметь доступ к одним и тем же документам в системе документооборота, запрет на общение пользователей, ассоциированных с узлами РИС, несовместимость разных команд в одном управляющем устройстве и т.п. Два пользователя РИС несовместимы, если они не имеют доступа к одной и той же информации. Более подробно семантика, основанная на несовместимости узлов, описана в [5].

В результате на множестве N задано бинарное коммутативное отношение $InCom \subseteq N \times N$ – описывающее отношение несовместимости. Узлы РИС можно представлять как обладающие определенными значениями конечной совокупности элементарных признаков. Поэтому узел мыслится как вектор, каждый его компонент указывает на наличие / отсутствие определенного признака. Признак однозначно соответствует компоненту вектора, разным компонентам соответствуют разные признаки и наоборот. Значения признаков, в свою очередь, бывают *совместимыми* и *несовместимыми*. Разные значения одного признака – *несовместимы*, поэтому они не встречаются в описании одного узла.

По множеству N и отношению $InCom$ построим *граф несовместимости* $GIncom$ с множеством узлов N . Узлы $b_1, b_2 \in GIncom$ смежные тогда и только тогда, когда выполняется отношение $InCom(b_1, b_2)$. Мы несколько усилим понятие кластера РИС: множество N^* – это *совокупность кластеров*, которые суть максимальные подмножества множества N , т.е. к условию $N^* \subseteq 2^N$ добавляется требование, что каждый кластер нельзя расширить за счет добавления какого-либо узла из N .

Введенное определение РИС назовем *содержательным*, так как отношение несовместимости *InCom* может носить неформальный характер и определяться, исходя из соображений, формулируемых содержательно. Предикат *InCom* введен, чтобы сделать изложение более формальным.

Нетрудно увидеть, что каждому кластеру $B \in N^*$ в графе *GIncom* соответствует максимальный пустой подграф G_B , и наоборот, каждому максимальному пустому подграфу графа *GIncom* соответствует единственный кластер. Таким образом, можно говорить не о кластерах РИС, а о пустых подграфах графа *GIncom*.

Логическая семантика РИС

В [5] введено понятие *логической матрицы*, представляющей КНФ или ДНФ логической функции, строки матрицы соответствуют переменным, а столбцы – соответственно дизъюнктам или конъюнктам. Элементы матрицы суть 0, 1, _ (обозначая соответственно отрицательную литеру, положительную литеру или ее отсутствие). Каждый столбец матрицы с n строками представляет интервал в единичном n -мерном кубе E^n , а совокупность M^* столбцов определяет интервал $I^* \subseteq E^n$: I^* есть пересечение интервалов, определяемых M^* . Поэтому интервал I^* называется *определяемым* подматрицей M^* . В [5] описано, как по графу *GInCom* построить матрицу с n строками и m столбцами, где $m = |N|$ и она обладает следующими свойствами.

1. Два разных столбца b_1 и b_2 матрицы помечены интервалами σ_1 и σ_2 единичного n -мерного куба, пересечение которых $\sigma_1 \cap \sigma_2$ пусто тогда и только тогда, когда *InCom*(b_1, b_2). В этом случае столбцы b_1 и b_2 называются *ортогональными*.

2. Каждый пустой подграф $G_0 \subseteq GIncom$, $G_0 = \{b_1, b_2, \dots, b_q\}$ определяет такую подматрицу $M(b_1, b_2, \dots, b_q)$, что определяемый ею интервал в n -мерном единичном кубе не пуст. Для непустых графов определяемые ими интервалы пусты.

Разные способы построения матрицы по графу *GIncom* дают различные матрицы с разным числом строк, тем самым поразному определяя интервалы единичного куба. Однако в любом случае отношение *InCom* элементов множества N описывается логическими терминами, и отношение несовместимости однозначно совпадает с отношением пустоты пересечения интервалов.

Как видно, знание такой матрицы позволяет на логическом уровне ввести кластеры РИС. Для этого достаточно несовместимым узлам поставить в соответствие ортогональ-

ные столбцы. Это можно делать разными способами, что приводит к различным матрицам, характеризующимся разным числом строк, однако с одинаковым отношением ортогональности на столбцах. Наиболее прозрачный метод построения ортогональности состоит в выделении максимальных двудольных подграфов исходного графа и введении соответствующих ортогональных меток.

Построение кластеров по заданным характеристикам

Пусть в РИС надо выделить кластер как совокупность узлов, обладающих заданными характеристиками, причем практически невозможно перечислить в точности все свойства узлов в силу того, что они составляют представительное множество. Однако в данном случае можно воспользоваться содержательным отношением несовместимости, не особенно озадачиваясь его точным формальным представлением. Это есть одно из достоинств предлагаемого метода, когда вместо анализа формальных показателей, используются содержательные соображения. Поэтому в процессе уточнения задача вполне решается, и в конечном итоге удастся получить искомым кластер. Для определенности можно считать, что на начальном этапе имеется перечисление некоторых свойств, которые позволяют выделять кластер, обладающий искомыми свойствами, и проектирование РИС состоит в том, чтобы описать все кластеры.

Сформулируем задачу в терминах РИС, для которой задано отношение *InCom*. Содержательно ее элементами могут трактоваться как конкретные сущности, так и признаки, позволяющие однозначно идентифицировать кластер. Поэтому можно полагать, что для элементов РИС задана содержательная семантика в терминах совместимости / несовместимости. Рассматриваемая РИС характеризуется графом *GIncom*, по которому в соответствии с описанной выше процедурой строится логическая матрица M . Теперь задача описания кластера состоит в построении максимального пустого подграфа, узлы которого суть элементы РИС, имеющие содержательную интерпретацию и удовлетворяющие содержательным критериям принадлежности искомого кластеру.

Пометим столбцами логической матрицы M соответствующие узла множества N . Метку узла n обозначим $I(n)$, а множество всех меток подмножества $N_1 \subseteq N$ обозначим $I(N_1)$. Пересечение всех меток множества $I(N_1)$ обозначим $\cap I(N_1)$. Введенные метки обладают следующим свойством. Для лю-

бого множества $N_1 \subseteq N$ пересечение $\cap I(N_1)$ не пусто тогда и только тогда, когда множество N_1 принадлежит одному кластеру. Как следствие, получаем, что множество меток кластера определяет непустой интервал. Теперь организуем информационное взаимодействие узлов лишь в случае, если пересечение их меток не пусто. Тем самым информационное взаимодействие узлов осуществляется только в случае принадлежности их одному кластеру. Передача данных между узлами разных кластеров исключается. И получение информации по запросу узла n к базам данных узлов $n_1, n_2, \dots, n_m$ возможно, лишь если пересечение $\cap I(n_1, n_2, \dots, n_m)$ не пусто.

Закключение

Предложена методика проектирования РИС, представляющей собой множество узлов с заданным на нем разбиением на кластеры. Основным требованием к узлам РИС является ограничение на взаимодействие узлов только в рамках одного кластера. Описан метод присвоения меток узлам РИС, представляющих собой интервалы единичного n -мерного куба, который позволяет организовать кластеры так, чтобы метки узлов одного кластера принадлежали одному непустому интервалу единичного куба, а метки узлов из разных кластеров представляли

собой непересекающиеся интервалы. Используя такую разметку узлов РИС, можно легко обеспечить искомую связь узлов лишь в рамках одного кластера. Тем самым легко строится «плоская» РИС, с изолированными друг от друга кластерами. С другой стороны, можно обобщить данную методику и построить «двухуровневую» РИС, когда передача информации между кластерами возможна через определенные фильтры, пропускающие только специфическую информацию. Метод реализуется программно, путем присвоения узлам описанных меток и проверке по ним произвольного подмножества узлов простого критерия непустоты пересечения интервалов. При переопределении структуры РИС, т.е. ином разбиении ее на кластеры, происходит лишь простая перенастройка параметров программы.

Список литературы

1. Миков А.И., Замятина Е.Б. Распределенные системы и алгоритмы. – М.: Интуит.ру, 2012. – 370 с.
2. Таненбаум Э., Ван-Стеен М. Распределенные системы. Принципы и парадигмы. – СПб.: Питер, 2013. – 877 с.
3. Федоров А., Мартынов Д. Windows Azure. Облачная платформа Microsoft. Microsoft. – 2012. – 96 с.
4. Черняк Л. Web-сервисы, grid-сервисы и другие // Открытые системы. СУБД. – 2014. – № 12. – С. 20–27.
5. Попов С.В. Синтез предметных областей. Решение одного класса переборных задач. – LAP LAMBERT Academic Publishing, 2017. – 96 с.