УДК 004:75

ВИКОРИСТАННЯ КОНЦЕПЦІЇ ЗВ'ЯЗАНИХ ДАНИХ У ПРОЦЕСІ ПРОЕКТУВАННЯ ТА РОЗРОБКИ РОЗПОДІЛЕНИХ ІНФОРМАЦІЙНИХ СИСТЕМ

А. С. Петрусь
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: petrus.andrew@ukr.net

Розглядається можливість використання технології LinkedData в процесі проектування і розробки розподілених інформаційних систем (РІС) різної спрямованості. Наведено основні компоненти РІС, особливості їх взаємодії, запропонована параметрична модель, що описує РІС у загальному вигляді, а також розглянуті основні положення та особливості технології пов'язаних даних LinkedData. Виявлено й обгрунтовано можливість використання даної технології при побудові РІС, наведено практичні переваги підходу.

Ключові слова: розподілена інформаційна система, розподілене сховище даних, пов'язані дані, LinkedData.

ИСПОЛЬЗОВАНИЕ КОНЦЕПЦИИ СВЯЗАННЫХ ДАННЫХ В ПРОЦЕССЕ ПРОЕКТИРОВАНИЯ И РАЗРАБОТКИ РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

А. С. Петрусь
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: petrus.andrew@ukr.net

Рассматривается возможность использования технологии LinkedData в процессе проектирования и разработки распределенных информационных систем (РИС) разной направленности. Приведены основные компоненты РИС, особенности их взаимодействия, предложена параметрическая модель, описывающая РИС в общем виде, а также рассмотрены основные положения и особенности технологии связанных данных LinkedData. Выявлена и обоснована возможность использования данной технологии при построении РИС, приведены практические преимущества подхода.

Ключевые слова: распределенная информационная система, распределенное хранилище данных, связанные данные, LinkedData.

АКТУАЛЬНОСТЬ РАБОТЫ. В процессе развития локальных и глобальных сетей появилась возможность совместного использования вычислительных мощностей и ресурсов отдельных сетевых компонентов для решения более сложных и ресурсоёмких задач. Ресурсы самых разнообразных компьютерных устройств, объединённых в сеть, доступны для распределения задач и решения некоторой ее части, объединяя впоследствии микрорешения в единое целое. Это направление решений задач объединения распределенных ресурсов компьютерных систем может успешно развиваться лишь при выполнении двух главных условий – адекватном развитии глобальной сетевой инфраструктуры и применении реальных технологий создания распределенных информационных систем.

На сегодняшний день существует множество различных веб-сервисов, аггрегаторов ресурсов, предоставляющих услуги для открытого доступа к данным, как специфичным для конкретных предметных областей, так и общего плана. Каждый из этих сервисов физически представляет собой распределенную информационную систему (РИС) [1], данные между элементами которой передаются через веб.

Одной из наиболее актуальных на сегодняшний день является задача повышения эффективности взаимодействия компонентов РИС с учетом возможных технологический проблем, связанных с транспортировкой данных через открытую гипертекстовую среду Web [2].

Среди множества задач, выполняемых РИС, особое место занимает задача эффективного представления данных, а также их обработки.

Представление данных, обрабатываемых РИС, имеет большое значение при проектировании системы. С увеличением количества и направленности сервисов, предоставляемых РИС, применяются все более абстрактные концепции представления данных.

Зачастую данные, хранимые РИС, организованы исходя из изначальных требований функционала самой информационной системы, что впоследствии сказывается на производительности при модернизации либо перепрофилированию РИС. Большинство систем, предоставляющих те или иные сервисы, спроектированы с учетом своей узкой специализации, что позволяет оптимизировать производительность системы на начальных этапах ее эксплуатации.

При объединении обособленных РИС в одну общую систему, либо при модернизации и расширения функционала существующей системы, важным аспектом является возможность связи предоставляемых данных. Если система спроектирована без учета возможности связей между хранимыми данными, то дальнейшая интеграция такой подсистемы в общую РИС вызовет трудности.

Задача поиска и выделения необходимой в определенный момент времени информации усложняется при отсутствии должных связей между хранимыми данными внутри РИС. Поиск информации приходится производить в определенных частях системы, либо во всем ее массиве физических носителей, что существенно увеличивает время доступа к данных, и, как следствие, время отклика системы в целом. В таком случае адекватное время обработки данных при возрастающем количестве пользователей и их потребностей в информации обеспечивает постоянная модернизация оборудования РИС.

Повышения производительности всей системы можно добиться без обновления оборудования благодаря реорганизации хранилища информации, и установив связи между данными. Подход, представляющий данные, хранимые РИС, в виде связанных знаний, является альтернативным решением проблемы быстродействия доступа и обработки данных в распределенных информационных системах.

Цель работы – возможности использования концепции связанных данных в контексте организации хранилищ данных в распределенных информационных системах различной направленности.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ. В качестве типичного примера реализации распределенной информационной системы, обрабатывающей пользовательскую информацию, рассмотрим такой тип РИС, как социальные сети. Общая схема архитектуры социальной сети приведена на рис. 1.

В общем виде компоненты приведенного типа распределённой информационной системы можно разделить на:

– балансировщик нагрузок (Lb);

– веб-сервер, предоставляющий необходимые для обработки пользовательской информации сервисы для использования как локально (внутри самой РИС), так и наружно (Ws, может быть представлен серверным массивом);

– сервер баз данных (DBs, может быть престален серверным массивом);

– дополнительные выделенные сервера, отвечающие за определенный функционал распределенной информационной системы. Например: сервер мгновенных сообщений, сервер для обработки видео- и аудиоинформации и т.д. (Ds, может быть представлен серверным массивом).

Исходя из вышесказанного, распределенную информационную систему (DIS), состоящую из вышеуказанных компонентов, можно представить в виде параметрической модели:

$$DIS\ \{Lb,\ Ws,\ DBs,\ Ds\}. \qquad (1)$$

Архитектура такой системы разрабатывается с учетом функциональных возможностей проектируемой системы, и, следовательно, может отличаться от приведенной выше.

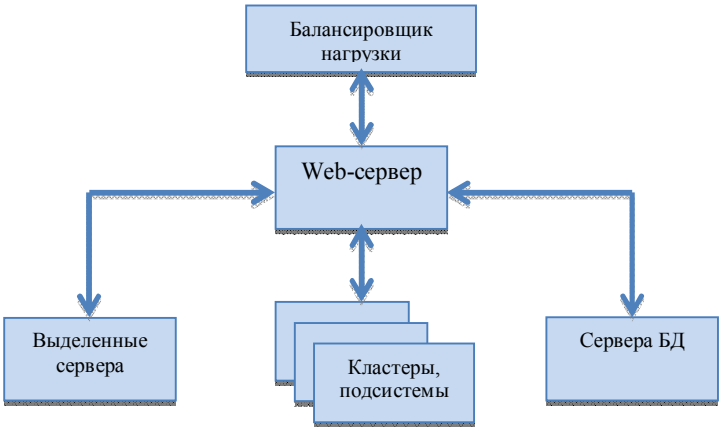


Рисунок 1 – Пример архитектуры РИС, предоставляющей услуги организации социальной сети

Поскольку РИС, представляющая сервисы социальной сети, является высоконагруженной благодаря большому количеству пользователей, возникает проблема масштабирования системы в связи с растущими нагрузками.

Стандартный подход к построению распределенных информационных систем дает определенные преимущества для поддержки и расширения готового проекта.

Одно из таких преимуществ – модульная распределенная архитектура, позволяющая производить горизонтальное масштабирование отдельных частей РИС, которые в этом нуждаются.

Горизонтальное масштабирование решает проблему повышения быстродействия распределенной информационной системы при растущих нагрузках, однако требует высоких затрат в связи с покупкой нового оборудования, дополняющего либо заменяющего старое.

Быстродействие распределенной системы напрямую зависит от того, как организованы знания, представляемые хранимыми данными. В таком случае, выбирая концепцию и принцип хранения данных на этапе проектирования РИС, можно добиться как увеличения производительности, так и решить проблему масштабирования системы с учетом роста числа клиентов, что, в свою очередь, обеспечит дополнительные удобства сопровождения и модернизации готовой системы.

С увеличением количества и направленности сервисов, предоставляемых РИС, развивается различные концепции абстрактного представления данных. Среди множества таких концепций следует выделить концепцию «связанных данных» LinkedData. В рамках концепции LinkedData каждый элемент распределенной информационной системы представляет собой базу знаний, описывающую хранимые данные, а также необходимые ссылки на данные, хранимые другими узлами внутренней сети.

Выборка искомой информации от различных компонентов осуществляется асинхронно.

Linked Data — это способ публикации данных в Интернете, который обладает следующими важнейшими чертами:

– все описываемые предметы имеют уникальные адреса (URI), что позволяет публиковать сведения об одном и том же предмете в разных местах;

– зная адрес предмета, можно получить его описание, сделав HTTP-запрос по этому адресу;

– описание имеет вид графа (RDF) и в нём указываются свойства предмета, в том числе — связи с другими предметами. Благодаря этому клиент может переходить от одного описания к другому;

– описания можно записывать в отдельных файлах с помощью разных форматов RDF, а также встраивать прямо в атрибуты HTML- или XML-документов;

– все свойства и отношения между предметами тоже имеют уникальные адреса. Это позволяет создавать общие термины в любой предметной области, а затем описывать предметы этими терминами так, чтобы клиенты могли «понимать» описания независимо от того, кто и где их опубликовал;

– с помощью логического вывода можно получать новые данные из имеющихся.

Связанные данные как элемент концепции LinkedData формально могут быть представлены с помощью выражения

$$t =< g, s, p, o >, \qquad (2)$$

где t – триплет; g – именованный граф; s – субъект; p – предикат; o – объект.

Множество структур t, определенных формулой (2), будем считать хранилищем данных:

$$T = \{t_i\}, i = \overline{(1,n)}, \qquad (3)$$

где t_i – i-й триплет, n – количество триплетов в хранилище.

Учитывая необходимость для поиска использовать информацию про контекст связанных данных, будем считать, что все t_i, в которых g однаковое, объеденены одним контекстом. Таким образом, определим контект следующим образом:

$$G = \{g_j\}, j = \overline{(1,m)}, \qquad (4)$$

где G – множество контекстов хранилища, g_j – j-контекст хранилища данных, m – количество контекстов хранилища.

Каждому контексту, исходя из формулы (2), поставим в соответствие трехэлементный набор <*s,p,o*>. Таким образом, контекст определим с помощью выражения:

$$\forall g \in G : g =< S, P, O >, \qquad (5)$$

где g – контекст хранилища данных, G – множество всех контекстов, S – множество субъектов, P – множество предикатов, O – множество объектов.

Расширим понятие объекта следующим образом (6):

$$\forall o \in O : o =< T, L, V >, \qquad (6)$$

где T – тип данных, L – язык представления, V – значение.

Учитывая особенность концпеции LinkedData, а конкретно необязательность определения схем, типов данных и языков представления значений объекта O в случае строкового типа, будем считать элементы объекта o необязательными для определения, а схему объекта исходя из формулы (6) будем считать частично-определенной схемой связанных данных, которые определяются формулами (2)–(4).

Учитывая, что на множестве G необходимо решать поисковые задачи, модифицируем формулу (5) путем введения вспомогательного элемента – множества функций, которые могут использоваться на множествах элементов контекстов G:

$$\forall g \in G : g =< S, P, O, F >, \qquad (7)$$

где g – контекст хранилища данных, G – множество всех контекстов, S – множество субъектов, P – множество предикатов, O – множество объектов, F – множество функций.

Практические аспекты реализации функций F формулы (7) возлагаются на сторонних разработчиков программного обеспечения.

Технологии, основанные на LinkedData, являются надстройкой над стандартными веб-технологиями, такими как HTTP и URLs, однако используются не для отображения информации конечному пользователю, а с целью обмена информацией между вычислительными системами. Это позволяет данным с разных источников быть взаимно доступными и связанными между собой.

Пирамида технологий и их надстроек LinkedData показана на рис. 2.

Рисунок 2 – Пирамида технологий и надстроек LinkedData

Физической реализацией концепции являются распределенные информационные системы, предоставляющие услуги хранения данных, или же распределенные хранилища связанных данных.

Под *распределенным хранилищем данных (РХД)* обычно понимают хранилище, декомпозированное и фрагментированное на несколько узлов вычислительной сети.

Распределенная информационная система, предоставляющая услуги хранения данных, физически состоит из массива компьютеров, объедененных в единую сеть.

Логические связи между хранимыми данными организовываются в процессе обработки входящей информации, что позволяет значительно ускорить доступ ко всем объектам, так или иначе относящимся к запрашиваемому (рис. 3).

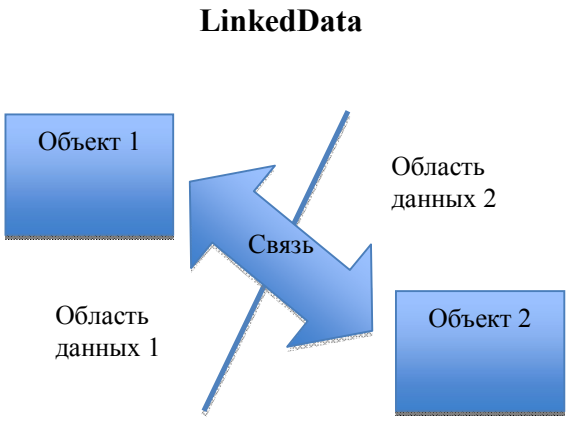


Рисунок 3 – Логические связи между объектами

Логические связи между хранимыми данными организовываются в процессе обработки входящей информации, что позволяет значительно ускорить доступ ко всем объектам, так или иначе относящимся к запрашиваемому.

Использование знание-ориентированных систем и интеграция технологий, связанных с LinkedData в веб – основное направление исследований и разработки на сегодняшний день в данной области. Среди вариантов реализации следует выделить подходы, использующие семантические РИС на основе SPARQL – языка запросов к данным, представленным в модели RDF (англ. *Resourse Description Framework*). Язык и протоколы данных РИС ориентированы на упрощение и увеличение производительности операций над связанными данными.

В связи с вышесказанным, нами были разработаны требования к архитектуре высоконагруженной РИС, выполнение которых на этапе проектирования архитектуры будущей РИС даст возможность снизить затраты на повышение быстродействия РИС, а также повышение эффективности ее работы:

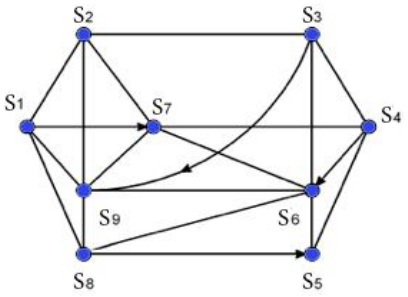
– *распределенность данных*. Задача поиска данных, хранимых массивом серверов распределенной системы, усложняется, если их представление либо каталогизация осуществлены не оптимально. Концепция LinkedData подразумевает представление данных в виде связанного графа. Задача поиска данных, в свою очередь, сводится до задачи поиска кратчайшего пути на графе (рис. 4), узлами которого являются данные, хранимые РИС, с ребрами различного веса. Вес ребер графа может быть оценен самыми различными критериями: скорость доступа к определенному серверу, максимальная пропускная способность сети, общая загруженность конкретного сервера и т.д.

– *единый язык и протокол доступа к данным среди серверных компонентов системы*. Поскольку высоконагруженные РИС рано или поздно сталкиваются с возрастающим количеством пользователей, и, как следствие, нагрузки, то решение задачи масштабирования такой РИС должно быть предусмотрено на этапе ее проектирования. Стандартизация протоколов доступа к данным позволит существенно упростить введение в строй новых серверов, а также гипотетическое использование ресурсов посторонних РИС, либо их отдельных компонентов без переработки существующих элементов.

– *эффективная маршрутизация запросов внутри РИС*. Представление данных в виде связанного графа не только ускоряет доступ к хранимым РИС данным, но и обеспечивает увеличение скорости их обработки за счет широко известных алгоритмов поиска пути на графе.

Разработка критериев оценки ребер графа с учетом возможностей и факторов эксплуатации распределенной системы играет ключевую роль в корректности работы этих алгоритмов, и, как следствие, существенно повлияет на общую производительность РИС на этапе ее проектирования.

– минимальное количество промежуточных звеньев при получении и компоновке ответа системы на пользовательский запрос;



C=

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1		5					3	6	12
S2	5		18				2		13
S3		18		25		20			7
S4			25		5	16	4		
S5				5		5			
S6			20		5		14	15	9
S7		2		4		14			24
S8	6				23	15			5
S9	12	13				9	24	5	

Рисунок 4 – Пример представления узлов и ребер графа организации связанных данных, хранимых РИС

– связанность данных, предоставляемых отдельными компонентами РИС вне зависимости от их типа.

Поскольку распределенная система представляет собой совокупность компьютерных ресурсов для обработки данных либо решения круга задач, может возникнуть необходимость внедрения еще одной РИС в качестве отдельного компонента общей системы (рис. 5).

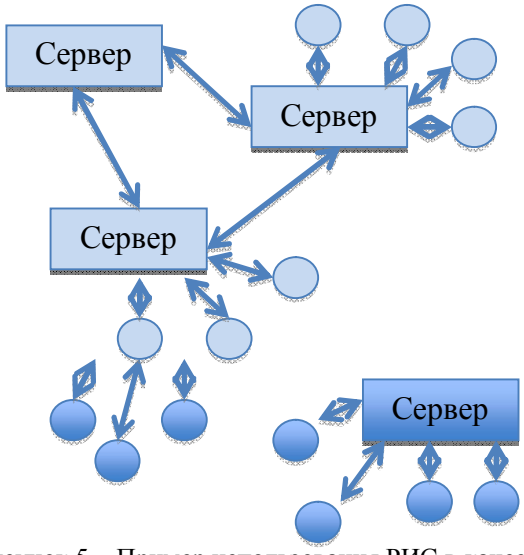


Рисунок 5 – Пример использования РИС в качестве компонентов общей системы

Соблюдение общих стандартов и концепций хранения данных гарантирует корректное функционирование системы в целом.

Надежные протоколы защиты данных, включающие соответствующие методы аутентификации и авторизации пользователей с целью разграничения прав на предоставление информации, должны быть предусмотрены в процессе проектирования системы. Уровни доступа пользователей системы к хранимой информации – важный аспект предоставления услуг РИС.

ВЫВОДЫ. В последнее время рынок онлайн-сервисов, предоставляющих услуги РХД, растет, потому разработка качественного и быстрого решения задачи распределенного хранения данных, в числе прочего, является залогом успеха нового сервиса в среде высокой конкуренции.

1. Предложена параметрическая модель распределенной информационной системы в общем виде.
2. В результате проведенного исследования были предложены требования, соблюдение которых при проектировании РИС для выполнения самого разнообразного круга задач существенно облегчит поддержку, модернизацию и управление системой в процессе эксплуатации.
3. Использование концепции связанных данных в процессе проектирования РИС позволит максимально эффективно структурировать хранимые данные, минимизировать время доступа к данным, а также внедрять новые подсистемы, расширяя функционал существующей РИС.

Мы полагаем, что использование концепции связанных данных, а также соблюдение приведенных требований позволит добиться повышения общего качества, производительности и удобства поддержки РИС.

ЛИТЕРАТУРА

1. William S. Davis, David C. Yen. The Information System Consultant’s Handbook. Systems Analysis and Design. CRC Press, 1998.
2. Малко А.А., Миронов А.С. Современные информационные технологии // Сборник научных трудов НГТУ. – 2011. – № 2 (64). – С. 75–84.
3. Таненбаум Э., Ван Стеен М. Распределенные системы. Принципы и парадигмы Серия «Классика Computer Science». – СПб.: Питер, 2003. – 877 с.

USING THE CONCEPT OF LINKEDDATA FOR DESIGN AND DEVELOPMENT OF DISTRIBUTED INFORMATION SYSTEMS

A. Petrus
Kremenchuk Mykhailo Ostrohradskyi National University
vul. Pershotravneva 20, Kremenchuk, 39600, Ukraine. E-mail: petrus.andrew@ukr.net

This paper describes the use of LinkedData technology for the design and development process of distributed information systems (DIS) of different directionalities. The main components of DIS are shown in this work, as well as particular features of their interactions. The parametric model describing the RIS in general, as well as the basic terms and features of the LinkedData data-related technology is proposed. The author has identified and demonstrated the possibility of using this technology for DIS building, practical advantages of the approach suggested.

Key words: distributed information system, a distributed data store, associated data, LinkedData.

REFERENCES

1. William S. Davis, David C. Yen. (1998), *The Information System Consultant’s Handbook. Systems Analysis and Design*. CRC Press.
2. Malko, A.A., Mironov, A.S. (2011), “Up-to-date information technologies”, *Collected works of Novosibirsk State Technical University*, no. 2 (64), pp. 75–84.
3. Tanenbaum, E., Van Steen, M. (2003), *Raspredelennye sistemy. Printsipy i paradigmy. Seriya “Klassika Computer Science”* [Distributed systems. Principles and paradigms. Series “Classics Computer Science”], Piter, St.-Petersburg, Russia.

Стаття надійшла 05.06.2013.