Модель тиражирования библиографических баз данных с использованием алгоритмических кодов записей

Model of replication of bibliographic database with use the identification codes of records

Карауш А.С.

Alexander. S. Karaush

Предложена новая модель для тиражирования (репликации) библиографических баз данных с использованием алгоритмического кода библиографических записей. Введено понятие алгоритмического кода в качестве идентификатора библиографической записи. На практике, строка алгоритмического кода - свертка библиографической записи, содержащая элементы обязательных полей библиографического описания. Описание представленной модели сделано с использованием дискретного анализа и теории множеств. Новая модель тиражирования позволяет создавать распределенные АБИС в системах библиотек без использования круглосуточного канала связи, обеспечив при этом минимальный объем передаваемых данных. В случае использования описываемой модели по каналам связи передаются только вновь созданные, и измененные повторения полей, а также алгоритмические коды удаленных библиографических записей.

The new model of replication of bibliographic database with use the algorithmic code of bibliographic record is offered. The notion of algorithmic code as identifier to bibliographic records is defined. The algorithmic code is folding of bibliographic record in practice, containing elements of obligatory floors of bibliographic description. The description of models is made with use the sampling analysis and theories of sets. The new model of replication allows to create distributed automated library systems without use the on-line Network or Internet channel to relationships, having provided herewith minimum volume sent data. In the event of use described models on channels of relationships are sent only newly created, and changed repetitions of floors, as well as algorithmic codes of deleted bibliographic records.

Задача распределения и территориального разнесения данных, а также тиражирования баз данных появляется в библиотеках не случайно. Каждая библиотека состоит в какой-либо системе или корпорации, а то и в нескольких. На уровне вузов это связка библиотека - филиалы - кабинеты курсового проектирования – методические кабинеты, на уровне научных или централизованных библиотечных систем – это связка центральная библиотека – филиал, где необходимо использование электронного каталога центральной библиотеки [6]. Не всегда для решения этих задач можно использовать информационные каналы связи между центральной библиотекой и филиалом, например, канал Интернет, по причинам:

1. постоянный канал связи организовать невозможно по экономическим или иным соображениям;
2. канал связи имеет недостаточную пропускную способность для работы библиотеки-филиала;
3. защита информационных потоков и баз данных.

Для поддержания библиографических баз данных (электронных каталогов библиотек) в актуальном состоянии используются различные модели тиражирования, или репликации, если использовать англоязычную терминологию данного технологического процесса [2, 5].

Основная задача тиражирования данных состоит в получении наиболее приемлемого решения для синхронизации изменений, выполненных в течение некоторого промежутка времени $(t_2 - t_1)$ и $(t_2 > t_1)$ в центральной библиотеке с тем, чтобы наилучшим образом привести базу данных в библиотеке-филиале в состояние равенства с базой данных центральной библиотеки.

При описании модели тиражирования будут использованы элементы теории множеств. Использование комбинаторики и теории множеств позволяет логично описать

процессы тиражирования в достаточно простом и понятном виде. Вводится несколько понятий, так библиографическая база данных определяется как множество библиографических записей [1]:

$$\boldsymbol{C} = \{\boldsymbol{B_k} \mid \boldsymbol{k} \in \boldsymbol{N}\}, \quad (1)$$

где $\boldsymbol{N} = \{0,\ldots\}$ - множество записей в библиографической базе данных.

В свою очередь, библиографическая запись, также определяется как множество полей, входящих в конкретную библиографическую запись:

$$\boldsymbol{B} = \{\boldsymbol{H_j} \mid \boldsymbol{j} \in \boldsymbol{N}\}, \quad (2)$$

где $\boldsymbol{N} = \{0,\ldots\}$ - множество полей в каждой библиографической записи,

причем, каждое поле представляет собой также множество $\boldsymbol{H}$ для повторений этого поля:

$$\boldsymbol{H} = \{\boldsymbol{A_i} \mid \boldsymbol{i} \in \boldsymbol{N}\}, \quad (3)$$

где $\boldsymbol{A}$ - поле библиографической записи, которое представляет также множество атомарных текстовых строк, называемых подполями, в которые заносится непосредственно информация, касающаяся библиографического описания информационного источника.

$$\boldsymbol{A} = \{\boldsymbol{a_i} \mid \boldsymbol{i} \in \boldsymbol{S}\}, \quad (4)$$

где $$\boldsymbol{S} = \{0,\ldots,9,\boldsymbol{a},\ldots,\boldsymbol{z}\} - \quad (5)$$

множество подполей в повторении поля библиографической записи, которые могут принимать значения меток букв латинского алфавита и одиночных цифровых символов.

Итак, пусть, $\boldsymbol{C_{t1}}$ – база данных $\boldsymbol{C}$ в момент времени $\boldsymbol{t}_1$, а $\boldsymbol{C_{t2}}$ – база данных $\boldsymbol{C}$ в момент времени $\boldsymbol{t}_2$, при условии $\boldsymbol{t}_2 > \boldsymbol{t}_1$.

Для уменьшения объема данных, передаваемых в процессе тиражирования, используется утверждение, что при достаточно большом объеме библиографической базы данных объем изменений за короткий период времени будет конечным и несравненно меньшим, чем весь объем базы данных. Основная задача для работы модели тиражирования - использование алгоритмов, которые позволяют определить измененные записи в библиографической базе данных за период времени.

Таким образом, для имеющихся баз данных $\boldsymbol{C_{t1}}$ и $\boldsymbol{C_{t2}}$ определяются множества $\boldsymbol{C_{const}}$ и $\boldsymbol{C}_\Delta$, такие что:

$$\boldsymbol{C_{const}} = \boldsymbol{C_{t1}} \cap \boldsymbol{C_{t2}} \quad (6)$$

$$\boldsymbol{C}_\Delta = \boldsymbol{C_{t1}} \Delta \boldsymbol{C_{t2}} \quad (7)$$

Пересечение для множеств $\boldsymbol{C_{t1}}$ и $\boldsymbol{C_{t2}}$ - есть множество $\boldsymbol{C_{const}}$, состоящее из библиографических записей, которые не изменялись в течение промежутка времени $\boldsymbol{t}_2 - \boldsymbol{t}_1$. Напротив, операция симметричной разности множеств $\boldsymbol{C_{t1}}$ и $\boldsymbol{C_{t2}}$ - множество $\boldsymbol{C}_\Delta$, состоящее из записей, которые были отредактированы, удалены или созданы заново в течение времени $\boldsymbol{t}_2 - \boldsymbol{t}_1$ и входят в одно из множеств $\boldsymbol{C_{t1}}$ или $\boldsymbol{C_{t2}}$. При этом выполняется предположение, что если библиографическая запись подверглась редактированию, то она представляет собой новое значение последовательности $\boldsymbol{C_{t2}}$ и с предыдущим своим состоянием в последовательности $\boldsymbol{C_{t1}}$ связь теряет.

Основываясь на вышеприведенных утверждениях можно определить множества:

$$\boldsymbol{C_{t1}} \cap \boldsymbol{C}_\Delta = \boldsymbol{C_{\Delta t1}} \quad (8)$$

$$\boldsymbol{C_{t2}} \cap \boldsymbol{C}_\Delta = \boldsymbol{C_{\Delta t2}} \quad (9)$$

$$\boldsymbol{C_{\Delta t1}} \cup \boldsymbol{C_{\Delta t2}} = \boldsymbol{C}_\Delta, \quad (10)$$

где $C_{\Delta t1}$ - множество библиографических записей, которые входят во множество C_{t1}, т.е. записи, которые были изменены или удалены во множестве C_{t1} за время $t_2 - t_1$;

$C_{\Delta t2}$ - множество записей, которые были изменены или добавлены в базу данных C_{t1} за время $t_2 - t_1$, но не входят во множество C_{t1}.

Используя вышеприведенные рассуждения и формулы (1)-(10), можно создать модель тиражирования библиографической базы данных, измененной в центральной библиотеке за период времени $t_2 - t_1$. Эта модель позволяет обеспечить существования «зеркального» электронного каталога в библиотеках-филиалах. Объем передаваемой информации по каналам передачи данных будет ограничен только множеством библиографических записей C_{Δ}, которое состоит из измененных, удаленных или добавленных записей. Такая модель уже существует и работает при тиражировании систем правовых баз данных. При этом объем информации, передаваемой по каналам связи, будет равен суммарному объему удаленных, добавленных записей и измененных записей базы данных, т.е. суммарному объему множеств $C_{\Delta t1}$ и $C_{\Delta t2}$.

Алгоритм, основанный на использовании модели тиражирования методом передачи измененных записей, в общем случае содержит следующие операции:

- сравнение библиографических записей в базах данных C_{t1} и C_{t2};
- определение записей, которые были удалены, отредактированы или добавлены за период времени $t_2 - t_1$, согласно (7)-(9);
- передача по каналам связи множеств библиографических записей $C_{\Delta t1}$ и $C_{\Delta t2}$;
- удаление в базе данных библиотеки-филиала записей, входящих во множество $C_{\Delta t1}$;
- добавление в базу данных библиотеки-филиала записей, входящих во множество $C_{\Delta t2}$.

При рассмотрении достоинств и недостатков данного алгоритма тиражирования следует обратить внимание на то, что ее использование оправдано в централизованных библиотечных системах, где ставится задача тиражирования библиографических баз данных в библиотеках-филиалах только для задач поиска и заказа, но не для редактирования и создания новых записей. Дальнейшее улучшение и усовершенствование модели создания и тиражирования библиографических баз данных возможно при использовании ключей библиографических записей. Подобная практика разработки ключей для библиографических записей существует для определения дублетных записей, а также для кодирования записей на основе имеющейся в ее полях информации.

Как было показано ранее на основе (8)-(10), можно определить множества записей, которые были изменены, удалены или добавлены в базы данных C_{t1} и C_{t2}.

На этом анализ данных, подвергшихся редактированию вести невозможно без определения алгоритмического кода (АЛКОДа) для библиографической записи B_k.

Алгоритмический код (АЛКОД) - идентификатор издания, который создается по определенным правилам (алгоритму) и однозначно идентифицирует конкретное издание (источник), позволяет хранить информацию об издании в удобном, компактном виде, а также позволяет осуществлять обслуживание пользователей в автоматизированном режиме.

Пусть G - множество алгоритмических кодов для базы данных C, которое может быть получено с использованием функции преобразования f для каждой записи B_k множества C:

$$G = \{X_k \mid k \in N\}, \qquad (11)$$

где X_k - строка алгоритмического кода для конкретной библиографической записи в базе данных C.

$$X_k = f(B_k) \quad (12)$$

В результате выполнения функции f построения АЛКОДа для каждой библиографической записи B_k, будет построена строка X_k, состоящая из символов, находящихся в полях A_{ik} этой записи.

Задача создания АЛКОДа сводится к условию получения одинаковых значений X_k для библиографических записей одного и того же ресурса, описанного разными каталогизаторами, но в то же время, к получению различных значений X_k для различных библиографических ресурсов. На практике, строка АЛКОДа - свертка библиографической записи, содержащая элементы обязательных полей библиографического описания.

В дальнейшем, операции по сравнению множеств библиографических записей C_{t1} и C_{t2} сводятся не только к получению множеств C_{const} и C_Δ, что определяют состояние библиографической базы данных на период времени t_2, но и к анализу множества алгоритмических кодов G для записей, входящих во множество C_Δ.

Для дальнейших рассуждений требуется ввести множества:

$G_{\Delta t1}$ - множество алгоритмических кодов для библиографических записей $C_{\Delta t1}$;

$G_{\Delta t2}$ - множество алгоритмических кодов для библиографических записей $C_{\Delta t2}$;

Такие, что:

$$G_{\Delta t1} = f(C_{\Delta t1}) \quad (13)$$

$$G_{\Delta t2} = f(C_{\Delta t2}) \quad (14)$$

Библиографическая запись считается отредактированной, если изменения данных, сделаны в полях, которые не участвуют в создании строки АЛКОДа для соответствующей записи. В противном случае запись считается удаленной за период времени $t_2 - t_1$ и вновь созданной с другими данными в полях, участвующих в построении АЛКОДа. Таким образом, множество АЛКОДов для измененных библиографических записей, при условии что значения АЛКОДа не изменялось, можно определить, как G_{const}, получаемое пересечением множеств $G_{\Delta t1}$ и $G_{\Delta t2}$:

$$G_{const} = G_{\Delta t1} \cap G_{\Delta t2} \quad (15)$$

При этом возможно определение АЛКОДов для библиографических записей, которые были удалены G_{del} за время $t_2 - t_1$:

$$G_{del} = G_{\Delta t1} \setminus G_{const}, \quad (16)$$

причем

$$G_{\Delta t1} = G_{del} \cup G_{const} \quad (17)$$

Множество АЛКОДов для библиографических записей G_Δ, которые были созданы за время $t_2 - t_1$:

$$G_{add} = G_{\Delta t2} \setminus G_{const}, \quad (18)$$

причем

$$G_{\Delta t2} = G_{add} \cup G_{const} \quad (19)$$

Для дальнейших исследований потребуется:

$$G_{const_{t1}} = G_{\Delta t1} \setminus G_{del} \quad (20)$$

$$G_{const_{t2}} = G_{\Delta t2} \setminus G_{add}, \quad (21)$$

причем должно выполняться:

$$\boldsymbol{G}_{\boldsymbol{const}_{t1}} = \boldsymbol{G}_{\boldsymbol{const}_{t2}} = \boldsymbol{G}_{\boldsymbol{const}}, \quad (22)$$

где $\boldsymbol{G}_{\boldsymbol{const}_{t1}}$ - множество АЛКОДов, входящих во множество $\boldsymbol{G}_{\Delta t1}$ и определяющих измененные библиографические записи во множестве $\boldsymbol{C}_{\Delta t1}$;

$\boldsymbol{G}_{\boldsymbol{const}_{t2}}$ - множество АЛКОДов, входящих во множество $\boldsymbol{G}_{\Delta t2}$ и определяющих измененные библиографические записи во множестве $\boldsymbol{C}_{\Delta t2}$;

Следует также определить действия, которые могут быть выполнены с данными, находящимися в полях библиографической записи:

- удаление повторения поля;
- добавление повторения поля;
- изменение данных в повторении поля.

Таким образом, для определения данных, необходимых для новой модели тиражирования требуется знание того, какие поля были удалены или (и) добавлены в библиографических записях за время $\boldsymbol{t}_2 - \boldsymbol{t}_1$. Удаленные и добавленные поля в библиографических записях, измененных целиком, определены ранее из (16) и (18). Для определения списка измененных полей в библиографических записях имеющих одинаковое значение АЛКОДа, и входящих во множества $\boldsymbol{G}_{\Delta t1}$ и $\boldsymbol{G}_{\Delta t2}$, потребуется провести сравнение полей и их повторений на равенство.

Для каждого АЛКОДа множества $\boldsymbol{G}_{\boldsymbol{const}_{t1}}$, ищется равный ему во множестве $\boldsymbol{G}_{\boldsymbol{const}_{t2}}$:

$$\boldsymbol{G}_{\boldsymbol{const}_{t1}} = \left\{\boldsymbol{X}_{\boldsymbol{const}_{\boldsymbol{k}}} \mid \boldsymbol{k} \in \boldsymbol{N}\right\}, \quad (23)$$

$$\boldsymbol{G}_{\boldsymbol{const}_{t2}} = \left\{\boldsymbol{X}_{\boldsymbol{const}_{\boldsymbol{p}}} \mid \boldsymbol{p} \in \boldsymbol{N}\right\}, \quad (24)$$

$$\boldsymbol{N} = \{0, \ldots\},$$

такой что:

$$\boldsymbol{X}_{\boldsymbol{const}_{\boldsymbol{k}}} = \boldsymbol{X}_{\boldsymbol{const}_{\boldsymbol{p}}}, \quad (25)$$

при этом для каждой пары $\boldsymbol{k}$ и $\boldsymbol{p}$ производится вычитание множеств полей библиографической записи (26), (27) с равным значением АЛКОДа, такие что:

$\boldsymbol{H}_{t1}$ - множество полей библиографической записи на момент времени $\boldsymbol{t}_1$;

$\boldsymbol{H}_{t2}$ - множество полей библиографической записи на момент времени $\boldsymbol{t}_2$;

В результате, для каждой библиографической записи, входящей во множество $\boldsymbol{C}_{\Delta}$, такой, что ее АЛКОД принадлежит множеству $\boldsymbol{G}_{\boldsymbol{const}}$, будет определено множество полей, которые были удалены $\boldsymbol{H}_{\boldsymbol{del}}$ или добавлены $\boldsymbol{H}_{\boldsymbol{add}}$ за время $\boldsymbol{t}_2 - \boldsymbol{t}_1$:

$$\boldsymbol{H}_{\boldsymbol{del}} = \boldsymbol{H}_{t1} \setminus \boldsymbol{H}_{t2}, \quad (26)$$

$$\boldsymbol{H}_{\boldsymbol{add}} = \boldsymbol{H}_{t2} \setminus \boldsymbol{H}_{t1} \quad (27)$$

Можно предложить новую модель тиражирования библиографических данных, основанную на использовании АЛКОДов библиографических записей со следующей последовательностью действий:

1. сравнение библиографических записей в базах данных $\boldsymbol{C}_{t1}$ и $\boldsymbol{C}_{t2}$, причем сравнению подлежат каждая запись в базе $\boldsymbol{C}_{t1}$ с каждой записью в базе $\boldsymbol{C}_{t2}$;
2. определение множеств записей $\boldsymbol{C}_{\Delta t1}$ и $\boldsymbol{C}_{\Delta t2}$, которые были удалены, отредактированы или добавлены, согласно (7)-(9);
3. построение АЛКОДов для множеств записей $\boldsymbol{C}_{\Delta t1}$ и $\boldsymbol{C}_{\Delta t2}$ согласно (11)-(14);

4. определение множеств G_{del} и G_{add} согласно (16) и (18);
5. определение множеств $G_{const_{t1}}$, $G_{const_{t2}}$ (20), (21);
6. определение множеств удаленных H_{del} и добавленных полей H_{add} за время $t_2 - t_1$ в библиографических записях баз данных, согласно (23)-(27).
7. передача по каналам связи:
   a. множества библиографических записей, которые имеют АЛКОДы, принадлежащие множеству G_{add}, для добавления этих записей в синхронизируемую базу данных филиала;
   b. множества АЛКОДов G_{del}, для удаления записей с равными АДКОДами, при этом вся запись не передается;
   c. множеств H_{del} и H_{add} с привязкой к значению АЛКОДа для каждой записи библиографической базы данных.
8. изменение синхронизируемой базы данных в филиале в соответствии со значением переданных множеств записей и АЛКОДов.

Можно определить достоинства и недостатки для модели тиражирования библиографических баз данных с использованием АЛКОДа.

Достоинства:

1. **Минимальный объем данных, передаваемых по каналам связи.** Действительно, по каналам передаются только изменения баз данных, причем если в библиографической записи проводилась редакция только одного поля, то только значение этого поля до и после редакции будет передано.
2. **Последовательность расположения записей в базах данных не важна.** В этом случае, если произведена реорганизация библиографической базы данных в каком-либо филиале, то это никак не отразиться на работе модели тиражирования. Таким образом, записи в центральной или синхронизуемой базе данных за время между моментами синхронизации могут быть «перемешаны».
3. **Возможность создания алгоритма двухстороннего тиражирования данных**, который позволяет производить изменение тиражируемой библиографической базы данных одновременно в двух библиотеках.

Недостатки:

1. **Сложность алгоритма построения АЛКОДа.** Действительно, основным условием работы данной модели является уникальность АЛКОДа для любой записи, входящей в любую базу данных. Для соблюдения этого условия приходится использовать дополнительные проверки на этапе ввода библиографического описания, когда АЛКОД может строиться автоматически.
2. **Возможность конфликтных ситуаций.** Такое может произойти при одновременном (в период времени $t_2 - t_1$, между тиражированием) изменении в разных библиотеках одной и той же записи.
3. **Дополнительные вычислительные ресурсы.** Сравнение библиографических записей попарно, находящихся в базах данных C_{t1} и C_{t2} требует дополнительных вычислительных ресурсов.

При анализе недостатков последней рассмотренной модели можно сказать, что конфликтные ситуации, возникающие в процессе тиражирования, можно существенным образом уменьшить, если ввести алгоритмы дополнительного анализа данных в полях библиографической записи. Затраты вычислительных ресурсов можно существенно сократить, если вместо сравнения библиографических записей целиком использовать для сравнения их АЛКОДы. И при утверждении, что в течение небольшого отрезка времени в

библиографической базе данных изменяется незначительное количество записей, относительно всего объема базы данных, возможно упрощение алгоритмов сравнения записей.

Выводы

Основные причины перехода библиотек от централизованного процесса создания и предоставления своего электронного каталога к распределенному, по видимому, кроются не столько в улучшении качества обслуживания, сколько в создании более эффективных схем управления и организации информационных и производственных ресурсов. Описанные в статье модели тиражирования библиографических баз данных нашли и находят свое применение в распределенных АБИС. На основе этих моделей возможно построение эффективной схемы предоставления электронного каталога библиотеки или системы библиотек, а также подготовка структуры организации для введения новых форм управления фондами и персоналом.

Список используемой литературы

1. UNIMARC Manual. Руководство по UNIMARC / Пер. на рус. яз. коллектива под рук. А.И. Земскова, Я.Л. Шрайберга. М.: ГПНТБ России, 1992. 319 с.
2. Когаловский М.Р. Энциклопедия технологий баз данных. – М.: Финансы и статистика, 2002. – 800 с.: ил. - ISBN 5-279-02276-4
3. Коннолли, Томас, Бегг, Каролин, Страчан, Анна Базы данных: проектирование, реализация и сопровождение. Теория и практика, 2-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2001. – 1120 с.: ил. – Парал. тит. англ.
4. Модели построения и функционирования корпоративной информационной сети муниципальных библиотек/ Карауш А.С., Левицкая Л.В.// Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: Материалы конф. – М., 2002. – Т.2. – С. 912-914.
5. Практическая репликация / Луковенко А., Фаритов А. // Открытые системы – 2001. – N12
6. Программное обеспечение для автоматической синхронизации баз данных системы "ИРБИС" / Карауш А.С., Копытков Д.Ю. // 10 -я Международная конференция "Крым 2003" Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества. ", Судак, 8-16 июня, 2003 - М.: Изд-во ГПНТБ России, 2003. - Т. 2.
7. Репликация данных как управленческая задача: подходы к решению/ Максименко Ю. // BYTE/Россия – 2001. – N2 – 4.