УДК 002:330.163

ОБЕСПЕЧЕНИЕ ОНЛАЙН-УПРАВЛЕНИЯ ЛОКАЛИЗАЦИЕЙ ИНФОРМАЦИОННЫХ РЕСУРСОВ С ОТКРЫТЫМ ИСХОДНЫМ КОДОМ И ПОДДЕРЖКОЙ ИНТЕГРАЦИИ С ОСНОВНЫМИ ОТКРЫТЫМИ ИСТОЧНИКАМИ КОНТЕНТА

Беневоленский С.Б., Кирьянов А.А., Анашкин Р.В.

ООО «Связь-Строй»

Приведены результаты разработки новых технических решений для предоставления онлайн интерфейса переводчика. При этом поддерживается обсуждение перевода по сегментам и использование памяти переводов. Получено оригинальное решение для применения в облачных технологиях при автоматизации переводов. Разработаны средства технического обеспечения онлайн-управления локализацией информационных ресурсов. В разработанный программный комплекс (ПК) встроена система управления базой данных с открытым кодом, создающая единую среду хранения данных. При этом ПК коллективного перевода текстов, работающий с тысячами активных пользователей (исполнители, менеджеры и переводчики), инициирует множество подключений и запросов к БД. Наряду с обеспечением подключений осуществляется обработка в реальном времени обращений переводчиков к памяти переводов. Это, в свою очередь, приводит к большому количеству запросов полнотекстового поиска по ключевым словам, фразам и предложениям. Эффективная работа ПК обеспечивается в условиях, когда система управления БД одновременно обрабатывает более 10 тыс. запросов. Проведен эксперимент, в ходе которого получены данные, с использованием которых выполнен анализ эффективности работы серверной и клиентской подсистем ПК (программного комплекса). При этом измерялись временные значения основных параметров, характеризующих производительность ПК (время начала рендеринга, время готовности документа и время получения первого байта ответа).

Ключевые слова: интерфейс переводчика, контент провайдер, локализация сайтов.

PROVIDING ON-LINE CONTROL LOCALIZATION INFORMATION RESOURCES WITH OPEN SOURCE SOFTWARE AND THE ABILITY TO INTEGRATE WITH THE MAIN OPEN-SOURCE CONTENT

Benevolenskiy S.B., Kiryanoff A.A., Anashkin R.V.

OOO “Svyaz-Stroy”

The results of the development of new technical solutions to provide online translators interface. Herewith a discussion of transfer to the segments and to use the translation memory. An original solution for use in the cloud technology in the automation of translation. Developed by means of technical support online control localization of information resources. In the developed program complex (PC) embedded database management system, open source, create a single storage environment. In this case, the PC collective translation of texts, working with thousands of active users (artists, managers, and translators) initiates multiple connections and queries to the database. Along with providing the connections will be processed in real-time applications for translators translation memory. This, in turn, leads to a large number of requests for full-text search by key words, phrases and sentences. Efficient operation of a PC is provided in an environment where database management system simultaneously handles more than 10 thousand requests. The experiment, in which the data were obtained with the use of which has evaluated the performance of the server and client subsystems PC (software). At the same time we measured values of key parameters characterizing the performance of the PC (start time rendering, time commitment and time that a document of the first byte of the response).

Keywords: interface translation, content providers, website localization.

Введение. Задачи перевода контента многочисленных информационных интернет-ресурсов на иностранные языки становятся все более актуальными в настоящее время в связи с интенсивным развитием информационных технологий во всем мире. В соответствии с этим разрабатываются решения таких задач.

Цель исследования. Повышение эффективности решения задач локализации информационных ресурсов, обеспечение масштабируемости таких систем и возможности предоставления ими онлайн интерфейса переводчика с поддержкой памяти переводов, организацией обсуждения перевода по сегментам и предоставлением интерфейса переводчика с глоссарием.

Материал и методы исследования. Актуальность задач перевода контента многочисленных информационных Интернет-ресурсов на иностранные языки продиктована интенсивным развитием информационных технологий во всем мире.

Для повышения эффективности решения задачи локализации информационных ресурсов [4; 5] с открытым программным кодом нами используется REST API [2], обеспечивающий масштабируемость системы и возможности предоставить онлайн интерфейс переводчика с поддержкой памяти переводов и организовать обсуждение перевода по сегментам предоставлением интерфейса переводчика с глоссарием.

На этой основе разработаны программные модули, позволяющие предоставлять онлайн интерфейс переводчика с поддержкой памяти переводов, организовать обсуждение перевода по сегментам и предоставить интерфейс переводчика с глоссарием.

На процесс перевода с иностранного языка, как известно [1; 3], оказывают сильное влияние культурологические аспекты исходного и целевого языков. В связи с этим упомянутый процесс непосредственно не поддается алгоритмизации из-за психологических особенностей. В данной работе содержатся результаты разработки программной архитектуры системы перевода, основанной на памяти значений слов и словосочетаний, хранящихся в символьном формате.

При этом в памяти переводов находятся сегменты, соответствующие по определенному условию искомому, т.е. создается анализатор тэгов «нечетких совпадений» с учетом текстуального и культурологического контекста перевода и фразеологических оборотов. При этом решаются такие задачи, как:

1) сегментация;

2) обработка специальных символов и информации о форматировании.

В данном анализаторе учтен опыт работы с подобными системами, такими как «Translator's Workbench» фирмы Trados, «Transit» фирмы Star, «IBM Translation Manager» (фирмы IBM) и «DejaVu» фирмы Atril, позволяющими:

- найти в переводимом тексте ранее переводившиеся фрагменты;
- вручную перевести текст с автоматической подстановкой перевода для ранее переводившихся фрагментов;

• автоматически проверить тождественность переводов для разных вхождений одинаковых слов, словосочетаний и фрагментов.

В основе работы анализатора лежит использование расстояния Дамерау-Левенштейна [6].

На этой основе языковая пара, из исходного и переведенного сегментов, записывается в память переводов, и, когда переводчик приступает ко второй фразе, система находит аналогию и выводит определенный таким образом сегмент. Переводчик при этом использует ранее сделанный перевод, основанный на полном использовании лингвистического аппарата (морфология, синтаксис, семантика). Далее сегмент считается переведенным, и в памяти переводов появляется еще одна языковая пара.

Таким образом, перевод выполняется собственно человеком, но с привлечением вычислительной техники. Интерфейс переводчика показан на рисунке 1.

В разработанный программный комплекс (ПК) встроена система управления базой данных с открытым кодом, создающая единую среду хранения данных.

При этом ПК коллективного перевода текстов, работающий с тысячами активных пользователей (исполнители, менеджеры и переводчики), инициирует множество подключений и запросов к БД. Наряду с обеспечением подключений осуществляется обработка в реальном времени обращений переводчиков к памяти переводов. Это, в свою очередь, приводит к большому количеству запросов полнотекстового поиска по ключевым словам, фразам и предложениям. Эффективная работа ПК обеспечивается в условиях, когда система управления БД одновременно обрабатывает более 10 тыс. запросов.

Решение задачи достижения максимального времени отклика в 1000 мс получено нами в рамках решения проблемы 10 000 соединений (c10k) как на уровне веб-сервера, обрабатывающего множество асинхронных запросов от клиентского веб-интерфейса, так и на уровне БД.

В связи с этим для снижения времени отклика и повышения доступности системы и ее сервисов было использовано распределенное хранение данных (шардинг) и создана подсистема управления БД, позволяющая выполнить интеграцию в ПК на различных этапах разработки.

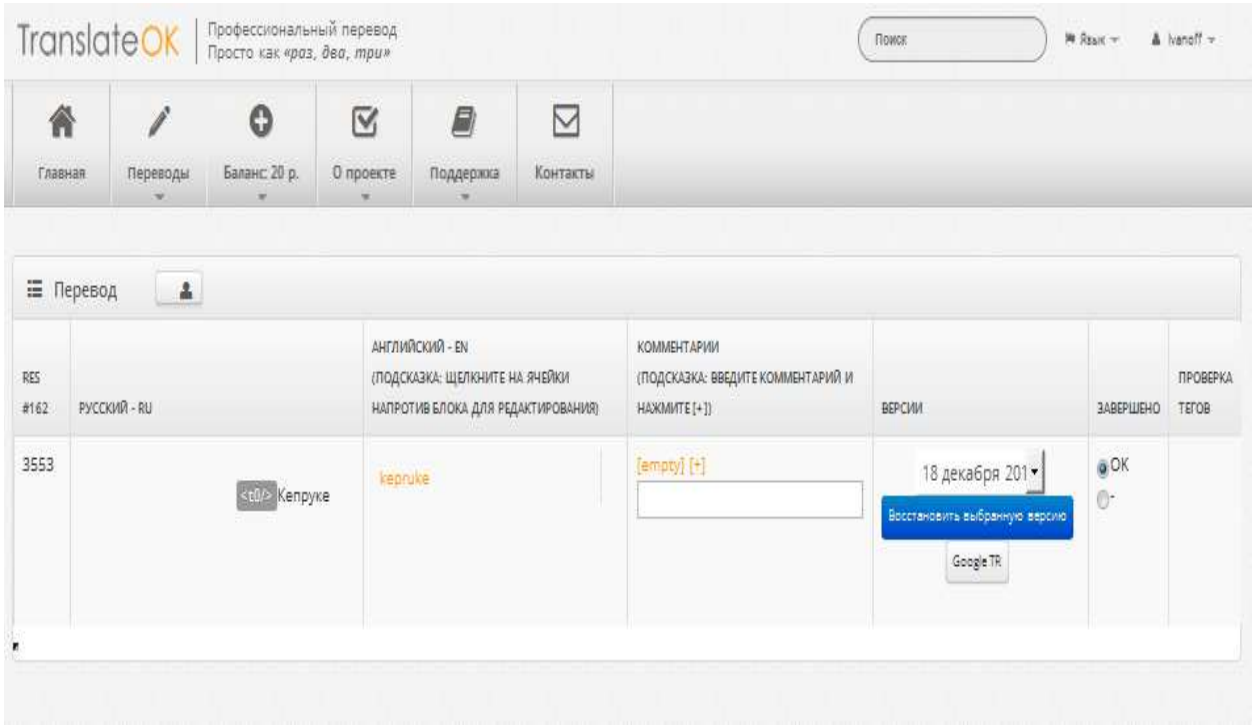

Рисунок 1 – Интерфейс переводчика.

При этом MySQL-proxy работает «прозрачно» для клиента и может передавать запросы между несколькими slave- и master-серверами, а также позволяет выполнять:

- балансировку нагрузки;
- failover;
- перехват, фильтрацию и модифицирование запросов;
- контроль доступа;
- обработку результатов.

При этом Mysql-proxy принимает входящие запросы от модуля взаимодействия с БД, модифицирует их и распределяет между менее нагруженными серверами, принимает ответы, обрабатывает и возвращает их клиенту. Встроенный в прокси-сервер скриптовый язык Lua позволяет увеличить функционал платформы в целом, добавляя новый функционал.

Параллельно с этим происходит оптимизация запросов (рис. 2) путем добавления нескольких запросов в общую очередь, при этом в случае отправки очереди запросов происходит возврат одного конечного результата от БД.

Также для снижения времени отклика системы применена технология полнотекстового поиска — Apache Solr, использующая API библиотеки The Apache Lucene. Это решение применяется для высокоскоростного полнотекстового поиска в Интернете и других источниках.

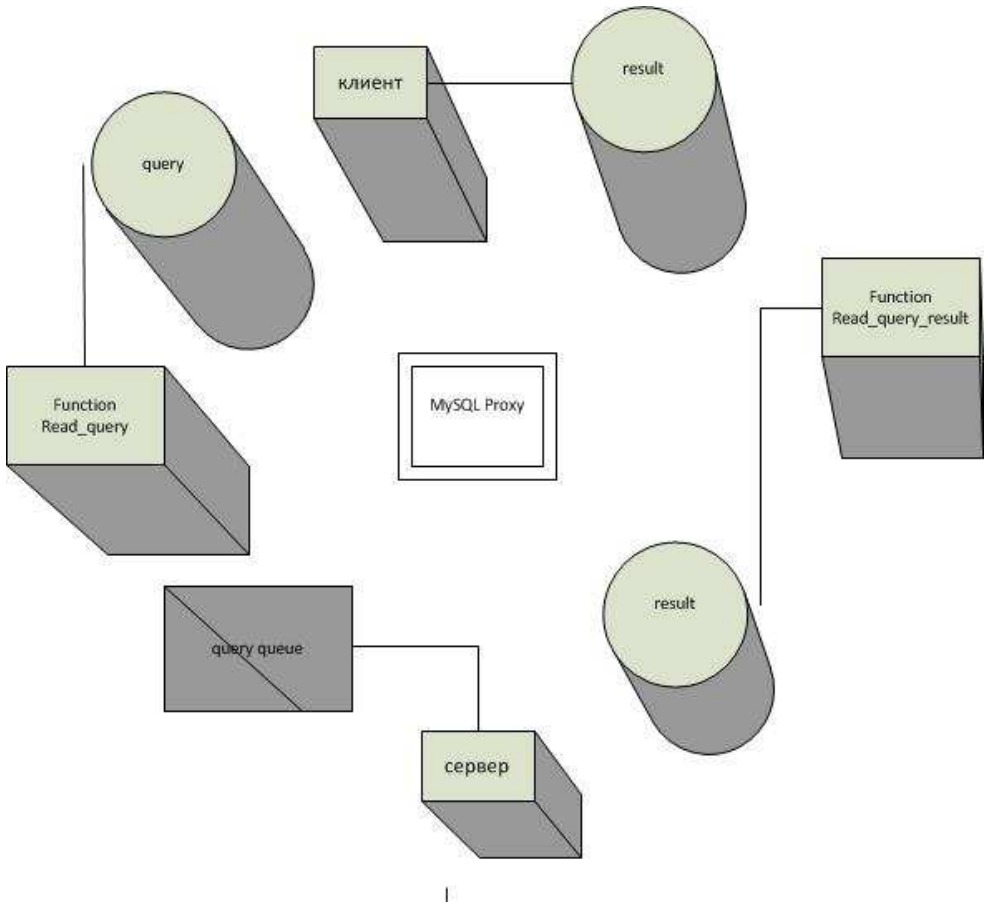

Рисунок 2 – Процесс оптимизации запросов

Данная подсистема является платформой, масштабируемой под различные задачи - от полнотекстового поиска на сайте до распределенной системы хранения/получения/аналитики текстовых и т.п. данных, обладающей развитым языком запросов. Apache Solr расположен над mysql-proxy.

Это ПО может быть расширено в рамках разрабатываемого комплекса при помощи встроенного в него языка программирования, через который реализуется вышеперечисленный и дополнительный функционал.

При этом ПК имеет некоторые специфические особенности системы, которые при проектировании СУБД необходимо учесть. В первую очередь, это большое число пользователей, которые будут активно пользоваться услугами переводов.

В связи с этим для снижения времени отклика и повышения доступности системы и ее сервисов была использована модель распределенного хранения и обработки данных (шардинг) и создана подсистема управления БД, позволяющая выполнить интеграцию в ПК на различных этапах разработки.

Суть метода заключается в том, что все пользовательские запросы данных попадают в модуль MySQL Proxy. Данный модуль при получении запроса на чтение данных производит

оценку доступных подчиненных (slave) серверов MySQL, и затем перенаправляет запрос на наиболее разгруженный сервер. А при получении запроса на запись направляет запрос главному серверу. Также распределение нагрузки на несколько серверов позволяет увеличить стабильность системы, а сама структура MySQL-proxy позволяет отделить пользователя от БД, что повышает защищенность. При этом идентичность интерфейсов MySQL и MySQL-proxy делает их взаимозаменяемыми, увеличивая стабильность и время реагирования на внештатные ситуации, а использование идентичного интерфейса работы при этом оптимизирует внутренние процессы и уменьшает время отклика.

На этой основе создан соответствующий программный комплекс [7].

Результаты исследования и их обсуждение. Экспериментальная оценка эффективности работы серверной и клиентской подсистем ПК основывается на измерении временных значений трех ключевых параметров, отражающих общую производительность системы ПК, а именно времени готовности документа, времени получения первого байта ответа и времени начала рендеринга.

В результате эксперимента, проведенного с использованием канала обмена информацией производительностью 100Мб/с, были получены данные, сведенные в табл. 1.

Таблица 1.

Местонахождение клиентского компьютера	**Время готовности документа, с**	**Время получения первого байта ответа, с**	**Время начала рендеринга, с**
Токио, Япония	0,3815	0,1264	0,3429
Буэнос-Айрес, Аргентина	0,2229	0,0743	0,2157
Москва, Россия	0,2025	0,0279	0,2157

Эти данные показывают, что в ПК со значительным «запасом» решается проблема с10к.

Выводы. Таким образом, обеспечено онлайн-управление локализацией информационных ресурсов с открытым исходным кодом и поддержкой интеграции с основными открытыми источниками контента. При этом повышена эффективность решения задач локализации информационных ресурсов, обеспечения масштабируемости таких систем и возможности предоставления ими онлайн интерфейса переводчика с поддержкой

памяти переводов, организацией обсуждения перевода по сегментам и предоставлением интерфейса переводчика с глоссарием.

Работа выполняется при поддержке Министерства образования и науки РФ (государственный контракт № 07.524.11.4020).

Список литературы

1. Вежбицкая А. Понимание культур через посредство ключевых слов / пер. с англ. А.Д. Шмелева. - М. : Языки славянской культуры, 2001. – 288 с.
2. Джон Фландерс Введение в службы RESTful с использованием WCF. MSDNMagazine (январь 2009).
3. Иванова С.В. Культурологический аспект языковых единиц / Башкирский университет. - Уфа, 2002. – 116 с.
4. Качественный и быстрый перевод текстов онлайн [Электронный ресурс]. - Режим доступа: http://mrtranslate.ru/perevod.php (дата обращения: 08.04.13).
5. Kochereshkin K. Transifex помогает пользователям всего мира локализовать продукты ROSA [Электронный ресурс]. - Режим доступа: http://www.rosalab.ru/blogs/transifex-pomogaet-polzovatelyam-vsego-m (дата обращения: 08.04.13).
6. Damerau F.A. Technique for Computer Detection and Correction of Spelling Errors // Communications of the ACM. - 1964. - Vol. 7. - No. 3. - P. 171–176.
7. Translateok [Электронный ресурс]. – Режим доступа: http://advisor.wmtransfer.com/SiteDetails.aspx?url=translateok.com&tab=rl (дата обращения: 08.04.13).

Рецензенты:

Галушкин А.И., д.т.н., профессор, начальник лаборатории «Интеллектуальные информационные системы» Федерального государственного научного учреждения «Центр информационных технологий и систем органов исполнительной власти», г. Москва.

Марсов В.И., д.т.н., профессор Московского автомобильно-дорожного государственного технического университета (МАДИ), кафедра «Автоматизация производственных процессов», г. Москва.