0.1. *Кононов Д.Д.* Средства автоматизированного мониторинга контента веб-пространства научно-образовательных ресурсов Красноярского края

Работа посвящена разработке средства автоматизированного мониторинга контента на примере веб-сообщества образовательных ресурсов Красноярского края. В рамках данной работы выполнен анализ связности сайтов данного сообщества за 2013 и 2014 годы.

Спроектирована концептуальная модель данных для хранения информации о сайтах как элементах веб-пространства. Полученная модель позволяет сохранять полную информацию о сайтах, страницах и связях между ними посредством ссылок. Ссылки имеют порядок: 1-й порядок — ссылка между страницами одного сайта, 2-й порядок — ссылка между разными сайтами внутри исследуемого сообщества, 3-й порядок — ссылка на сайты, не входящие в исследуемое сообщество. Также модель позволяет отслеживать динамику изменения элементов модели во времени с сохранением истории изменения. Временные срезы осуществляются посредством создания снимков сайтов. Снимок включает полную карту связности страниц сайтов в рамках заданного сообщества. На основе концептуальной модели создана реляционная модель данных.

Разработаны параллельные алгоритмы обхода сайтов на основе модели гипервизора [1]. Гипервизор управляет обработчиками, осуществляя их запуск, остановку и контроль выполнения заданий. Разработаны алгоритмы обработки страниц сайтов с вычленением метаинформации и ключевых слов. Также реализованы методы снижения нагрузки на веб-сервера при обходе сайтов сообщества.

На основе разработанных алгоритмов реализована автоматизированная система мониторинга контента веб-ресурсов. В системе используется расширенная ролевая модель безопасности [2]. Система реализована на языке PHP с использованием библиотек CURL, PDO, PCRE, phpQuery. В качестве СУБД для хранения данных используется PostgreSQL.

Система использовалась для анализа связности научно-образовательных ресурсов Красноярского края за 2013 и 2014 годы. В настоящее время в системе зарегистрировано 27 научно-образовательных организаций, имеющих 70 сайтов. В системе хранится 315145 страниц с общим объемом 10.5 Гб, 170678 ресурсов (файлов) с общим объемом 100.52 Гб, 6552857 ссылок 1-го уровня, 3822 ссылок 2-го уровня, 2585509 ссылок 3-го уровня и 4021767 ссылок на ресурсы. Также зарегистрировано 6065 внешних сайтов, имеющих 148258 уникальных страниц.

В рамках проведенного исследования выявлена тенденция к увеличению объема сайтов и количества размещаемых электронных документов. Установлено малое число ссылок 2-го уровня, что свидетельствует о слабой связности между членами данного сообщества. Одновременно высокое число ссылок 3-го уровня показывает высокую степень связности с внешними сайтами. Также установлено активное использование средств социальных сетей на страницах сайтов сообщества.

Работа выполнялась в рамках междисциплинарного интеграционного проекта СО РАН № 21 "Исследование закономерностей и тенденций развития самоорганизующихся систем на примере веб-пространства и биологических сообществ".

Список литературы

[1] Yang X. Liu L. Principles, Methodologies, and Service-Oriented Approaches for Cloud Computing / ISI Global, 2013. 452 p.

[2] Кононов Д.Д. Исаев С.В. Модель безопасности веб-приложений на основе мандатного ролевого разграничения доступа // Вестник БГУ. — 2012. — № 9, С. 29–32.