Наукометрические измерения в электронных библиотеках на основе рубрикаторов научной информации

М.Р. Когаловский, С.И. Паринов

Аннотация

Ряд научных систем электронных библиотек располагает средствами статистических измерений востребованности (количества просмотров и загрузок) содержащихся в них информационных объектов. Результаты этих измерений могут агрегироваться по их авторам и организациям, в которых созданы эти информационные объекты. Вместе с тем, большой интерес представляет также тематически структурированная статистика востребованности. Для идентификации тематики информационных объектов могут использоваться широко признанные научные классификационные системы или рубрикаторы научной информации. В данной статье рассматриваются функции сервиса системы Соционет, обеспечивающего указанные тематически структурированные статистические измерения.

Ключевые слова: электронная библиотека, наукометрия, рубрикатор научной информации, система Соционет, наукометрический сервис, тематический запрос, ГРНТИ, JEL.

1. Введение

Одной из важных наукометрических задач в области электронных библиотек является получение статистической информации о востребованности представленных в библиотеке научных публикаций. Эта задача решается в ряде библиотечных систем. В развитых системах может быть получена указанная статистика, агрегированная по авторам публикаций, а также по организациям, в которых они работают. Однако наряду с этим значительный интерес представляет статистика востребованности, ассоциированная с тематическими направлениями в интересующих областях знаний. Представляет также не меньший интерес и анализ тематической структуры состояния контента библиотеки.

При добросовестном рубрицировании информационных объектов библиотеки на стадии создания (или обновления) их метаданных и при достаточной представительности контента указанные статистические данные позволяют более или менее достоверно ранжировать уровни активности научных исследований в тех или иных разделах науки в рамках научного сообщества в целом, а также в рамках отдельных крупных научных коллективов, например, в национальных или международных научных центрах, университетах или исследовательских институтах. Можно также выявлять исследователей, наиболее активно публикующихся в интересующем научном направлении, или работы которых по данной тематике наиболее востребованы научным сообществом.

Для решения указанных задач необходимо иметь средства идентификации научных направлений в различных областях знаний. Для этой цели могут использоваться широко принятые национальными и/или международными научными сообществами универсальные или специально разработанные для какой-либо отдельной области знаний научные классификационные системы (Scientific Classification Systems) или рубрикаторы научной и научно-технической информации. В нашей стране создан универсальный рубрикатор – Государственный рубрикатор научно-технической информации (ГРНТИ) [1], определяющий иерархическую структуру различных областей знаний, которая предназначена для систематизации всего потока научно-технической информации. Для рубрикации публикаций в области экономики Журналом экономической литературы (Journal of Economic Literature) Американской экономической ассоциацией создан классификатор Journal of Economic Literature Classification System (JEL) [2], получивший международное признание. Универсальные и специализированные рубрикаторы обычно имеют иерархическую структуру, как правило, небольшой глубины. Так, ГРНТИ и JEL имеют глубину в три уровня.

Рубрикаторы используются в ряде электронных библиотек для поиска представленных в них информационных объектов, которые описываются метаданными, включающими атрибут рубрикации. Значение этого атрибута представляет собой набор рубрик используемого рубрикатора, к которым можно отнести данный информационный объект в соответствии с его содержанием. Вместе с тем, рубрики, соответствующие информационным объектам, могут использоваться и для целей, обсуждаемых в данной работе – для формирования в электронных библиотеках статистических данных востребованности информационных объектов интересующей тематики, а также и наличия их в библиотеке с агрегацией по рубрикам. Насколько известно авторам, такие возможности в настоящее время в известных научных электронных библиотеках, к сожалению, не реализуются.

Нужно заметить, что существующие широко распространенные рубрикаторы научной и научно-технической информации позволяют лишь весьма агрегированно характеризовать тематику информационных объектов электронных библиотек. Тем не менее, они являются своего рода стандартами де-факто, и привлекательной стороной их использования является именно их широкое распространение для рубрицирования информационных объектов.

2. Система Соционет – среда реализации

Адекватную среду для реализации сервиса, генерирующего тематически структурированные данные востребованности и наличия информационных ресурсов электронной библиотеки, обеспечивает система Соционет [3, 4], основанная на технологиях открытых архивов [5].

В настоящее время Соционет фактически обладает статусом институционального информационного пространства Отделения общественных наук РАН, которое включает информационные ресурсы более 30 институтов отделения. Помимо этого, система обеспечивает доступ к информационным ресурсам ряда других академических и образовательных учреждений, а также научных библиотек. Соционет выступает, таким образом, в роли интегратора научно-образовательных информационных ресурсов из многих источников. Репозиторий метаданных системы, интегрирует метаданные около 1500 открытых архивов (по данным на сентябрь 2012 г.). В нем содержатся описания более 1.5 млн информационных объектов, составляющих около 4500 электронных коллекций. Соционет обеспечивает доступ к ним и поддерживает сложную структуру связей между информационными объектами (более 840 тысяч связей). Имеются средства мониторинга изменений в ее информационном пространстве и уведомления пользователей об этих изменениях. Система располагает сервисами для сбора статистики доступов к представленным в ней публикациям и на этой основе может

генерировать дифференцированные с точностью до одного дня разнообразные наукометрические характеристики.

В Соционет поддерживаются рубрикаторы ГРНТИ и JEL. В репозитории метаданных системы для каждого информационного объекта этих открытых архивов имеется его описатель (карточка) – набор метаданных, описывающих его свойства и связи с другими информационными объектами системы. Один из атрибутов описателя содержит совокупность кодов рубрик какого-либо из поддерживаемых системой рубрикаторов, которым соответствует содержание данного информационного объекта. Система обеспечивает также возможности поиска информационных объектов по соответствующим им рубрикам.

Пользователям Соционет доступны наукометрические сервисы, генерирующие статистику востребованности представленных в системе информационных объектов в заданном интервале времени. Можно получить также статистику востребованности информационных объектов, агрегированную по их авторам, по организациям (местам работы авторов), по коллекциям информационных объектов, по открытым архивам, составляющим контент Соционет, а также по пользователям системы, которые при этом идентифицируются IP-адресами их компьютеров.

В Соционет также формируются механизмы и специальные пользовательские интерфейсы, обеспечивающие многослойное семантическое структурирование ее контента на основе онтологии связей между представленными в системе информационными объектами. Технология децентрализованного решения этой задачи авторами информационных объектов и пользователями системы при модерировании администратором Соционет подробно описана в работах [6, 7]. Поддержка семантической структуры контента системы позволяет осуществлять ряд весьма полезных нетрадиционных наукометрических измерений (*см. там же*).

В настоящее время, как уже отмечалось, наукометрический инструментарий системы включает новый сервис, который по запросам пользователей позволяет получать статистику востребованности представленных в ней информационных объектов, агрегированную по рубрикам ГРНТИ или JEL, а также некоторую другую необходимую информацию. Рассмотрим возможности указанного сервиса подробнее.

3. Функциональные возможности сервиса, основанного на рубрикаторах

3.1. Спецификация статистических запросов

В системе Соционет все статистические функции запрашиваются пользователями через единый системный интерфейс. Интерфейсная веб-страница статистических сервисов системы (рис. 1) становится доступной, если на домашней странице системы выбрать в главном системном меню альтернативу «Статистика Соционет». В таблице, выполняющей функции меню на интерфейсной странице, каждому статистическому сервису (запрашиваемому виду наукометрической статистики), соответствует отдельная строка. Такая строка предусмотрена и для обращения к сервису, генерирующему статистику на основе тематических рубрикаторов (последняя строка таблицы).

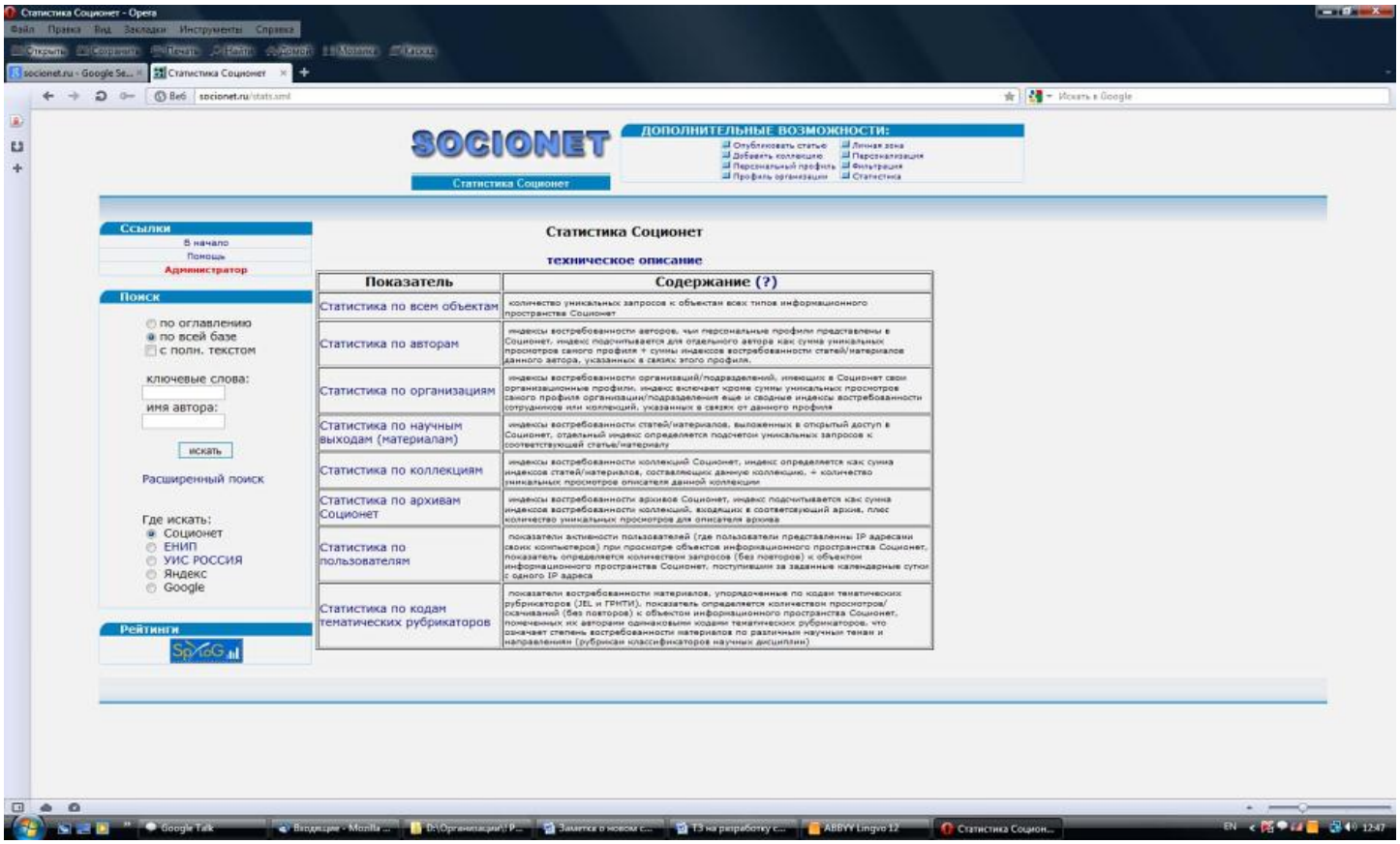


Рис. 1. Интерфейсная веб-страница статистических сервисов Соционет

После обращения к этому сервису пользователю выдаются результаты запроса с принятыми по умолчанию параметрами. Фрагмент веб-страницы с результирующими данными показан на рис. 2. По умолчанию принимаются следующие значения для параметров этого первичного запроса:

- *Оцениваемый период времени доступов:* предыдущие сутки
- *Используемый рубрикатор:*оба поддерживаемые рубрикатора – ГРНТИ и JEL
- *Состав интересующих рубрик:* рубрики всех уровней обоих рубрикаторов, для которых отлично от нуля число загрузок и/или просмотров, с агрегированием результатов для рубрик более высоких уровней

- *Упорядочение строк результирующей таблицы:* по количеству загрузок информационных объектов, соответствующих рубриками включаемым в таблицу.

Следует еще раз подчеркнуть, что в результирующие данные включаются данные для тех рубрик, по которым ненулевое значение имеет количество загрузок или просмотров относящихся к ним информационных объектов.

Далее пользователь имеет возможности уточнить свой первичный запрос, изменяя значения перечисленных параметров. Рассматриваемый сервис предоставляет для этого следующие возможности:

- Задание иного оцениваемого периода времени доступов
- Выбор нужного тематического рубрикатора (JEL или ГРНТИ)
- Выбор конкретной рубрики для получения более детальной статистики
- Задание иного способа упорядочения результирующих данных.

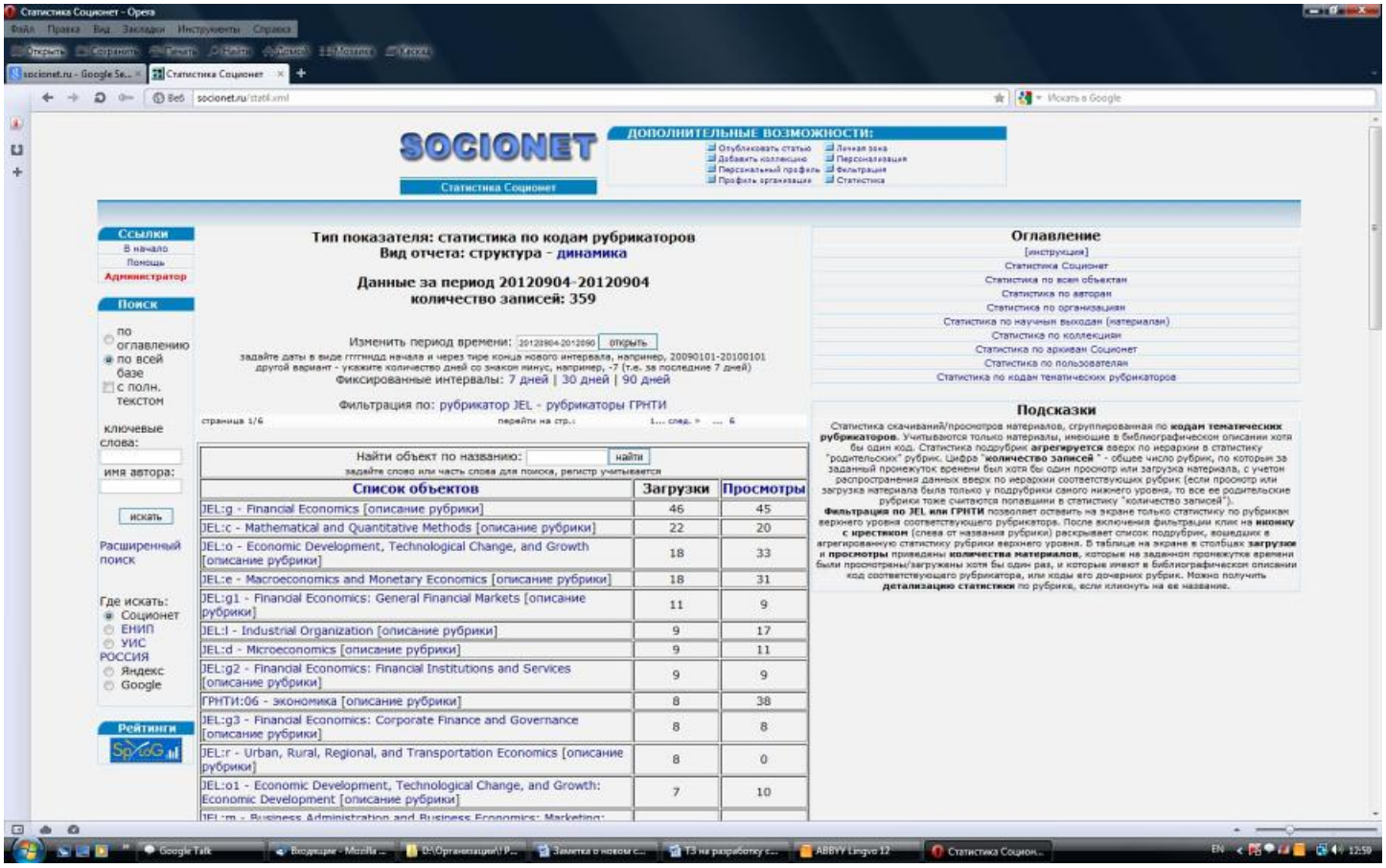


Рис. 2. Фрагмент веб-страницы табличных результирующих данных

Для задания оцениваемого периода времени можно выбрать один из вариантов в меню периодов (последние 7 дней, 30 дней, 90 дней) либо задать произвольное нужное число последних дней (целое число со знаком минус) или интервал дат в окне периода времени (даты задаются в формате ггггммдд).

Нужный рубрикатор выбирается в меню рубрикаторов. В настоящее время в системе Соционет, как уже указывалось, поддерживаются пока только два рубрикатора – ГРНТИ и JEL.

Для выбора нужной рубрики необходимо кликнуть указателем мыши на строке нужной рубрики в списке рубрик, включенных в результирующие данные, которые выдаются по умолчанию в результате первичного запроса. Если для этой рубрики выдано агрегированное значение, можно детализировать его по рубрикам более низких уровней. Для этого нужно кликнуть на соответствующем ей флажке в начале ее строки. После того, как строка агрегированной рубрики «раскроется» и появятся строки подчиненных рубрик, входящих в ее состав, следует выбрать конкретную рубрику более низкого уровня.

Упорядочить результирующие данные можно несколькими способами в случае, если они запрашиваются в виде статистической таблицы со строками, соответствующими конкретным рубрикам тематических рубрикаторов. Можно упорядочить таблицу по именам рубрик (кликнув на имени столбца «Список объектов»), по количеству загрузок или просмотров информационных объектов, соответствующих данным рубрикам. Для этого также необходимо кликнуть курсором мыши на имени соответствующего столбца.

3.2. Результаты обработки тематических статистических запросов

По запросам к рассматриваемому сервису можно получить результаты нескольких разновидностей:

- Статистическую таблицу загрузок и просмотров информационных объектов, рубрицированных всеми рубриками верхнего уровня тематического рубрикатора и/или подрубриками более низких уровней.
- Описание интересующей рубрики.
- График, характеризующий динамику (временные ряды) востребованности (просмотров/загрузок) информационных объектов, рубрицированных конкретной рубрикой тематического рубрикатора или всеми его рубриками, с таблицей временных рядов дневной периодичности показателей загрузки и просмотра информационных объектов за последний двухмесячный период, на основе которых

этот график построен.
- Список востребованных в заданном периоде времени информационных объектов, рубрицированных какой-либо отдельной рубрикой, представленной в статистической таблице.

Рассмотрим эти разновидности результатов обработки запросов несколько подробнее.

Статистическая таблица. **Фрагмент страницы результирующих данных, представленных в виде статистической таблицы, показан на рис. 2. Таблица состоит из трех столбцов. В левом столбце содержатся коды и имена рубрик используемого в запросе тематического рубрикатора. В двух остальных столбцах таблицы указываются количества загрузок и просмотров информационных объектов соответствующих рубрик. В результирующие таблицы для первоначальных запросов (с заданными по умолчанию значениями параметров) включаются рубрики всех уровней обоих рубрикаторов системы, для которых хотя бы одно из соответствующих им количеств загрузок или просмотров отлично от нуля. При этом на основе данных рубрик каждого из двух нижних уровней генерируются данные непосредственно предшествующего верхнего уровня.**

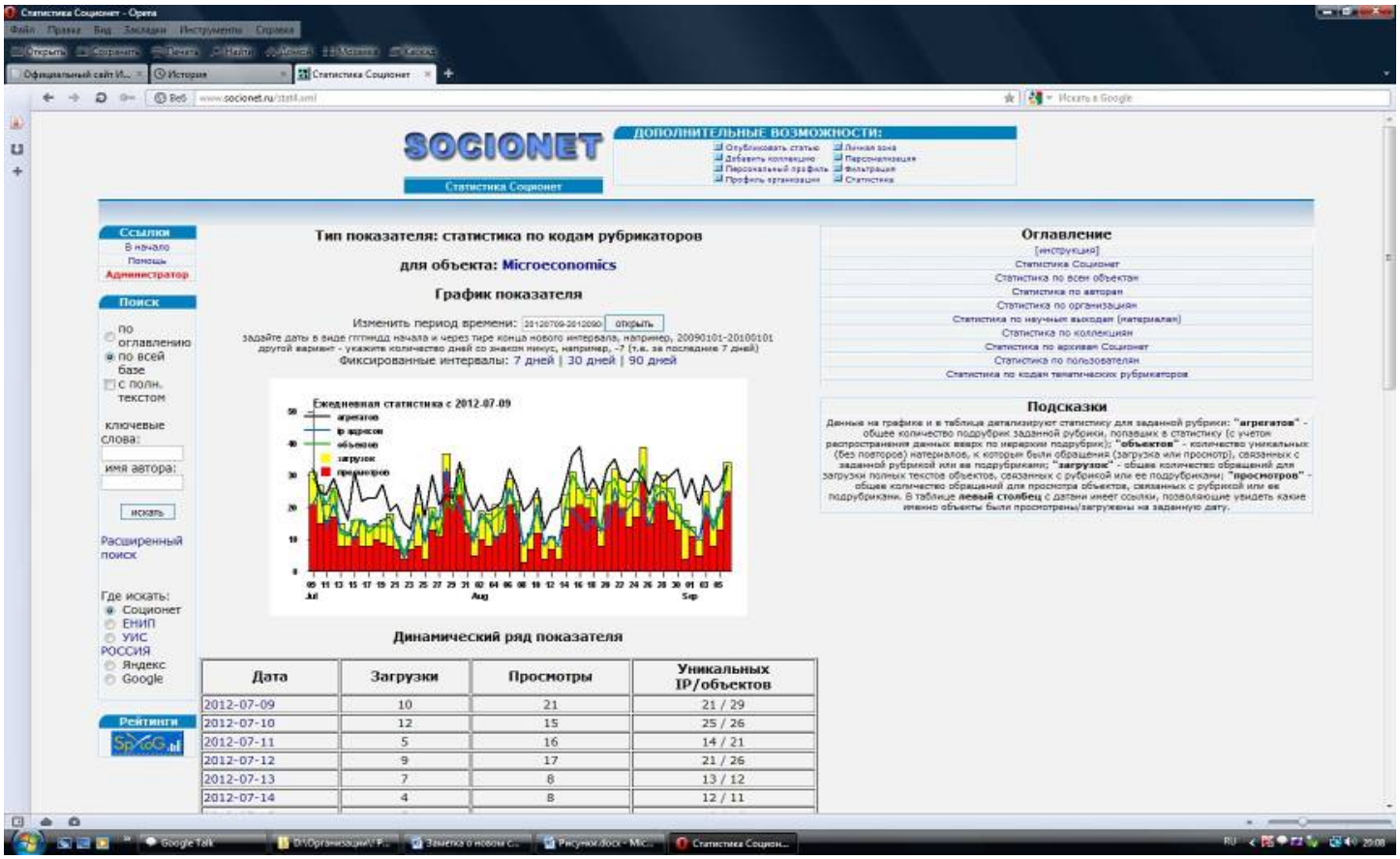
В результирующей таблице строки рубрик, для которых имеются ненулевые значения загрузок или просмотров по каким-либо их подрубрикам (агрегированные рубрики), снабжены флажком слева от кода рубрики. Флажок со значком «+» позволяет разворачивать (детализировать рубрику - показать данные ее подрубрик), флажок со знаком «-» позволяет свернуть такую рубрику – отказаться от ее детализации.

При выборе рубрики для детализации будут выведены как агрегированные статистические данные для этой рубрики, так и данные по ее подрубрикам. Упорядочение подрубрик осуществляется при этом по убыванию числа загрузок относящихся к ней информационных объектов. Одновременно в таблице результатов сохраняется показ значений по остальным рубрикам и детализирующим их подрубрикам, выведенным в таблице на предыдущих шагах. После этого при необходимости можно детализировать данные для выбранной детализированной подрубрики, запросив ее детализацию, т.е. осуществить детализацию статистики до самого нижнего (третьего) уровня рубрикатора.

Если после вывода результирующей таблицы по первичному запросу пользователь выбирает нужный ему рубрикатор, то система выведет таблицу, включающую только рубрики верхнего уровня выбранного рубрикатора, для которых агрегированные для них значения количеств загрузок или просмотров отличны от нуля.

Описания рубрик. **Просмотреть описание какой-либо из представленных в таблице рубрик можно, перейдя по имеющейся в каждой строке таблицы гиперссылке на это описание.**

Графические результаты. **Если кликнуть на имени рубрики в любой строке таблицы, независимо от того, сгенерирована ли для нее детализированная по ее подрубрикам статистика (т.е. указан ли флажок «+» слева от кода этой рубрики в соответствующей ей строке), то можно увидеть динамику обращений пользователей к информационным объектам, соответствующим этой рубрике, т.е. в графическом и в табличном виде дневные ряды за последние два месяца показателей просмотра и загрузки, количеств востребованных информационных объектов, ip-адресов, с которых они запрашивались для просмотра или загрузки, а также агрегатов.**



Дата	Загрузки	Просмотры	Уникальных IP/объектов
2012-07-09	10	21	21 / 29
2012-07-10	12	15	25 / 26
2012-07-11	5	16	14 / 21
2012-07-12	9	17	21 / 26
2012-07-13	7	8	13 / 12
2012-07-14	4	8	12 / 11

Рис. 3. График и временные ряды для рубрики *JEL:d Microeconomics*

Таблица временных рядов состоит из четырех столбцов. В первом из них содержатся последовательные даты анализируемого двухмесячного периода, завершающегося вчерашней датой. Во втором столбце указываются соответствующие этим датам количества загрузок информационных объектов. В третьем столбце содержатся количества просмотров. В четвертом столбце приводятся количества

уникальных (без повторений) IP-адресов и информационных объектов.

Фрагмент страницы результатов запроса с графиком и несколькими начальными строками таблицы временных рядов показан на рис. 3.

Следует отметить, что графики и таблицы временных рядов, аналогичные описанным, можно получить также после вывода статистической таблицы – ответа на первичный запрос, если перейти в режим «Динамика». В этом случае данные на графике и в таблице временных рядов соответствуют всем рубрицированным по обоим рубрикаторам информационным объектам системы. Если же после вывода результирующей таблицы первичного запроса выбрать желаемый рубрикатор и, получив соответствующую ему статистическую таблицу, перейти в режим «Динамика», то график и временные ряды будут относиться только ко всем информационным объектам системы, прорубрицированным по этому рубрикатору.

Рис. 4. Фрагмент веб-страницы со списком востребованных работ

4. Планируемые доработки сервиса

В планах развития функциональных возможностей обсуждаемого сервиса предусматривается включение в него дополнительных функциональных возможностей. Наиболее существенные из них следующие:

Предусматривается возможность генерации по запросам пользователей статистики, аналогичной рассмотренной выше, оценивающей не только востребованность информационных объектов системы, но и наличие их в системе.

Представляется важным также получение статистики обоих видов (востребованности и наличия), агрегированной по рубрикам тематических рубрикаторов, не на контенте всей системы Соционет, а на ресурсах отдельных организаций или отдельных открытых архивов, составляющих ее контент.

Предполагается, наконец, обеспечить для заданного периода времени и выбранного рубрикатора ранжирование рубрик не только верхнего, но и более низких уровней тематических рубрикаторов в рамках заданной рубрики более высокого уровня, по количеству загрузок/просмотров востребованных или по количеству имеющихся в Соционет в момент запроса информационных объектов, соответствующих этим рубрикам.

Благодарность

Авторы выражают благодарность В.М. Ляпунову, выполнившему программную реализацию описанного в статье инструментария.

Литература

1. ГРНТИ – рубрикатор научно-технической информации. Редакция 2007 года. – (Рус.). – URL: http://www.grnti.ru/ [10 сентября 2012]

2. Journal of Economic Literature (JEL) Classification System. – (Engl.). – URL: http://www.aeaweb.org/jel/jel_class_system.php [10 сентября 2012]

3. Онлайновая научная инфраструктура Соционет. – (Рус.). – URL: http://www.socionet.ru/ [10 сентября 2012]

4. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Электронные библиотеки. 2003. Выпуск 1. – (Рус.). – URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP [10 сентября 2012]

5. Open Archives Initiative. - (Engl.). – URL: http://www.openarchives.org/ [10 сентября 2012]

6. Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек /Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011. Воронеж, 19-22 октября 2011 г.». – г. Воронеж: Воронежский государственный университет, 2011.

7. Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках //Информатика и ее применения, 2012. Т. 6. Вып. 3. С. 31-41.

Об авторах

Когаловский Михаил Рувимович – ведущий научный сотрудник Института проблем рынка РАН, личная страница: http://www.cemi.rssi.ru/mei/win1251/kogalov.htm, e-mail: kogalov@cemi.rssi.ru

Паринов Сергей Иванович – главный научный сотрудник Центрального экономико-математического института РАН,

личная страница: http://socionet.ru/pub.xml?h=repec:rus:ecoper:parinov_sergey.56054-1, e-mail: sparinov@gmail.com

Работа поддержана грантом РГНФ (проект 11-02-12026-в).