DOI: 10.18137/RNU.V9187.23.02.P.167 УДК 027.06:004

А.В. Фролов, Е.А. Верещагина, А.Л. Золкин

BIG DATA В БИБЛИОТЕКАХ И НАУЧНЫХ ИССЛЕДОВАНИЯХ

Аннотация. Целью данного исследования является системный анализ цифрового менеджмента библиотек (особенно его инструментария) на основе больших данных и их анализа (Big Data, Data Mining), а также других аналитических инструментов решения актуальных задач эволюции цифровых библиотек, площадок для читателей-исследователей. Используя методы системного подхода, в частности моделирования, проводится классификация подходов и задач управляемости библиотечными ресурсами. Представлен сравнительный анализ возможностей библиотек (темпов, объемов и др.) в условиях наличия белого (гауссова) шума, мешающего консолидации данных, их фильтрации для оценки процессов. Описана формальная модель библиотечного мониторинга, позволяющая эффективно исследовать глубинные библиотечные связи. Предложенная модель используется для преодоления разрывов связей.

Ключевые слова: библиотека, цифровая библиотека, библиотечная экосистема, аналитика, вебометрика, менеджмент, Big Data.

A.V. Frolov, E.A. Vereshchagina, A.L. Zolkin

BIG DATA IN LIBRARIES AND RESEARCH

Abstract. The purpose of this article is a systematic analysis of the digital management of libraries (data steward, researcher, digital data, content manager, etc.), especially its tools, in particular, based on big data and their analysis (Big Data, Data Mining). Other analytical tools are also used to solve urgent problems of the evolution of digital libraries, platforms for readers and researchers. Not only the interest in the analytical management tools of "Libraries 4.0" was used, but also the possibilities for the evolution of relations with librarians-researchers, team information and library activities. Using the methods of a systematic approach, in particular, modeling, an analysis (classification) of approaches, problems of library resource management is carried out. A comparative analysis of the capabilities of libraries (rates, volumes, etc.) is carried out in the presence of "white" (Gaussian) noise that interferes with data consolidation, their filtering to evaluate processes, etc. A formal model of library monitoring is described that makes it possible to effectively explore "deep" library connections. Our analysis and the proposed model are used to overcome the "breaks" of links.

Keywords: library, digital library, library ecosystem, analytics, webometrics, management, Big Data.

Введение

В библиотечном деле количество документов, источников информации возрастает динамически не только по объему, но и по характеру данных, их важности. Большие данные позволяют обрабатывать и управлять как структурированными, так и неструктурированными массивами данных.

Библиотекам необходимы адаптивные информационно-технологические решения для изменившихся потребностей читателей и своих возможностей. Цель исследования – системный анализ задач цифровых библиотек, центров и площадок для читателей-исследователей.

Методы исследования

Библиотекам нужны ИТ-ресурсы сопровождения потребительских запросов, приложений и сервисов [1], а также библиотекари, владеющие цифровым инструментарием ав-

Фролов Александр Владимирович
системный администратор отдела информационных технологий, Морской государственный университет имени адмирала Г.И. Невельского, город Владивосток. Сфера научных интересов: сетевые технологии, программирование, моделирование систем и процессов. Автор более 30 опубликованных научных работ. SPIN-код: 7889-3026, AuthorID: 887541.
Электронный адрес: vip.al75@list.ru

Верещагина Елена Александровна
кандидат технических наук, доцент, доцент департамента информационной безопасности Института математики и компьютерных технологий, Дальневосточный федеральный университет, город Владивосток. Сфера научных интересов: сетевые технологии, программирование, цифровая трансформация, цифровая экономика. Автор более 40 опубликованных научных работ. SPIN-код: 1607-7454, AuthorID: 287436.
Электронный адрес: eretr11@mail.ru

Золкин Александр Леонидович
кандидат технических наук, доцент кафедры информатики и вычислительной техники, Поволжский государственный университет телекоммуникаций и информатики, город Самара. Сфера научных интересов: автоматика и автоматизация, информатика и вычислительная техника, прикладная информатика, программирование, транспорт. Автор 450 опубликованных научных работ. SPIN-код: 1685-2404, AuthorID: 540174.
Электронный адрес: alzolkin@list.ru

томатизации и интеллектуализации библиотечных процессов, – библиотекари-исследователи, помогающие в интеллектуальном поиске данных [2; 3].

Современные библиотеки активны в развитии компетенций специалистов библиотечного дела. Появились новые менеджеры в библиотеках: стюард (data steward) данных (data manager), исследований (management librarian), цифровых данных (digital data specialist), контента (content curator) и др. [4]. Они отвечают за управление в процессах сбора, обработки и хранения данных, их актуализацию (использование), мониторинг и организацию, правовую поддержку [5; 6].

Эти специалисты – цифровые посредники (intermediators), поддерживающие знания, фильтрующие их от информационного шума (белого, гауссова) с целью повышения качества поиска и его релевантности.

Мягкие навыки (soft skills) библиотекаря способствуют развитию библиотечной экосистемы «Библиотека 4.0», а именно:

- развитию отношений с исследователями;
- повышению ИТ- и гуманитарных мультикомпетенций библиотекарей;
- повышению навигационных возможностей в информационной среде;
- работе в команде;
- повышению отдачи от информационных продуктов, сервисов (аналитики, дайджестов и др.) и востребованности услуг по их переработке и упорядочиванию (примеры – ЭБ РГБ [7], ГПНТБ СОРАН [8]);
- развитию индивидуальных технологий информационно-библиотечной деятельности.

Решение указанных задач требует библиотечной аналитики, стратегии и особых инструментов, которые можно классифицировать по следующим критериям:

– вебметрический – счетчики, лог-анализаторы, внутренняя аналитика, маркетинговая аналитика;

– вебаналитический – анализ сайтов, соцмедиа;

– условия доступа – бесплатный, условно-бесплатный, коммерческий;

– защищенность – открытый, закрытый (инвайты), смешанный;

– оптимизация – контента, репутации, маркетинга;

– функциональность – юзабилити, карты поведения, дерево целей.

Результаты

Мониторинг ситуаций в библиотечной экосистеме стал динамичным, но одновременно и сложным. BigData (термин введен в 2008 г. изданием *Nature*) – системный термин для технологий анализа и использования больших и неструктурированных массивов данных, обработка которых ранее была невозможной [1; 9]. Базируется на реальном режиме и соблюдении требований «5V»:

1) Volume – применение инновационных технологий;

2) Variety – параллельность обработки данных (независимо от формата, источника данных);

3) Velocity – темп, скорость обработки потока данных;

4) Value – извлечение полезных свойств из данных;

5) Veracity – релевантность данных и связей.

В РФ актуальна мобилизация библиотечных знаний.

Система Big Data позволяет сосредоточиться на технологиях обработки данных, их анализе, Data Mining. Потребуется сравнительный анализ с реальными показателями библиотек, их использованием (темп, объем, воронка, отклонения). Временным рядам свойственен белый шум, мешающий консолидации данных (корреляция, гипотезы, избыточность и др.), но их фильтрация дает ценную информацию.

Без релевантной аналитики невозможно принятие решения, ситуационное прогнозирование. Библиотекари применяют различный инструментарий – свой на каждом этапе цикла актуальности данных.

Big Data [10] в библиотечном деле и исследованиях имеет следующие преимущества:

- упрощение и оперативность мониторинга;
- рост читательской аудитории;
- взаимодействие с потребителями данных и их поставщиками, рост его темпа.

Big Data помогает создавать профиль модели потребителя, что снижает издержки обработки данных, сложность прогноза и мониторинга. Модели больших библиотечных данных многомерны и сложны. Но распределенная обработка в библиотечных узлах (кластерах) данных позволяет лучше понимать контекст данных. Здесь необходимы вычисления в памяти (IMC), обработка «на лету», применение предикативной аналитики.

Рассмотрим информационную библиотечную экосистему, опирающуюся на профили читателей, объектов, представляемых тематическими связями и ориентированных на группу пользователей минимальной достаточности.

Пусть $I = \{i_1, \ldots, i_n\}$ – элемент экосистемы (определенная группа задач, потребителей). Каталог библиотечных ресурсов $K = I_1, \ldots, I_N$, *где* $i_j \in I_i$ – i-й ресурс, а j_i принадлежит классу информации S (предметно ориентированных по данным D).

Востребованность ресурсов по различным группам аудитории за период $T = t_1, \ldots, t_k$ различна. Частота запросов $f_k\left(i_j\right)$:

$$f_{max} = \max_T \left(f_{t_1}(i_j), \ldots, f_{t_k}\left(i_j\right) \right),$$

где f_{max} – максимум из всех частот запросов к библиотечным ресурсам; f_{t_k} – частота запросов к i_j на промежутке t_k.

Ранжирование по востребованности производим по индексу, который позволяет фильтровать данные *K*:

$$w_k = \frac{f_k}{f_{max}},$$

Определим шкалу востребованности элементов $W_K = \left\{ w_{k_i} \right\}$, чтобы построить классы востребованности библиотечных ресурсов.

Модель инфологической обработки (вычислений) позволяет обрабатывать большие данные в оперативной памяти, что дает следующие преимущества:

- упрощение анализа (сокращение уровней структурирования) данных;
- повышение адаптивности и релевантности модели;
- активность структур запросов, их фильтрация в структурах памяти.

Сначала идентифицируем классы атрибутов сущностей, которые соответствуют потребностям анализа. Детализация данных – задача библиотекаря, предметного аналитика.

Ключевые базы данных и знаний (далее – БД и БЗ) библиотечной экосистемы определим следующим образом:

- БД запросов, профилей, задач и ситуаций;
- БЗ сценариев;
- БД-БЗ отслеживания функций, сценариев, устойчивости;
- БЗ интерфейсной поддержки;
- БД-БЗ мониторинга;
- БЗ-БД оценки и решений.

Привлекаются экспертные, эвристические, статистические и другие методы.

Оценивать устойчивость можно по формуле

$$K = S / \sum_{i=1}^{n} \alpha_i S_i,$$

где *S* – инфологический показатель потенциала экосистемы; S_i – аналогичный структурный (модульный) показатель; *n* – число подсистем; α_i – коэффициент важности (вес) фактора *i*.

Заключение и выводы

Идентификация библиотечных событий основывается на Big Data. Без Data Analytics эффективно решать идентификационные задачи в реальном режиме невозможно. Исследование глубинных связей, Data Mining помогают в решении этой проблемы. Таким образом, можно сделать вывод, что выполненный анализ и предложенная модель могут быть использованы для преодоления «разрывов» связей.

Литература

1. *Редькина Н.С.* «Надпрофессиональные» навыки и профессиональные знания библиотечного специалиста: требования времени // Библиотековедение. 2019. Т. 68. № 6. С. 647–658. EDN RGDZRW. DOI: 10.25281/0869-608X-2019-68-6-647-658
2. *Фролов А.В.* Машинное обучение: типы и модели // Системный администратор. 2021. № 4 (221). С. 94–95. EDN IPIPML.
3. *Yumashev A., Koneva E., Borodina M., Lipson D., Nedosugova A.* Electronic apps in assessing risk and monitoring of patients with arterial hypertension // Prensa Medica Argentina. 2019. Vol. 105. No. 4. P. 235–245. EDN VWYKCH.
4. *Brochu L., Burns J.* Librarians and Research Data Management. Review: Commentary from a Senior Professional and a New Professional Librarian // New Review of Academic Librarianship. 2019. Vol. 25. No. 1. P. 49–58. DOI: http://dx.doi.org/10.1080/13614533.2018.1501715
5. *Казиев В.М., Казиев К.В., Казиева Б.В.* Основы правовой информатики и информатизации правовых систем. М. : ИНФРА-М, 2011. Сер. «Вузовский учебник». ISBN 978-5-9558-0157-5. EDN QRSUOT.
6. *Чирков М.С., Лачинина Т.А., Чистяков М.С.* Знания и информация как синергия платформенного подхода цифровизации глобального развития // Свободная мысль. 2020. № 5 (1683). С. 37–44. DOI: 10.24411/0869-4435-2020-00003
7. *Давыдова Н.Р.* Электронная библиотека РГБ: этапы развития и особенности формирования цифровых коллекций // Библиотековедение. 2019. Т. 68. № 2. С. 144–154. EDN EHJUTF. DOI: 10.25281/0869-608X-2019-68-2-144-154
8. *Ударцева О.М.* Менеджмент библиотечных веб-ресурсов // Научные и технические библиотеки. 2020. № 2. С. 105–124. EDN JXWAAM. DOI: 10.33186/1027-3689-2020-2-105-124
9. *Золкин А.Л., Мунистер В.Д.* Проектирование цифровых экосистем окружающего интеллекта, сенсорных и компьютерных сетей: монография. М. : Русайнс, 2022. 148 с. ISBN 978-5-4365-9267-1. EDN LZYEEM.
10. *Фролов А.В., Титова А.А., Верещагина Е.А.* BigData и виртуальные ЦОД // Промышленные АСУ и контроллеры. 2022. № 2. С. 25–29. EDN AJUXPV. DOI: 10.25791/asu.2.2022.1347.

References

1. Redkina N.S. (2019) Over-professional skills and professional knowledge of library specialist: Demands of the time. *Bibliotekovedenie* [Library and Information Science (Russia)]. Vol. 68. No. 6. Pp. 647–658. DOI: 10.25281/0869-608X-2019-68-6-647-658 (In Russian).
2. Frolov A.V. (2021) Machine learning: Types and models. *Systemnyi administrator*. No. 4 (221). Pp. 94–95. EDN IPIPML. (In Russian).
3. Yumashev A., Koneva E., Borodina M., Lipson D., Nedosugova A. (2019) Electronic apps in assessing risk and monitoring of patients with arterial hypertension. *Prensa Medica Argentina*. Vol. 105. No. 4. Pp. 235–245. EDN VWYKCH.
4. Brochu L., Burns J. (2019) Librarians and Research Data Management. Review: Commentary from a Senior Professional and a New Professional Librarian. *New Review of Academic Librarianship*. Vol. 25. No. 1. Pp. 49–58. DOI: http://dx.doi.org/10.1080/13614533.2018.1501715
5. Kaziev V.M., Kaziev K.V., Kazieva B.V. (2011) *Osnovy pravovoi informatiki i informatizatsii pravovykh sistem* [Fundamentals of legal informatics and informatization of legal systems]. Moscow: INFRA-M Publ. Ser. "University textbook". ISBN 978-5-9558-0157-5. (In Russian).

6. Chirkov M.S., Lachinina T.A., Chistyakov M.S. (2020) Knowledge and information as a synergy of the platform approach to digitalization of global development. *Svobodnaya mysl'*. No. 5 (1683). Pp. 37–44. DOI: 10.24411/0869-4435-2020-00003 (In Russian).
7. Davydova N.R. (2019) Electronic Library of the RSL: Development Stages and Features of Formation of Digital Collections. *Bibliotekovedenie* [Library and Information Science (Russia)]. Vol. 68. No. 2. Pp. 144–154. DOI: 10.25281/0869-608X-2019-68-2-144-154 (In Russian).
8. Udartseva O.M. (2020) Managing library www-resources. *Scientific and Technical Libraries*. No. 2. Pp. 105–124. DOI: 10.33186/1027-3689-2020-2-105-124 (In Russian).
9. Zolkin A.L., Munister V.D. (2022) *Proektirovanie tsifrovykh ekosistem okruzhayushchego intellekta, sensornykh i komp'yuternykh setei* [Designing digital ecosystems of ambient intelligence, sensor and computer networks: Monograph]. Moscow : Ruscience Publ. 148 p. ISBN 978-5-4365-9267-1. (In Russian).
10. Frolov A.V., Titova A.A., Vereshchagina E.A. (2022) Big Data and virtual data centers. *Industrial Automatic Control Systems and Controllers*. No. 2. P. 25–29. DOI: 10.25791/asu.2.2022.1347 (In Russian).