

библиографического описания. Каждый элемент списка одновременно является гиперссылкой на полное библиографическое описание, включающее аннотацию и ссылку на полный текст, если такой присутствует в базе данных.

Сервис избирательного распространения информации (ИРИ) представляет собой типовую услугу информирования пользователей о выходе новых изданий или публикаций по интересующей их теме. Информирование осуществляется в автоматизированном режиме с определенной периодичностью. Абоненты ИРИ могут самостоятельно через Интернет заказать тематику запросов к базе данных, используя предложенный рубрикатор, а также указать периодичность информирования. Сведения о выполненных запросах автоматически отправляются абонентам по электронной почте. Существуют также средства обратной связи абонента с системой, при которой возможна уточняющая коррекция запросов пользователя.

Разработанная автоматизированная система способствует выполнению одной из задач библиотеки – эффективному информационному обеспечению научных исследований в области экологии, охраны окружающей среды и природопользования.

УДК 004.912 : 021

**А. Г. Федорчук,**  
научный сотрудник НБУВ

### **СОЗДАНИЕ БИБЛИОТЕЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ МОНИТОРИНГА СОДЕРЖАНИЯ ПУБЛИКАЦИЙ В СЕТИ ИНТЕРНЕТ**

Публикация посвящена вопросам технологии мониторинга содержания текстов публикаций средств массовой информации, размещенных в сети Интернет, и созданию на основе этих исследований новых информационных ресурсов. Рассмотрена история и преимущества внедрения технологии на основе стандартных решений, которые предлагает система автоматизации библиотек «ИРБИС64».

*Ключевые слова:* средства массовой информации, информационные ресурсы, мониторинг содержания, библиографическая база данных.

A publication is devoted to the questions of technology content-monitoring of mass media and creation of new information resources on the basis of these researches. History and advantages of introduction of technology is considered on the basis of standard decisions, the library automation system «IRBIS64» offers.

*Keywords:* mass media, information resources, content monitoring, a bibliographic database.

### **Актуальность исследования сетевых СМИ**

Средства массовой информации (СМИ) на сегодняшний день являются одними из крупнейших коммуникативных средств общества. Именно СМИ во всем мире воплощают и представляют общественное мнение, осуществляют функции наблюдателя и контролера законодательной, исполнительной и судебной власти. Одновременно СМИ играют важную роль в гражданском обществе, создавая общественную трибуну, которая может служить для выражения и лоббирования общественных и частных интересов.

Значение СМИ в обществе и их роль в формировании взглядов и вкусов граждан закономерно вызывает повышенный интерес исследователей различных отраслей науки. Информация социально-политической тематики, которая появляется в СМИ, наиболее точно отражает состояние и структуру политической и общественной системы любой страны и взаимосвязи основных институтов государственной власти и гражданского общества. СМИ, с одной стороны, в любом обществе являются инди-

катором, который позволяет определить степень зрелости и структурирования общества, с другой, в методологическом смысле могут быть тем дескриптором, с помощью которого возможно описание, интерпретация и оценка перспектив политической и общественной системы в целом.

В свою очередь, современный этап развития общества требует пристального внимания со стороны государства, общественности и науки к тем процессам и феноменам, которые могут активно влиять на стратегические основы и тактическое наполнение пути развития государства. В процессе исследования общественно-политических преобразований СМИ в последние годы становятся как источником, так и объектом анализа. Методологическое и технологическое развитие СМИ в современном мире все сильнее влияет не только на степень информированности граждан, но и на характер, направления и темпы развития самого общества.

### **Особенности сетевых СМИ**

Сегодня границы между сетевыми СМИ и традиционной прессой постепенно стираются, происходит их взаимопроникновение. Информационные и развлекательные порталы, сетевые информационные агентства широко используют информационные потоки, рождаемые печатными публикациями, совершенствуют журналистскую подготовку своих сотрудников, заказывают аналитику у солидных издательств и известных корреспондентов. Печатные издания, в свою очередь, с каждым днем расширяют свое присутствие в сети, обзаводятся новостными порталами, совершенствуют интерактивное общение с читателем. То, что происходит в этой области, многие исследователи характеризуют как «четвертую коммуникационную революцию», характеризующуюся развитием коммуникаций, созданием компьютерных сетей и программных средств, способных обеспечить накопление и передачу огромных массивов информации в глобальном масштабе. В результате сеть Интернет сама по себе превращается в одно из важнейших СМИ, что обеспечивает, с одной стороны, дополнительный канал для распространения информации, с другой — поступление различных данных для всех СМИ. Кроме того, конвергенция (взаимное проникновение) информационных Интернет-ресурсов и традиционных СМИ создает предпосылки расширения аудитории СМИ, объединяя ее во времени и пространстве с аудиторией Интернета.

Миграция различных форм СМИ в Интернет демонстрирует развитие принципиально новых видов информационных процессов, которые интегрируют исторический опыт традиционных СМИ и новые возможности сети. Другими словами, Интернет становится новым открытым пространством для сбора, хранения и распространения информации, важным фактором развития публичной сферы и гражданского общества. С точки зрения качества любых научных исследований в обществе, фактор взаимного проникновения электронных и традиционных СМИ выдвигает дополнительные требования к технологиям мониторинга информационных потоков, которые они порождают.

Существует мнение, что в сети Интернет есть всё или почти всё. Однако этим богатством надо уметь воспользоваться, что представляет собой достаточно сложную задачу как для рядовых пользователей сети, так и для профессионалов. Во-первых, объём данных, размещённых в сети, очень велик, а сама эта информация практически не структурирована. Во-вторых, Интернет изменчив: ежедневно в нём появляются новые данные — страницы или целые сайты, что-то исчезает, а часть ресурсов меняет адрес. Наконец сетевые публикации в целом отличаются меньшей достоверностью, чем публикации бумажные, так что информацию, размещённую в сети, ещё нужно каким-то образом оценить или проверить. Всё это затрудняет задачу поиска и превращает выбор стратегии исследования информационных потоков сети в сложную проблему, не имеющую общего решения.

С появлением сети Интернет возникло множество информационно-технических проблем. Одним из основных недостатков является протокол HTTP, который используется для передачи информации и хорошо подходит для навигации по сайтам, но в то же время не очень удобен для поиска информации в сети. То же самое можно сказать и о протоколе передачи файлов FTP, который более примитивен, чем HTTP, и предназначен только для передачи данных.

Основная проблема современной Сети состоит в том, что эти протоколы не позволяют отслеживать изменение информации. Учитывая, что на сегодняшний день в Сети находятся миллиарды страниц, отыскать быстро или вообще найти нужную и актуальную информацию уже не представляется возможным. В результате даже самые технологически оснащенные поисковые системы способны осуществлять поиск лишь в 20–40 % потенциально доступных для данной системы ресурсах сети.

Проблема отчасти решается тем, что каждая поисковая система выбирает те доменные зоны, в которых она может обеспечить быстрое обновление информации и выстраивать соответствующую стратегию обходов Интернет-ресурсов в них так, чтобы отследить как можно большее количество часто обновляющихся сайтов, не растратив свои мощности на старые, неизменяющиеся сайты.

Так, российские поисковые системы Яндекс и Рамблер индексируют сайты в доменных зонах .ru, .ua, .by постсоветских республик и не индексируют зарубежных сайтов; украинские Мета и Аванпорт индексируют исключительно украинские ресурсы и т. д.

Таким образом, каждая поисковая система имеет свое собственное, ограниченное ее ресурсами, множество доступных для поиска документов. Ни одна из подобных систем не имеет и не может иметь достаточных мощностей, чтобы охватывать все ресурсы в сети Интернет, поэтому в любой момент может возникнуть ситуация, когда информационные потребности пользователя не смогут быть удовлетворены. Как правило, в этом случае пользователь переходит на другую поисковую систему и пытается искать то, что ему нужно, там.

В результате – сегодня государственным и крупным коммерческим структурам становится все труднее следить за динамично меняющимся вокруг них информационным полем, которое содержит массу новостных и обзорных материалов. Кроме того, регулярное ознакомление с публикациями СМИ для любой серьезной деятельности необходимо, но не всегда достаточно. Большие массивы информации должны подвергаться качественному анализу, поскольку на основе информации из открытых источников можно не только анализировать состояние дел, но и строить прогнозы развития ситуации, что жизненно важно для принятия верных решений.

В условиях небольшого количества информации аналитики могли бы работать вручную или с помощью типовых офисных программ. Но практика показывает, что уже через несколько дней работы такая служба начинает понимать, что существующий в сегодняшнем мире объем доступной информации невозможно обрабатывать по старинке. Очевидно, что такие службы необходимо вооружать современными поисковыми и аналитическими инструментами.

### **Проблемы поиска в сети Интернет**

Современные поисковые системы в Интернете сегодня столкнулись с целым рядом проблем [3]. Среди наиболее часто упоминаемых можно назвать: разрастание общего объема материалов, расположенных в

Интернете, переход к новым форматам, увеличение количества нетекстовых материалов. Каждая из этих проблем серьезна, в этих направлениях ведутся достаточно продуктивные работы с привлечением лучших мировых специалистов. Однако основной проблемой, с которой встретились современные поисковые системы (ПС), присутствующие на современном рынке средств текстового поиска, и с которой сталкивается практически каждый пользователь сети, – это проблема избытка информации.

Проявилась эта проблема при наращивании объема индексируемых текстовых документов. Рост объема текстов, проиндексированных поисковой системой, привел к тому, что практически любой запрос выдает выборку из многих тысяч, а то и десятков тысяч подходящих документов. Естественно, это превышает максимум, который способен обработать один человек за разумное время. Такой максимум для профессионального аналитика находится в пределах нескольких сотен документов, для непрофессионала, естественно, границы существенно ниже, как минимум, на порядок. Проблема эта не является новой, для ее решения и создавалась технология ПС. Но на сегодняшнем этапе эта технология уже не может адекватно справиться с соответствующими задачами.

Конечно, эта проблема коснулась не только ПС, занимающихся индексированием материалов в Интернете, но и ПС, работающих с любыми другими текстовыми базами, например, электронными библиотеками. К таким системам относятся известные библиотеки НЭБ-НСН, Интегрум-Техно в России, Лексис-Нексис, Рейтер на Западе. Пользователей у таких ПС, естественно, существенно меньше, чем у ПС в Интернете, хотя объем документов, доступных в ПС Лексис-Нексис, не только существенно выше объема Интернет, но даже скорость поступления новых документов в эту библиотеку до сих пор превосходит скорость поступления информации в Интернет. Однако, даже гораздо более строгая организация данных в таких библиотеках, почти полное единство форматов внутри одной библиотеки не являются панацеей от общей проблемы современных ПС.

Сегодня рынок предлагает довольно большое количество поисковых систем. Но работа многих из них построена достаточно примитивно. Система ищет слово или словосочетание. В результате человек получает сотни, а чаще тысячи ненужных документов, содержащих это слово. При этом документ, относящийся к теме, но не содержащий в явном виде этого термина, будет пропущен. Зачастую существует и обратная задача – узнать, о чем пишется в наборе документов или статей, выявить важные темы. С этими задачами или проблемами аналитические центры сталкиваются каждый день. И центров, и информации становится все больше. Недавно

возникнув, спрос на новые интеллектуальные поисково-аналитические системы растет довольно быстро.

### **Инструменты мониторинга сетевых ресурсов**

Под «удобными инструментами для работы с информацией сети», как правило, понимаются компьютерные программы, помогающие собирать и сортировать материалы сетевых СМИ. Таких программ на рынке около сотни, но все они, как правило, занимаются организацией хранения средних или больших объемов информации, обладают простыми поисковыми возможностями и/или тематическим рубрикаторм, не предлагая какого-либо механизма качественного анализа. И это не случайно, ибо качественный анализ текста предполагает оценку таких нюансов, как эмоции, угрозы, характер отношений между объектами... В этих сферах человеческий мозг, способный накапливать опыт и обладающий интуицией, предпочтительнее. Хотя скорость оценки при этом и невелика, но выигрывает качество анализа информации.

Таким образом, при решении интеллектуальных задач компьютерная программа призвана максимально облегчить труд человека: во-первых – обеспечить его ограниченной выборкой документов, отсеив лишнее, во-вторых – предоставить условия для проведения анализа.

### **Аналитические программы для работы с текстами**

Аналитические системы различаются по виду обрабатываемых данных – полнотекстовых или фактографических. Методы обработки фактографических данных известны достаточно давно. Среди них в последнее время особой популярностью пользуются OLAP-анализ и Data Mining (выявление последовательностей, ассоциаций, дерева решений и т. д.). Эти методы в той или иной мере сейчас поддерживаются всеми современными системами. Частично они реализованы в MS OLAP Services и в продуктах компании Business Objects. Наиболее полно – в системе PolyAnalyst компании Megaruter.

Методы анализа текстов распространены гораздо меньше. Это, в основном, тематическое рубрицирование входящего потока документов и подсчет статистики встречаемых слов и словосочетаний. Наиболее известными производителями этих систем и отдельных компонентов являются канадская фирма Hummingbird (продукт Hummingbird Knowledge Management), а также российские компании Media Lingva («Классификатор»), Megaruter (TextAnalyst) и «Гарант-Парк-Интернет» (ее продукт реализован на основе технологий американской фирмы InterMedia).

Для успешного применения такого рода методов входящий поток всегда подвергается предварительной обработке, включающей просмотр оператором, контроль орфографии, нормализацию регистра и т. д. Для последующего контекстного поиска или сложных процедур анализа, как правило, проводится полнотекстовое индексирование содержимого документов.

Главное, что объединяет все системы контент-мониторинга, – это сочетание качественного и количественного анализа текстов. Доверяя в рамках проведения контент-анализа компьютеру количественные оценки, аналитические системы предоставляют человеку возможность качественной оценки исследуемых текстов, помогающей фиксировать и структурировать новый слой знаний для последующего его анализа. Таким образом, данные системы являются новым шагом в плане работы с информацией, и недостатки избытка информации превращаются в достоинство. Ведь чем большее количество разнообразных источников будет проанализировано, тем полнее и качественнее будет информационный портрет исследуемого объекта. Контент-мониторинг, таким образом, дает возможность объективации субъективных мнений, выраженных в большом количестве текстовых материалов СМИ.

По всей видимости, настоящий прорыв в обработке материалов СМИ будет достигнут лишь тогда, когда сами авторы станут сопровождать текст некоторой информационной структурой, описывающей смысл статьи и «знания», в ней изложенные, другими словами, когда возобладает подход, основанный на разделении данных, характеризующих содержание, представление и смысловое значение. Адептом этого подхода является один из основателей современного Интернета Тим Бернерс-Ли. Он предлагает объединить документы подобного рода в единую сеть знаний, которая будет называться Semantic Web.

### **Системы мониторинга содержания публикаций**

Отдел организации и использования документального фонда Фонда Президентов Украины Национальной библиотеки Украины имени В. И. Вернадского уже 20 лет (1993–2013) осуществляет мониторинг содержания газетной периодики [10–12].

У автоматизированной технологии содержательного мониторинга публикаций существуют несколько важных особенностей. Прежде всего – за единицу формирования текстового информационного массива используется ключевой фрагмент публикации. Кроме того – формирование банка ключевых фрагментов публикаций является

объединением двух взаимосвязанных автоматизированных процессов: аналитико-синтетической переработки и многоуровневой процедуры анализа содержания текстов публикаций. Индексация ключевых фрагментов публикаций происходит при помощи многофасетной классификации.

Уникальность предложенной технологии состоит в объединении содержательных и количественных методов анализа текстов. Последовательность этапов содержательного анализа проблемы, которая исследуется конкретной информационной системой, условно можно поделить на содержательный (качественный) анализ совокупности публикаций и формализованный (количественный) анализ информационных массивов: индексного, библиографического и массива текстов ключевых фрагментов публикаций.

При создании и функционировании системы мониторинга по ключевым фрагментам публикаций используются следующие принципы:

- обработка больших потоков газетной периодики;
- выделение из публикаций компактных фрагментов, которые по содержанию соответствуют теме исследования;
- сортировка этих фрагментов и объединение их в однотипные по содержанию подгруппы.

Особенности технологии обусловлены, прежде всего, необходимостью получения результатов отбора материалов, независимых от взглядов или впечатлений информационного аналитика. Объективность при этом достигается тем, что личные убеждения исследователя нивелируются процедурой формализации и квантификации текста. В рамках публикации выделяются отдельные фрагменты, отображающие разные аспекты исследуемой проблемы, которые потом в системе сортируются и при необходимости подсчитываются. Реализация этой процедуры требует специально разработанного детального классификатора, по которому аналитик отбирает и индексирует материал в БД [8].

Для формального описания документов при формировании БД системы по ключевым фрагментам публикаций были выбраны классификаторы фасетно-иерархического типа, где каждому элементу исследуемой проблемы отвечает свой фасет с фиксированным месторасположением в фасетной формуле, а совокупности значений в рамках каждого элемента – соответствующий фасетный класс. Кроме того, с помощью отдельного фасета (фасет модальности) передается отношение автора публикации (позитивное, нейтральное, негативное) к описанному им событию или персоне.

Особенность данной технологии мониторинга газетной периодики состоит еще и в том, что анализ и синтез ключевых цитат публикаций из широкого круга источников осуществляется по продуманным и заданным схемам на основании разработанного классификатора. Иначе говоря, цитаты, лаконично передающие заложенную в первоисточнике идею, могут объединяться разными способами в рамках параметров, представленных в фасетной формуле.

### **Тенденции и перспективы**

Автоматизированная технология контент-мониторинга в течение многолетнего периода успешной эксплуатации постоянно совершенствовалась как в технологическом и методическом плане, так и в плане программного обеспечения. В качестве инструмента для хранения информационных массивов сначала была использована программа Abmagic, которая позволяет сохранять и управлять БД в формате MARC. Для аналитико-синтетической переработки, автоматического анализа и формирования информационно-аналитических материалов использовались собственные программные средства.

Улучшение компьютерного и программного обеспечения, создание локальной сети и появление новых задач по созданию электронных информационных ресурсов, а также обеспечению доступа к ним читателей выявили необходимость адаптации технологии контент-мониторинга новых общеобразовательных библиотечных программных средств и систем автоматизации библиотеки.

### **Система автоматизации библиотек «ИРБИС»**

Система автоматизации библиотек (СAB) «ИРБИС» разработана Государственной публичной научно-технической библиотекой России (ГПНТБ) и является типичным интегрированным решением в области автоматизации библиотечных технологий. Система предназначена для использования в библиотеках любого типа и профиля, полностью отвечает международным требованиям, предъявляемым к таким системам, и поддерживает все отечественные библиографические стандарты и форматы [4, 5]. Кроме того, она позволяет описывать все виды изданий, а интерфейсы системы максимально приближены к потребностям пользователя и легко осваиваются.

Новая версия системы автоматизации библиотек – «ИРБИС64» – начала внедряться в НБУВ с 2009 г., что создало перспективы автоматизации практически всех отраслей библиотечно-информационной деятельности.

Система включает средства, позволяющие использовать как иллюстративный материал любые внешние по отношению к библиографическому документу объекты, такие, как полные тексты, графика, таблицы, аудио-и видеоматериалы, а также ресурсы Интернет. Система работает в архитектуре клиент-сервер, обеспечивая взаимодействие клиентских приложений (пользовательских АРМ) и сервера БД на основе протокола ТСР/ IP как в локальных, так и в глобальных сетях [13].

САБ «ИРБИС64» является современным проблемно-ориентированным программным продуктом. Ее работа основана на клиент-серверной платформе, что дает возможность организовать одновременную работу с библиографической базой данных многих пользователей, наладить клиентские профили в соответствии со специализированными задачами сотрудников.

Система имеет возможности поддержки каталогизации и описания документов любого вида, в том числе документов специализированных фондов: аудио-и видеоматериалы, газеты, электронные ресурсы, карты, ноты. В том числе удобный интерфейс для обработки периодических изданий и газетной прессы, а также автоматизированные механизмы аналитической росписи изданий по их содержанию с возможностью вводить рефераты и аннотации публикаций.

Имеются также развитые механизмы поддержки систематизации и тематического упорядочения собраний документов: рубрикаторы, библиотечные классификации, словари предметных рубрик.

Библиографическое описание документов может сопровождать любое количество внешних объектов, таких как иллюстративный материал, полные тексты, графика, аудио-или видеоресурсы Интернета.

Выходные форматы библиотечно-библиографических баз данных можно получить как в формате RTF для Microsoft Word, так и в формате HTML, предназначенном для онлайн-публикации обработанных материалов.

Форматы представления данных, рабочие листы, вспомогательно-справочный аппарат баз данных, поисковые словари открыты и доступны для настройки и перестройки в соответствии с проблемно-ориентированными задачами.

Все эти преимущества программного библиотечного продукта нового поколения создают предпосылки для профессиональной организации библиографической и информационно-аналитической деятельности библиотеки.

В 2010 г. САБ «ИРБИС64» успешно была внедрена и Отделом

организации и использования документального фонда Фонда Президентів України в качестве среды технологии контент-мониторинга СМИ.

#### Список использованных источников

1. Антонов А. Современные проблемы поисковых систем и некоторые пути их преодоления [Текст] / А. Антонов, В. Мешков // Аналитика – Капитал / под науч. ред. А. В. Рудского). – М. ВИНТИ РАН – ИПКИР Минпромнауки России – Академия аналитики и информатики, 2000. – С. 229–233.

2. Бродовский А. И. Интегрированная библиотечно-информационная система ИРБИС – современное средство для автоматизации малых и средних библиотек [Текст] / А. И. Бродовский // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества : материалы конф., Судак, Автоном. Респ. Крым, Украина, 6–14 июня 1998 г. / 5-я юбилейная междунар. конф. «Крым 98». – М., 1998. – Т. 1. – С. 114–118; Науч. и техн. б-ки. – 1999. – № 1. – С. 87–94

3. Бродовский А. И. Новое поколение системы автоматизации библиотек ИРБИС – ИРБИС64: от электронного каталога к полнотекстовым базам данных [Текст] / А. И. Бродовский, К. О. Сбойчаков // Науч. и техн. б-ки. – М., 2005. – № 2. – С. 107

4. Липинский Ю. В. Средства информационного поиска и навигации в больших массивах неструктурированной информации [Электронный ресурс] / Ю. В. Липинский. – Режим доступа: <http://www.fep.ru/text/dataarrays04.html>. – Назва з екрана.

5. Моніторинг діяльності органів виконавчої влади із застосуванням комп'ютерної системи контент-аналізу електронних ЗМІ / Г. Леліков, В. Сороко, О. Григор'єв, Д. Ланде // Вісн. держ. служби України. – 2002. – № 2. – С. 21–38.

6. Опарин А. Системы мониторинга и анализа СМИ / А. Опарин // PCWEEK. – 2003. – № 47. – С. 19–24.

7. Самойлов Ю. Система мониторинга и анализа СМИ «Медиалогия» [Электронный ресурс] – Режим доступа: <http://integration.ibs.ru/content/rus/rubr65/tubr-652.asp>. – Назва з екрана.

8. Танатар Н. В. Інформаційно-аналітичні системи за ключовими фрагментами публікацій як основа для створення електронної бібліотеки соціально-політичного спрямування / Н. В. Танатар // Наук. пр. Нац. б-ки України ім. В. І. Вернадського. – К. : НБУВ, 2001. – Вип. 6. – С. 190–200.

9. Федорчук А. Г. Контент-мониторинг информационных потоков / А. Г. Федорчук // Библиотеки национальных академий наук: проблемы функционирования, тенденции развития : науч.–практ. и теорет. сб. – 2005. – Вип. 3. – С. 141–150.

10. Федорчук А. Г. Теоретико-методичні засади аналізу інформаційного потоку соціально-політичного спрямування [Текст] / А. Г. Федорчук, Н. В. Танатар // Бібліотекознавство, документознавство, інформологія. – 2004. – № 2. – С. 33–38.