Федорчук А. Г.

Контент-мониторинг информационных потоков

Одной из главных особенностей нашего времени есть постоянный рост темпов производства информации. Этот процесс объективен и в целом, безусловно, позитивен. Однако на сегодняшний день человечество встретилось с парадоксальной, на первый взгляд, ситуацией: прогресс в области производства информации ведет к снижению общего уровня информированности.

Кроме увеличения объемов информации до масштабов, которые делают невозможным ее непосредственную обработку, возник целый ряд специфических проблем, связанных с быстрым развитием информационных технологий.

Ситуация резкого роста темпов производства информации породила ряд проблем:

- непропорциональный рост информационного шума из-за слабой структурированности информации;
- появление паразитной информации (получаемой в качестве приложений);
- несоответствие формально релевантной информации (тематически соответствующей) действительным потребностям ее потребителей;
- многократное дублирование информации (типичный пример — публикация одного сообщения в разных изданиях).

Вследствие перечисленных обстоятельств, традиционные информационно-поисковые системы постепенно стали терять свою актуальность. Причина этого кроется не столько в физических объемах информационных потоков, сколько в их динамике, то есть в постоянном систематическом обновлении информации, которая далеко не всегда имеет очевидную регулярность. Охват и обобщение больших динамических информационных потоков, которые непрерывно генерируются в средствах массовой информации (СМИ), требует качественно новых подходов.

Выход может быть найден только в средствах автоматизации выявления наиболее важных составляющих в информационных потоках. Именно поэтому в последние годы все чаще стали использовать системы мониторинга ресурсов, тесно связанные с контент-анализом. Данное перспективное направление получило название контент-мониторинг. Его появление было вызвано, прежде всего, задачами систематического отслеживания тенденций и процессов в постоянно обновляющейся информационной среде. Под контент-мониторингом чаще всего понимают содержательный анализ информационных потоков с целью получения необходимых качественных и количественных срезов, который ведется постоянно на протяжении не определенного заранее промежутка времени. Важнейшей методологической составляющей контент-мониторинга есть контент-анализ — понятие, достаточно заезженное социологами.

Метод контент-анализа

Контент-анализ начинался как количественно-ориентированный метод анализа текстов для изучения массовых коммуникаций. Впервые он был применен в 1910 году социологом Максом Вебером (Max Weber) для оценки охвата печатью политических акций в Германии. Американский исследователь средств коммуникации Гарольд Лассвелл (Harold Lasswell) в 30-40-е годы использовал подобную методику для изучения содержания пропагандистских сообщений военного времени.

С появлением средств автоматизации, текстов в электронном виде, начиная с 60-х годов прошлого века, начальное развитие получил контент-анализ информации больших объемов — баз данных и интерактивных медиа-источников. Традиционное политическое использование современных технологий контент-анализа было дополнено неограниченным перечнем рубрик и тем, охватывающих производственную и социальную сферы, бизнес и финансы, культуру и науку. Этот процесс, в свою очередь, сопровождался большим количеством разнородных программных систем.

Развитие технологических систем невозможно без стандартизации. Понятие же контент-анализа, берущего свое начало в психологии и социологии, сегодня пока не имеет однозначного определения. Это порождает ряд проблем, важнейшая из которых состоит в том, что программные системы, построенные на основе разнообразных подходов к контент-анализу, в общем случае несовместимы.

Приведем лишь некоторые определения контент-анализа:

- Контент-анализ — это методика объективного качественного и систематического изучения содержания средств коммуникации (Д. Джери, Дж. Джери).
- Контент-анализ — это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Рич).
- Контент-анализ — это качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и представляет собой квантификационную обработку текста с дальнейшей интерпретацией результатов (В. Иванов).
- Контент-анализ состоит из поиска в тексте определенных содержательных понятий (единиц анализа), выявления частоты их появления и соотношения с содержанием всего документа (Б. Краснов).
- Контент-анализ — это исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Э. Таршис).

Большинство из приведенных выше определений конструктивные, т.е. процедурные. Через разные начальные подходы они порождают разнообразные алгоритмы, которые временами противоречат друг другу. Существующие разнообразные подходы к пониманию контент-анализа поддаются целиком оправданной критике. Наибольшие сомнения, на наш взгляд, вызывает игнорирование роли контекста. Тем не менее, несмотря на многообразие трактовок контент-анализа, большое прикладное значение методологии все же позволяет избежать многих противоречий. Объединение средств и методов, их естественный отбор путем многократной оценки полученных результатов открывают возможность выделения и подтверждения знаний, а также фактической силы и полезности данного инструментария.

Диапазон методов и процедур, которые касаются самого процесса контент-анализа, очень широкий. Но наиболее важным при подготовке исследования, как показал опыт, есть выполнение следующих действий:

- описание проблемной ситуации, поиск цели исследования;
- точное определение объекта и предмета исследования;
- предварительный анализ объекта;
- содержательное уточнение и эмпирическая интерпретация понятий;
- описание процедур регистрации свойств и явлений;

- определение общего плана исследования;
- определение типа выборки, круга источников и т.п.

Интересной особенностью контент-анализа есть и то, что эту методологию до последнего времени связывали с определенной сферой человеческой деятельности (политикой и социологией). Тем не менее, на сегодня контент-анализ все шире применяется во многих областях политической и экономической жизни, что способствует большему прикладному значению использованных в методологии контент-анализа философских категорий, социологии и лингвистики. Контент-анализ в рамках исследования информационных потоков — новое направление, которое предусматривает анализ массива текстовых документов — результатов мониторинга информационного пространства.

Общепризнанным есть распределение методологии контент-анализа на две ветви: качественную и количественную. Основа количественного контент-анализа - частота появления в документах определенных характеристик содержания. Метод качественного контент-анализа базируется на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.

Метод качественного контент-анализа основан на том, что в любой фазе количественного контент-анализа для оценок результатов может быть привлечен эксперт. Таким образом, этот метод призван обеспечить эксперта необходимыми средствами для выводов и дополнительных результатов. Эксперт с помощью таких средств может обнаружить определенные свойства части информации и проверить их относительно общего текстового потока, а общие свойства текстового потока распространить на его конкретную тематическую часть.

Процесс метода качественного контент-анализа состоит из трех основных стадий.

Первая — сведение большого количества текстовой информации к конечному числу интегрированных блоков текста — единиц содержания, которые кодируются для дальнейшей обработки этих блоков. Основными единицами содержания являются категории, последовательности и темы.

Вторая стадия качественного контент-анализа — реконструкция субъективных составных текстового потока — системы значений, мыслей, взглядов и доказательств каждого источника текста.

Третья стадия — формирование выводов и обобщений путем сравнения индивидуальных систем значений.

Метод количественного контент-анализа, в свою очередь, как правило, состоит из трех основных этапов. На первом этапе выделяются единицы анализа и переводятся в форму, приемлемую для обработки (сегодня — в электронный вид). Второй этап состоит в подсчете частот единиц анализа с применением разнообразного математического аппарата для выявления взаимосвязей между ними. Суть третьего этапа состоит в интерпретации полученных результатов. При этом без привлечения искусственного интеллекта, объемных семантических формализаторов, даже экспертов как таковых, с использованием только математических методов могут быть получены содержательные, семантически наполненные результаты.

В качестве примеров можно привести автоматическое формирование дайджестов, автоматическое выявление взаимосвязи понятий (категорий), автоматическую

кластеризацию взаимосвязей для выявления наиболее важных, автоматическое выявление окраски взаимосвязей, в простейшем случае — определения положительных и отрицательных взаимосвязей.

Одной из важнейших проблем в методологии контент-анализа есть процесс категоризации. Использование набора категорий задает концептуальную сетку, в терминах которой анализируется текстовый поток.

Исследования текстового потока, если он достаточно большой, можно проводить двумя путями.

Первый путь — определение конечной, но заведомо избыточной, совокупности категорий для получения количественных данных о встречаемости некоторых из них. При этом предполагается и автоматическая или полуавтоматическая кластеризация (деление на группы и классы) неупорядоченной последовательности категорий и, соответственно, получение на ее основе новых обобщенных категорий.

Второй путь — выявление в потоке с помощью количественных многоразовых оценок новых знаний с последующей квалификацией их как категорий. Это направление контент-анализа получило название Data Mining — дословно "раскопка данных".

Заметим, что при любом из двух подходов происходит ни что иное, как генерация новых категорий.

Контент-мониторинг

Таким образом, в простейшем виде идею контент-мониторинга можно сформулировать как постоянное выполнение узко очерченного своими задачами контент-анализа беспрерывных информационных потоков. Подчеркнем, что именно беспрерывное воспроизведение во времени процесса обработки входных данных есть самой характерной особенностью контент-мониторинга. Собственно контент-анализ выступает здесь как составная, а контент-мониторинг имеет собственную проблематику и собственные пути решения прикладных задач.

Методы контент-мониторинга, как эволюция идеологии контент-анализа, получили значительное развитие на территории бывшего СССР. Так, наиболее интересными сегодня являются проекты М. Г. Крейнеса "Ключи от текста", Д. А. Поспелова "Интерактивное выявление семантических структур текста", проект "Оружие аналитика" компании "Инвента", проект ВААЛ и прочие.

Контент-мониторинг прессы

Рассмотрим использование метода контент-мониторинга на конкретном примере мониторинга информационного потока социально-политического направления, которое порождает газетная печать Украины. Это исследование было начато в 1993 году и ведется с целью информационно-аналитического обеспечения аппарата государственной власти отделом организации и использования документального фонда ФПУ Национальной библиотеки Украины имени В. И. Вернадского. Контент-мониторинг выполняется беспрерывно, начиная с 1993 года по сегодняшний день. Таким образом, накоплен более чем 10-летний опыт мониторинга значительных по объему информационных потоков на основе использования метода контент-анализа и создания автоматизированных информационно-аналитических систем.

Отличительная особенность такой работы состоит, прежде всего, в обслуживании узкого круга потребителей со специфической сферой задач, которые требуют оперативного решения. Это, в свою очередь, требует четкой постановки информационно-аналитических задач и тесного контакта между заказчиками и службами поиска и анализа информации. К числу наиболее сложных проблем в этой связи следует отнести определение специфических информационных интересов, которые довольно часто до конца не осознают и сами потребители аналитических материалов. Поэтому уточнение информационных потребностей в большинстве случаев происходит уже в процессе работы. Возникают трудности как с определением исследуемых параметров и их предельных значений, так и с установлением частоты оценки этих параметров. Процесс мониторинга усложняет необходимость разнопланового воспроизведения результатов наблюдений в текстовых обзорах и в статистической форме. Для решения поставленных задач нужно в полном объеме предусмотреть набор параметров и их количественные характеристики, обнаружить причинно-следственные связи между ними и включить методологические принципы исследования в описание функционирования технических систем. Сама подготовка информационно-аналитических материалов состоит из ряда последовательных процедур, начиная от упорядочения списка первоисточников для просмотра, разработки методик отбора и классификации материалов, их автоматической обработки и заканчивая анализом занесенной в базы данных (БД) информации и формированием результатов мониторинга прессы.

У автоматизированной технологии контент-мониторинга существует несколько важных особенностей:

- использование ключевого фрагмента публикации как единицы формирования текстового информационного массива;
- формирование банка ключевых фрагментов публикаций является объединением двух взаимосвязанных автоматизированных процессов: аналитико-синтетической переработки и многоуровневой процедуры контент-анализа текстов публикаций;
- индексация ключевых фрагментов публикаций происходит при помощи многофасетной классификации.

Уникальность предложенной технологии состоит в объединении содержательных и количественных методов контент-анализа. 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.

Для контент-мониторинга используются монотематические по наполнению системы с возможностями для многоаспектного использования информации при анализе и подготовке материалов.

Процедура контент-анализа публикаций направлена на выделение из текста фрагментов, которые отвечают наименьшему, но целостному модулю информации в границах исследуемой проблемы. В рамках такого модуля определяются элементы проблемы, адекватные конкретным значениям классификатора и между ними устанавливаются связи для последующей формальной передачи содержания фрагмента публикации с помощью фасетной формулы. Каждой из выделенных цитат после ее тщательного анализа присваивается совокупность определенных индексов (кодов), которые отвечают конкретным значениям классификатора и располагаются в строго фиксированной последовательности в фасетной формуле. Таким образом, на основе разработанного

классификатора с помощью фасетной формулы ведется формальное описание выделенного фрагмента текста, а совокупность фасетных формул всех выделенных фрагментов обеспечивает формальное описание документа в целом (в контексте исследуемой проблемы). Информация, которая не касается проблемы, не выделяется из текста и не заносится в БД. Введенный в информационную систему документ представляет собой совокупность ключевых фрагментов текста (каждый из которых заиндексирован в соответствии с его содержанием). Процедура обработки публикаций — своеобразное информационное сито, которое пропускает лишь релевантную теме информацию в виде фрагментов текста. Среди преимуществ данной технологии, следует отметить сведение к минимуму информационного шума. Сформированные БД, сохраняя текст оригинала, — достаточно компактны и удобны в работе. Среди недостатков технологии — большие затраты интеллектуальной работы как при обработке первоисточника, так и при наполнении БД. Трудоемкость технологии обусловлена главным образом тем, что системы создаются на базе газетных изданий преимущественно регионального происхождения, электронные версии которых в данное время не получили распространения. Более того, чрезвычайно низкое качество печати таких изданий делает невозможным использование процедуры электронного распознавания текста.

Система обработки информационных массивов представляет собой совокупность информационных файлов, которые аккумулируют информацию о фасетах, значениях фасетных индексов и порядке сортировки информационных модулей, исполнительного файла и файла конфигуратора. Специально написанные программные средства разбивают входной документ на отдельные независимые фрагменты (информационные модули), автоматически обеспечивая каждый из них ссылкой на библиографическую информацию, которая была внесена при описании документа. В результате таких технологических преобразований формируется массив ключевых фрагментов публикаций, который является информационным отображением исследуемой проблемы. Кроме того, каждый информационный модуль такой сети при работе технологических программ разбивается на три независимые составляющие, каждая из которых может функционировать самостоятельно. Первая из них представляет собой цитату документа (содержательная информация), вторая — полное библиографическое описание документа и третья — фасетную формулу (структурная характеристика фрагмента текста).

Предложенный способ представления и расчленения входной информации позволяет многопланово использовать ее в процессе формирования выходных текстовых файлов и при получении количественных показателей.

Технологически предусмотрены многоаспектные комбинации цитат в рамках параметров классификатора. В качестве доминантного может быть выбран любой из элементов фасетной формулы или его отдельное значение. Проблему можно представлять как комплексно, так и избирательно — в случае, когда пользователя интересуют только отдельные аспекты проблемы.

Таким образом, в технологии проведения контент-мониторинга информационного потока социально-политической направленности на всех этапах, как в технологии обработки первоисточников и формирования БД, так и в программных средствах аналитико-синтетической обработки информации, предусмотрено разнообразное и многоаспектное использование тематически очерченной входной информации. Анализ и синтез включенных в систему фрагментов разрешают получать оригинальные информационные продукты. Выходные файлы могут быть представлены в виде автоматически сформированных дайджестов (структурированных цитат публикаций с соблюдением

определенной иерархии и порядка сортировки) или статистических таблиц, то есть частотных характеристик исследуемой проблемы.

Как показал многолетний опыт, информация в таком виде очень удобна для аналитиков. Пользователи получают максимально краткую, но полную и объективную информацию по проблеме, которая интересует их на конкретном этапе. Таким образом, удается решать основные информационные проблемы нашего времени: отсеивать информационный шум, обеспечивать информационный поиск и отбор релевантной информации из довольно значительных по объему массивов документов и вносить в информационный поток необходимую структурированность и системность при сохранении его непрерывности.

Информационные технологии позволяют вводить новые элементы, необходимые в отдельных случаях конкретным потребителям. Технически это решается путем добавления в фасетную формулу новых фасет и соответствующих им совокупностей фасетных классов и их индексов — в классификатор. Таким образом, в процессе развития проблемы пополняется и модифицируется начальный вариант классификатора, и, соответственно, совершенствуется сама система в плане возможностей анализа проблемы.

Уникальность методологии контент-мониторинга состоит также в том, что она не привязана ни к конкретной СУБД, ни к конкретным видам источника информации, ни к тематике информации. Хотя технологии были более тщательно апробированы на материалах журнальной и газетной печати социально-политического характера, теоретически они могут быть использованы для мониторинга любых информационных потоков, в том числе и для информационных сообщений в сети интернет.

На сегодняшний день информационно-аналитические системы политического направления охватывают широкий спектр вопросов, обеспечивая мониторинг освещения печатью деятельности и имиджа политических деятелей и партий. При этом используется единая классификационная схема, которая разрешает совместно использовать и взаимно дополнять разные по тематическим наполнениям системы мониторинга. Также приведенная методология успешно используется для мониторинга в печати имиджа украинских библиотек. Результаты последнего, начиная с января 2003 года, представлены на сайте Национальной библиотеки Украины имени В. И. Вернадского в виде ежемесячных аналитических обзоров. Многолетний опыт позволяет утверждать, что возможности и круг использования данной методологии еще далеко не исчерпаны.

Література

1. *Федорчук А. Г.,* Танатар Н. В. Теоретико-методичні засади аналізу інформаційного потоку соціально-політичного спрямування // Бібліотекознавство, документознавство, інформологія. — 2004. — № 2. — С. 33-38.
2. *Танатар Н. В.* Информационно-аналитическое обслуживание высших органов государственной власти с использованием автоматизированных систем ключевых фрагменов публикаций // Б-ки нац. акад. наук: проблемы функционирования, тенденции развития: Науч.-практ. и теорет. сб. — 2003. — Вып. 2. — С. 91-100.
3. *Танатар Н. В.* Бібліотеки України — у дзеркалі преси (Контент-аналіз газетних публікацій, липень-вересень 2002 року) // Бібл. планета. — 2003. — № 2. — С. 11-14.
4. *Сорока М., Танатар Н.* Використання методу контент-аналізу при створенні автоматизованих інформаційних систем // Наук. пр. НБУВ. — 1998. — Вып. 1. — С. 318-323.

5. *Танатар Н., Федорчук А.* Сучасні інформаційні технології прес-моніторингу передвиборних кампаній // Українська періодика: Історія і сучасність: Доп. та повід. шостої Всеукр. наук. — теорет. конф., 11-13 трав., 2000 р. — Л., 2000. — С. 342-344.
6. *Федорчук А., Танатар Н.* Президентські вибори 1999 року: контент-аналіз матеріалів преси // Там само. — С. 351- 354.